Estimating Causal Additive Interaction using Linear Odds-Based Methods

Yiliu Cao

Abstract

Causal additive interaction is a measure of synergistic or antagonistic effects between two or more exposures, with wide applications in epidemiologic research such as drug resistance. A common metric to quantify such interaction is the relative excess risk due to interaction (RERI), which measures the extent to which the joint effect of two exposures exceeds the sum of their individual effects. While the RERI is typically derived from the logistic regression by a non-linear transformation of the fitted odds ratio, it becomes problematic when the outcome is rare, as the estimated coefficients are unstable with high variance. This paper will revisit the linear odds model and extend it to the marginal structural linear odds model (MSLOM). Both methods yield closed-form expressions for marginal RERI and its standard errors. Through simulation studies, this paper will evaluate the performance of each model under various causal structures and confounding effects. The results show that the MSLOM consistently yields smaller bias when the outcome is rare. Besides, the higher the confounding effect, the better the performance of MSLOM.

1 Introduction

2 Framework

In the usual setting where we have a single binary treatment A, we use $P(Y = 1|A = a)$ to denote the probability of outcome under treatment a. However, in causal interactions, there are at least two exposures in casual interactions. We will only focus on two binary exposures, denoted as G and E. We will introduce causal interactions across different sections, as outlined below.

2.1 Measurements and Scales

Let Y denotes the binary outcome of interest, and let $\mathcal{T}$ denote the exposures of interest. In this paper, we assume there are only two binary exposures such that $\mathcal{T} = \{G, E\}$. Denote Y^t as the potential outcome of receiving $\mathcal{T} = t$. Under the assumption of consistency ($Y = Y^t$) and exchangebility ($T \perp Y^t$), $\mathbb{E}[Y^t] = \mathbb{E}[Y^t|G = g, E = e] = \mathbb{E}[Y|G = g, E = e] = p(Y = 1|G = g, E = e)$. Denote $p_{ge} = p(Y = 1|G = g, E = e)$ be the probability of outcome of receiving $G = g$ and $E = e$, then the risk difference, RD, is defined as

$$RD = (p_{11} - p_{00}) - [(p_{10} - p_{00}) + (p_{01} - p_{00})]$$
$$= p_{11} - p_{10} - p_{01} + p_{00}$$

where $p_{11} - p_{10} - p_{01} + p_{00}$ is referred as additive interaction. If $RD > 0$, the additive interaction is said to be "super-additive", and "sub-additive" when $RD < 0$. Correspondingly, risk ratios RR and odds ratio OR are defined as multiplicative interaction as $RR = \frac{RR_{11}}{RR_{10}RR_{01}}$, $OR = \frac{OR_{11}}{OR_{10}OR_{01}}$ where $RR_{ij} = p_{ij}/p_{00}$, $OR_{ij} = \frac{p_{ij}/(1-p_{ij})}{p_{00}/(1-p_{00})}$ [@vanderweelebook, p.254].

In addition, if we divide p_{00} from above, we have

$$\frac{1}{p_{00}}(p_{11} - p_{10} - p_{01} + p_{00}) = RR_{11} - RR_{10} - RR_{01} + 1$$

which is referred as "relative excess risk due to interaction" in relative risk scale or "$RERI_{RR}$". It can be used to conduct the same inference but use relative risks instead of risks. For example, $RERI_{RR} > 0$ if and only if $p_{11} - p_{10} - p_{01} + p_{00} > 0$. In addition, under the rare diseases, $OR_{ge} = \frac{p_{11}/(1-p_{11})}{p_{00}/(1-p_{00})} \approx \frac{p_{11}}{p_{00}} = RR_{ge}$, then

$$RERI_{RR} = RR_{11} - RR_{10} - RR_{01} + 1$$
$$\approx OR_{11} - OR_{10} - OR_{01} + 1 = RERI_{OR}$$

As noted above, additive interaction is often considered more important than multiplicative interaction. This is because the additive interaction incorporates the difference scale, which is particularly useful for assessing the impact of public health interventions [@rothman1980]. For instance, in a subgroup of the population, a larger RD indicates a greater number of individuals whose disease is prevented or cured. This provides more information for identifying the most effective interventions [@greenland2008]. Moreover, additive interaction is closer related to the test for mechanistic interaction than the multiplicative one [@greenland2008]. It can also provide us with information when detecting synergism between two component exposure in a sufficient cause framework [@rothman1976]. We will discuss this in more detail in Section 3.

