Effect modification in anchored indirect treatment comparisons: Comments on "Matching-adjusted indirect comparisons: Application to time-to-event data"

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

[1]Department of Statistical Science, University College London, London, United Kingdom

[2]Medical Affairs Statistics, Bayer plc, Reading, United Kingdom

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

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

Correspondence

*Antonio Remiro Azócar, Department of Statistical Science, University College London, London, United Kingdom. Email: antonio.remiro.16@ucl.ac.uk. Tel: (+44 20) 7679 1872. Fax: (+44 20) 3108 3105

Present Address

Antonio Remiro Azócar, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom

This commentary regards a recent simulation study conducted by Aouni, Gaudel-Dedieu and Sebastien, evaluating the performance of different versions of matching-adjusted indirect comparison (MAIC) in an anchored scenario with a common comparator. The simulation study uses survival outcomes and the Cox proportional hazards regression as the outcome model. It concludes that using the LASSO for variable selection is preferable to balancing a maximal set of covariates. However, there are no treatment effect modifiers in imbalance in the study. The LASSO is more efficient because it selects a subset of the maximal set of covariates but there are no cross-study imbalances in effect modifiers inducing bias. We highlight the following points: (1) in the anchored setting, MAIC is necessary where there are cross-trial imbalances in effect modifiers; (2) the standard indirect comparison provides greater precision and accuracy than MAIC if there are no effect modifiers in imbalance; (3) while the target estimand of the simulation study is a conditional treatment effect, MAIC targets a marginal or population-average treatment effect; (4) in MAIC, variable selection is a problem of low dimensionality and sparsity-inducing methods like the LASSO may be problematic. Finally, data-driven approaches do not obviate the necessity for subject matter knowledge when selecting effect modifiers. R code is provided in the Appendix to replicate the analyses and illustrate our points.

KEYWORDS:
Matching-adjusted indirect comparison, effect modification, variable selection, indirect treatment comparison

We read with interest the recent paper by Aouni, Gaudel-Dedieu and Sebastien[1] (AGS), in which they examine the statistical performance of matching-adjusted indirect comparison (MAIC) in an "anchored" scenario with a common comparator treatment across studies. MAIC is a relatively novel methodology used to compare the efficacy or effectiveness of treatments in a two-study scenario, often seen in health technology assessment, where there is access to individual patient data (IPD) on covariates and outcomes from one study, and only access to published aggregate-level data (ALD) — typically a limited set of covariate moments ("Table 1") and summary outcomes — from the other study. AGS denote the IPD study, comparing intervention A to common comparator C, as "study A" (population $S = 1$), and the ALD study, comparing intervention B to common comparator C as "study B" (population $S = 2$). In this sparse network, with only one study per active treatment, indirect comparisons are vulnerable to bias induced by cross-trial differences in baseline covariates.

AGS compare several implementations of MAIC in a simulation study, using time-to-event outcomes and the Cox proportional hazards regression as the outcome model. AGS conclude that using the LASSO technique "is better" than balancing a maximal set of covariates. We raise the following points about AGS' conclusions:

1. In the anchored setting, MAIC is necessary where there are cross-study imbalances in effect modifiers that induce bias. If this is the case, MAIC requires accounting for *all* effect modifiers in order to satisfy the *conditional constancy of relative effects* assumption and be unbiased. [2,3]

2. In the data-generating mechanism of the simulation study, there is no treatment effect modification. Therefore, there is no bias to remove and accuracy is driven entirely by precision. LASSO has lower variance because it selects a subset of the maximal set of covariates. In the simulation study setting, MAIC is not necessary — the standard indirect comparison would provide the greatest precision and accuracy.

3. The target estimands of the simulation study are conditional treatment effects (a ratio of conditional hazard ratios) but MAIC targets marginal or population-average treatment effects (a ratio of marginal hazard ratios). These measures of effect may not coincide for the hazard ratio, even under covariate balance and no confounding, as the measure of effect is non-collapsible. The bias observed is potentially due to this conflation of effects.

4. In the context of MAIC, variable selection is typically a problem of low dimensionality. A sparsity-inducing method like LASSO that shrinks the regression coefficients may be problematic and induce bias, due to breaking the conditional constancy of relative effects. Data-driven approaches do not obviate the necessity for substantive subject matter knowledge when selecting effect modifiers.

These points are discussed in more detail in the below paragraphs. In the Appendix, we attempt to replicate the analyses of AGS using the information provided in the original article and illustrate some of the aforementioned points.

EFFECT MODIFICATION AND METHOD ASSUMPTIONS

Effect modification

Using the notation of AGS, covariate X is an *effect measure modifier* (*effect modifier* for short) if the effect of treatment T on outcome $\mathcal{O}$, on a specific scale, varies by level or strata of X. [4]

Studies A and B are both randomized controlled trials (RCTs). In an ideal or perfectly-executed scenario with perfect measurement and no missing data, the nature of randomization implies that, in expectation, there is covariate balance (i.e., the covariates are similarly distributed) and the active treatment and control groups are exchangeable. The random assignment of units to treatment eliminates confounding by design, [5,6] providing protection from the confounding of the treatment effect by baseline covariates. Hence, "within-study" inferences are internally valid — one can produce unbiased estimates for the marginal A vs. C treatment effect in the $S = 1$ population (we shall denote this estimand $\Delta_{AC}^{(1)}$), and for the marginal B vs. C treatment effect in the $S = 2$ population (we denote this estimand $\Delta_{BC}^{(2)}$).

