S1 APPENDIX. SIMULATION RESULTS COMPARING METHOD 1 AND METHOD 2 WHEN THE OUTCOME MODEL IS CORRECTLY OR INCORRECTLY SPECIFIED

(for the paper *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*)

The simulations here use the same setup as in [1]. In fact, results regarding bias have been reported in [1]. In the current simulations, we track not only bias, but also variance, RMSE, model estimated variance, and confidence interval coverage proportion.

Data generation. All the scenarios include a X variable, a Z variable and a V variable. X is standard normal. Z and V are first generated as multivariate normal with correlations ranging from 0 to ± 0.5, then each is either kept in continuous form or dichotomized. When either Z or V is binary, its prevalence is 0.25 in the trial sample and 0.5 in the target population. When either Z or V is continuous, it has mean 0 in the trial sample and 0.5 in the target population, and variance 1 in both. In the trial, A is randomly assigned to 0 and 1 with equal probability. With regards to the outcome, for the continuous Z and V combination, we use a base model with Z and V as effect modifiers, plus three other models, each with one additional effect modifier from among Z^2, V^2 or ZV:

$$\begin{array}{ll} A. & Y = X + A + Z + V + ZA + VA + \epsilon_Y, \\ B. & Y = X + A + Z + V + ZA + VA + Z^2A + \epsilon_Y, \\ C. & Y = X + A + Z + V + ZA + VA + V^2A + \epsilon_Y, \\ D. & Y = X + A + Z + V + ZA + VA + ZVA + \epsilon_Y, \end{array} \quad \epsilon_Y \sim \mathrm{N}(0, 4).$$

For the continuous Z and binary V combination, we use models A, B and D. For the binary Z and continuous V combination, we use A, C and D. For the binary Z and V combination, we use A and D. For each scenario (combining Z and V types and outcome model), we generate 100,000 pairs of datasets including an $n = 400$ trial sample and an $n = 5000$ target population sample.

Outcome model specification in method implementation. For both methods 1 and 2, in all scenarios we implement the method with the correct outcome model. For scenarios including Z^2, V^2 or ZV as effect modifiers, we also implement the methods with the misspecified outcome model that leaves out these terms and retains only Z and V as effect modifiers; this misspecified model is perhaps the most commonly encountered in practice.

For method 2, the weighting is with respect to X, Z using weights based on a logistic regression of sample membership. Continuous predictors are included using natural splines.

Results. The findings from these simulations are already summarized in the text of the paper. Here we include all the plots of the results, starting on the next page.

References

[1] Trang Quynh Nguyen, Cyrus Ebnesajjad, Stephen R. Cole, and Elizabeth A. Stuart. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Annals of Applied Statistics*, 11(1):225–247, 2017.