2.2 Statistical and Causal Interaction

In practice, interactions between two variables are evaluated using statistical models by including a product term between them. For additive interaction, it can be estimated using generalized linear models (GLMs) by selecting an appropriate canonical link function g. Considering the simplest case where there is no covariates, the GLM is expressed as:

$$g(P(Y = 1|G = g, E = e)) = \eta = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 ge$$

The interpretation of coefficients varies across canonical link. If we choose the identity link, then it is clear that $\beta_0 = p_{00}$, $\beta_1 = p_{10} - p_{00}$, $\beta_2 = p_{01} - p_{00}$ and $\beta_3 = p_{11} - p_{10} - p_{01} + p_{00}$ which is our target estimate. Similarly, under log link, $e^{\beta_0} = p_{00}$, $e^{\beta_1} = RR_{10}$, $e^{\beta_2} = RR_{01}$, $e^{\beta_3} = RR_{11}/(RR_{10}RR_{01})$, and $RERI_{RR} = \exp(\beta_1 + \beta_2 + \beta_3) - \exp(\beta_1) - \exp(\beta_2) + 1$. In case of logit link, $e^{\beta_3} = OR_{11}/(OR_{10}OR_{01})$ and $\beta_3 = p_{11} - p_{10} - p_{01} + p_{00}$, respectively [@vanderweelebook, p.258], and $RERI_{OR} = \exp(\beta_1 + \beta_2 + \beta_3) - \exp(\beta_1) - \exp(\beta_2) + 1$.

However, it is often that we want to control the covariates by including the covaraites $\mathbf{X}$ in the function. The GLM now becomes

$$g(P(Y = 1|G = g, E = e, \mathbf{X} = \mathbf{x})) = \eta_{ge}(\mathbf{X}) = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 ge + \boldsymbol{\gamma}^\top \mathbf{x}$$

Comparing to the simplest case, the identity and log link does not work here as the probability may run out of boundry 0 and 1, especially when continuous covariates. The logit link is the optimal choice.

Notably, under the logit link, even though the risk is conditional on covaraites, β_3 is still interpreted as the (marginal) $RERI_{OR}$. Specifically,

$$OR_{ge} = \frac{\text{Odds}_{ge}}{\text{Odds}_{00}} = \frac{\exp(\eta_{ge}(\mathbf{X}))}{\exp(\eta_{00}(\mathbf{X}))} = \exp(\beta_1 g + \beta_2 e + \beta_3 ge)$$

and hence $RERI_{OR} = \exp(\beta_1 + \beta_2 + \beta_3) - \exp(\beta_1) - \exp(\beta_2) + 1$ does not depends on covariates and hence marginal. The proof can be found at Appendix.

However, even though the logistic regression can help us to find the estimates of $RERI_{OR}$, the estimates using maximum likelihood estimation (MLE) are unstable and less accurate when the outcome is rare, known as sparse data bias. Correspondingly, its variance will be huge as well. Mathematically, the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is

$$\text{Var}(\hat{\boldsymbol{\beta}}) \approx I(\boldsymbol{\beta})^{-1}$$
$$= (\mathbf{Z}^\top \mathcal{W} \mathbf{Z})^{-1}$$

where $\mathcal{W} = \text{diag}\{p_1(1 - p_1), \ldots, p_n(1 - p_n)\}$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top \in \mathbb{R}^{n \times p}$ such that $\mathbf{z}_i = (1, g_i, e_i, g_i e_i, \mathbf{x}^\top)^\top$. Under the rare outcome, $p_i(1 - p_i) \approx p_i$ will become small and hence make $\mathcal{W}^{-1}$ becoming very large. These will make the estimates of parameters to be unstable.

In addition, it is hard to conduct statistical inference of $RERI_{OR}$ by fitting the logistic regression. Even though we can easily get the estimates, its standard error is hard to compute. To compute its se, A suggests that we can use bootstrap to estimate the non-parametric se. In addition, B suggests that we can use delta methods.

As illustrated above, although the logistic regression is easy to fit and give us the estimates of $RERI_{OR}$ directly, it restricts that the outcome is not so rare and can not lead us to the standard error of estimated parameters. To solve this issue, this research will discuss an alternative method, called Linear Odds Model as well as its extensions, to fit the data, which has better performance than the traditional logistic regression in estimating $RERI_{OR}$ when the outcome is rare. Moreover, it can efficiently give us the standard errors. More detailed will be discussed in the next section.

However, using statistical interaction is far not enough to estimate the additive causal interaction. One significant issue is that the models are unconstrained with identity and log-linear links, with the predicted probabilities falling outside the valid range of $(0, 1)$ [@vanderweelebook, pp.258-259]. Additionally, these two links always turn into the convergence problem in MLE, especially when incorporating the covariates. This may also result in inaccurate estimates and increase the computational cost. Furthermore, as noted above, even though we can easily estimate OR using statistical interaction, OR only approximates RR when the outcome is rare. These all make estimating the additive causal interaction more complex than simply fitting a statistical interaction model.