In the context of study A, different levels of the effect modifiers of treatment A are associated with differential marginal or population-average treatment effects for A vs. C. Hence, the set of effect modifiers is the source of treatment effect variation, fully explaining the heterogeneity of the effect of treatment A vs. C. If there are differences in effect modifiers between the $S = 1$ population and the $S = 2$ population, implicitly assumed to be the target population by MAIC, the marginal treatment effect for A vs. C in $S = 1$ is expected to differ from the marginal treatment effect for A vs. C in $S = 2$. [7,8,9,10,11,12] Namely, if covariate X modifies the effect of treatment A, i.e., such treatment effect is heterogeneous, and the distribution of the factor X in $S = 1$ differs from the distribution of that factor in $S = 2$, we typically have $\Delta_{AC}^{(1)} \neq \Delta_{AC}^{(2)}$. [a] The difference between $\Delta_{AC}^{(1)}$ and $\Delta_{AC}^{(2)}$ is a function of the treatment effect heterogeneity induced by the effect modifiers and the imbalance (difference in means) in effect modifiers across populations. This heterogeneity is the primary motivation for the use of MAIC in the anchored setting.

---

[a]Except in the pathological case where the bias induced by different effect modifiers is in opposite directions and cancels out.

Method assumptions

Standard indirect comparison methods such as the Bucher method[13] do not explicitly produce an estimate for any target population in particular — it is not typically stated whether the target population is $S = 1$, $S = 2$ or otherwise, regardless of whether the analysis is based on ALD or on IPD from each study.[14] The Bucher method estimates the marginal treatment effect for A vs. B as:

$$\hat{\Delta}_{AB} = \hat{\Delta}_{AC} - \hat{\Delta}_{BC},$$

where $\hat{\Delta}_{AC}$ is the estimated marginal treatment effect of A vs. C, and $\hat{\Delta}_{BC}$ is the estimated marginal treatment effect of B vs. C. This comparison is only valid for any target population where treatment effects are not heterogeneous or all potential effect modifiers (for A vs. C and B vs. C) are equidistributed across studies.

In the context of the article by AGS, one only has access to published ALD for study B. Therefore, the indirect comparison is conducted by necessity in population $S = 2$. The Bucher method would estimate the marginal treatment effect for A vs. B in the $S = 2$ population as:

$$\hat{\Delta}_{AB}^{(2)} = \hat{\Delta}_{AC}^{(1)} - \hat{\Delta}_{BC}^{(2)}.$$

Here, $\hat{\Delta}_{AC}^{(1)}$ is the estimated marginal treatment effect of A vs. C in the $S = 1$ population, using the IPD for study A, and $\hat{\Delta}_{BC}^{(2)}$ is the estimated marginal treatment effect of B vs. C in $S = 2$, available in the published ALD or calculated from published aggregate outcomes. In order for this comparison to be unbiased, a requirement is the *constancy of relative effects*:[2,3] $\Delta_{AC}^{(1)} = \Delta_{AC}^{(2)}$. This means that all covariates, measured or unmeasured, modifying the effect of treatment A versus C must be balanced across the populations.[b]

The objective of MAIC is to weight the IPD for study A so that it resembles the population of study B ($S = 2$), with respect to the distribution of selected baseline covariates. The weighted IPD is then used to estimate the direct target of MAIC: $\Delta_{AC}^{(2)}$, the marginal A vs. C treatment effect in the $S = 2$ population. Then, an adjusted indirect comparison is performed in the $S = 2$ population, where the marginal A vs. B treatment effect is estimated as:

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

where $\hat{\Delta}_{AC}^{(2)}$ is the estimated marginal treatment effect of A vs C in $S = 2$.

AGS imply that the weighting procedure should include all the variables that explain the *absolute* outcome $\mathcal{O}$ under active treatment A. Conditional on these prognostic covariates, the distribution of the absolute outcomes is independent of the population. An important point is that this assumption can be relaxed in the anchored scenario. In an anchored MAIC, a comparison of *relative* outcomes or effects, not absolute outcomes under each treatment, is of interest. An anchored comparison only requires identifying and balancing the effect modifiers, which are the covariates that explain the heterogeneity of the A vs. C treatment effect. This is likely a smaller set of variables than that of variables explaining the absolute outcomes.

An anchored MAIC assumes the *conditional constancy of relative effects*[2,3] across populations. Namely, given the selected effect-modifying covariates, the (weighted) covariate-adjusted marginal treatment effect for A vs. C in population $S = 1$ is equal to the unadjusted marginal treatment effect for A vs. C in $S = 2$.[c] That is, the weighting model needs to include *all* the variables that modify the effect of treatment A, whether measured or unmeasured. AGS posit that an "influential covariate" can be ignored if it is balanced between the populations. However, one must account for *all* effect modifiers, regardless of balance between studies, because excluding balanced covariates from the weighting procedure does not ensure balance after the weighting.[2,3] The MAIC estimate $\hat{\Delta}_{AC}^{(2)}$ for the marginal A vs. C treatment effect in population $S = 2$ will be biased if any pre-treatment covariates that modify the treatment effect are unobserved (either in study A or study B) or unaccounted for in the weighting model.

The conditional constancy of relative effects is an untestable assumption because the A vs. C trial has not been conducted in the $S = 2$ population, where outcomes under treatment A are unobserved. It is also a very demanding assumption in practice due to a variety of reasons. Firstly, it requires that all effect modifiers of treatment A are measured in the IPD for study A. In addition, the set of published baseline characteristics for study B must be sufficiently rich, such that all effect modifiers of treatment A are available. Unfortunately, this is a key challenge as data on the $S = 2$ population are often limited, and it can sometimes be difficult to find common measures between different studies.[11,18] Moreover, only marginal moments of the covariates are published

---

[b]If ALD were available for study A and IPD for study B, the comparison would have to be conducted in $S = 1$, and would require the cross-study balance of all covariates modifying the effect of treatment B.