Figure 1: Bias

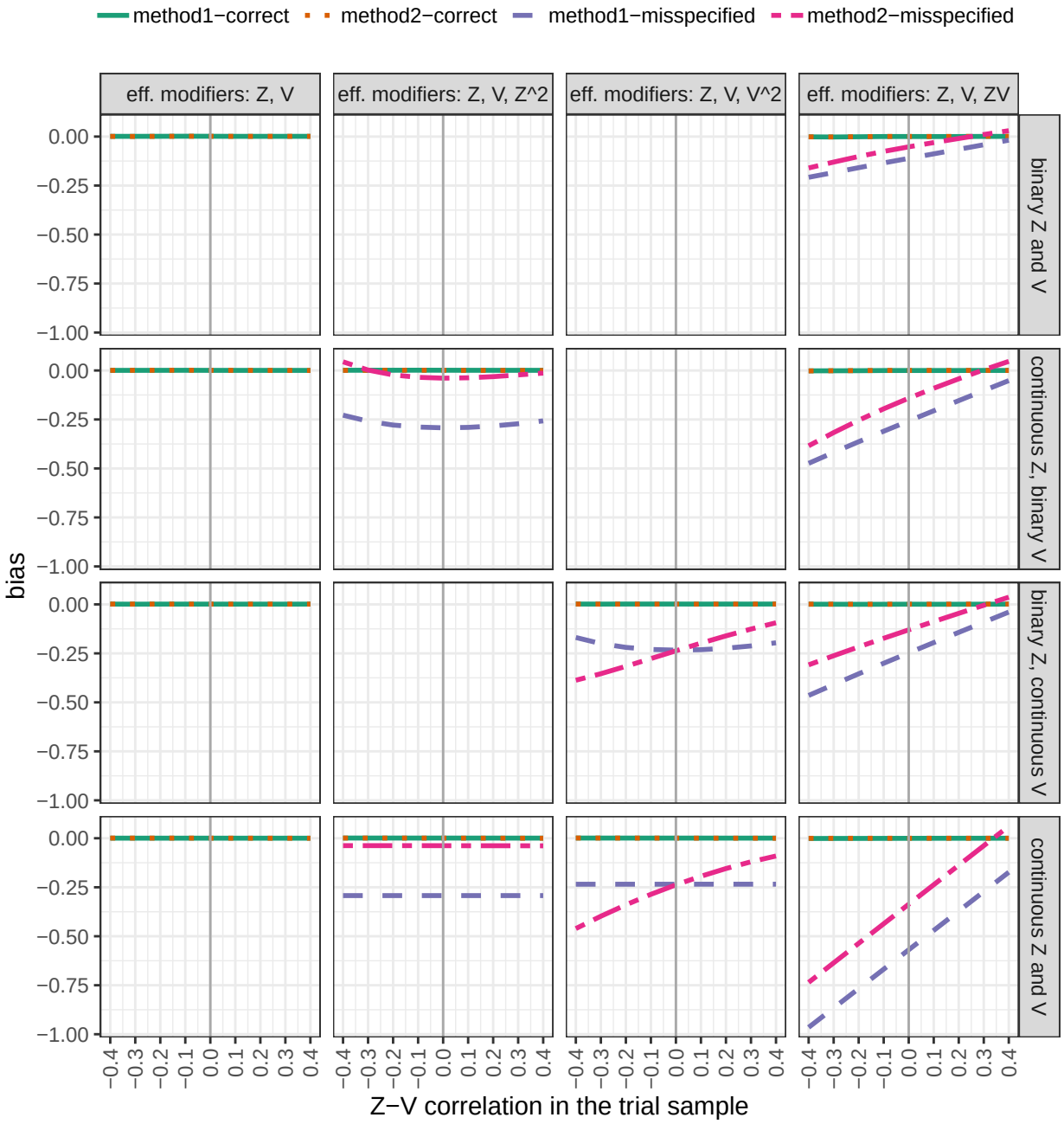


Figure 2: RMSE

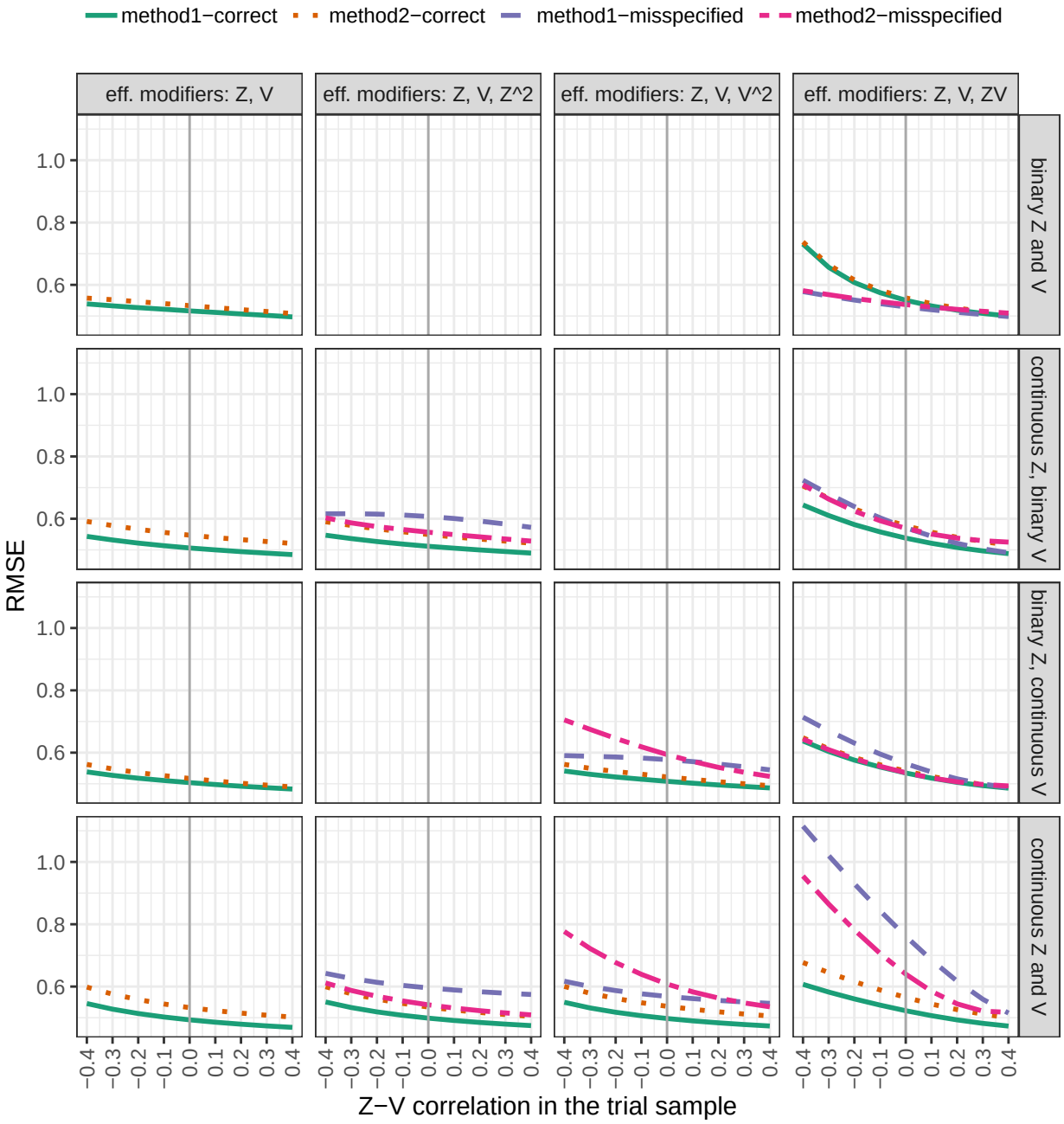


Figure 3: Standard Error

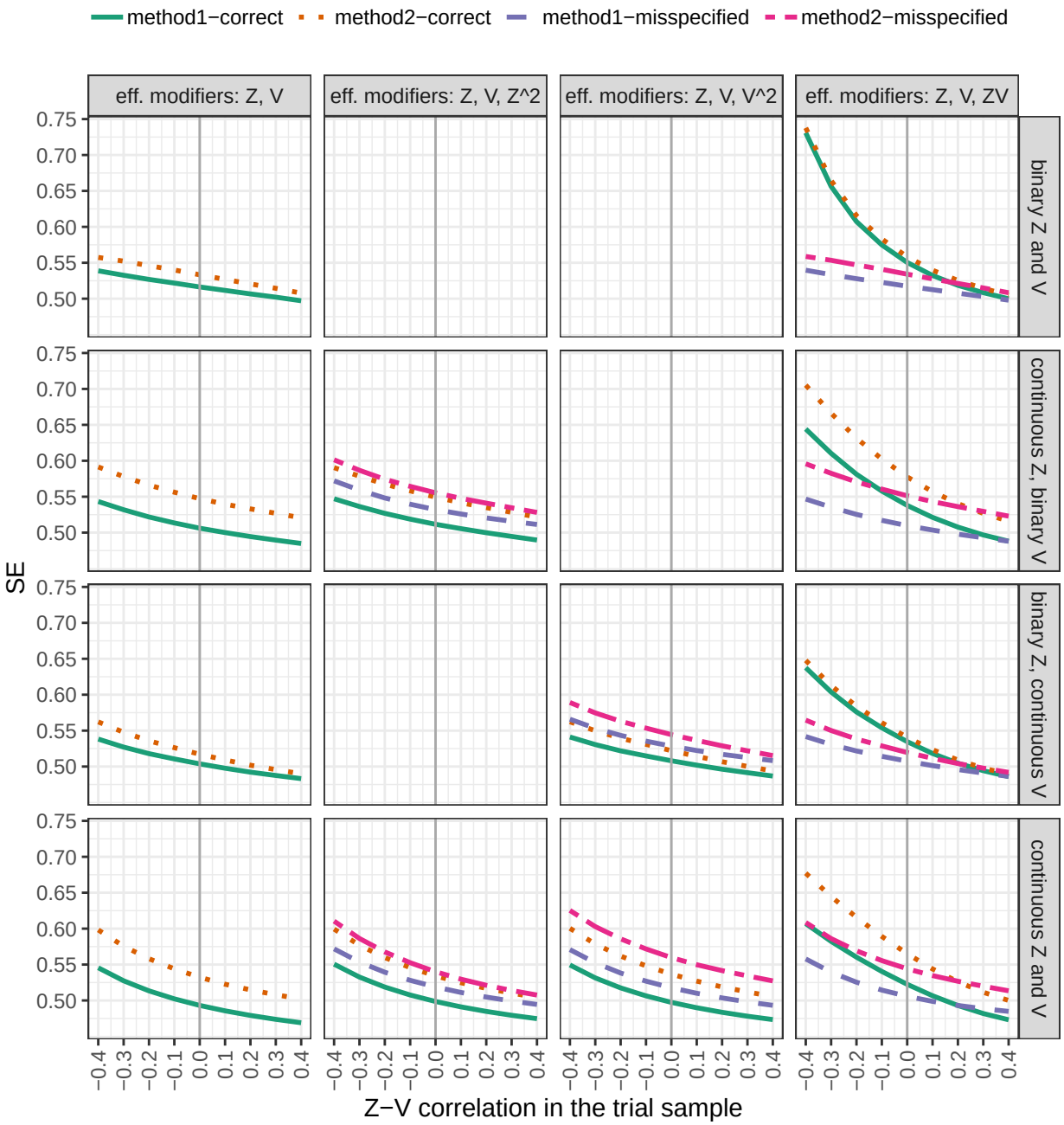