Lastly, we want to determine whether the estimated effects reflect causal relationships rather than mere associations. This requires to consider the set of confounders for each exposure separately. Let's consider we have controlled the confounders of our primary exposure G but not control for the secondary exposure E. If the effect of G differs from the strata classified by E, then we conclude with the effect of the modification. Nevertheless, we can not claim that E is indeed the effect modifier as we do not know whether the effect of the modification is caused by E or some other factors relating to E [@vanderweele2007]. In contrast, if we control for two sets of confounding factors, then the coefficient of the interaction term can, therefore, be interpreted as the causal interaction as the observed interaction is not confounded by external factors that may independently influence the relationship between the exposures and the outcome [@vanderweele2009_interaction].

3   Methods

To solve the limitation of logistic regression, we provide three linear odds based methods. The details are shown below.

3.1   Linear Odds Model

We will firstly consider the simplest case where there are no covaraites. Recall in the traditional logistic regression, odds $= \exp\{\theta_0 + \theta_1 g + \theta_2 e + \theta_3 ge\}$, which models the odds for each observation via log-linear model. In contrast, the linear odds model directly model the odds as a linear function of exposures. By this model, the effects of both exposures are no longer linear on an exponential scale, which will have a better accomplishment on detecting the additive interaction we mentioned earlier. The foundamental linear odds model has the form

$$\text{odds}(Y^t) = f(t)^\top \boldsymbol{\beta}$$

where $f(t)$ is a vector function of exposures t, and $\boldsymbol{\beta}$ is a vector of parameters. Since we assume there are only two exposures, G and E, we have $f(t) = (1, g, e, ge)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$. Then the linear odds model can be also written as

$$\text{odds}(Y^t) = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 ge$$

To see how the linear odds model is linked to the logistic regression, recall the logistic regression odds $= \exp\{\theta_0 + \theta_1 g + \theta_2 e + \theta_3 ge\}$ and $G, E \in \{0, 1\}$, we can re-write it as

$$
\begin{aligned}
\text{odds} = {} & \exp\{\theta_0\} + \\
& (\exp\{\theta_0 + \theta_1\} - \exp\{\theta_0\})g + \\
& (\exp\{\theta_0 + \theta_2\} - \exp\{\theta_0\})e + \\
& (\exp\{\theta_0 + \theta_1 + \theta_2 + \theta_3\} - \exp\{\theta_0 + \theta_1\} - \exp\{\theta_0 + \theta_2\} + \exp\{\theta_0\})ge
\end{aligned}
$$

We can also parametrize the linear odds model as $\text{odds}(Y^t) = \exp\{\beta_0\}(1 + \beta_1 g + \beta_2 e + \beta_3 ge)$. That is

$$
\begin{aligned}
\text{odds} = {} & \exp\{\theta_0\}(1 + \\
& (\exp\{\theta_1\} - 1)g + \\
& (\exp\{\theta_2\} - 1)e + \\
& (\exp\{\theta_1 + \theta_2 + \theta_3\} - \exp\{\theta_1\} - \exp\{\theta_2\} + 1)ge
\end{aligned}
$$

which implies

$$
\begin{cases}
\beta_0 = \theta_0 \\
\beta_1 = \exp\{\theta_1\} - 1 \\
\beta_2 = \exp\{\theta_2\} - 1 \\
\beta_3 = \exp\{\theta_1 + \theta_2 + \theta_3\} - \exp\{\theta_1\} - \exp\{\theta_2\} + 1
\end{cases}
$$

Under the linear regression shown above, by substituting $G, E \in \{0, 1\}$, we have

$$
\begin{cases}
\text{odds}_{00} = \exp\{\beta_0\} \\
\text{odds}_{10} = \exp\{\beta_0\}(1 + \beta_1) \\
\text{odds}_{01} = \exp\{\beta_0\}(1 + \beta_2) \\
\text{odds}_{11} = \exp\{\beta_0\}(1 + \beta_1 + \beta_2 + \beta_3)
\end{cases}
$$

β_0 represents the log odds when both exposures are absent. However, β_1 and β_2 are now the excess odds ratio when each exposure is present, respectively. β_3 represents the $RERI_{OR}$ as

$$
\begin{aligned}
\text{RERI}_{OR} &= OR_{11} - OR_{10} - OR_{01} + 1 \\
&= 1 + \beta_1 + \beta_2 + \beta_3 - 1 - \beta_1 - 1 - \beta_2 + 1 \\
&= \beta_3
\end{aligned}
$$

Predictions of odds that might be negative is another problem we should consider for the linear odds model. When both exposures are present, the odds is negative if $\beta_1 + \beta_2 + \beta_3 + 1 < 0$. However, $\beta_1 + \beta_2 + \beta_3 = \exp\{\theta_1 + \theta_2 + \theta_3\}$ which is impossible to be negative. Similarly, if only either or neither of exposures are present, the predicted odds can not be less than 0, and hence the model is valid.

