eAppendix for *Mediational E-values: Approximate Sensitivity Analysis for Unmeasured Mediator-Outcome Confounding*

CONTENTS

1 BACKGROUND

Motivation

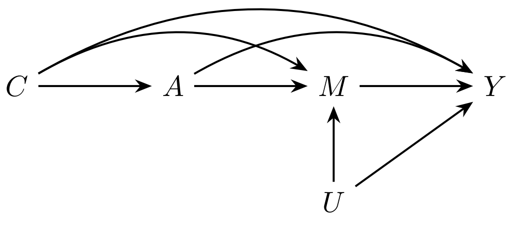
Estimates of mediated effects can be biased in the presence of unmeasured mediator-outcome confounding, even when the total effect can be estimated without bias. Sensitivity analysis allows researchers to assess the possible extent of the bias with some assumptions about the strength of the unmeasured confounding. Often assumptions are also made about the functional forms of the relationships between the variables or the type or number of unmeasured confounders.

A recent method for sensitivity analysis in this context requires none of the latter type of assumptions. Ding and VanderWeele showed that the magnitude of bias can be bounded with just two parameters that describe the strength of confounding.[1] It is intuitive to think about the magnitude of unmeasured mediator-outcome confounding in terms of the respective strengths of the relationships between confounder and outcome and confounder and mediator. However, this method requires the latter relationship to be defined by the association between confounder and exposure that is induced when conditioning on the mediator (i.e., collider bias). The magnitude of this association is likely not readily known or estimated, but previous work has suggested that it is similar in magnitude to either the effect of the confounder on the mediator or the exposure on the mediator.[1,2]

The purpose of this study was to investigate to what extent an alternative parameter could be used in Ding and VanderWeele's method to perform approximate instead of exact sensitivity analysis for natural direct and indirect effects.

Notation and definitions

We assumed a causal structure as represented in the following diagram. We use A to denote exposure, M mediator, and Y outcome. Measured confounders of any of the relationships are denoted C, and the unmeasured mediator-outcome confounder(s) U.



In the above graph, the total effect of A on Y is identifiable conditional on C, but direct and indirect effects are only identifiable if U is measured and also conditioned on. We assume $A \in \{0, 1\}$ but could also consider any two levels a and a^*. Similarly, U is assumed to be a single discrete random variable, but the results hold for any continuous U or vector-valued U. (See VanderWeele and Ding, 2016.[1])

We use counterfactual notation in which Y_{am} denotes the value of the outcome under exposure $A = a$ and mediator $M = m$, and M_a the value of the mediator under exposure $A = a$. The counterfactual $Y_{aM_{a^*}}$ denotes the value of the outcome under exposure $A = a$ and under the natural counterfactual value of the mediator under $A = a^*$.

For simplicity we omit C from the notation in this text; however, all statements hold conditional on measured C.

The natural direct effect on the risk ratio scale is defined as

$$\mathrm{NDE}_{\text{true}} = \frac{\mathrm{E}[Y_{1M_0}]}{\mathrm{E}[Y_{0M_0}]} ,$$

which, under identifiability assumptions, can be defined in terms of probabilities as

$$\mathrm{NDE}_{\text{true}} = \frac{\sum_u \sum_m \mathrm{P}(Y = 1 \mid A = 1, M = m, U = u)\mathrm{P}(M = m \mid A = 0, U = u)\mathrm{P}(U = u)}{\sum_u \sum_m \mathrm{P}(Y = 1 \mid A = 0, M = m, U = u)\mathrm{P}(M = m \mid A = 0, U = u)\mathrm{P}(U = u)} .$$

Similarly, the natural indirect effect on the risk ratio scale is defined as

$$\mathrm{NIE}_{\text{true}} = \frac{\mathrm{E}[Y_{1M_1}]}{\mathrm{E}[Y_{1M_0}]} ,$$

which we can express as

$$\mathrm{NIE}_{\text{true}} = \frac{\sum_u \sum_m \mathrm{P}(Y = 1 \mid A = 1, M = m, U = u)\mathrm{P}(M = m \mid A = 1, U = u)\mathrm{P}(U = u)}{\sum_u \sum_m \mathrm{P}(Y = 1 \mid A = 1, M = m, U = u)\mathrm{P}(M = m \mid A = 0, U = u)\mathrm{P}(U = u)} .$$

If we don't measure U, however, we don't directly observe all of these probabilities and cannot calculate the true direct and indirect effects. We therefore calculate $\mathrm{NDE}_{\text{obs}}$ as

$$\begin{aligned}
\mathrm{NDE}_{\text{obs}} &= \frac{\sum_m \mathrm{P}(Y = 1 \mid A = 1, m)\mathrm{P}(m \mid A = 0)}{\sum_m \mathrm{P}(Y = 1 \mid A = 0, m)\mathrm{P}(m \mid A = 0)} \\
&= \frac{\sum_m \left\{\sum_u \mathrm{P}(Y = 1 \mid A = 1, m, u)\mathrm{P}(u \mid A = 1, m)\right\} \left\{\sum_u \mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)\right\}}{\sum_m \left\{\sum_u \mathrm{P}(Y = 1 \mid A = 0, m, u)\mathrm{P}(u \mid A = 0, m)\right\} \left\{\sum_u \mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)\right\}} \\
&= \frac{\displaystyle\sum_m \Big\{\sum_u \mathrm{P}(Y = 1 \mid A = 1, m, u)\frac{\mathrm{P}(m \mid A = 1, u)\mathrm{P}(u)}{\sum_u \mathrm{P}(m \mid A = 1, u)\mathrm{P}(u)}\Big\}\Big\{\sum_u \mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)\Big\}}{\displaystyle\sum_m \Big\{\sum_u \mathrm{P}(Y = 1 \mid A = 0, m, u)\frac{\mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)}{\sum_u \mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)}\Big\}\Big\{\sum_u \mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)\Big\}} \\
&= \frac{\displaystyle\sum_m \Big\{\sum_u \mathrm{P}(Y = 1 \mid A = 1, m, u)\frac{\mathrm{P}(m \mid A = 1, u)\mathrm{P}(u)}{\sum_u \mathrm{P}(m \mid A = 1, u)\mathrm{P}(u)}\Big\}\Big\{\sum_u \mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)\Big\}}{\sum_m \sum_u \mathrm{P}(Y = 1 \mid A = 0, m, u)\mathrm{P}(m \mid A = 0, u)\mathrm{P}(u)} .
\end{aligned}$$

Similarly, we observe the NIE as

$$\begin{aligned}
\mathrm{NIE}_{\mathrm{obs}} &= \frac{\sum_m \mathrm{P}(Y=1 \mid A=1, m)\mathrm{P}(m \mid A=1)}{\sum_m \mathrm{P}(Y=1 \mid A=1, m)\mathrm{P}(m \mid A=0)} \\
&= \frac{\sum_m \left\{\sum_u \mathrm{P}(Y=1 \mid A=1, m, u)\mathrm{P}(u \mid A=1, m)\right\}\left\{\sum_u \mathrm{P}(m \mid A=1, u)\mathrm{P}(u)\right\}}{\sum_m \left\{\sum_u \mathrm{P}(Y=1 \mid A=1, m, u)\mathrm{P}(u \mid A=1, m)\right\}\left\{\sum_u \mathrm{P}(m \mid A=0, u)\mathrm{P}(u)\right\}} \\
&= \frac{\displaystyle\sum_m \Big\{\sum_u \mathrm{P}(Y=1 \mid A=1, m, u)\frac{\mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}{\sum_u \mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}\Big\}\Big\{\sum_u \mathrm{P}(m \mid A=1, u)\mathrm{P}(u)\Big\}}{\displaystyle\sum_m \Big\{\sum_u \mathrm{P}(Y=1 \mid A=1, m, u)\frac{\mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}{\sum_u \mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}\Big\}\Big\{\sum_u \mathrm{P}(m \mid A=0, u)\mathrm{P}(u)\Big\}} \\
&= \frac{\sum_m \sum_u \mathrm{P}(Y=1 \mid A=1, M=m, U=u)\mathrm{P}(M=m \mid A=1, U=u)\mathrm{P}(U=u)}{\displaystyle\sum_m \Big\{\sum_u \mathrm{P}(Y=1 \mid A=1, m, u)\frac{\mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}{\sum_u \mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}\Big\}\Big\{\sum_u \mathrm{P}(m \mid A=0, u)\mathrm{P}(u)\Big\}}\,.
\end{aligned}$$

Our assumption of the unbiasedness of the total effect is illustrated by the composition of the NIE and the NDE. Because we assume the unmeasured confounder is independent of the exposure, it doesn't bias the total effect, and both the true and the observed values of the mediation effects can be multiplied together to obtain the same total effect. The bias of the NDE and NIE is therefore the extent to which the observed value in the numerator of the NDE and the denominator of the NIE deviates from the true value. On the ratio scale, we can express the bias of these effects as

$$\begin{aligned}
\text{bias}_{\mathrm{RR}} &= \frac{\mathrm{NDE}_{\mathrm{obs}}}{\mathrm{NDE}_{\mathrm{true}}} = \frac{\mathrm{NIE}_{\mathrm{true}}}{\mathrm{NIE}_{\mathrm{obs}}} \\
&= \frac{\displaystyle\sum_m \Big\{\sum_u \mathrm{P}(Y=1 \mid A=1, m, u)\frac{\mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}{\sum_u \mathrm{P}(m \mid A=1, u)\mathrm{P}(u)}\Big\}\Big\{\sum_u \mathrm{P}(m \mid A=0, u)\mathrm{P}(u)\Big\}}{\sum_m \sum_u \mathrm{P}(Y=1 \mid A=1, m, u)\mathrm{P}(m \mid A=0, u)\mathrm{P}(u)}\,.
\end{aligned}$$

The NIE and NDE can also be calculated on the risk difference scale, where $\mathrm{NDE}_{\mathrm{true}} = \mathrm{E}[Y_{1M_0}] - \mathrm{E}[Y_{0M_0}]$ and $\mathrm{NIE}_{\mathrm{true}} = \mathrm{E}[Y_{1M_1}] - \mathrm{E}[Y_{1M_0}]$. (We use the same notation to refer to the concepts on both scales; in the rest of the text, context will make it clear whether we refer to the risk ratio or difference scale.) Expressions for the true and unobserved effects in terms of the joint distribution of M, Y, and U can be derived similarly. On this scale we are concerned with the absolute bias:

$$\text{bias}_{\mathrm{RD}} = \mathrm{NDE}_{\mathrm{obs}} - \mathrm{NDE}_{\mathrm{true}} = \mathrm{NIE}_{\mathrm{true}} - \mathrm{NDE}_{\mathrm{obs}}\,.$$