Figure 4: Model-estimated standardeError

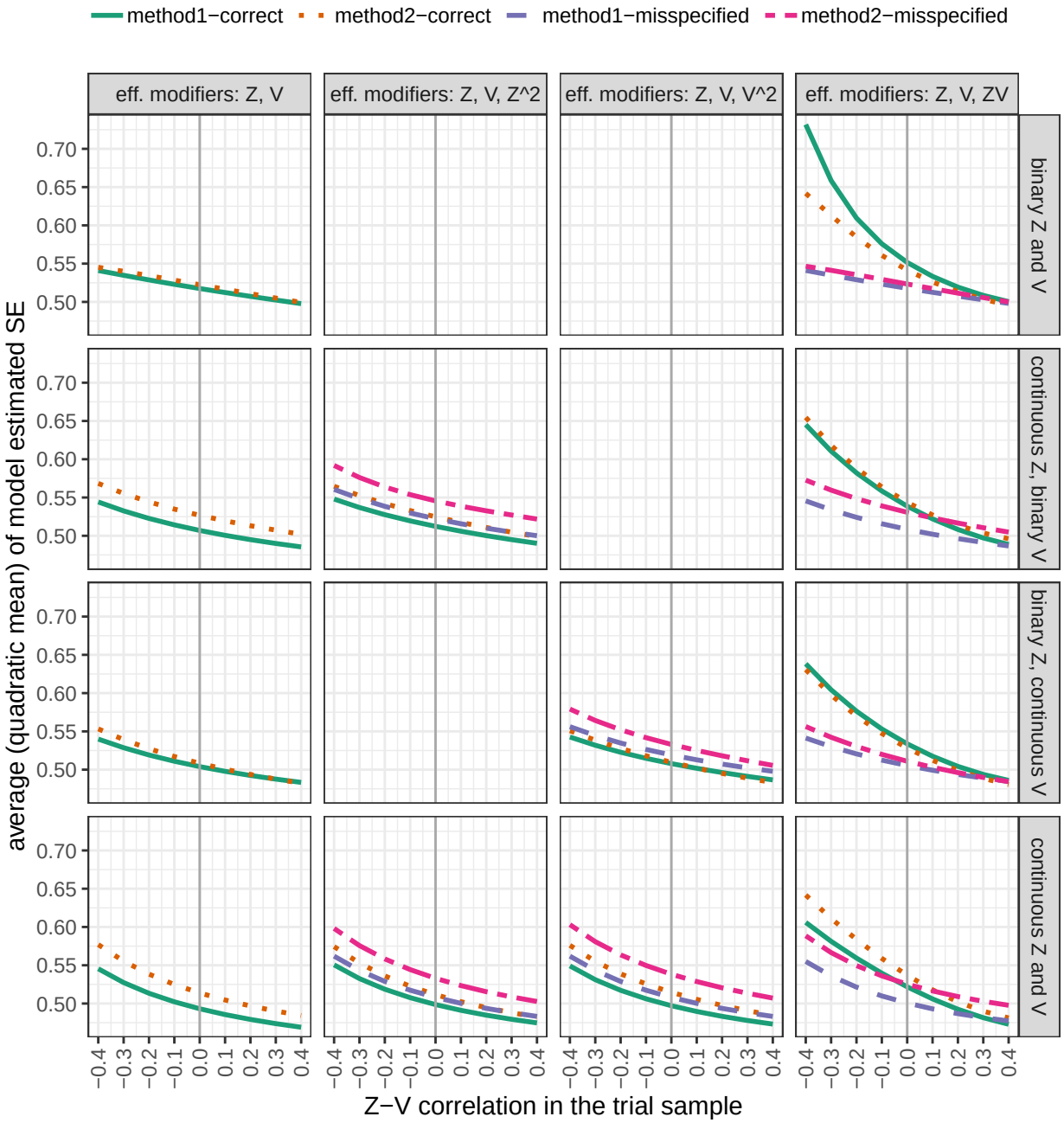
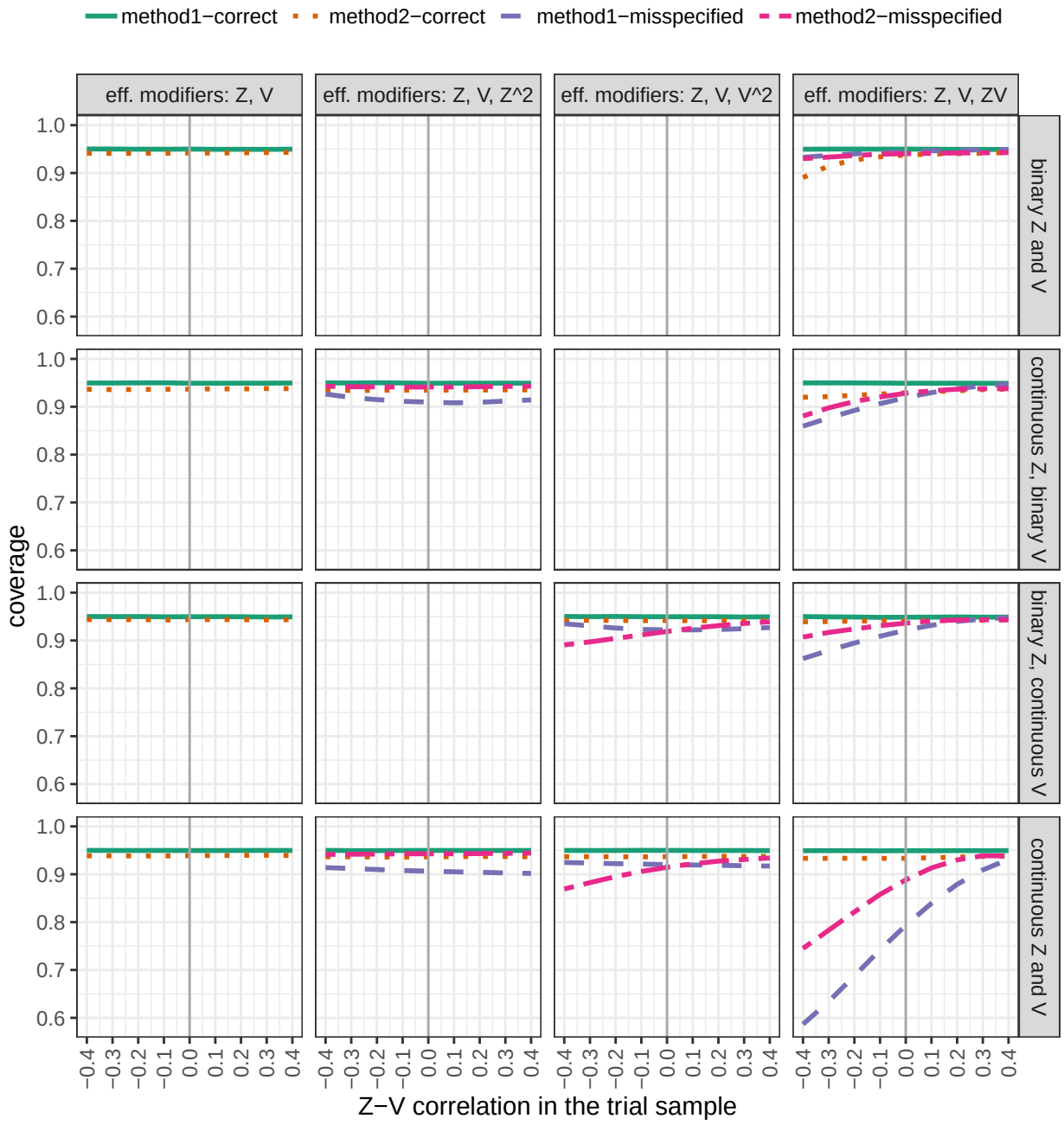


Figure 5: Coverage proportion of 95% confidence interval



S2 APPENDIX. A UNIFIED EXPLANATION OF WEIGHTING PROCEDURES FOR THE DIFFERENT DATA SCENARIOS

(for the paper *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*)