Comparing to the traditional logistic regression, the estimates of $RERI_{OR}$ from the linear odds model, β_3, is exactly the same as the one using the logistic regression model, $\exp\{\theta_1 + \theta_2 + \theta_3\} - \exp\{\theta_1\} - \exp\{\theta_2\} + 1$. However, since the latter one is a function of $\boldsymbol{\theta}$, the estimation of standard error needs to use delta method or non-parametric methods such as bootstrap. In addition, the variance of estimated parameters get higher under the rare outcome. By implementing the linear odds model, the asymptotic variance $RERI_{OR}$ can be computed as the inverse of fisher infomaiton matrix, $\mathcal{I}^{-1}(\boldsymbol{\beta})$. This provides a very convenient way when conducting the inference of $RERI_{OR}$. On the other hand, since the linear odds model does not incorporate any covariates into the model, it may have good performance when there are no confounders or weak confoudning effects but will introduce bias on β_3 when the confounders are present. To solve this drawback, this paper will also introduce the conditional linear odds model on the next section.

3.2 Marginal Structural Linear Odds Model

Instead of having a treatment-only logistic regression model, suppose the logistic regression includes that covariates such that odds $= \exp\{\beta_0 + \beta_1 g + \beta_2 e + \beta_3 ge + \boldsymbol{\gamma}^\top \mathbf{x}\}$. Comparing to the linear odds model, the marginal structural linear odds model is a weighting approach to estimate $RERI_{OR}$ when there are covariates. It has the exactly the same model form as linear odds model but adding weights by inverse probability of exposures to each observation.

For each individual i, the weight for the first exposure, called w_G, is computed by the inverse of the predicted probability that $G = 1$ given some covaraites $\mathbf{X}$. Similarly, the weights for the second exposure, w_E, is computed by the inverse of predicted probabilities of $E = 1$ given both exposure G and same covariates $\mathbf{X}$. The product of w_G and w_E gives the overall weight for that individual. Specifically, the weights are expressed as

$$w^i = \frac{1}{p(G = g_i, E = e_i | \mathbf{X} = \mathbf{x}_i)}$$
$$= \frac{1}{p(G = g_i | \mathbf{X} = \mathbf{x}_i)} \times \frac{1}{p(E = e_i | G = g_i, \mathbf{X} = \mathbf{x}_i)}$$

In addition to inverse probability of weights, there are two assumptions when implement MSLOM: rare outcome and unconfounding conditional on $\mathbf{X}$. Under the rare outcome, the odds for each individual, $\frac{p_i}{1-p_i}$, is now approximated as risks, p_i as $1 - p_i \approx = 1$. That is, the model becomes $p = \exp\{\beta_0\}(1 + \beta_1 g + \beta_2 e + \beta_3 ge)$. Therefore, the coefficients and confidence interval β_3 is approximately estimating the $RERI$ in risk ratio scale instead of odds ratio scale adjusted for the covariates, which is our desired quantity. Additionally, no confounding conditional on $\mathbf{X}$ is also crucial, as it will make the conditional exchangebility holds such that $Y^{ge} \perp \{G, E\} | \mathbf{X}$. Failure of conditional exchangebility will cause the estimation of $RERI_{OR}$ to be biased. However, even though the second assumption fails, the above procedure will still give the consistent estimates as $p(Y = 1 | G = g, E = e) = \sum_x p(Y = 1 | G = g, E = e, X = x)$. In the simulation session, we will not limit the $\mathbf{X}$ on the above two logistic regressions are same; instead, we will generalize the covariates to the confounder of each exposure, that is

$$w^i = \frac{1}{p(G = g_i | \mathbf{X}^G = \mathbf{x}_i^G)} \times \frac{1}{p(E = e_i | G = g_i, \mathbf{X}^E = \mathbf{x}_i^E)}$$

Comparing to the previous two, the score and information are a little bit different, which are

$$S(\boldsymbol{\beta}) = \mathbf{Z}^\top (\mathbf{y} - \boldsymbol{\mu})$$

where $\boldsymbol{\mu} = (p_1, \ldots, p_n)^\top$, and

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbf{Z}^\top \mathcal{W} \mathbf{Z}$$

where $\mathcal{W} = \text{diag}\{w_1 p_1 (1 - p_1), \ldots, w_n p_n (1 - p_n)\}$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top \in \mathbb{R}^{n \times p}$ such that $\mathbf{z}_i = (1, g_i/z_i, e_i/z_i, g_i e_i/z_i)^\top$.