Results of Ding & VanderWeele, 2016

The bias due to unmeasured confounding, as defined above, was shown to have the sharp upper bound, which we refer to as the bounding factor ($\mathrm{BF}_{\mathrm{true}}$):

$$\text{bias}_{\text{ RR}} \leq \frac{\mathrm{RR}_{AU|M} \times \mathrm{RR}_{UY|(A=1,M)}}{\mathrm{RR}_{AU|M} + \mathrm{RR}_{UY|(A=1,M)} - 1} \; .$$

On the risk difference scale, this means that the true effects can be bounded as follows:

$$\mathrm{NDE}_{\mathrm{true}} \geq \sum_m \mathrm{P}(Y=1 \mid A=1,m)\mathrm{P}(M=m \mid A=0) \Big/ \mathrm{BF}_{\mathrm{true}} - \Pr(Y=1 \mid A=0)$$

and

$$\mathrm{NIE}_{\mathrm{true}} \geq \Pr(Y=1 \mid A=0) - \sum_m \mathrm{P}(Y=1 \mid A=1,m)\mathrm{P}(M=m \mid A=0) \Big/ \mathrm{BF}_{\mathrm{true}} \; .$$

The bounding factor depends on two values. The first is straightforward:

$$\mathrm{RR}_{UY|(A=1,M)} = \max_m \mathrm{RR}_{UY|(A=1,m)} = \max_m \frac{\max_u \mathrm{P}(Y=1 \mid A=1,m,u)}{\min_u \mathrm{P}(Y=1 \mid A=1,m,u)} \; .$$

The second value in the bounding factor is the one we are interested in finding an alternative for, due to its unintuitive nature:

$$\mathrm{RR}_{AU|M} = \max_m \max_u \frac{\mathrm{P}(u \mid A=1,m)}{\mathrm{P}(u \mid A=0,m)} \; .$$

As with the rest of the text, the above parameters are also conditional on measured C if such confounders have indeed been included in the analysis.

Since the bounding factor is guaranteed to be greater than 1, it is of little use when the bias is less than 1; that is, when $\mathrm{NIE}_{\mathrm{obs}} > \mathrm{NIE}_{\mathrm{true}}$ and $\mathrm{NDE}_{\mathrm{obs}} < \mathrm{NDE}_{\mathrm{true}}$. In those cases, $\mathrm{RR}_{AU|M}$ can be replaced by the reciprocal of the minimum risk ratio relating A and U, conditional on M; this is equivalent to recoding the exposure.

It was further shown that, for the bias to be of large enough magnitude to explain away one of $\mathrm{NIE}_{\mathrm{obs}}$ or $\mathrm{NDE}_{\mathrm{obs}}$ (that is, the true value is 1), $\mathrm{RR}_{UY|(A=1,M)}$ and $\mathrm{RR}_{AU|M}$ must be at least as great as $\mathrm{RR} + \sqrt{\mathrm{RR} \times (\mathrm{RR} - 1)}$, where RR refers to $\mathrm{NIE}_{\mathrm{obs}}$ or $\mathrm{NDE}_{\mathrm{obs}}$. This is analogous to the E-value for unmeasured confounding,[3] and we refer to it as the mediational E-value. If the bounding factor successfully bounds the bias using some other parameter than $\mathrm{RR}_{AU|M}$, then the mediational E-value can also be used to describe the necessary magnitude of that alternative parameter as well.

Note that if any such bounding factor successfully bounds the bias on the risk ratio scale, it will also do so on the risk difference scale, and vice versa.

Possible alternative parameters for the bounding factor

We considered a number of other parameters as possible alternatives for $\mathrm{RR}_{AU|M}$. If an alternative parameter is greater than $\mathrm{RR}_{AU|M}$, then the bounding factor constructed with that parameter will also be greater than the true bounding factor, and will therefore bound the bias. In a similar setting but on the odds ratio scale, Greenland showed that the magnitude of the bias due to conditioning on M tended to be smaller than the conditional relationships between A and M given U and between U and M given A.[2] This result led us to investigate the feasibility of using in the bounding factor risk ratios relating A and M, both marginally and conditionally on U, as well as risk ratios relating U and M. For both the $A-M$ and $U-M$ values, we chose the maximum of four subsets: first, the marginal risk ratios; second, the marginal risk ratios for $M=1$ (i.e., excluding those relating probabilities for $M=0$ in different strata); third, the stratified risk ratios; fourth, the stratified risk ratios for $M=1$.

$$\begin{aligned}
\mathrm{RR}_{AM} &= \max_m \frac{\max_a \mathrm{P}(M=m \mid A=a)}{\min_a \mathrm{P}(M=m \mid A=a)} \\
\mathrm{RR}_{AM=1} &= \frac{\max_a \mathrm{P}(M=1 \mid A=a)}{\min_a \mathrm{P}(M=1 \mid A=a)} \\
\mathrm{RR}_{AM|U} &= \max_m \max_u \frac{\max_a \mathrm{P}(M=m \mid A=a, U=u)}{\min_a \mathrm{P}(M=m \mid A=a, U=u)} \\
\mathrm{RR}_{AM=1|U} &= \max_u \frac{\max_a \mathrm{P}(M=1 \mid A=a, U=u)}{\min_a \mathrm{P}(M=1 \mid A=a, U=u)} \\
\mathrm{RR}_{UM} &= \max_m \frac{\max_u \mathrm{P}(M=m \mid U=u)}{\min_u \mathrm{P}(M=m \mid U=u)} \\
\mathrm{RR}_{UM=1} &= \frac{\max_u \mathrm{P}(M=1 \mid U=u)}{\min_u \mathrm{P}(M=1 \mid U=u)} \\
\mathrm{RR}_{UM|A} &= \max_m \max_a \frac{\max_u \mathrm{P}(M=m \mid A=a, U=u)}{\min_u \mathrm{P}(M=m|A=a, U=u)} \\
\mathrm{RR}_{UM=1|A} &= \max_a \frac{\max_u \mathrm{P}(M=1 \mid A=a, U=u)}{\min_u \mathrm{P}(M=1 \mid A=a, U=u)}
\end{aligned}$$

Though we considered these eight options, our focus was on $\mathrm{RR}_{UM=1|A}$ as it intuitively describes the strength of confounding via the relationship between U and M, with obvious analogy to $\mathrm{RR}_{UY|(A=1,M)}$. If the bounding factor using $\mathrm{RR}_{UM=1|A}$ and $\mathrm{RR}_{UY|(A=1,M)}$ bounds the bias, we can use the mediational E-value to describe the relative increase in risk of the mediator comparing any two values of the unmeasured confounder, within strata of exposure (and measured confounders), that would be necessary to explain away $\mathrm{NIE}_{\mathrm{obs}}$ or $\mathrm{NDE}_{\mathrm{obs}}$ (along with an equivalent increase in risk of the outcome).

In the case that M is not binary, the $\mathrm{RR}_{UM|A}$ parameter could be used instead. However, we chose to focus on $\mathrm{RR}_{UM=1|A}$ as a more conservative (as well as intuitive) choice; any time a bounding factor that uses it bounds the bias, one using $\mathrm{RR}_{UM|A}$ would as well.

2 METHODS

Data generation

We generated the necessary conditional probabilities of binary M, Y, and U to calculate the NIE_{true}, NDE_{true}, and the bias of their observed values (see Section 6 for a list of these probabilities). As in the text above, the joint distribution of the variables can be further assumed to be conditional on measured confounders C. For our main analysis we drew each of these probabilities independently from a Uniform(0,1) distribution. We then used these probabilities to compute the NIE_{true}, NDE_{true}, NIE_{obs}, NDE_{obs}, and the bias on both the risk ratio and risk difference scales. We repeated this 5,000,000 times. We refer to the data generated in this way as probabilities drawn from uniform distribution with no interaction restriction.

Other data-generating distributions

Allowing the probabilities to vary over any range results in many unrealistic situations, so we considered several possibly more plausible restrictions.

Because the direction of the bias due to unmeasured confounding is influenced by the direction of the various effects (see Section 4), we subsetted the data to those scenarios for which the direction of effects was constant across strata. That is, we restricted the data to the $\approx 11\%$ with no qualitative interaction by U of the $A-M$ relationship, by A of the $U-M$ relationship, by U of the $M-Y$ relationship, or by M of the $U-Y$ relationship. We refer to this data as probabilities drawn from uniform distribution, restricted to no qualitative interaction.

Next, we redrew probabilities assuming a log-linear model for M with no interaction term. This corresponds to the following model:

$$\log\big(\mathrm{P}(M=1 \mid A=a, U=u)\big) = \beta_0 + \beta_1 a + \beta_2 u \ .$$

We forced the probabilities for M to fit the model by drawing three out of the four necessary values $\mathrm{P}(M = 1 \mid A = a, U = u)$ from a Uniform(0,1) and then calculating the fourth based on the relationship

$$\frac{\mathrm{P}(M=1 \mid A=1, U=0)}{\mathrm{P}(M=1 \mid A=0, U=0)} = \frac{\mathrm{P}(M=1 \mid A=1, U=1)}{\mathrm{P}(M=1 \mid A=0, U=1)} \ .$$

Which of the four probabilities was to be missing was chosen randomly, and a set of probabilities was removed if any $\mathrm{P}(M = 1 \mid A = a, U = u) > 1$. We repeated this process so that we would have approximately 5,000,000 sets of probabilities. We then computed the true and observed effects and the bias, as above. We refer to this data as that with a log-linear model for M.

As with the original data, we also separately restricted the dataset to the $\approx 33\%$ with no qualitative interaction of any of the relationships (no interaction of U and A on M is implied by the model). We refer to this data as that with a log-linear model for M, restricted to no qualitative interaction.