[c]There are many ways to articulate the assumption of conditional constancy of relative effects. Other formulations appear in the "generalizability" or "transportability" literature.[8,15,16,17] One can consider that being in study A ($S = 1$) or study B ($S = 2$) does not carry over any information about the marginal A vs C treatment effect, once we condition on the treatment effect modifiers. Namely, we can assume that trial assignment/selection is conditionally "ignorable", unconfounded or exchangeable for such marginal treatment effect, i.e., conditionally independent of the treatment effect, given the selected effect modifiers. This means that after adjusting for these effect modifiers, the treatment effect and trial assignment are conditionally independent.

for the ALD study, with data on joint covariate distributions typically unavailable. Marginal balance across study populations does not guarantee multidimensional balance across the full joint distributions without further distributional assumptions. [19] Conditional constancy could be broken if there are higher-order treatment-by-covariate interactions, involving two or more covariates. Importantly, the conditional constancy assumption is tied to the scale used to define the treatment effects and effect modifiers. For a given covariate, treatment effect modification may exist on one scale but not on another. [20] Furthermore, the effect modifier status of a variable can itself be "modified" by simultaneously accounting for other variables. [4]

SIMULATION STUDY PROTOCOL

As explained earlier, the divergence between $\Delta_{AC}^{(1)}$ and $\Delta_{AC}^{(2)}$ is a function of: (1) the extent to which each covariate modifies the treatment effect; and (2) the imbalance in effect modifiers across studies, which is the extent to which the effect modifiers are related to trial assignment. In the simulation study protocol of AGS we face the latter threat but not the former.

Two covariates, Age and ISS, govern trial assignment through a logistic regression model: $\text{logit}(\mathbb{P}(S = 1|\text{covariates})) = \theta_1 \times (\text{Age} - 65) + \theta_2 \times \text{ISS}$, with the parameter values set to $\theta_1 = \theta_2 = 0.1$. Age and ISS, a binary disease indicator, determine the trial selection mechanism by which subjects come to be assigned to one study or the other. This creates some covariate imbalance between the populations of the two studies. However, while Age and ISS are associated with trial selection, they do not modify the A vs. C marginal treatment effect. Therefore, any covariate imbalances do not induce bias in the standard indirect comparison.

As there are no covariates modifying the effect of treatment A vs. C in the data-generating mechanism, the marginal treatment effects $\Delta_{AC}^{(1)}$ and $\Delta_{AC}^{(2)}$ are identical. In our Appendix, these are computed as log hazard ratios of $\Delta_{AC}^{(1)} = \log(0.76)$ and $\Delta_{AC}^{(2)} = \log(0.76)$. Because there is no treatment effect variation or heterogeneity, the assumptions of the standard indirect comparison hold. This will be unbiased because the naive IPD estimate $\hat{\Delta}_{AC}^{(1)}$ is an unbiased estimate of $\Delta_{AC}^{(2)}$, and is applicable to the $S = 2$ population. MAIC will also be unbiased because there are no effect modifiers to account for. Therefore, no effect modifiers have been unaccounted for, and the MAIC estimate $\hat{\Delta}_{AC}^{(2)}$ is an unbiased estimate of $\Delta_{AC}^{(1)}$ across the simulation scenarios. In this case, the use of a more complex method like MAIC with a larger number of assumptions is not warranted — weighting induces an increase in variance without the potential for bias reduction.

These points are illustrated in our Appendix, which can be viewed as a supplement to this section. In the outcome-generating mechanism of AGS, when accounting for different covariate sets in the weighting model, MAIC provides unbiased estimates of $\Delta_{AC}^{(1)} = \log(0.76)$. When we add an imbalanced effect modifier, Age, to the outcome-generating process and account for Age in the weighting model, the MAIC estimate in the $S = 2$ population is biased for $\Delta_{AC}^{(1)}$. Similarly, the naive IPD Bucher estimate would be biased for $\Delta_{AC}^{(2)}$.

CLARIFICATION OF ESTIMANDS

Further clarification of the treatment effect targeted by the simulation study, and of the estimand targeted by MAIC, is required. Consider the time-to-event setting presented by the original article. Using our own notation, each randomized trial has survival endpoint Y, subject to censoring, a dichotomous treatment indicator T (coded 1 for active treatment and 0 for control), and a set of baseline covariates $\boldsymbol{X}$ which are prognostic for Y. In the simulation study design, the hazard function for each subject at follow-up time y follows the Cox model:

$$h(y \mid \boldsymbol{X}, T) = h_0(y) \exp(\boldsymbol{\beta}^\top \boldsymbol{X} + \beta_T T), \tag{1}$$

where $h_0(\cdot)$ is the baseline hazard function, assumed to be constant, and $\boldsymbol{\beta}$ and β_T denote coefficients for the prognostic covariates and treatment, respectively. The hazard function above is conditional on $\boldsymbol{X}$ and T. The *conditional* hazard ratio comparing active to control treatment at y is $h(y \mid \boldsymbol{X}, T = 1)/h(y \mid \boldsymbol{X}, T = 0) = \exp(\beta_T)$. The *marginal* or *population-average* hazard ratio at y is $h(y \mid T = 1)/h(y \mid T = 0)$.