3.3 Conditional Linear Odds Model

Comparing to the simplest linear odds model as shown above, the conditional linear odds model is originated from the logistic regression that includes the covariates, odds $= \exp\{\beta_0 + \beta_1 g + \beta_2 e + \beta_3 ge + \boldsymbol{\gamma}^\top \mathbf{x}\}$. Following the similar derivations above, the odds can be written as

$$\text{odds} = \exp\{\beta_0 + \boldsymbol{\gamma}^\top \mathbf{x}\}(1 +$$
$$(\exp\{\theta_1\} - 1)g +$$
$$(\exp\{\theta_2\} - 1)e +$$
$$(\exp\{\theta_1 + \theta_2 + \theta_3\} - \exp\{\theta_1\} - \exp\{\theta_2\} + 1)ge$$

then conditional linear odds model is specified as

$$\text{odds}_i = \frac{p_i}{1 - p_i} = \exp\{\beta_0 + \boldsymbol{\gamma}^\top \mathbf{x}\}(1 + \beta_1 g + \beta_2 e + \beta_3 ge)$$

5

where the expression of $\boldsymbol{\beta}$ are the same as above. Similar to th above, β_1 and β_2 are interpreted as the excess odds ratio when each exposure is present, and β_3 represents the $RERI_{OR}$.

3.4 Fitting the models

Unfortunately, there are not so many available packages in R that to fit the above three models. The only one available package is called `interactionR`, which However, since this package is relatively new and there are not so many literatures using this package, this paper will estimate $RERI_{OR}$ manually using Newton-Raphson by `optim()` function from package `stats` in R to estimate each parameter as well as their standard errors. Since the outcome follows a bernoulli distribution, the score function and fisher information can be derived as

$$S(\boldsymbol{\beta}) = \mathbf{Z}^\top (\mathbf{y} - \boldsymbol{\mu})$$
$$\mathcal{I}(\boldsymbol{\beta}) = \mathbf{Z}^\top \mathcal{W} \mathbf{Z}$$

where $\boldsymbol{\mu} = (p_1, \ldots, p_n)^\top$, $\mathcal{W} = \mathrm{diag}\{p_1(1 - p_1), \ldots, p_n(1 - p_n)\}$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top \in \mathbb{R}^{n \times p}$ such that $\mathbf{z}_i = (1, g_i/z_i, e_i/z_i, g_i e_i/z_i)^\top$.

Comparing to the linear odds model, the conditional linear odds model adds covariates directly to the regression model. Under the same likelihood function, the score function and information matrix can be written as

$$S(\boldsymbol{\beta}) = \widetilde{\mathbf{Z}}^\top (\mathbf{y} - \boldsymbol{\mu})$$
$$\mathcal{I}(\boldsymbol{\beta}) = \widetilde{\mathbf{Z}}^\top \mathcal{W} \widetilde{\mathbf{Z}}$$

where $\boldsymbol{\mu} = (p_1, \ldots, p_n)^\top$, $\mathcal{W} = \mathrm{diag}\{p_1(1 - p_1), \ldots, p_n(1 - p_n)\}$, $\widetilde{\mathbf{Z}} = (\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n)^\top \in \mathbb{R}^{n \times p}$ such that $\tilde{\mathbf{z}}_i = (1, g_i/z_i, e_i/z_i, g_i e_i/z_i, \mathbf{x}_i^\top)^\top$.

4 Simulation

There are not too many simulation studies on causal interactions; most literatures just simply apply the above methods to a real data set and draw empirical conclusions. The only simulation we saw is from the original paper of MSLOM, where they draw the simulation based on a case-control study of lung cancer. Specifically, from each of 1000 simulations they draw, they generated a large number of samples ($1e^6$) of covariates following the empirical distribution of the real data sets. The two treatments are generated by two logistic regression conditioning on covaraites, with coefficients chosen from the empirical distribution. Finally, the outcome is generated by logistic regression as well, conditional on both exposures and their interactions, and the main effects of all covariates. The number of case and control of final data for each simulation corresponds to the real data set, in their study, 1836 cases and 1452 controls. The results of estimated RERI is computed as the average of the 1000 simulations, as well as the wald-based confidence intervals. They also repeat the entire process under five different prevalence rates.

However, since their simulation design is deliberately mis-specified as their data-generation process follows the empirical distribution from the real data. Their conclusion of excellent performance under rare outcome but getting worse when prevalence rises may not be valid. More importantly, as we show later, the target of RERI from the above three methods are not quite the same. Simply comparing MSLOM and conditional linear odds model many not give the comprehensive conclusion. To address these limitations, we will conduct a better simulation design by generate the data explicitly. The goal of this simulation design is to show that the target of RERI from the above methods are not the same, and to test the performance each model with corresponding targets. In addition, this simulation will also extend to different scenarios such as confounding, prevalence rates, to see how the performance varies.