Comparisons

For each of the samples in the simulation, we computed the bias of the risk ratio as $\mathrm{NDE}_{\mathrm{obs}}/\mathrm{NDE}_{\mathrm{true}}$. We know the bounding factor $\mathrm{BF}_{\mathrm{true}}$ is greater than this value; we wished to see what proportion of the alternative bounding factors were also greater than this bias. To do so we then computed the bounding factor with the true $\mathrm{RR}_{AU|M}$ parameter and each of the possible alternatives. The true $\mathrm{RR}_{UY|(A=1,M)}$ was used in this calculation for both the true and alternative bounding factors.

Finally, we compared how well the alternative bounding factors worked, evaluating in each case whether the bounding factor was greater than the bias (which we inverted if less than 1), or, on the risk difference scale, whether the correction with the bounding factor was greater than the absolute value of the bias. We also assessed whether each of the alternative RRs was greater than $\mathrm{RR}_{AU|M}$ (which would guarantee that the bounding factor too was greater than $\mathrm{BF}_{\mathrm{true}}$).

3 RESULTS

Overall results

As expected, there was no alternative parameter that could be used to create a BF that provided a bound to the bias in all scenarios. However, the bound failed in only a small proportion of cases for several of the alternative parameters (see eTable 1 for exact proportions), and under certain conditions for the data-generating distribution, there were two that led to a true bound (see Section 4 for some analytic results).

Results using $\mathrm{RR}_{UM=1|A}$ in bounding factor

Although it did not consistently bound the bias, the BF constructed using $\mathrm{RR}_{UM=1|A}$ ($\mathrm{BF}_{UM=1|A}$) is of particular interest, as it more intuitively corresponds to the magnitude of the unmeasured confounding. eTable 1 shows that the bound failed in only 0.65% of cases (0.46% in situations with no qualitative interaction). The number of failures was slightly higher among data that conformed to a log-linear model for M, at 2.18% overall and 2.11% when restricted to no qualitative interactions of U and M on Y.

eFigure 1 shows the distribution of the true and alternative bounding factors in relation to the magnitude of the bias. On average, the alternative bound was larger than the true on average, and the discrepancy appeared larger when the joint distribution of the variables did not allow for qualitative interaction. However, the direction of that relationship was not consistent; the alternative bound was of course smaller than the true bound in the cases in which it was also smaller than the bias. There did not seem to be a particular magnitude of bias at which the alternative bounding factor was more or less likely to fail under any of the data-generating distributions.

That the strength of the bound was similar when constructed with the alternative parameter compared to the true parameter is also demonstrated by eFigure 2. Overall, $\mathrm{BF}_{\mathrm{true}}$ and $\mathrm{BF}_{UM=1|A}$ corrected the mediated effects to a similar extent on both the risk ratio scale. Ideally, the ratio of the corrected to true effect would be 1; for bias of small magnitudes, $\mathrm{BF}_{UM=1|A}$ tended to be weaker, so that the correction went too far. When the bias was larger, it was more likely that the correction was not sufficient (i.e., it did not always successfully bound the bias). The latter was more likely to be the case when M was generated according to a log-linear model.

As shown in eTable 2, undercorrection (failure) by $\mathrm{BF}_{UM=1|A}$ occurred less than 1% of the time when the bias was less than 1.5 (on the ratio scale), less than 5% of the time when the bias was between 1.5 and 2, and less than 10% of the time when the bias was greater than 2. On the additive scale (eTable 3), this corresponded to less than 1% of the time when the bias was greater than 0.1.

When M was generated according to a log-linear model, undercorrection occurred slightly more frequently and at smaller magnitudes of bias. On the risk difference scale, this meant that up to 10% of the time the bound failed when the bias was greater than 0.1. In all scenarios, when $\mathrm{BF}_{UM=1|A}$ did succeed in bounding the bias, it was slightly more likely to be weaker than $\mathrm{BF}_{\mathrm{true}}$, resulting in greater overcorrection.

Even when $\mathrm{BF}_{UM=1|A}$ failed, most of the time correcting with it left little residual bias (eFigure 3). In each of the analyses, 80% of the time the bias remaining after correction was less than 1.2 (on the risk ratio scale). Over 60% of failures resulted in residual bias of less than 1.1.

Similarly, on the risk difference scale, where the true mediated effect was subtracted from that corrected by $\mathrm{BF}_{UM=1|A}$, the residual bias was minor. In situations with no qualitative interaction, over 70% of the time the remaining bias was less than 0.05 on the difference scale when the bound failed, and over 90% of the time it was less than 0.1 (eTable 4).

True bounds using alternative parameters under certain conditions

Under somewhat mild conditions, both $\mathrm{RR}_{UM|A}$ and $\mathrm{RR}_{AM|U}$ can be shown to be greater than $\mathrm{RR}_{AU|M}$. Since the BF is monotonically increasing in both RR parameters, replacing $\mathrm{RR}_{AU|M}$ with either $\mathrm{RR}_{UM|A}$ or $\mathrm{RR}_{AM|U}$ can in those cases result in an adequate, if weaker, bound.

We can rewrite

$$\begin{aligned}\mathrm{RR}_{AU|M} &= \max_m \max_u \frac{\mathrm{P}(u \mid A=1,m)}{\mathrm{P}(u \mid A=0,m)} \\ &= \max_m \max_u \left\{\frac{\mathrm{P}(m \mid A=1,u)}{\mathrm{P}(m \mid A=0,u)}\right\} \Bigg/ \left\{\frac{\mathrm{P}(m \mid A=1)}{\mathrm{P}(m \mid A=0)}\right\} .\end{aligned}$$

For the BF with $\mathrm{RR}_{UM|A}$ to always work, we need to show that

$$\max_m \max_u \left\{\frac{\mathrm{P}(m \mid A=1,u)}{\mathrm{P}(m \mid A=0,u)}\right\} \Bigg/ \left\{\frac{\mathrm{P}(m \mid A=1)}{\mathrm{P}(m \mid A=0)}\right\} \leq \frac{\mathrm{P}(M=m' \mid A=a', U=u')}{\mathrm{P}(M=m' \mid A=a', U=u'')}$$

for some m', a', u', and u''.

Assume that m and u indicate the values at which the left-hand side is maximized. In the case of the binary U, letting u^* denote the other value of U, which occurs with probability $1-\mathrm{P}(u)$, we can rewrite the expression:

$$\begin{aligned}&\left\{\frac{\mathrm{P}(M=m \mid A=0, U=u)}{\mathrm{P}(M=m \mid A=0)}\right\} \Bigg/ \left\{\frac{\mathrm{P}(M=m \mid A=1)}{\mathrm{P}(M=m \mid A=1, U=u)}\right\} \\ &= \left\{\frac{\mathrm{P}(u)\mathrm{P}(M=m \mid A=0, U=u) + (1-\mathrm{P}(u))\mathrm{P}(M=m \mid A=0, U=u^*)}{\mathrm{P}(M=m \mid A=0, U=u)}\right\} \Bigg/ \\ &\qquad \left\{\frac{\mathrm{P}(u)\mathrm{P}(M=m \mid A=1, U=u) + (1-\mathrm{P}(u))\mathrm{P}(M=m \mid A=1, U=u^*)}{\mathrm{P}(M=m \mid A=1, U=u)}\right\} \\ &= \left\{\mathrm{P}(u) + (1-\mathrm{P}(u))\frac{\mathrm{P}(M=m \mid A=0, U=u^*)}{\mathrm{P}(M=m \mid A=0, U=u)}\right\} \Bigg/ \\ &\qquad \left\{\mathrm{P}(u) + (1-\mathrm{P}(u))\frac{\mathrm{P}(M=m \mid A=1, U=u^*)}{\mathrm{P}(M=m \mid A=1, U=u)}\right\} .\end{aligned}$$

Because we have said that the values m and u are those which maximize this expression, this implies that

$$\frac{\mathrm{P}(M=m \mid A=0, U=u^*)}{\mathrm{P}(M=m \mid A=0, U=u)} \geq \frac{\mathrm{P}(M=m \mid A=1, U=u^*)}{\mathrm{P}(M=m \mid A=1, U=u)} .$$

Denote with RR_0 and RR_1 the values $\frac{\text{P}(m|A=a,U=u^*)}{\text{P}(m|A=a,U=u)}$ for $A = 0, 1$, and RR^* the value of $\text{RR}_{UM|A}$. We are therefore trying to show that

$$\frac{\text{P}(u) + (1 - \text{P}(u))\text{RR}_0}{\text{P}(u) + (1 - \text{P}(u))\text{RR}_1} \leq \text{RR}^* \ .$$

We can rewrite that expression

$$\text{P}(u) + (1 - \text{P}(u))\text{RR}_0 \leq \big(\text{P}(u) + (1 - \text{P}(u))\text{RR}_1\big)\text{RR}^*$$

so that

$$\text{P}(u)(1 - \text{RR}^*) + (1 - \text{P}(u))\left(\text{RR}_0 - \text{RR}_1\text{RR}^*\right) \leq 0 \ .$$

We can guarantee that the first term is not positive because there is clearly some $\text{RR}_{UM|A} > 1$. In order for the second term to be as well, so that their sum ≤ 0, we need that $\text{RR}_0 \leq \text{RR}_1\text{RR}^*$.

In the original notation, we can write that expression

$$\frac{\text{P}(M = m' \mid A = a', U = u')}{\text{P}(M = m' \mid A = a', U = u'')} \geq \frac{\dfrac{\text{P}(M = m \mid A = 0, U = u^*)}{\text{P}(M = m \mid A = 0, U = u)}}{\dfrac{\text{P}(M = m \mid A = 1, U = u^*)}{\text{P}(M = m \mid A = 1, U = u)}}$$

The condition required for this to hold is that the direction of the effect of $\text{RR}_{UM|A}$ is the same across values of A; that is, that there is no qualitative interaction for the effect of U on M by A. In that case, we know that the risk ratios in the numerator and the denominator of the right-hand side are either both > 1 or both < 1. In the case where they are both > 1, take the left-hand side to be the numerator for the inequality to hold. In the case where they are both < 1, take the left-hand side to be the inverse of the denominator. Under the condition of no qualitative interaction by A, there is always a value of $\text{RR}_{UM|A}$ such that $\text{RR}_{UM|A} > \text{RR}_{AU|M}$.