We return to the notation used in the original article (Section 2.3). Because the treatment coefficients in the outcome-generating model, $b_A = \log(0.53)$ and $b_B = \log(0.55)$, AGS state that the "true hazard ratio of treatment A vs. treatment B is equal to $0.53/0.55 = 0.964$". It is worth highlighting that b_A and b_B (β_T in Equation 1 above) are conditional treatment effects because they are coefficients of the outcome-generating Cox regression, conditional on the effects of the prognostic variables that have also been included in the model. That is, b_A represents the true average log hazard ratio, at the individual level, of changing a

subject's treatment from C to A (the average treatment effect conditioned on the average combination of covariates in $S = 1$, or the average effect across sub-populations of subjects who share the same covariates). Similarly, b_B denotes the true average effect in $S = 2$, at the unit level, of changing a subject's treatment from C to B.

The target estimands of the simulation study are conditional or unit-level treatment effects. However, MAIC targets an estimand that is calibrated at a different hierarchical level. This estimand is a marginal treatment effect, which has a different interpretation than the conditional treatment effect. The marginal treatment effect is the average effect, at the *population level*, of moving everyone in the population from one treatment to another. MAIC is a method based on propensity score weighting, where the A vs. C treatment effect estimate targets a *marginal* log hazard ratio, rather than a *conditional* log hazard ratio. This is because the outcome model is a univariable weighted regression of outcome on treatment assignment. The estimated treatment effect is the fitted treatment coefficient of this weighted regression. This coefficient estimates a relative effect between subjects that, on expectation, have the same distribution of covariates, corresponding to population $S = 2$.

AGS may have derived the B vs. C treatment effect in $S = 2$ by fitting a simple Cox regression of outcome on treatment assignment. This is regularly the case for treatment effects reported in RCT publications, where evidence is often stated at the population level. In that case, the estimated (log) hazard ratio for B vs. C also targets a marginal treatment effect. It is assumed that the objective of methodologies such as MAIC in health technology assessment is to help make inferences and policy decisions at the population level, not the individual level.

The (log) hazard ratio is a non-collapsible measure of effect. Therefore, marginal and conditional (log) hazard ratios may not coincide, even where there is covariate balance and the absence of confounding.[21] In this particular simulation study, the observed relative bias (Tables 5 to 9 of the original article by AGS) may be due to a conflation of marginal and conditional measures of effect. While the true target estimand is defined as a ratio of conditional treatment effects (hazard ratios as opposed to log hazard ratios, in this case), MAIC targets a ratio of marginal treatment effects. In our Appendix, the ratio of marginal hazard ratios is computed as 0.984, which does not coincide with the ratio of conditional hazard ratios, $\exp(b_A)/\exp(b_B) = 0.53/0.55 = 0.964$, reported by AGS as the true target estimand. Bias has also been induced in other simulation studies where the measure of effect is non-collapsible, due to targeting the wrong measure of effect.[22,23] Collapsibility does not hold for most measures of interest in population-adjusted indirect comparisons, such as (log) hazard ratios or (log) odds ratios in oncology applications, when investigating time-to-event or binary outcomes, respectively.

CONCLUDING REMARKS

Our commentary has some implications for the conclusions of AGS. In the original article, AGS explore use of the LASSO regression for covariate selection. This approach is perceived to be "better" and "more efficient" than selecting a maximal set of covariates because this "reduces the bias as well as the SE".

In our opinion, the bias for both approaches is about the same in the simulation study. This is because marginal effects are constant across studies — any implementation of MAIC is likely unbiased, as is the standard indirect treatment comparison. Secondly, LASSO has greater precision because it will select a subset of the maximal set of covariates. As none of the covariates are effect modifiers, including more of these in the weighting model does not remove bias further but reduces the effective sample size and inflates the standard error. This may explain why the LASSO is more accurate or efficient. The standard indirect comparison would be even more accurate in these simulation settings because it does not adjust for any covariates.

These results confirm that covariates that do not modify the A vs. C treatment effect should be excluded from an anchored MAIC, as recommended by current guidance.[2,3] There is a disadvantage to unnecessary or excessive adjustment. Considering bias-variance trade-offs, accounting for a covariate that is not an effect modifier is a sub-optimal use of information because it increases variance and affects the precision of the treatment effect estimate negatively. Similarly, failure to include an effect-modifying covariate is also a sub-optimal use of information that results in increased bias.

Further research should evaluate the performance of the LASSO in a scenario where there are cross-study imbalances in effect modifiers inducing bias. Nevertheless, we highlight two caveats that are important to bear in mind:

1. **Sparsity-inducing methods.** The LASSO imposes a penalty which shrinks all regression coefficients towards zero and directly sets some to zero, thereby providing a variable selection procedure. While AGS perceive this to be an advantage, it may be a disadvantage in the typical application of MAIC, where the covariate selection problem is one of low dimensionality. In practice, there may be little comparability of measures across data sources. We are likely only capable of selecting a subset of the true effect modifiers because we only have access to a few covariates from the published ALD

from study B. Anchored MAIC is susceptible to bias where effect modifiers are omitted because the conditional constancy of relative effects is invalid. Therefore, use of a sparsity-promoting method may be problematic; unless there is a large number of potential effect modifiers and poor overlap inducing imprecision (outweighing the potential for bias reduction).