4.1 Simulation Metrics

As this simulation study has two main objectives on 1) quantifying the accuracy of the estimated RERI produced by different models, and 2) to evaluate how that accuracy changes across a range of outcome

prevalences and confounding structures. We will adopt below metrics to assess the performance of each method.

The accuracy is assessed by the bias from each model. To compute the bias of each method, the critical first step is to clarify the target estimand of each model. Although all models are modelling the odds, they differ in how the risk is defined. From the earlier discussion, the linear odds model and marginal-structural linear odds model (MSLOM) work with marginal risks, $p(Y = 1|G = g, E = e)$, whereas the conditional linear odds model is build on risks conditioning on $\mathbf{X}$, $p(Y = 1|G = g, E = e, \mathbf{X} = \mathbf{x})$. Due to the discrepancy on defining risks, the interpretations of odds are different as well. As a result, the β_3 in the former two models estimate the marginal RERI, $RERI^m$, while the conditional linear odds model estimates the conditional RERI, $RERI^c$.

Given the true data generation process, $\text{logit}(p(Y = 1 \mid G, E, \mathbf{X})) = \eta_{ge}(X) = \theta_0 + \theta_1 g + \theta_2 e + \theta_3 ge + \boldsymbol{\gamma}^\top \mathbf{x}$, $RERI^c$ and $RERI^m$ can be shown to be different. As discussed in the previous section, $RERI^c$ is derivated from conditional odds ratio and can be written as the closed-form constant, $e^{\theta_1+\theta_2+\theta_3} - e^{\theta_1} - e^{\theta_2} + 1$. Note that this expression does not involve any covariates explicitly; however, it is still considered a conditional estimand as it is based on conditional risks rather than marginal ones. In contrast, marginal risks are obtained by integrating over the distribution of covariates conditional on the exposure. Specifically,

$$p(Y = 1|G, E) = \mathbb{E}_{X|G,E}[p(Y = 1|G, E, X)]$$
$$= \mathbb{E}_{X|G,E}\left[\frac{\exp\{\eta_{ge}(X)\}}{1 + \exp\{\eta_{ge}(X)\}}\right]$$

and the corresponding marginal $odds_{ge}$ are given by

$$\text{odds}_{ge} = \frac{p^c_{ge}}{1 - p^c_{ge}}$$
$$= \frac{\mathbb{E}_{X|ge}\left[\frac{\exp\{\eta_{ge}(X)\}}{1+\exp\{\eta_{ge}(X)\}}\right]}{\mathbb{E}_{X|ge}\left[\frac{1}{1+\exp\{\eta_{ge}(X)\}}\right]}$$

Because these expectations are taken with respect to the conditional distribution of X given exposures, the marginal odds ratios generally differ from their conditional counterparts. This discrepancy implies that the $RERI^m$, as estimated by the linear odds model or MSLOM, does not equal to $RERI^c$. An example of OR_{11} is presented on the Appendix. However, since there is no closed-form expression for marginal RERI from the data-generation process, we will estimate the true marginal RERI by large number of simulation. More details will be discuss on the next session.

In addition to examining bias, this study evaluates the standard errors and confidence intervals associated with each estimator. Specifically, we report both the empirical standard error which is calculated as the sample standard deviation of the estimates across simulations, and the average model-based standard error, computed as the mean of the estimated standard errors from each simulated dataset. Based on the model-based standard error from each simulation, we assess the coverage probability of the 95% confidence intervals. Given the true conditional and marginal RERI, the coverage probability is defined as the proportion of simulations in which the confidence interval for the estimated RERI includes the corresponding true value. Let $\hat{\sigma}^{(b)}$ represent the estimated standard error in the bth simulation, and $\hat{\beta}_3^{(b)}$ the corresponding point estimate. The empirical coverage probability is then estimated as

$$\hat{\rho} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ \theta \in \left[\hat{\beta}_3^{(b)} \pm 1.96 \times \hat{\sigma}^{(b)} \right] \right\}$$

where $\theta \in \{\Phi, \phi\}$ depending on whether the marginal or conditional RERI is being evaluated, and B denotes the total number of simulation replications. The R codes can be found at Appendix.