Similarly, such a value of $\text{RR}_{AM|U}$ is guaranteed when there is no qualitative interaction by U on the effect of A on M. We can see this by ordering

$$1 \leq \frac{\text{P}(M = m \mid A = 1, U = u^*)}{\text{P}(M = m \mid A = 0, U = u^*)} \leq \frac{\text{P}(M = m \mid A = 1)}{\text{P}(M = m \mid A = 0)} \leq \frac{\text{P}(M = m \mid A = 1, U = u)}{\text{P}(M = m \mid A = 0, U = u)}$$

or

$$1 \geq \frac{\text{P}(M = m \mid A = 1, U = u)}{\text{P}(M = m \mid A = 0, U = u)} \geq \frac{\text{P}(M = m \mid A = 1)}{\text{P}(M = m \mid A = 0)} \geq \frac{\text{P}(M = m \mid A = 1, U = u^*)}{\text{P}(M = m \mid A = 0, U = u^*)} \ .$$

One of these must hold if the no qualitative interaction condition holds, and the values m and u are those which maximize

$$\text{RR}_{AU|M} = \max_m \max_u \left\{\frac{\text{P}(m \mid A = 1, u)}{\text{P}(m \mid A = 0, u)}\right\} \Bigg/ \left\{\frac{\text{P}(m \mid A = 1)}{\text{P}(m \mid A = 0)}\right\} \ .$$

If the numerator and denominator are both ≥ 1, clearly $\mathrm{RR}_{AM|U} = \frac{\mathrm{P}(M=m|A=1,U=u)}{\mathrm{P}(M=m|A=0,U=u)}$ satisfies $\mathrm{RR}_{AM|U} \geq \mathrm{RR}_{AU|M}$. If the numerator and denominator are both ≤ 1, then $\frac{\mathrm{P}(M=m|A=0,U=u^*)}{\mathrm{P}(M=m|A=1,U=u^*)}$ can be chosen so that $\mathrm{RR}_{AM|U} \geq \mathrm{RR}_{AU|M}$.

In conclusion, replacing $\mathrm{RR}_{AU|M}$ in the bounding factor with $\mathrm{RR}_{UM|A}$ will be guaranteed to bound the bias if the direction of the U-M effect is consistent across levels of A, and the same is true with $\mathrm{RR}_{AM|U}$ as a replacement if the A-M effect is consistent across levels of M. However, both of these parameters are less intuitive with binary M than $\mathrm{RR}_{UM=1|A}$.

Direction of confounding

Unlike with a simple exposure-outcome relationship, the direction in which the NIE and the NDE are biased due to unmeasured confounding of a mediator-outcome relationship is not necessarily intuitive.

However, we can look at the directions of the $A-M$, $U-M$, and $U-Y$ relationships to predict the direction of the bias. For simplicity, we will consider only situations in which the directions of the effects are constant across strata of the other variables (qualitative interactions are generally more unpredictable in terms of direction). If the $A-M$ relationship is a “positive” relationship (that is, $\mathrm{P}(M=1 \mid A=1, U=u) > \mathrm{P}(M=1 \mid A=0, U=u)$), the $\mathrm{NIE}_{\mathrm{true}}$ is generally overestimated (on an absolute scale – that is, $\mathrm{NIE}_{\mathrm{obs}} > \mathrm{NIE}_{\mathrm{true}}$) when the direction of the $U-M$ and $U-Y$ effects are in the same direction (either both positive or both negative) and underestimated with the two U effects are in opposite directions. If the $A-M$ relationship is negative and the two U effects are in the same direction, the $\mathrm{NIE}_{\mathrm{true}}$ is generally underestimated; it tends to be overestimated if they are in different directions.

However, because the bias depends on the probabilities for $M=0$ as well as those for $M=1$, in order for the bias to be guaranteed to go in the expected direction, we also must require that the “interaction” terms for both $M=1$ and $M=0$ are in the same direction. In other words, if

$$\frac{\mathrm{P}(M=1 \mid U=1, A=1)}{\mathrm{P}(M=1 \mid U=0, A=1)} < \frac{\mathrm{P}(M=1 \mid U=1, A=0)}{\mathrm{P}(M=1 \mid U=0, A=0)}$$

then we need that

$$\frac{\mathrm{P}(M=0 \mid U=1, A=1)}{\mathrm{P}(M=0 \mid U=0, A=1)} < \frac{\mathrm{P}(M=0 \mid U=1, A=0)}{\mathrm{P}(M=0 \mid U=0, A=0)}$$

or vice versa with respect to the inequalities. We could also write the requirement as:

$$\text{if } RR_{UM|A=1} > RR_{UM|A=0}\text{ , then } RR_{U\bar{M}|A=1} > RR_{U\bar{M}|A=0}$$

and

$$\text{if } RR_{UM|A=1} < RR_{UM|A=0}\text{ , then } RR_{U\bar{M}|A=1} < RR_{U\bar{M}|A=0}$$

where $RR_{U\bar{M}|A=a}$ refers to the risk ratio of U on $M = 0$ for a given value of A. If neither of the two conditions holds, the bias is likely but not guaranteed to be in the expected direction.

However, if the $M = 0$ relationship is in the opposite direction as the one that guarantees the confounding direction as expected, then the BF using $\text{RR}_{UM|A=1}$ is guaranteed to hold.

For a given $U - M$ direction of effect, only one of the sets of inequalities is possible. If the $RR_{U\bar{M}|A=a}$ is in the opposite direction as the one that guarantees the expected direction, even though the direction of the bias is not guaranteed, the $\text{BF}_{UM=1|A}$ is guaranteed to hold.

Whether the over- or underestimation of the NIE_{true} results in a bias toward or away from the null value of 1 depends of course on whether $\text{NIE}_{\text{true}} < 1$ or $\text{NIE}_{\text{true}} > 1$. The direction of NIE_{true} depends on the $A - M$ relationship and the $M - Y$ relationships; if they are in the same direction, $\text{NIE}_{\text{true}} > 1$, and if different directions, $\text{NIE}_{\text{true}} < 1$.

Continuous M and U

We have considered binary M and U in the rest of the text and in the numerical analysis, as coarsened versions of variables are often measured and analyzed in epidemiology and because it is often easier to conceptualize an unmeasured and hypothetical U as the variable that distinguishes between the highest and lowest risk groups. We will show that when these variables are continuous, the bound may be less useful, and furthermore that a bounding factor with an alternative parameter is trivially equal to that constructed from the true $\text{RR}_{AU|M}$ parameter.

Let

$$\begin{bmatrix} M \\ U \end{bmatrix} \sim \text{Multivariate Normal} \left(\begin{bmatrix} A \\ 0 \end{bmatrix}, \begin{bmatrix} \rho & 1 \\ 1 & \rho \end{bmatrix} \right)$$

Then

$$U \mid m \sim \text{Normal}\left(\rho(m - A), 1 - \rho^2\right)$$

and

$$M \mid u \sim \text{Normal}\left(A + \rho u, 1 - \rho^2\right) \ .$$

Because M and U are continuous, for the $\text{RR}_{AU|M}$ and $\text{RR}_{UM|A}$ parameters we consider the generalized relative risks

$$\text{RR}_{AU|M} = \max_m \max_u \frac{\text{P}(U \leq u \mid A = 1, m)}{\text{P}(U \leq u \mid A = 0, m)}$$

and

$$\mathrm{RR}_{UM|A} = \max_m \max_a \frac{\max_u \mathrm{P}(M \leq m \mid A = a, U = u)}{\min_u \mathrm{P}(M \leq m | A = a, U = u)} \ .$$

In each case, the RR parameters are ratios of cumulative distribution functions of normal random variables with different means but the same variance. In the numerator of $\mathrm{RR}_{AU|M}$ the mean is $\rho(m-1)$ and in the denominator it is ρm. In the numerator of $\mathrm{RR}_{UM|A}$ the mean is $a + \rho u$ and in the denominator it is $a + \rho u^*$ where $u \neq u^*$. Unless $\rho = 0$ (in which case U and M are independent, and U is not a confounder), there is no finite maximum for either of the ratios. In this case, as one of the parameters in each is infinite, both the true bounding factor and its alternative are simply $\mathrm{RR}_{UY|(A=1,M)}$. However, further statements about approximate bias can be made in this context by considering all possible coarsenings of confounders that reduce bias to within a small threshold of the true effect, with reasoning analogous to what is possible with E-values for total effects (cf. VanderWeele et al., 2019[4]).