Method 2 presented in this paper call for weighting the trial sample so that it resembles the target population with respect to the distribution of the observed baseline covariates X, Z. Depending on the source of target population data – a full population ($P = 1$) dataset or a representative ($S = 2$) sample, and how it relates to the trial sample (see example scenarios in Fig. 1) – the specific weighting procedures vary. All these weighting procedures, however, relate to one idea we call "ratio-of-probability weighting". We borrow this term from [1], who used it in a different context (mediation analysis), but the term is appropriate for our current purpose. The idea is simple: to weight a sample E so that it resembles sample/population F with respect to the distribution of variables C, we use weights that are ratios of sample membership probabilities conditional on C, $W_i = \frac{\mathrm{P}(\mathrm{E}|C = C_i)}{\mathrm{P}(\mathrm{F}|C = C_i)}$. We now elaborate how this plays out in several data scenarios.

In scenario 1(b), the weights for the $S = 1$ sample to make it resemble the $S = 2$ sample with respect to baseline covariates X, Z are

$$W_i = \frac{\mathrm{P}(S = 2|X = X_i, Z = Z_i)}{\mathrm{P}(S = 1|X = X_i, Z = Z_i)}. \quad \text{(w1)}$$

With data only from these two samples (assumed to be disjoint), we estimate these weights by stacking the two datasets and fitting a model for sample membership S with X, Z as predictors (e.g., using logistic or another model deemed appropriate), and obtaining for each trial participant a weight that is the model-predicted odds of $S = 2$ vs. $S = 1$ given their X, Z values. Formally, this is an estimate of $\frac{\mathrm{P}\left(S = 2 \middle| X = X_i, Z = Z_i, (S = 1 \text{ or } S = 2)\right)}{\mathrm{P}\left(S = 1 \middle| X = X_i, Z = Z_i, (S = 1 \text{ or } S = 2)\right)}$, which is equivalent to (w1). This weighting-by-the-odds method [2,3] is analogous to the propensity score weighting version for estimating the average treatment effect on the treated, where control units are weighted by their predicted odds of being in the treatment vs. control condition [4].

In scenario 1(a), the weights for the $S = 1$ sample to make it resemble the $P = 1$ dataset with respect to baseline covariates X, Z are

$$W_i = \frac{\mathrm{P}(P = 1|X = X_i, Z = Z_i)}{\mathrm{P}(S = 1|X = X_i, Z = Z_i)}. \quad \text{(w2)}$$

If we know which units in the target population dataset are the specific units in the trial, we can fit to the target population dataset a model for trial participation with X, Z as predictors, and use inverse-trial-participation-probability weighting to weight the trial sample up to the population. The weights are estimates of $\frac{1}{\mathrm{P}(S = 1|X = X_i, Z = Z_i, P = 1)} = \frac{\mathrm{P}(P = 1|X = X_i, Z = Z_i, P = 1)}{\mathrm{P}(S = 1|X = X_i, Z = Z_i, P = 1)}$, which are equivalent to (w2). This weighting is analogous to inverse-probability-of-selection weighting in complex survey design [5]. If, on the other hand, the trial participants cannot be linked to their records in the population dataset, we can still estimate these weights by treating

the population dataset as an $S = 2$ dataset, stacking it with the trial dataset and using weighting-by-the-odds as in scenario 1(b). In this case, these artificial "odds" are estimates of $\frac{\text{P}(P = 1|X = X_i, Z = Z_i)/[\text{P}(P = 1|X = X_i, Z = Z_i)+\text{P}(S = 1|X = X_i, Z = Z_i)]}{\text{P}(S = 1|X = X_i, Z = Z_i)/[\text{P}(P = 1|X = X_i, Z = Z_i)+\text{P}(S = 1|X = X_i, Z = Z_i)]}$, which are equivalent to (w2). Figure 3: Standard Error

In a slightly different scenario where the trial sample has been drawn from within an observational sample that represents the target population, we proceed as in scenario 1(a), but treating the observational sample as if it were the population.

When the problem is one of *transportation*, where the trial sample is not part of the target population, the same weight formula (w2) or (w1) applies. Given a $P = 1$ or $S = 2$ dataset for the target population, we need to stack it with the $S = 1$ dataset, and use weighting-by-the-odds.

In the description of the sensitivity analyses, we mentioned that weighting-based sensitivity analyses are used only if a target population dataset (either $P = 1$ or $S = 2$) is available. To be precise, in a special case where there is no target population dataset, but information is available on the target population distribution of $\{X, Z\}$ (e.g., from a census or a prior population estimation exercise that reported these variables' joint distribution), weighting may also be implemented, using $W_i = \frac{\text{P}(X = X_i, Z = Z_i|P = 1)}{\text{P}(X = X_i, Z = Z_i|S = 1)}$, which are proportional to (w2). Here the numerator and denominator are the prevalences/densities of the $\{X_i, Z_i\}$ pattern in the target population and in the trial sample, respectively. This is only recommended for discrete $\{X, Z\}$ with a small number of combined categories, because beyond this situation, it is generally hard to estimate the denominator and the available estimates for the numerator may not be reliable.

To sum up, in most data scenarios where a dataset for/representing the target population is available, weighting the trial sample to make it resemble the target population involves data stacking and weighting-by-the-odds. The exception is when the trial sample is part of and can be identified within the population dataset, in which case inverse-probability weighting is used. If only summary statistics are available for the target population, the sensitivity analyses that involve weighting will generally not be used, except the very special case mentioned in the previous paragraph.