4.2 Simulation Setup

Following VanderWheele's motivated simulation, we adapt a design rooted in multi-level treatment studies and matching algorithms for multiple exposures. Our primary objective is to understand how the bias of the estimated RERI changes as causal structures become more complex. We therefore begin with the simplest configuration, where there are confounders only. Following the ideas from variables section in propensity scores, we then extend the data-generating mechanism to allow each treatment and the outcome depending on distinct sets of covariates, to see which covaraites has influenced the estimation dominantly.
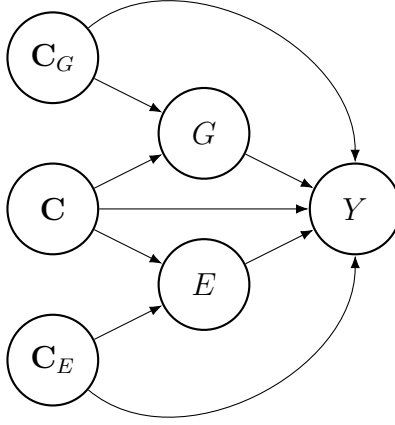
4.2.1 Confounder-only

Figure 1: Directed Acyclic Graphs for causal interactions under two treatments G and E.

Let G and E denote the two binary treatments and Y be the binary outcome. Let $\mathbf{C}_G$, $\mathbf{C}_E$ be the set of covariates that confound G and E individually, and let $\mathbf{C}$ be the set of common confounders to both treatments. The resulting causal Directed Acyclic Graphs are shwon in Figure 1. In this design, we set $\mathbf{C}_G = \{X_1, X_2\}$, $\mathbf{C}_E = \{X_3, X_4\}$ and $\mathbf{C} = \{X_5, X_6\}$. Suppose $\mathbf{X} = (X_1, \ldots, X_6)^\top$ is generated from a standard multivariate normal distribution such that $\mathbf{X} \sim \mathcal{N}_6(\mathbf{0}, \mathbf{I}_6)$.

Treatment assignment is generated in two steps. Firstly, we compute the propensity scores for each exposure:

$$\pi_G = p(G = 1 \mid \mathbf{Z}_G) = \text{logit}^{-1}\{f_G(\mathbf{Z}_G)\}$$
$$\pi_E = p(E = 1 \mid \mathbf{Z}_E) = \text{logit}^{-1}\{f_E(\mathbf{Z}_E)\}$$

where $\mathbf{Z}_G = \{X_1, X_2, X_5, X_6\}$ and $\mathbf{Z}_E = \{X_3, X_4, X_5, X_6\}$. To give X_5 and X_6 equal confounding effects on the two exposures and allow exposure-specific effects for the remaining covariates, we specify $f(\mathbf{X}_G) = 0.3X_1 + 0.4X_2 + 0.5X_5 + 0.5X_6$ and $f(\mathbf{X}_E) = -0.4X_3 + 0.3X_4 + 0.5X_5 + 0.5X_6$. In addition, as each covariate has mean zero, the marginal probabilities $p(G = 1)$ and $p(E = 1)$ are all closed to 0.5, which are less likely to have extreme weights. With the above setups, the probability of assignment for each combination of treatments is computed as

$$p_{00} = p(G = 0, E = 0) = (1 - \pi_G)(1 - \pi_E)$$
$$p_{10} = p(G = 1, E = 0) = \pi_G(1 - \pi_E)$$
$$p_{01} = p(G = 0, E = 1) = (1 - \pi_G)\pi_E$$
$$p_{11} = p(G = 1, E = 1) = \pi_G\pi_E$$

For each observation, the probability of receiving each pair of exposures then follows a multinomial distribution such that $T_i \mid \mathbf{X}_i \sim \text{Multinomial}(1; p_{00}, p_{10}, p_{01}, p_{11})$, where $T_i \in \{(0,0), (1,0), (0,1), (1,1)\}$. After assigning

the treatments, the binary outcome is generated as

$$\Pr(Y = 1 \mid \mathbf{X}, G, E) = \text{logit}^{-1}(\beta_0 +$$
$$0.4X_1 - 0.5X_2 + 0.7X_3 + 0.9X_4 - 0.2X_5 + 0.4X_6 +$$
$$0.3G + 0.4E + 0.8GE)$$

where β_0 is chosen from -6 to -1 by 0.1 to control the prevalence rate, with prevalence ranging from 1% to 60%. By this design, no matter what the prevalence rate is, the true conditional RERI is always the constant, which can be computed as $\exp\{0.3 + 0.4 + 0.8\} - \exp\{0.3\} - \exp\{0.4\} + 1 \approx 2.64$. 2.64 is hence the target value of conditional RERI. In contrast, the true value of marginal RERI is computed by the average of 2000 times simulation of 1×10^6 by compute the marginal RERI directly.

Given the true values of marginal RERI, the complete simulation runs as follows.