2. **Interaction testing.** Within the biostatistics literature, effect modification is usually referred to as interaction, because effect modifiers are considered to alter the effect of treatment by interacting with it on a specific scale,[24] and are often identified by evaluating statistical interaction terms in regression models fitted to the IPD.[3,25,26,27,28] While the standard LASSO does not include interactions and is restricted to look for main linear effects only,[29] there are extensions specifically designed to consider interaction terms.[30,31] Nevertheless, any data-driven approach based on statistical testing will be hindered by the relatively modest sample sizes of individual RCTs. These are, almost invariably, underpowered to assess treatment effect heterogeneity,[32,33] with statistical tests only detecting very large and possibly implausible interactions.[34] Meta-analyses of multiple trials, using IPD or ALD, may provide greater power.[33,35]

Ultimately, the selection of effect modifiers to include in the weighting model requires thorough care and investigation. Treatment effect heterogeneity is a complex and largely unknown process, likely to require leveraging both statistical and clinical expertise. It may be reasonable to balance a variable if there is a strong biological reasoning for effect modification, even if the interaction is statistically weak. The identification of effect modifiers will likely depend on a combination of data exploration and analysis, prior subject matter knowledge, firm clinical or biological hypotheses, and literature reviews. Novel causal inference methods, typically relying on directed acyclic graphs,[27,36] also show promise, but note that these themselves are often based upon background knowledge and expert opinion. Admittedly, strong theory on treatment effect modification is often not available, particularly for novel therapies with unknown mechanisms of action. Finally, the inclusion of all effect modifiers is an unverifiable assumption. Sensitivity analyses are crucial and should be conducted under alternative effect modifier specifications to explore the dependence of inferences on this selection and the robustness of results.[37,38]

One of the reviewers raises the problem of valid inference post-model selection with the LASSO[39] and potential sample-splitting solutions.[39,40] The LASSO's biased significance tests and the reduced power that comes with sample-splitting are relevant for the statistical detection of interactions. Nevertheless, we highlight that the propensity score model for the weights is a nuisance model fitted before the final step of marginal effect estimation in MAIC. We do not necessarily seek valid inferences for the "pre-processing" step used to select the model, focusing instead on valid inference in the final step. One can view estimation of the weights or of alternative nuisance models as a "best prediction" problem, for which statistical learning or data-driven methods show some potential.[41,42,43]

APPENDIX

In this appendix, we attempt to replicate the analyses by AGS using the information provided in their article. Some of the points discussed in our commentary are exemplified. We perform the analyses using R software version 3.6.3.[44]

Data-generating mechanism

We simulate large populations of 100,000 subjects for both $S = 1$ (study A) and $S = 2$ (study B), with 50,000 units under each treatment. Tables 1 and 2 of the article by AGS report descriptive statistics of the simulated studies, where covariate means across the simulations are included. Not much information is provided on the underlying joint covariate distributions used to generate the study populations, e.g. on the moments (means and standard deviations) and forms of the marginal distributions, and on the correlation structure of the real database, which AGS resample via the bootstrap to simulate the covariates.

We use the information on overall covariate means, ignoring sampling variability, and make certain parametric assumptions about the covariate distributions to simulate the populations. It is assumed that the studies are appropriately randomized; hence it should make no difference to simulate the arms of each study jointly or separately, using the arm-specific covariate means. We assume that the covariates are uncorrelated.

```
rm(list=ls())
set.seed(555) # set random seed for reproducibility
```

```
N <- 100000 # number of simulated subjects per population

trt <- c(rep(1, N/2), rep(0, N/2)) # assume a 1:1 treatment allocation ratio

# population S=1 (study A)
age_S1 <- rnorm(n=N, mean=69.3, sd=5) # assume Age is normally-distributed, SD=5
plnen_S1 <- rpois(n=N, lambda=3.4) # assume PLNEN is Poisson-distributed
iss_S1 <- rbinom(n=N, size=1, prob=0.74) # assume ISS is Bernoulli-distributed
refr_S1 <- rbinom(n=N, size=1, prob=0.92) # assume Refr is Bernoulli-distributed

# population S=2 (study B)
age_S2 <- rnorm(n=N, mean=62.1, sd=5)
plnen_S2 <- rpois(n=N, lambda=3.4)
iss_S2 <- rbinom(n=N, size=1, prob=0.77)
refr_S2 <- rbinom(n=N, size=1, prob=0.92)
```

In Section 2.3 of AGS, it is stated that "for each patient, whatever the population, occurrence of the event of interest is governed by a proportional hazard model with a constant hazard", such that:

$$h = h_0 \times \exp\left((b_A \times (\text{"study A"}) + b_B \times (\text{"study B"})) \times \text{arm} + b_1 \times \text{Refr} + b_2 \times \text{ISS} + b_3 \times \text{PLNEN}\right),$$