References:

1 Hong G. Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proc Am Stat Assoc Biometrics Sect* 2010;:2401–15.

2 Kern HL, Stuart EA, Hill JL, *et al.* Assessing methods for generalizing experimental impact estimates to target samples. *J Res Educ Eff* 2016;**9**:103–27. doi:10.1080/19345747.2015.1060282

3 Westreich D, Edwards JK, Lesko CR, *et al.* Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol*

4 Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 2003;**71**:1161–89. doi:10.1111/1468-0262.00442

5 Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952;**47**:663–85.

S3 APPENDIX. ADDITIONAL MATERIAL ON EFFECT MODIFIERS NOT OBSERVED IN THE TRIAL

(for the paper *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*)

Denote the effect modifier not observed in the trial by U. As stated in the text, U can be a specific variable (e.g., addiction severity) or a generic representation of unknown factors. Mimicking the V case, assume the causal model

$$\mathrm{E}[Y_i(a)] = \beta_o + \beta_a a + \beta_x X_i + \beta_z Z_i + \beta_{za} Z_i a + \beta_u U_i + \beta_{ua} U_i a.$$

The TATE formula copied from the V case (replacing V with U),

$$\mathrm{TATE} = \beta_a + \beta_{za}\mathrm{E}[Z|P=1] + \beta_{ua}\mathrm{E}[U|P=1],$$

is not helpful, as it requires estimates for $\beta_a, \beta_{za}, \beta_{ua}$, all of which are not identified from trial data because we do not observe U. Let's try to see if we can do something else that might work better. The following (essentially a bias formula) is obtained by comparing the formulas for TATE and SATE.

$$\mathrm{TATE} = \mathrm{SATE} + \beta_{za}\underbrace{\{\mathrm{E}[Z|P=1] - \mathrm{E}[Z|S=1]\}}_{\Delta_Z} + \beta_{ua}\underbrace{\{\mathrm{E}[U|P=1] - \mathrm{E}[U|S=1]\}}_{\Delta_U}.$$

Here SATE is identified from trial data. So is Δ_Z as Z is observed in both the trial and the target population. Suppose we are willing to treat β_{ua} (effect modification by U) and Δ_U (the difference in mean U between the target population and the trial) as sensitivity parameters for which we will specify ranges, as these two parameters are meaningful and somewhat imaginable. We are still stuck with an unidentified parameter, β_{za}.

If we use the weighting approach, and manage to equate mean Z between the trial and the target population, the second term in the formula vanishes, so we no longer have to deal with β_{za}. However, the other terms also change. Instead of SATE, we now can estimate a weighted ATE in the weighted trial sample, which is fine. The sensitivity parameter β_{ua} retains its meaning, so it does not pose a problem. However, in the place of $\mathrm{E}[U|S=1]$, we now have the weighted trial mean U, so instead of the Δ_U above, we now have the difference in mean U between the target population and the weighted trial sample, an obscure quantity that is not as imaginable and meaningful as Δ_U, so it is not suitable to serve as a sensitivity parameter.

The conclusion then is that the sensitivity analyses developed for the V case do not extend to the U case!

Correction of previously published results: Our previous paper [1] claimed that the methods do extend to the U case if we consider a special U that is the *remaining composite effect modifier after accounting for* Z, i.e., it captures all effect modification forces other than Z and it is independent of X, Z (intuitively it is a combination of all the remaining effect modifiers and X, Z have been "regressed out" of it), then due to this independence, a regression model without U fit to the trial sample can recover β_{za}, so the TATE formula above can be used. Also due to this independence, weighting based on X, Z does not change the distribution of U, so after weighting, we still have the simple Δ_U in the TATE formula, without having to deal with a weighted trial sample mean U that is different from the original trial sample mean U. This reasoning is flawed. Both parts of this reasoning hangs on the idea of a composite U independent of X, Z. The problem is with Z and U both differentially distributed between the trial sample and the target population (the motivating factor for sensitivity analysis for U), the association of Z and U is generally different between the trial sample and the target population due to collider bias when conditioning on sample membership. Thus independence of U and Z does not exists in both places. It is independence in the trial sample that would give the result of recovering β_{za} and weighting not changing the distribution of U, but it needs to be independence in the target population to make the notion of U meaningful as we are interested in the universe

that is the target population, not just one specific piece of it that is the trial sample. In addition, there is another flaw, that regressing X, Z out of U results in U being uncorrelated with X, Z, not independence. If we replace independence with uncorrelatedness, then we also lose the claim that weighting based on X, Z does not change the distribution of U.

References

[1] Trang Quynh Nguyen, Cyrus Ebnesajjad, Stephen R. Cole, and Elizabeth A. Stuart. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Annals of Applied Statistics*, 11(1):225–247, 2017.

S4 APPENDIX. THE SENSITIVITY ANALYSES WHEN RANDOM INTERCEPTS MODELS ARE USED

(for the paper *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*)

When the trial data are analyzed using random intercepts models, it is natural to define treatment effects as differences in pre-to-post change in outcome between treatment and control (or effect on 'potential outcome change'). It turns out that this is equivalent to effect on potential outcome post-treatment. Take any individual i. Let Y_{i1} denote individual i's pre-treatment outcome measure, and $Y_{i2}(a)$ denote individual i's potential outcome post-treatment if treatment is set to a. Denote potential outcome change under treatment a as $H_i(a)$. Then $H_i(a) = Y_{i2}(a) - Y_{i1}$. The individual treatment effect defined as difference in potential outcome change is formally $\mathrm{TE}_i = \mathrm{E}[H_i(1) - H_i(0)]$, which is equal to $\mathrm{E}[Y_{i2}(1) - Y_{i2}(0)]$.