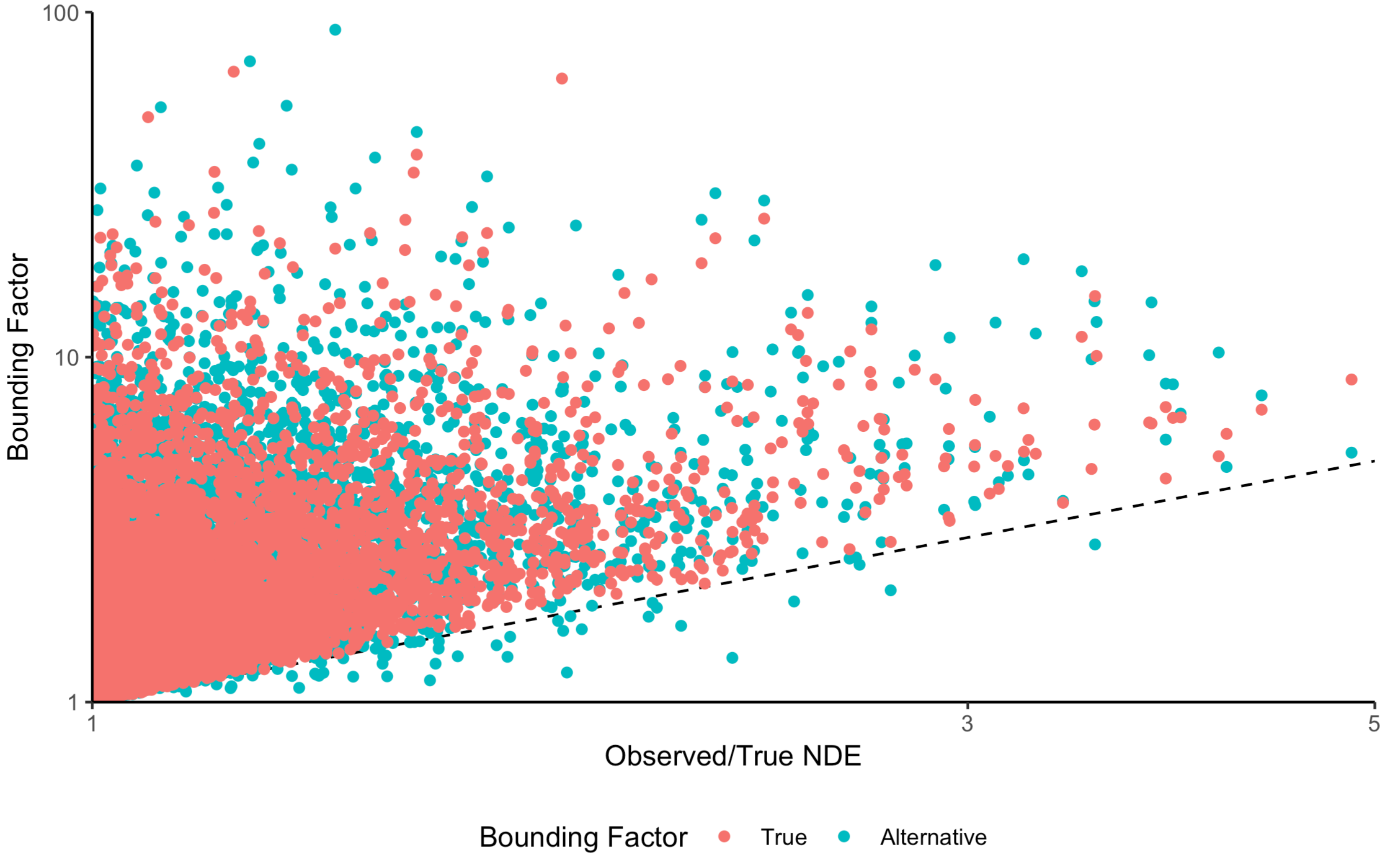
5 TABLES AND FIGURES

eTable 1: Frequencies in the various datasets with which the possible alternative parameters compared adequately with the true $\mathrm{RR}_{AU|M}$ parameter, and with which the bounding factor constructed with the alternative parameter failed to bound the bias. The four columns refer to the four sets of data-generating assumptions: uniform probabilities, a log-linear model for M, as well as a restriction to no qualitative interaction within each.

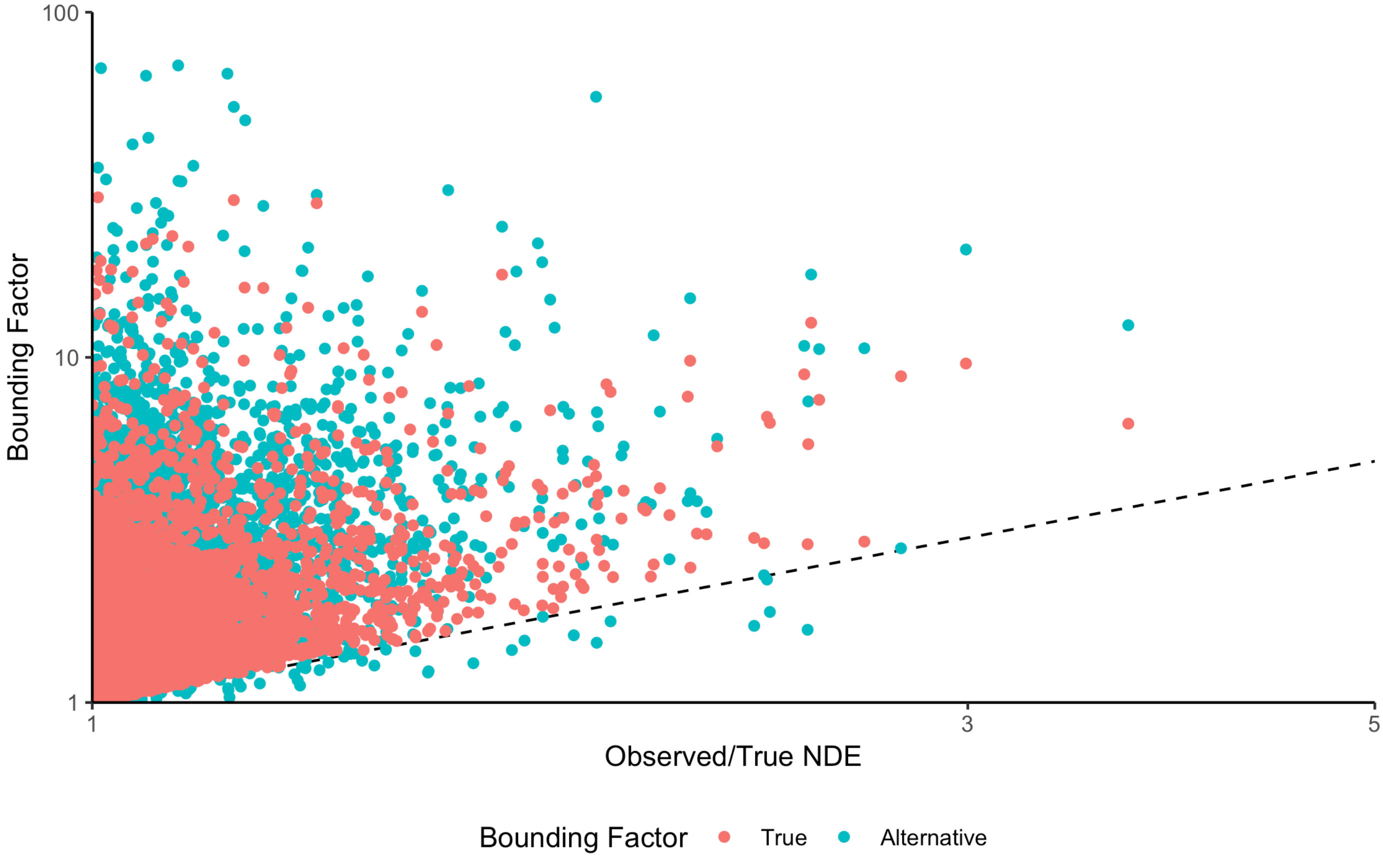
	Alternative RR $> \mathrm{RR}_{AU|M}$				Bias > bound			
	$\mathrm{P}(M|A,U)$ uniform		$\mathrm{P}(M|A,U)$ log-linear		$\mathrm{P}(M|A,U)$ uniform		$\mathrm{P}(M|A,U)$ log-linear	
Alternative	Full	Restricted	Full	Restricted	Full	Restricted	Full	Restricted
$\mathrm{RR}_{AU|M}$					0.000000	0.000000	0.000000	0.000000
RR_{AM}	0.427670	0.772424	0.869336	0.869013	0.106455	0.000581	0.001158	0.000376
$\mathrm{RR}_{AM=1}$	0.335555	0.643411	0.759864	0.759529	0.147954	0.007304	0.005953	0.004846
$\mathrm{RR}_{AM|U}$	0.856301	1.000000	1.000000	1.000000	0.000003	0.000000	0.000000	0.000000
$\mathrm{RR}_{AM=1|U}$	0.670154	0.858100	0.759864	0.759529	0.003230	0.001315	0.005953	0.004846
RR_{UM}	0.418313	0.724908	0.828754	0.828197	0.120022	0.012376	0.012299	0.011695
$\mathrm{RR}_{UM=1}$	0.331758	0.603611	0.747872	0.747164	0.165460	0.033009	0.021784	0.021057
$\mathrm{RR}_{UM|A}$	0.865605	1.000000	1.000000	1.000000	0.000004	0.000000	0.000000	0.000000
$\mathrm{RR}_{UM=1|A}$	0.663726	0.854724	0.747872	0.747164	0.006508	0.004635	0.021784	0.021057

eFigure 1: 10,000 randomly sampled pairs of bounding factors ($\mathrm{BF}_{\mathrm{true}}$and $\mathrm{BF}_{UM=1|A}$ for the same distribution) relative to the actual bias (on the risk ratio scale). The further a point is from the diagonal line, the weaker that bound is. The alternative bounding factors below the diagonal are smaller than the bias, so did not successfully bound it.

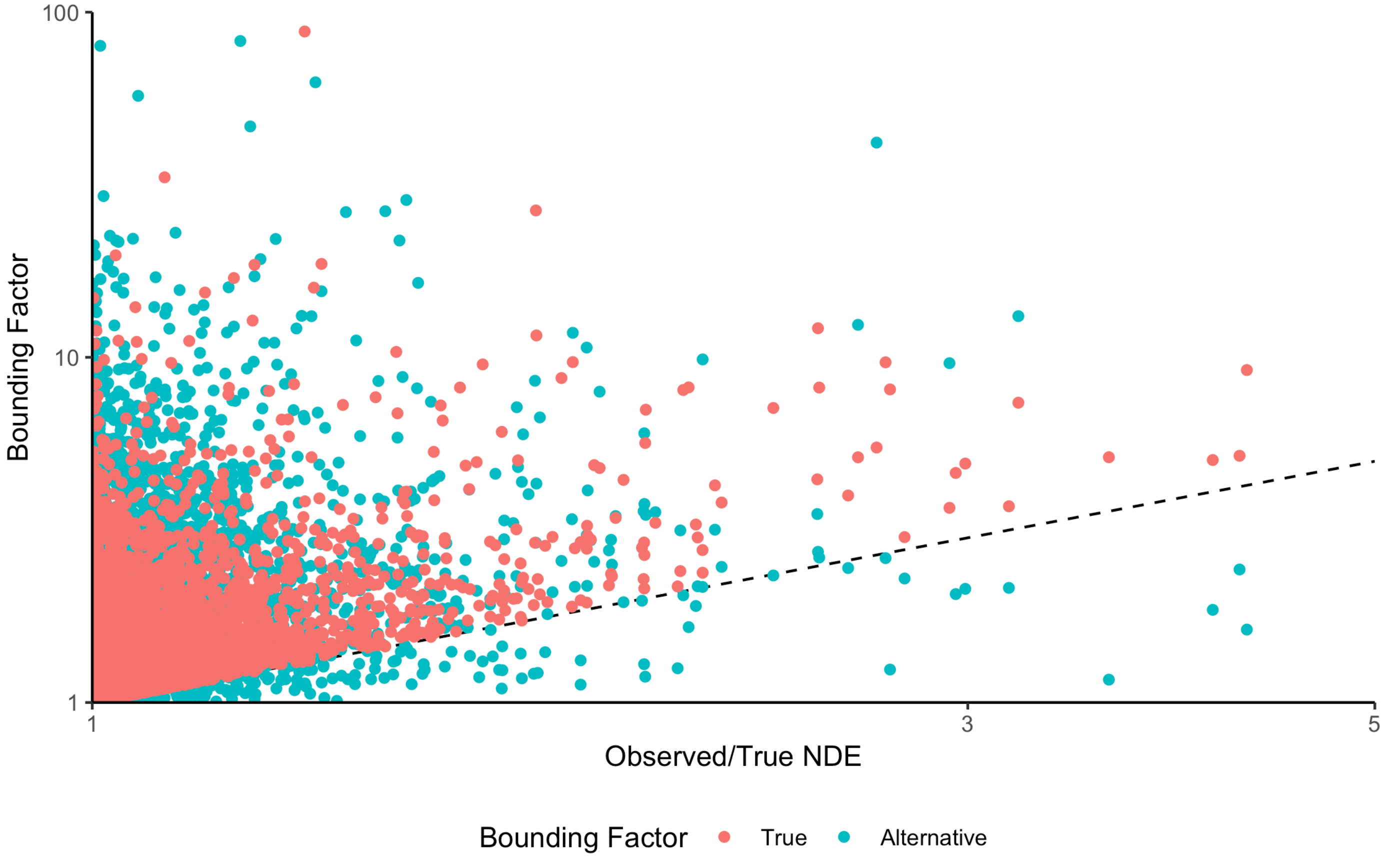
eFigure 1A. Probabilities drawn from uniform distribution