where h_0 denotes the baseline hazard function, assumed constant; and b_1, b_2, and b_3 represent the conditional prognostic effects of Refr (a binary variable indicating "refractory to some drug class"), ISS and PLNEN (the number of prior treatment lines), respectively. The true *conditional* treatment effect (log hazard ratio) for A vs. C (in both $S = 1$ and $S = 2$) is $b_A = \log(0.53)$ and the true *conditional* treatment effect for B vs. C in both study populations is $b_B = \log(0.55)$. Note that, due to the non-collapsibility of the (log) hazard ratio, these conditional effects are specific to the adjustment set of covariates used in the outcome-generating process. AGS mention that latent event times for both studies, without accounting for censoring, follow an exponential distribution but do not provide a value for the rate parameter. We set this to $0.5/365$. We use the data-generating process of Bender et al.[45] to simulate exponentially-distributed survival times under proportional hazards.

```
# outcome model parameters
b_A <- log(0.53) # conditional treatment effect (log HR) for A vs. C in study A
b_B <- log(0.55) # conditional treatment effect (log HR) for B vs. C in study B
b_1 <- 1.0682 # conditional effect of covariate Refr
b_2 <- -0.6651 # conditional effect of covariate ISS
b_3 <- 0.0825 # conditional effect of covariate PLNEN
rate <- 0.5/365 # rate of latent time distribution (not specified by AGS)
cens_rate <- 0.1/365 # rate of censoring distribution as set by AGS

# latent event times simulated according to Bender et al. (2005)
surv.sim <- function(N, LP, rate, cens_rate) {
  U <- runif(n=N)
  Tlat <- -log(U)/(rate*exp(LP)) # latent event times (exponential)
  C <- rexp(n=N, rate=cens_rate) # censoring distribution (exponential)
  time <- pmin(Tlat, C) # final follow-up times
  status <- as.numeric(Tlat<=C) # final event indicators
  return(cbind(time, status))
}

# simulate survival times and event indicators under A and C in S=1
LP_AC <- b_1*plnen_S1 + b_2*iss_S1 + b_3*refr_S1 + b_A*trt # linear predictor
survival_AC <- surv.sim(N=N, LP=LP_AC, rate=rate, cens_rate=cens_rate)
time_AC <- survival_AC[,1]
```

```
status_AC <- survival_AC[,2]

# simulate survival times and event indicators under B and C in S=2
LP_BC <- b_1*plnen_S2 + b_2*iss_S2 + b_3*refr_S2 + b_B*trt
survival_BC <- surv.sim(N=N, LP=LP_BC, rate=rate, cens_rate=cens_rate)
time_BC <- survival_BC[,1]
status_BC <- survival_BC[,2]
```

True marginal and conditional treatment effects for each study

We calculate the true value of the marginal or population-average treatment effect $\Delta_{AC}^{(1)}$ for A vs. C in the $S = 1$ population, and the true value of the marginal treatment effect $\Delta_{BC}^{(2)}$ for B vs. C in the $S = 2$ population. The marginal effect is the expected difference in the potential outcomes on the log hazard ratio scale if all members of a given population were under active treatment and if all members of the population were under the common comparator. Each simulated population can be conceptualized as the population of a very large randomized experiment or as two potential cohorts of 50,000 subjects, with one cohort under the active treatment and the other under the common comparator.

As the populations are sufficiently large to minimize sampling variability, the true marginal effects are computed by fitting simple univariable Cox regressions, regressing the simulated survival times on the indicator variable denoting treatment status. The estimated treatment coefficient of each regression represents the average difference in the potential outcomes on the log hazard ratio scale, and serves as the log of the true marginal hazard ratio for the two interventions under consideration. This is because the survival times have been generated according to the true data-generating mechanism of AGS, where the true conditional effects are explicit, and which uses the correct conditional model by definition. Due to the non-collapsibility of the hazard ratio, this simulation-based approach has been adopted in previous research to determine the true marginal effect.[21,46,47]

```
library("survival") # to fit Cox proportional hazards models
unadjusted_model_AC <- coxph(Surv(time_AC, status_AC)~trt)
exp(unadjusted_model_AC$coefficients[1]) # marginal hazard ratio for A vs. C (S=1)
```

```
##       trt
## 0.7575748
```

```
unadjusted_model_BC <- coxph(Surv(time_BC, status_BC)~trt)
exp(unadjusted_model_BC$coefficients[1]) # marginal hazard ratio for B vs. C (S=2)
```

```
##       trt
## 0.7697989
```

We know the true conditional treatment effects for A vs. C in $S = 1$ and for B vs. C in $S = 2$ to be $b_A = \log(0.53)$ and $b_B = \log(0.55)$, respectively. Indeed, these values can be recovered by fitting multivariable regressions of outcome on treatment and the covariates to the simulated data.

```
adjusted_model_AC <- coxph(Surv(time_AC, status_AC)~trt+plnen_S1+iss_S1+refr_S1)
exp(adjusted_model_AC$coefficients[1]) # conditional hazard ratio for A vs. C (S=1)
```

```
##       trt
## 0.5294677
```

```
adjusted_model_BC <- coxph(Surv(time_BC, status_BC)~trt+plnen_S2+iss_S2+refr_S2)
exp(adjusted_model_BC$coefficients[1]) # conditional hazard ratio for B vs. C (S=2)
```

```
##       trt
## 0.5500948
```

As expected, conditional and marginal hazard ratios do not coincide, due to the non-collapsibility of the hazard ratio. With non-collapsible measures of effect, conditional treatment effects will vary depending on the covariates used for adjustment in the regression and on the model specification. This may explain why the mean conditional hazard ratios in Table 1 and Table 2 of AGS overestimate b_A and b_B, respectively. This is not bias but a by-product of using a different adjustment set where the measure of effect is non-collapsible.