1. Generate the sample of $\{y_i, g_i, e_i, \mathbf{x}_i\}_{i=1}^n$ with sample size $n = 3000$, using the procedure described above.
2. Fit the three models, and estimated either marginal or conditional RERI.
3. Compute the confidence interval of each estimates using the theoretical standard errors, and check whether it includes the true value.
4. Repeat step 1 to 3 for 2000 times. Report the mean estimates and the bias, empirical standard errors, the average of theoretical errors, and the coverage (the proportion of simulations with confidence interval including the true value).
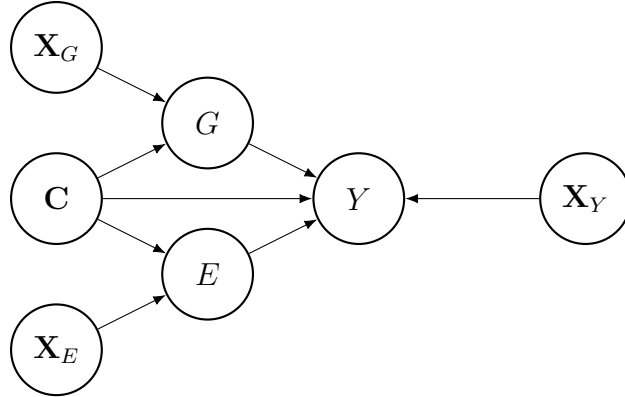
4.2.2 Scenario 2

Figure 2: Directed Acyclic Graphs for causal interactions under two treatments G and E.

In addition to the confounder-only design, we consider a second simulation scenario in which each exposure and the outcome have their own exclusive covariates. Specficially, let $\mathbf{X}_G$, $\mathbf{X}_E$ be the set of instrumental variables for the exposure G and E, respectively. Let $\mathbf{X}_Y$ be the outcome-only covariates, and $\mathbf{C}$ be the set of common confounders to both treatments. The causal DAG is shown in Figure 2. In this design, we set $\mathbf{X}_G = \{X_1, X_2\}$, $\mathbf{X}_E = \{X_3, X_4\}$, $\mathbf{X}_Y = \{X_5, X_6\}$ and $\mathbf{C} = \{X_7, X_8\}$. $\mathbf{X} = \{X_1, \dots, X_8\}$ is generated to follow a multivariate normal distribution such that $\mathbf{X} \sim \mathcal{N}_8(\mathbf{0}, \mathbf{I}_8)$. Propensity scores are obtained using both the exposure-specific instruments and the common confounders:

$$f(\mathbf{Z}_G) = 0.8X_1 + 0.5X_2 + 0.3X_7 + 0.5X_8$$
$$f(\mathbf{Z}_E) = -0.5X_3 + 0.2X_4 + 0.3X_7 + 0.5X_8$$

where $\mathbf{Z}_G = \{X_1, X_2, X_7, X_8\}$ and $\mathbf{Z}_E = \{X_3, X_4, X_7, X_8\}$. Conditional on $\mathbf{X}_i$, the probability of receiving each pair of treatment follows a multinomial distribution such that $T_i \mid \mathbf{X}_i \sim \text{Multinomial}(1; p_{00}, p_{10}, p_{01}, p_{11})$.

Beyond to the first design, this setting examines how the estimation of RERI varies to different confounding effects. We manipulate the confounding strength by varying the coefficients of X_7 and X_8 in the outcome

9

model while holding all other coefficients fixed. Let $\mathbf{Z}_Y = \{\mathbf{X}_Y, \mathbf{C}\} = \{X_5, X_6, X_7, X_8\}$, the outcome is generated as

$$\Pr(Y = 1 \mid \mathbf{Z}_Y, G, E) = \text{logit}^{-1}(\beta_0 +$$
$$0.8X_5 + 0.5X_6 + \gamma_7 X_7 + \gamma_8 X_8 +$$
$$0.3G + 0.4E + 0.8GE)$$

where (γ_7, γ_8) are set to $(-0.6, 0.4)$, $(-0.05, 0.05)$. The intercept β_0 is again tuned to control the overall prevalence. The true conditional RERI remains approximately 2.64 in this setup, whereas the marginal RERI is obtained via a large-scale simulation to approximate the population value.

4.3 Simulation Results

β_0	Model	β_3 ( )	$\widehat{SE}_{\beta_3}$	coverage	Abs. Bias	Bias(%)
$\beta_0 = -6$	Default LOM	3.948 (12.288)	4.326	0.968	1.362	52.67
	CLOM(all)	4.017 (9.252)	3.984	0.955	1.377	52.16
	MSLOM(corresp.)	3.502 (8.868)	3.078	0.946	0.916	35.42
	DR	4.148 (10.823)	2.440	0.734	1.508	57.12
$\beta_0 = -5.5$	Default LOM	2.748 (1.661)	1.533	0.964	0.274	11.08
	CLOM(all)	3.126 (2.415)	1.933	0.951	0.486	18.41
	MSLOM(corresp.)	2.713 (2.122)	1.767	0.948	0.239	9.66
	DR	3.262 (3.367)	1.054	0.693	0.622	23.56

5 Discussion