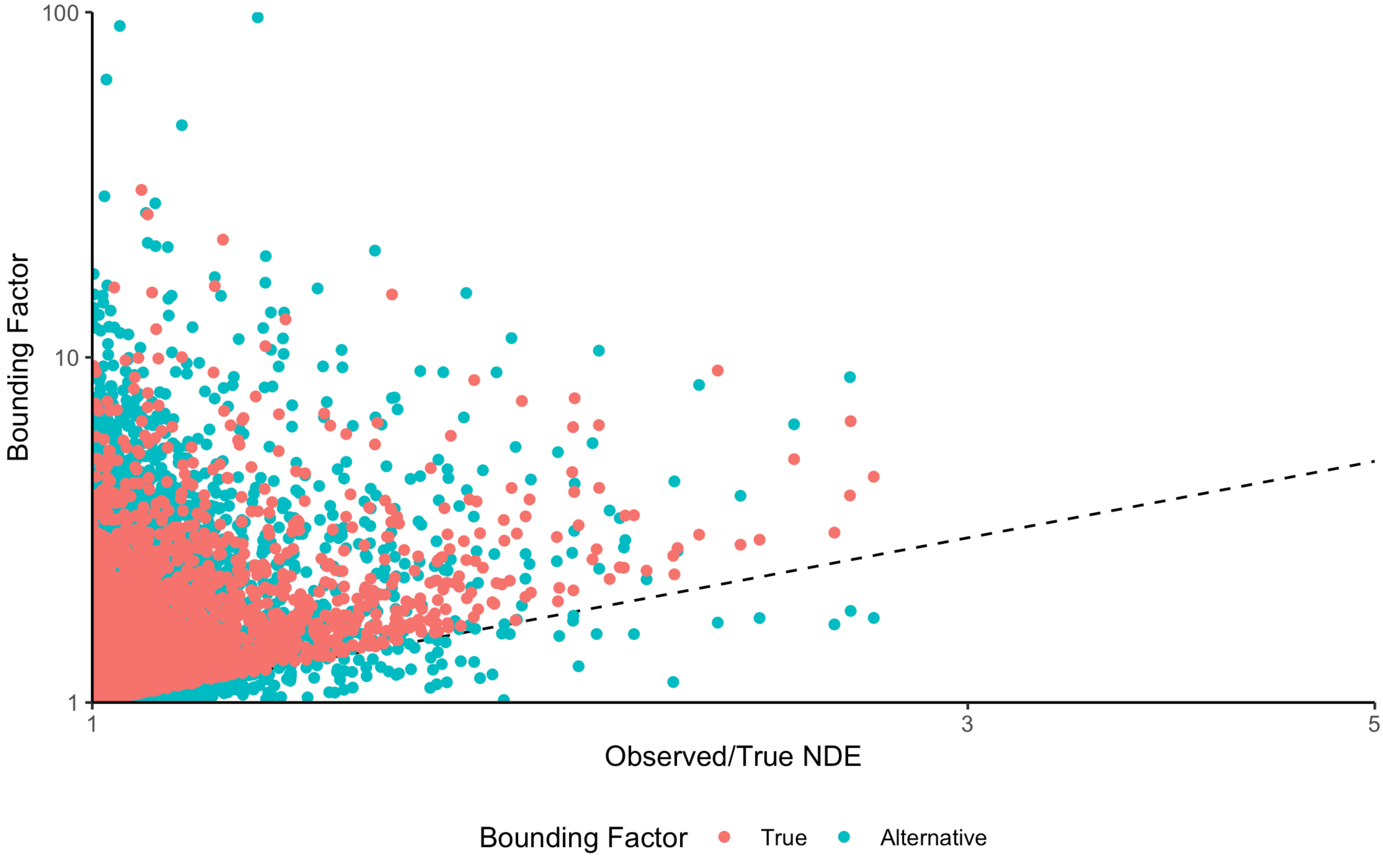
eFigure 1B. Uniform probabilities, restricted to no qualitative interaction



eFigure 1C. Log-linear model for M



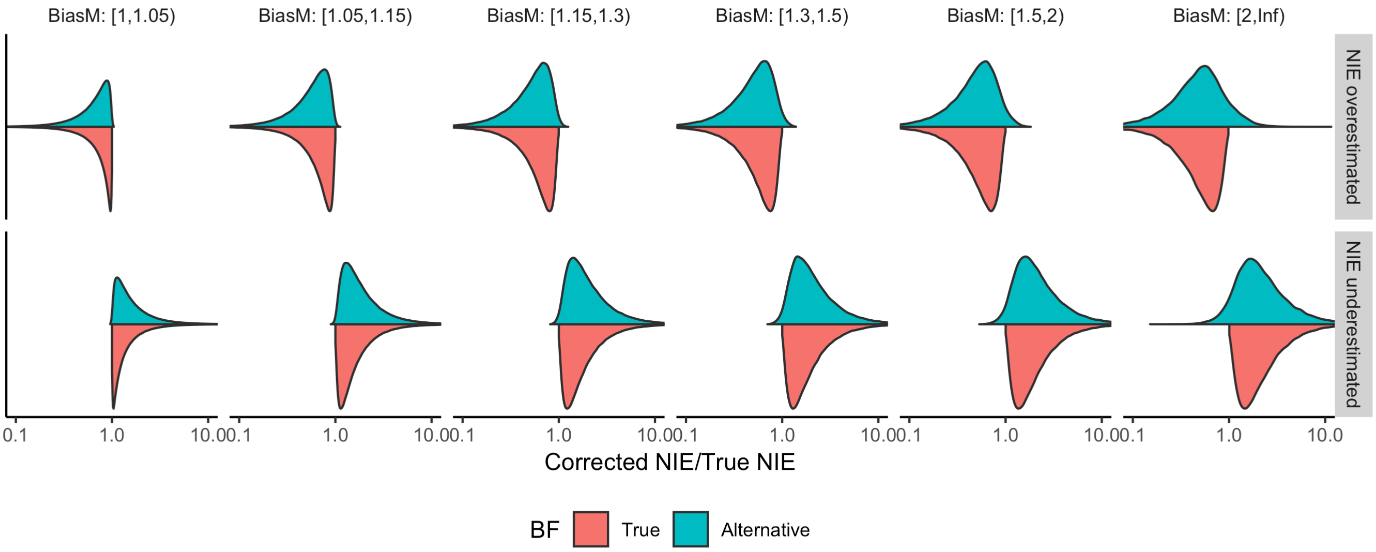
eFigure 1D. Log-linear model, restricted to no qualitative interaction



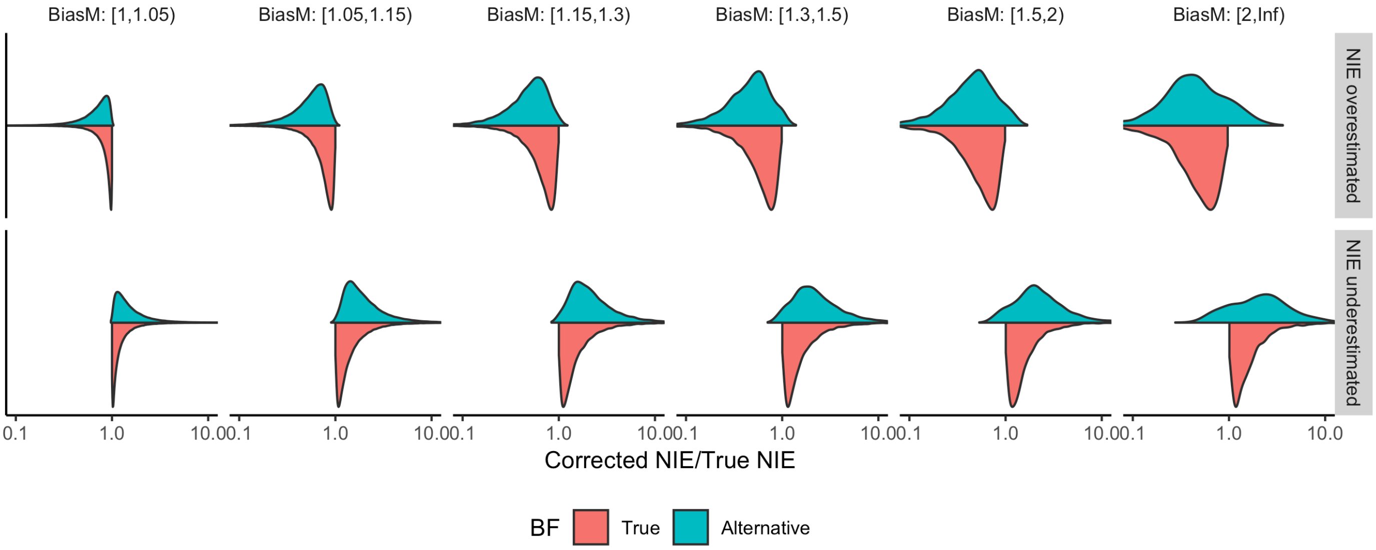
eFigure 2: The extent of the overcorrection of the true bias by $\mathrm{BF}_{\mathrm{true}}$ is shown in red as the ratio of the corrected to true NIE (on the risk ratio scale); values of 1 indicate perfect correction (note axis on log scale).