MAIC analyses

We now conduct MAIC, estimating the marginal treatment effect for A vs.C in the $S = 2$ population. We consider the selection of the balancing set of covariates.

```
# center the S=1 covariates on the S=2 means
cent_age_S1 <- age_S1 - mean(age_S2)
cent_plnen_S1 <- plnen_S1 - mean(plnen_S2)
cent_iss_S1 <- iss_S1 - mean(iss_S2)
cent_refr_S1 <- refr_S1 - mean(refr_S2)

# MAIC function
maic <- function(X) { # X: centered S=1 covariates
  # objective function to be minimized for weight estimation
  Q <- function(alpha, X) {
    return(sum(exp(X %*% alpha)))
  }
  X <- as.matrix(X)
  N <- nrow(X)
  K <- ncol(X)
  alpha <- rep(1,K) # arbitrary starting point for the optimizer
  # objective function minimized using BFGS
  Q.min <- optim(fn=Q, X=X, par=alpha, method="BFGS")
  hat.alpha <- Q.min$par # finite solution is the logistic regression parameters
  log.hat.w <- rep(0, N)
  for (k in 1:K) {
    log.hat.w <- log.hat.w + hat.alpha[k]*X[,k]
  }
  hat.w <- exp(log.hat.w) # estimated weights
}
```

In the first scenario, Scenario 1, we consider balancing the prognostic variables PLNEN, Refr and ISS. According to AGS, this is the correct set of covariates.

```
# centered covariates
cent_X_S1_prognostic <- cbind(cent_plnen_S1, cent_iss_S1, cent_refr_S1)
weights_prognostic <- maic(cent_X_S1_prognostic)
# fit weighted Cox proportional hazards model, robust=TRUE for robust sandwich variance
maic_outcome_model_1 <- coxph(Surv(time_AC, status_AC)~trt, robust=TRUE,
                              weights=weights_prognostic)
# marginal hazard ratio for A vs. C in S=2 population
exp(maic_outcome_model_1$coefficients)
```

```
##       trt
## 0.7575572
```

The marginal or population-average log hazard ratio for A vs. C in the $S = 2$ population ($\Delta_{AC}^{(2)} = \log(0.76)$) is virtually identical to that in the $S = 1$ population ($\Delta_{AC}^{(1)} = \log(0.76)$). There is no bias to remove by MAIC as there is no effect modification. Note that we have assumed that MAIC estimates are unbiased under no failures of assumptions, as recently shown by a simulation study in the context of survival outcomes and the Cox proportional hazards model.[22]

In a second scenario, Scenario 2, we consider balancing a sparse set of covariates. We only balance PLNEN. This selection is questionable because: (1) PLNEN is already balanced; and (2) it has the lowest explanatory power of the prognostic variables — according to AGS, it is the prognostic covariate most often discarded by LASSO.

```
cent_X_S1_sparse <- cent_plnen_S1
weights_sparse <- maic(cent_X_S1_sparse)
maic_outcome_model_2 <- coxph(Surv(time_AC, status_AC)~trt, robust=TRUE,
                              weights=weights_sparse)
# marginal hazard ratio for A vs. C in S=2 population
exp(maic_outcome_model_2$coefficients)
```

```
##       trt
## 0.7575059
```

Again, the marginal or population-average hazard ratio for A vs. C in the $S = 2$ population is practically equal to that in the $S = 1$ population. There is no bias to remove because there are no treatment effect modifiers in imbalance. Hence, only precision and not bias drives accuracy in this setup. Accounting for a lesser number of covariates will always result in greater accuracy or efficiency due to lower reductions in effective sample size and greater precision, i.e., decreased standard error. This explains why the LASSO approach is more efficient than selecting the maximal set of covariates in the article by AGS.

AGS report relative bias for each simulation scenario in Tables 5 to 9 of the original article. In this particular simulation study, the observed relative bias may arise due to a conflation of marginal and conditional measures of effect. For instance, in our Scenario 1, we can compute the marginal hazard ratio for A vs. B in the $S = 2$ population by dividing the marginal hazard ratio for A vs. C in $S = 2$ by the marginal hazard ratio for B vs. C in $S = 2$.

```
exp(maic_outcome_model_1$coefficients)/exp(unadjusted_model_BC$coefficients[1])
```

```
##       trt
## 0.9840976
```

This does not coincide with the true estimand defined by AGS, which is a ratio of conditional hazard ratios, $\exp(b_A)/\exp(b_B) = 0.53/0.55 = 0.964$. The relative bias observed by AGS is likely due to this conflation of effects and the non-collapsibility of the hazard ratio.

MAIC analyses with effect modification

We now consider a scenario, Scenario 3, where Age is a treatment effect modifier for A vs. C (and also B vs. C) on the (additive) log hazard ratio scale. Interaction terms with coefficient $b_{int} = 0.005$ are added to the linear predictor in the outcome-generating model for both studies.

```
b_int <- 0.005 # interaction coefficient for Age
# simulate outcomes under treatments A and C in S=1
LP_AC = b_1*plnen_S1 + b_2*iss_S1 + b_3*refr_S1 + b_A*trt + b_int*age_S1*trt
survival_AC <- surv.sim(N=N, LP=LP_AC, rate=rate, cens_rate=cens_rate)
time_AC <- survival_AC[,1]
status_AC <- survival_AC[,2]

# simulate outcomes under treatments B and C in S=2
LP_BC <- b_1*plnen_S2 + b_2*iss_S2 + b_3*refr_S2 + b_B*trt + b_int*age_S2*trt
```

```
survival_BC <- surv.sim(N=N, LP=LP_BC, rate=rate, cens_rate=cens_rate)
time_BC <- survival_BC[,1]
status_BC <- survival_BC[,2]
```

We consider balancing the maximal set of covariates and use MAIC to estimate the marginal treatment effect for A vs. C in the $S = 2$ population.

```
# centered covariates
cent_X_S1_interactions <- cbind(cent_plnen_S1, cent_iss_S1, cent_refr_S1, cent_age_S1)
weights_interactions <- maic(cent_X_S1_interactions)
# fit weighted Cox proportional hazards model, robust=TRUE for robust sandwich variance
maic_outcome_model_3 <- coxph(Surv(time_AC, status_AC)~trt, robust=TRUE,
                              weights=weights_interactions)
# marginal hazard ratio for A vs. C in S=2 population
exp(maic_outcome_model_3$coefficients[1])
```

```
##      trt
## 0.8765244
```