In this case, the same sensitivity analyses apply, with minor modifications to the models fit and the estimates used. The explanation below references the simpler case where treatment effect is the difference in potential outcomes post-treatment that we used to describe the methods in the main text of the paper; for brevity, we will refer to that case as "the simple case".

To estimate the ATE from a trial with both baseline and post-treatment measures of the outcome, the simplest model that can be used is the model without covariates,

$$\mathrm{E}\left[Y_{ij}|F_{ij}, A_i\right] = c_{0i} + \gamma_0 + \gamma_a A_i + \gamma_f F_{ij} + \gamma_{fa} F_{ij} A_i$$

where i indexes the person, j indexes the observation (each person has two observations), T indicates treatment condition assigned, F indicates that the outcome is post- versus pre-treatment, and c_{0i} is the departure of individual i's intercept from the mean intercept γ_0. In this model, γ_f and $(\gamma_f + \gamma_{fa})$ respectively estimate the average pre-to-post change in outcome in the control group and in the treatment group; γ_{fa} estimates the difference between these two average changes, i.e., SATE. This estimator is analogous to estimating SATE using the difference in mean outcome between the two conditions in the simple case.

Note that in this model the stand-alone treatment term A is included just to allow the pre-treatment outcome to differ between the two treatment conditions. Its coefficient (γ_a) is usually small because due to randomization, its expectation is zero.

Another way to estimate SATE is fit a model that adjusts for baseline covariates but not letting the covariates interact with treatment; this is analogous to the regression of the outcome on treatment and covariates in the simple case. With X and Z, the model is

$$\mathrm{E}\left[Y_{ij}|F_{ij}, A_i, X_i, Z_i\right] = c_{0i} + (\gamma_0 + \gamma_x X_i + \gamma_z Z_i) + \gamma_a A_i + (\gamma_f + \gamma_{xf} X_i + \gamma_{zf} Z_i) F_{ij} + \gamma_{fa} F_{ij} A_i.$$

Note that this model includes the possibility that some baseline covariates may influence change in outcome that is not due to treatment, via the interaction terms of X and Z with F; this does not

mean they modify treatment effect. Treatment effect, again, is represented by γ_{pa} and is not allowed to vary as a function of baseline covariates, i.e., it is SATE. The model can be written in a more conventional form,

$$\mathrm{E}\left[Y_{ij}|F_{ij}, A_i, X_i, Z_i\right] = c_{0i} + \gamma_0 + \gamma_a A_i + \gamma_f F_{ij} + \gamma_{fa} F_{ij} A_i + \gamma_x X_i + \gamma_z Z_i + \gamma_{xf} X_i F_{ij} + \gamma_{zf} Z_i F_{ij}.$$

The calibration of TATE and the sensitivity analyses rely on a model that captures treatment effect heterogeneity. With X and Z, the potential outcomes model is

$$\mathrm{E}\left[Y_{ij}(a)\right] = b_{0i} + (\beta_0 + \beta_x X_i + \beta_z Z_i) + \beta_a A + \left(\beta_f + \beta_{xf} X_i + \beta_{zf} Z_i\right) F_{ij} + \left(\beta_{fa} + \beta_{zfa} Z_i\right) F_{ij} a,$$

in which treatment effect modification by Z is represented by β_{zfa}. Written in a more conventional form,

$$\mathrm{E}\left[Y_{ij}(a)\right] = b_{0i} + \beta_0 + \beta_a a + \beta_f F_{ij} + \beta_{fa} F_i a + \beta_x X_i + \beta_z Z_i + \beta_{xf} X_i F_{ij} + \beta_{zf} Z_i F_{ij} + \beta_{zfa} Z_i F_{ij} a.$$

The individual treatment effect has expectation $\beta_{fa} + \beta_{zfa} Z_i$, and $\mathrm{TATE} = \beta_{fa} + \beta_{zfa}\mathrm{E}[Z|P = 1]$. The same sensitivity analyses as in the simple case apply, with the following changes in the regression model and the TATE formula.

With effect modifier V observed in the trial but not the target population (and effect modifier Z observed in both samples), the effect modification regression model is

$$\begin{aligned}\mathrm{E}\left[Y_{ij}|F_{ij}, A_i, X_i, Z_i, V_i\right] &= b_{0i} + \beta_0 + \beta_a A_i + \beta_f F_{ij} + \beta_{fa} F_{ij} A_i + \beta_x X_i + \beta_z Z_i + \beta_v V_i + \beta_{xf} X_i F_{ij} + \beta_{zf} Z_i F_{ij} \\ &\quad + \beta_{vf} V_i F_{ij} + \beta_{zfa} Z_i F_{ij} A_i + \beta_{vfa} V_i F_{ij} A_i\end{aligned}$$

(interaction terms of X, Z, V variables with F may be removed if their coefficients are zero). The formula for TATE is

$$\mathrm{TATE} = \beta_{fa} + \beta_{zfa}\mathrm{E}[Z|P = 1] + \beta_{vfa}\mathrm{E}[V|P = 1].$$

S5 APPENDIX. ADDITIONAL DETAILS ON EXTENSION 2 FOR MULTIPLICATIVE EFFECTS AND LOG/LOGIT LINK MODELS

(for the paper *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*)

The main text of the paper presents the case of a binary outcome and an assumed logit causal model. The reasoning for other cases with log/logit link models and multiplicative effects is similar. While it is somewhat repetitive, for clarity, we list the models and effects for a couple of other cases.