There is no case in which the ratio of the corrected NIE to true NIE is on the wrong side of 1, because $\mathrm{BF}_{\mathrm{true}}$ bounds the bias. However, because $\mathrm{BF}_{UM=1|A}$ (whose corrected/true ratio is in blue) is not a true bound for the bias, in some cases it fails to correct for it fully, particularly when the bias is large. Also, when the bias is small, it tends to overcorrect relative to the true $\mathrm{BF}_{\mathrm{true}}$.

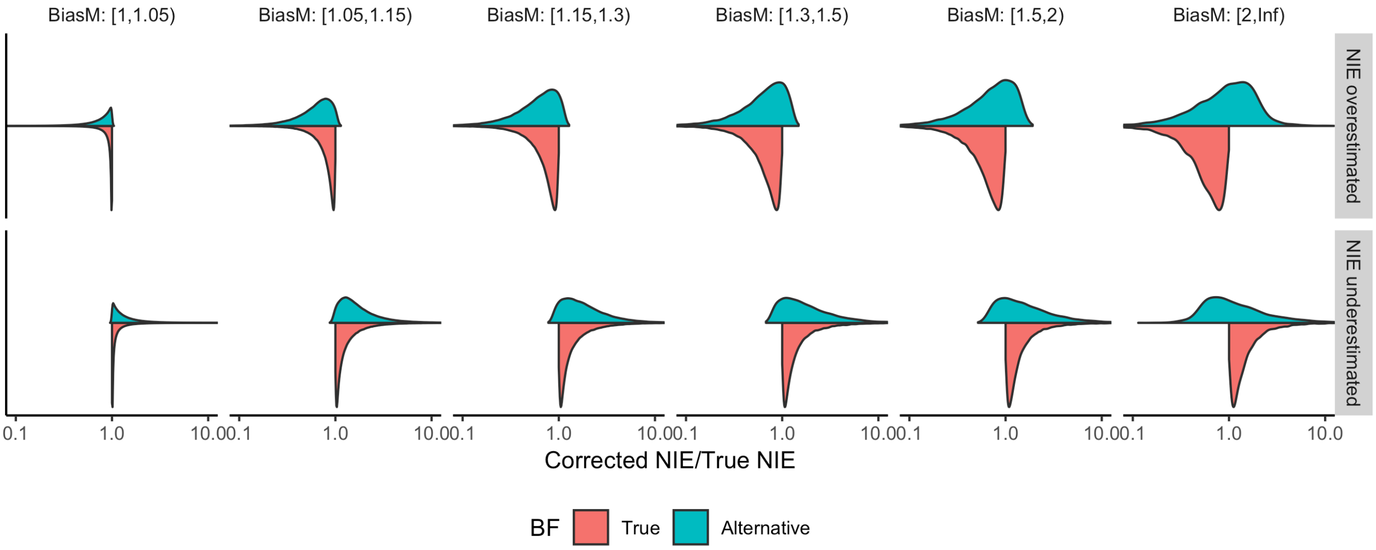
eFigure 2A. Probabilities drawn from uniform distribution (RR scale)



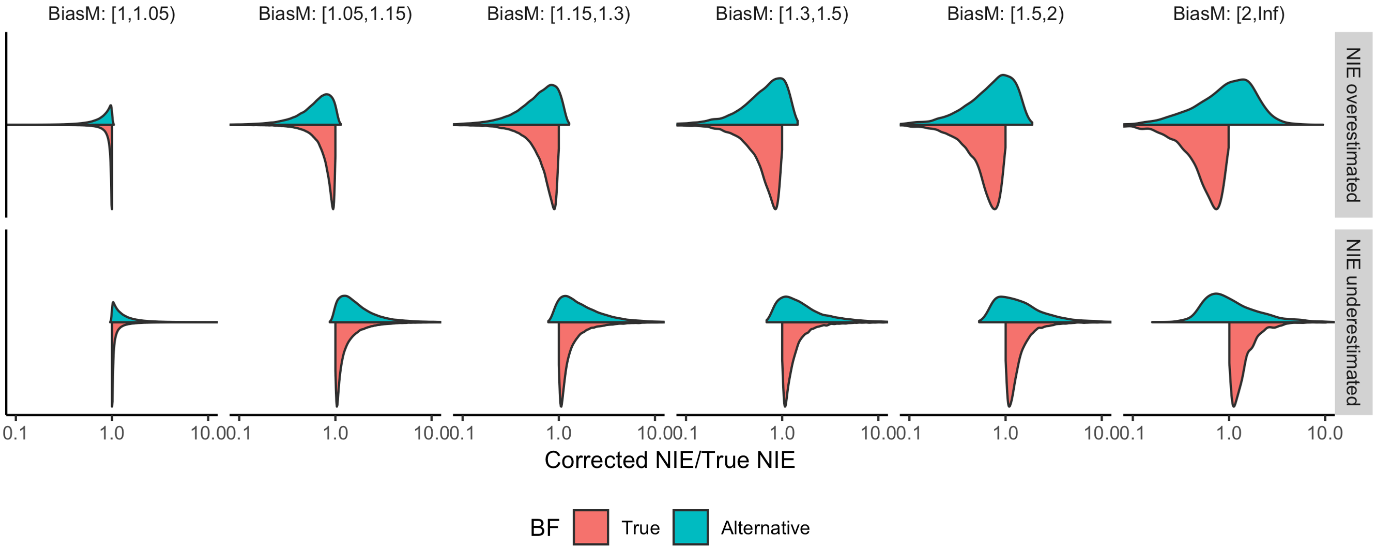
eFigure 2B. Uniform probabilities, restricted to no qualitative interaction (RR scale)



eFigure 2C. Log-linear model for M (RR scale)



eFigure 2D. Log-linear model, restricted to no qualitative interaction (RR scale)



eTable 2: Distribution of the ratios of corrected to true NIE on the risk difference scale (inverted if NIE overestimated) overall and in strata defined by the magnitude of the bias (intervals in top row). NIEs have been corrected by both $\mathrm{BF}_{\mathrm{true}}$ and the alternative $\mathrm{BF}_{UM=1|A}$. Values below 1 indicate that the corrected NIE did not fully correct for the magnitude of the unmeasured confounding. Values much greater than 1 indicate that the NIE has been extremely overcorrected.

eTable 2A. Probabilities drawn from uniform distribution (RR scale)

	Overall		[1, 1.05)		[1.05, 1.15)		[1.15, 1.3)		[1.3, 1.5)		[1.5, 2)		[2, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	1.01	1.01	1.00	1.02	1.02	1.03	1.04	1.00	1.05	0.94	1.06	0.83	1.08	0.58
5th	1.03	1.07	1.02	1.05	1.06	1.11	1.09	1.12	1.12	1.10	1.14	1.05	1.16	0.86
10th	1.06	1.12	1.03	1.09	1.09	1.16	1.14	1.20	1.17	1.20	1.21	1.18	1.24	1.05
25th	1.15	1.26	1.09	1.19	1.19	1.32	1.27	1.39	1.33	1.43	1.39	1.45	1.45	1.41
50th	1.37	1.58	1.24	1.42	1.42	1.65	1.56	1.79	1.67	1.88	1.77	1.95	1.90	1.99
75th	1.87	2.25	1.60	1.94	1.91	2.34	2.17	2.61	2.37	2.78	2.56	2.92	2.82	3.08
90th	2.86	3.53	2.35	2.93	2.88	3.66	3.35	4.18	3.73	4.48	4.08	4.77	4.56	5.14
95th	3.97	4.97	3.19	4.04	3.98	5.13	4.68	5.90	5.21	6.40	5.80	6.84	6.46	7.36
99th	8.63	10.96	6.82	8.68	8.60	11.26	10.21	13.18	11.41	14.47	12.94	15.28	14.33	16.46

eTable 2B. Uniform probabilities; restricted to no qualitative interactions (RR scale)

	Overall		[1, 1.05)		[1.05, 1.15)		[1.15, 1.3)		[1.3, 1.5)		[1.5, 2)		[2, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	1.00	1.01	1.00	1.02	1.01	1.03	1.02	0.97	1.03	0.88	1.03	0.74	1.04	0.48
5th	1.01	1.06	1.01	1.05	1.03	1.14	1.05	1.13	1.07	1.04	1.08	0.92	1.10	0.66
10th	1.02	1.10	1.01	1.08	1.06	1.22	1.09	1.25	1.11	1.19	1.13	1.08	1.15	0.83
25th	1.06	1.22	1.04	1.17	1.12	1.39	1.18	1.50	1.23	1.54	1.27	1.52	1.30	1.30
50th	1.18	1.49	1.12	1.37	1.27	1.73	1.38	1.96	1.48	2.08	1.57	2.14	1.68	2.25
75th	1.44	2.05	1.32	1.78	1.60	2.41	1.83	2.84	2.03	3.12	2.18	3.28	2.44	3.57
90th	1.98	3.11	1.71	2.56	2.25	3.71	2.70	4.51	3.08	5.03	3.34	5.34	3.93	5.69
95th	2.60	4.30	2.16	3.42	2.98	5.16	3.67	6.34	4.26	7.14	4.67	7.82	6.00	8.02
99th	5.25	9.25	4.05	6.97	6.25	11.24	7.70	13.97	8.89	16.58	10.93	18.66	13.51	20.30

eTable 2C. Log-linear model for M (RR scale)