Because there is a treatment effect modifier, Age, that is in imbalance between the populations, the marginal hazard ratio for A vs. C in $S = 2$ differs from that in $S = 1$. A virtually identical marginal hazard ratio for A vs. C in $S = 2$ is estimated if we only balance Age (Scenario 4), such that balancing the other covariates makes no difference in terms of bias reduction.

```
cent_X_S1_age <- cent_age_S1
weights_age <- maic(cent_X_S1_age)
maic_outcome_model_4 <- coxph(Surv(time_AC, status_AC)~trt, robust=TRUE,
                              weights=weights_age)
exp(maic_outcome_model_4$coefficients[1])
```

```
##      trt
## 0.8769922
```

ACKNOWLEDGMENTS

Financial disclosure

Funding agreements ensure the authors' independence in writing and publishing this article.

Conflict of interest

The authors declare no potential conflicts of interest.

Data Availability Statement

The code required to reproduce the analyses is available in the Appendix.

References

1. Aouni J, Gaudel-Dedieu N, Sebastien B. Matching-adjusted indirect comparisons: Application to time-to-event data. *Statistics in Medicine* 2021; 40(3): 566–577.

2. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical Decision Making* 2018; 38(2): 200–211.

3. Phillippo D, Ades T, Dias S, Palmer S, Abrams KR, Welton N. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE. 2016.

4. VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology* 2007; 18(5): 561–568.

5. Pocock SJ. *Clinical trials: a practical approach*. John Wiley & Sons . 2013.

6. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. *Fundamentals of clinical trials*. Springer . 2015.

7. Weisberg HI, Hayden VC, Pontes VP. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality?. *Clinical trials* 2009; 6(2): 109–118.

8. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American journal of epidemiology* 2010; 172(1): 107–115.

9. Olsen RB, Orr LL, Bell SH, Stuart EA. External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management* 2013; 32(1): 107–121.

10. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International journal of epidemiology* 2016; 45(6): 2184–2193.

11. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. *Prevention Science* 2015; 16(3): 475–485.

12. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass.)* 2017; 28(4): 553.

13. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of clinical epidemiology* 1997; 50(6): 683–691.

14. Manski CF. Meta-analysis for medical decisions. 2019.

15. Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness* 2016; 9(1): 103–127.

16. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2011; 174(2): 369–386.

17. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2015; 178(3): 757–778.

18. Stuart EA, Rhodes A. Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation review* 2017; 41(4): 357–388.

19. Hong JL, Webster-Clark M, Jonsson Funk M, et al. Comparison of methods to generalize randomized clinical trial results without individual-level data for the target population. *American journal of epidemiology* 2019; 188(2): 426–437.

20. Brumback B, Berg A. On effect-measure modification: Relationships among changes in the relative risk, odds ratio, and risk difference. *Statistics in Medicine* 2008; 27(18): 3453–3465.

21. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine* 2013; 32(16): 2837–2849.

22. Remiro-Azócar A, Heath A, Baio G. Methods for Population Adjustment with Limited Access to Individual Patient Data: A Review and Simulation Study. *Research Synthesis Methods* 2021 (in press).

23. Remiro-Azócar A, Heath A, Baio G. Conflating marginal and conditional treatment effects: Comments on "Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study". *Statistics in Medicine* 2021; 40(11): 2753–2758.

24. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009; 20(6): 880–883.

25. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *American journal of epidemiology* 1980; 112(4): 467–470.

26. Simon R. Patient subsets and variation in therapeutic efficacy.. *British Journal of Clinical Pharmacology* 1982; 14(4): 473–482.

27. VanderWeele TJ. Principles of confounder selection. *European journal of epidemiology* 2019; 34(3): 211–219.

28. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *American journal of epidemiology* 1993; 138(11): 923–936.

29. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996; 58(1): 267–288.

30. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics* 2015; 24(3): 627–654.

31. Shah RD. Modelling interactions in high-dimensional data with backtracking. *The Journal of Machine Learning Research* 2016; 17(1): 7225–7255.

32. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a $1 \times k$ factorial design when time-to-failure is the outcome. *Controlled clinical trials* 1993; 14(6): 511–522.

33. Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?. *bmj* 2017; 356: j573.

34. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Statistics in medicine* 1983; 2(2): 243–251.

35. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. *Network meta-analysis for decision-making*. John Wiley & Sons . 2018.

36. Ferguson KD, McCann M, Katikireddi SV, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. *International journal of epidemiology* 2020; 49(1): 322–329.

37. Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics* 2017: 225–247.

38. Nguyen TQ, Ackerman B, Schmid I, Cole SR, Stuart EA. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PloS one* 2018; 13(12): e0208795.

39. Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression. *Journal of the American Statistical Association* 2009; 104(488): 1671–1681.

40. Wasserman L, Roeder K. High dimensional variable selection. *Annals of statistics* 2009; 37(5A): 2178.

41. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *International journal of epidemiology* 2020; 49(6): 2058–2064.

42. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 2018; 113(523): 1228–1242.

43. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology* 2021; 32(3): 393–401.

44. Team RC, others . R: A language and environment for statistical computing. 2013.

45. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine* 2005; 24(11): 1713–1723.

46. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in medicine* 2016; 35(30): 5642–5655.

47. Lesko CR, Lau B. Bias due to confounders for the exposure-competing risk relationship. *Epidemiology (Cambridge, Mass.)* 2017; 28(1): 20.