Binary outcome, log-probability model, risk ratio scale effects

The causal model:

$$\log\{\Pr[Y_i(a)=1]\} = \beta_0 + \beta_a a + \beta_x X_i + \beta_z Z_i + \beta_{za} Z_i a + \beta_v V_i + \beta_{va} V_i a.$$

The regression model:

$$\log\{\Pr[Y=1|A,X,Z,V]\} = \beta_0 + \beta_a a + \beta_x X + \beta_z Z + \beta_{za} Z_i A + \beta_v V + \beta_{va} V A.$$

Individual effect on the risk ratio (RR) and log RR scales:

$$\begin{aligned}
\mathrm{TE}_i^{\mathrm{RR}} &:= \frac{\Pr[Y_i(1)=1]}{\Pr[Y_i(0)=1]} = \exp(\beta_a + \beta_{za} Z_i + \beta_{va} V_i), \\
\mathrm{TE}_i^{\log_\mathrm{RR}} &:= \log(\mathrm{TE}_i^{\mathrm{RR}}) = \beta_a + \beta_{za} Z_i + \beta_{va} V_i.
\end{aligned}$$

TATE as arithmetic mean of individual effects on the log RR scale and geometric mean of individual effects on the RR scale:

$$\begin{aligned}
\mathrm{TATE}^{\log_\mathrm{RR}} &= \beta_a + \beta_{za}\mathrm{E}[Z|P=1] + \beta_{va}\mathrm{E}[V|P=1], \\
\mathrm{TATE}^{\mathrm{RR}} &= \exp(\beta_a + \beta_{za}\mathrm{E}[Z|P=1] + \beta_{va}\mathrm{E}[V|P=1]).
\end{aligned}$$

Count outcome, log link model, mean/rate ratio scale effects

The causal model:

$$\log\{\mathrm{E}[Y_i(a)]\} = \beta_0 + \beta_a a + \beta_x X_i + \beta_z Z_i + \beta_{za} Z_i a + \beta_v V_i + \beta_{va} V_i a.$$

The regression model:

$$\log\{\mathrm{E}[Y|A,X,Z,V]\} = \beta_0 + \beta_a a + \beta_x X + \beta_z Z + \beta_{za} Z_i A + \beta_v V + \beta_{va} V A.$$

Individual effect on the mean ratio (MR) (or rate ratio) and log MR (or log rate ratio) scales:

$$\begin{aligned}
\mathrm{TE}_i^{\mathrm{MR}} &:= \frac{\mathrm{E}[Y_i(1)]}{\mathrm{E}[Y_i(0)]} = \exp(\beta_a + \beta_{za} Z_i + \beta_{va} V_i), \\
\mathrm{TE}_i^{\log_\mathrm{MR}} &:= \log(\mathrm{TE}_i^{\mathrm{MR}}) = \beta_a + \beta_{za} Z_i + \beta_{va} V_i.
\end{aligned}$$

TATE as arithmetic mean of individual effects on the log MR (or log rate ratio) scale and geometric mean of individual effects on the MR (or rate ratio) scale:

$$\begin{aligned}
\mathrm{TATE}^{\log_\mathrm{MR}} &= \beta_a + \beta_{za}\mathrm{E}[Z|P=1] + \beta_{va}\mathrm{E}[V|P=1], \\
\mathrm{TATE}^{\mathrm{MR}} &= \exp(\beta_a + \beta_{za}\mathrm{E}[Z|P=1] + \beta_{va}\mathrm{E}[V|P=1]).
\end{aligned}$$

Relating the average causal OR and the conditional OR estimated by logistic regression with main effects only

This discussion about average causal effects is more general than the specific case of the trial sample or target population in this paper. Therefore we drop the reference to the population/sample, and talk about the ATE in a generic way.

Before considering multiplicative effects, let's refer back to the case of additive effects based on a linear model. Clearly, the individual effects vary, as they depend on the individual's Z_i and V_i. We can fit a correct linear regression model (with A, X, Z, ZA, V, VA as predictors, predict the individual treatment effects using $\beta_a + \beta_{za} Z_i + \beta_{va} V_i$ and averaging those to estimate the ATE, which equals $\beta_a + \beta_{za} \mathrm{E}Z + \beta_{va} \mathrm{E}V$. On the other hand, if we fit a linear regression model with A, X, Z, V as predictors (the model with main effects only), then the regression coefficient γ_a of A in this model is equivalent to $\beta_a + \beta_{za} \mathrm{E}Z + \beta_{va} \mathrm{E}V$, which happens to be the ATE. This equivalence is a feature of linear models. Another way to think about this is that the coefficient of A in the model with main effects only estimates the effect of treatment on the outcome with a constraint that the treatment effect is the same for every individual. While for each individual, the estimate is off by some degree, on average, it is right, as it is equal to the average of the true individual effects.

That is, the linear regression model with main effects only is an unbiased estimate of the ATE – a fact that we already know and have used again and again in the paper for the estimation of SATE.

Now let's translate this reasoning to the OR case.

The average causal OR is the average (= geometric mean) of individual ORs which vary as they depend on Z_i, V_i. The logistic regression model with main effects only estimates treatment effects under a constraint (assumption) that treatment effects do not vary across individuals. Since some individuals have higher OR and some have lower, the estimate under this constraint is almost guaranteed to be off for the individuals, but reflects some sort of average over them. Like in the linear model case above, we can think of the OR estimated by this model as an estimate of the average of the individual effects, i.e., the average causal OR. However, it is only an approximate estimate because the model with main effects only has fewer predictors than the correct model with interaction effects, and with logistic regression dropping predictors leads to less variation in the outcome being explained, which tends to deflate the log OR; this is a problem with ORs called non-collapsibility. Therefore, the conditional OR estimated by logistic regression with main effects is in the spirit of estimating the average causal OR, but due to this reduction in variance explained, it tends to underestimate the average causal OR.