	Overall		[1, 1.05)		[1.05, 1.15)		[1.15, 1.3)		[1.3, 1.5)		[1.5, 2)		[2, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	1.00	0.95	1.00	1.00	1.00	0.95	1.01	0.86	1.01	0.76	1.01	0.61	1.02	0.27
5th	1.00	1.01	1.00	1.01	1.01	1.02	1.02	0.94	1.03	0.83	1.04	0.70	1.06	0.43
10th	1.00	1.03	1.00	1.03	1.03	1.07	1.04	1.00	1.06	0.90	1.07	0.76	1.10	0.51
25th	1.01	1.09	1.01	1.08	1.07	1.23	1.11	1.19	1.13	1.08	1.17	0.94	1.22	0.67
50th	1.07	1.28	1.04	1.23	1.18	1.53	1.27	1.59	1.33	1.48	1.39	1.33	1.51	1.00
75th	1.24	1.70	1.16	1.57	1.46	2.13	1.65	2.37	1.77	2.31	1.89	2.14	2.12	1.74
90th	1.64	2.56	1.45	2.27	2.03	3.27	2.44	3.82	2.68	3.93	2.89	3.71	3.32	3.16
95th	2.11	3.51	1.80	3.05	2.68	4.54	3.32	5.40	3.72	5.72	4.02	5.44	4.75	4.78
99th	4.14	7.52	3.29	6.32	5.44	9.84	6.95	12.03	7.92	12.69	8.93	12.19	10.42	11.58

eTable 2D. Log-linear model, restricted to no qualitative interaction (RR scale)

	Overall		[1, 1.05)		[1.05, 1.15)		[1.15, 1.3)		[1.3, 1.5)		[1.5, 2)		[2, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	1.00	0.96	1.00	1.00	1.01	0.95	1.01	0.86	1.01	0.76	1.01	0.61	1.03	0.30
5th	1.00	1.01	1.00	1.01	1.02	1.01	1.03	0.93	1.04	0.83	1.04	0.70	1.06	0.44
10th	1.00	1.02	1.00	1.02	1.03	1.06	1.05	0.99	1.06	0.89	1.07	0.76	1.10	0.51
25th	1.01	1.08	1.01	1.07	1.07	1.21	1.11	1.16	1.14	1.06	1.17	0.92	1.22	0.66
50th	1.06	1.25	1.04	1.21	1.19	1.50	1.26	1.52	1.32	1.41	1.39	1.28	1.47	0.96
75th	1.22	1.63	1.14	1.51	1.46	2.07	1.63	2.25	1.73	2.13	1.83	1.99	2.02	1.62
90th	1.58	2.37	1.40	2.12	2.03	3.12	2.38	3.59	2.56	3.59	2.78	3.31	3.10	2.86
95th	2.00	3.21	1.71	2.80	2.68	4.30	3.24	5.05	3.53	5.30	3.83	4.79	4.27	4.24
99th	3.81	6.72	2.99	5.64	5.37	9.09	6.77	10.93	7.62	12.13	8.08	10.52	8.80	8.93

eTable 3: Distribution of the absolute value of the differences between corrected and true NIE overall and in strata defined by the magnitude of the bias (intervals in top row). NIEs have been corrected by both BF_{true} and the alternative $\text{BF}_{UM=1|A}$. Values below 0 indicate that the corrected NIE did not fully correct for the magnitude of the unmeasured confounding.

eTable 3A. Probabilities drawn from uniform distribution (RD scale)

	Overall		[0, 0.005)		[0.005, 0.01)		[0.01, 0.025)		[0.025, 0.05)		[0.05, 0.1)		[0.1, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.02	-0.05
5th	0.01	0.03	0.00	0.02	0.01	0.03	0.02	0.03	0.02	0.04	0.03	0.04	0.04	0.03
10th	0.03	0.05	0.01	0.03	0.02	0.04	0.03	0.05	0.04	0.06	0.05	0.07	0.06	0.06
25th	0.06	0.10	0.02	0.06	0.04	0.08	0.05	0.10	0.07	0.11	0.09	0.13	0.11	0.13
50th	0.13	0.19	0.07	0.13	0.10	0.16	0.12	0.18	0.15	0.21	0.17	0.23	0.20	0.24
75th	0.27	0.37	0.18	0.26	0.21	0.30	0.24	0.34	0.28	0.39	0.33	0.43	0.41	0.51
90th	0.55	0.76	0.37	0.51	0.41	0.57	0.46	0.65	0.55	0.78	0.67	0.94	0.96	1.21
95th	0.96	1.30	0.60	0.81	0.67	0.93	0.77	1.09	0.92	1.30	1.14	1.57	1.62	2.03
99th	2.80	3.59	1.87	2.31	2.02	2.62	2.29	3.01	2.65	3.54	3.18	4.22	4.36	5.46

eTable 3B. Uniform probabilities; restricted to no qualitative interactions (RD scale)

	Overall		[0, 0.005)		[0.005, 0.01)		[0.01, 0.025)		[0.025, 0.05)		[0.05, 0.1)		[0.1, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.02	0.01	0.01	0.01	-0.01	0.01	-0.08
5th	0.01	0.03	0.00	0.01	0.01	0.03	0.01	0.04	0.02	0.05	0.02	0.05	0.02	0.00
10th	0.01	0.04	0.00	0.02	0.01	0.04	0.02	0.06	0.02	0.08	0.03	0.09	0.03	0.06
25th	0.03	0.09	0.01	0.05	0.02	0.08	0.03	0.10	0.05	0.14	0.06	0.16	0.07	0.15
50th	0.07	0.17	0.04	0.11	0.06	0.14	0.08	0.18	0.10	0.23	0.13	0.27	0.15	0.28
75th	0.16	0.32	0.10	0.21	0.13	0.27	0.15	0.33	0.20	0.41	0.24	0.48	0.32	0.53
90th	0.30	0.60	0.21	0.40	0.24	0.48	0.28	0.58	0.35	0.76	0.44	0.99	0.67	1.25
95th	0.44	0.98	0.31	0.59	0.35	0.73	0.40	0.92	0.50	1.25	0.67	1.63	1.08	2.08
99th	1.18	2.75	0.78	1.56	0.85	1.96	0.99	2.52	1.27	3.45	1.86	4.37	3.02	5.85

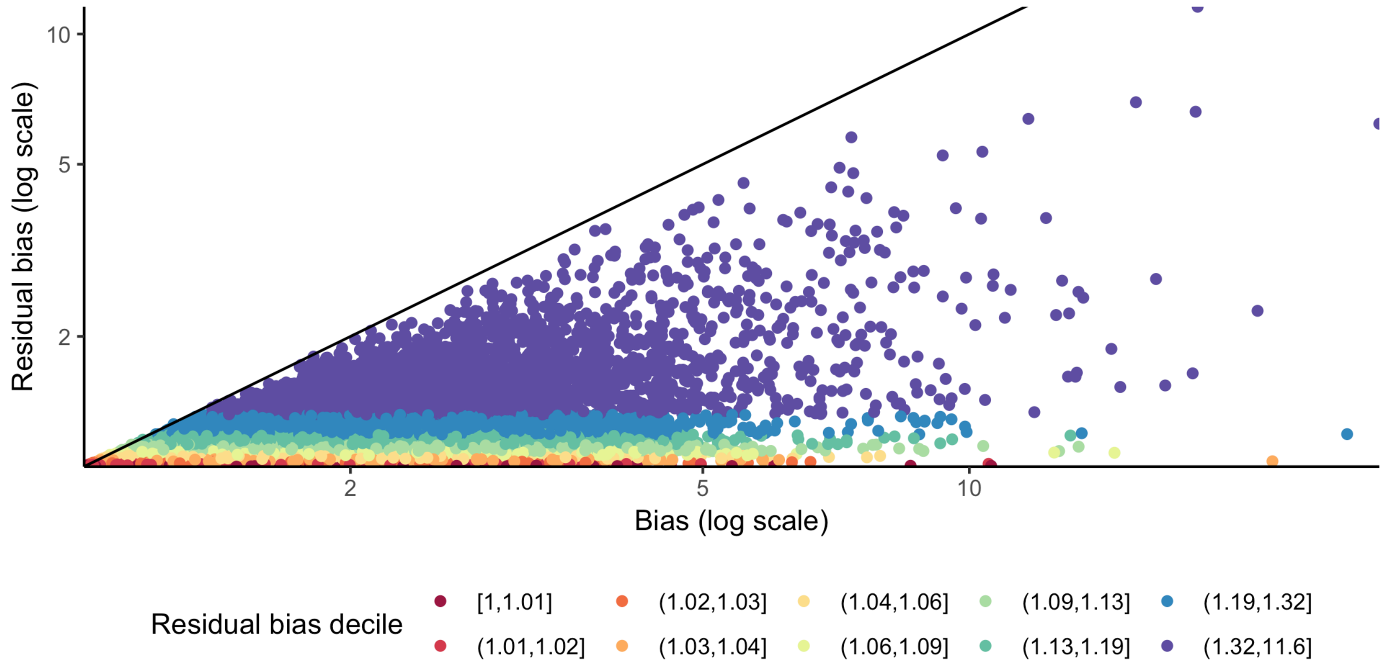
eTable 3C. Log-linear model for M (RD scale)

	Overall		[0, 0.005)		[0.005, 0.01)		[0.01, 0.025)		[0.025, 0.05)		[0.05, 0.1)		[0.1, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	0.00	-0.02	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.02	0.00	-0.05	0.00	-0.23
5th	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.01	0.01	-0.02	0.01	-0.12
10th	0.00	0.01	0.00	0.01	0.00	0.02	0.01	0.03	0.01	0.03	0.01	0.01	0.02	-0.08
25th	0.01	0.04	0.00	0.02	0.01	0.06	0.02	0.07	0.03	0.09	0.04	0.08	0.04	0.01
50th	0.03	0.11	0.01	0.07	0.03	0.12	0.05	0.15	0.07	0.18	0.09	0.19	0.09	0.13
75th	0.09	0.23	0.04	0.16	0.09	0.23	0.12	0.28	0.16	0.34	0.19	0.38	0.21	0.31
90th	0.21	0.46	0.12	0.34	0.19	0.44	0.24	0.52	0.29	0.65	0.35	0.79	0.45	0.65
95th	0.32	0.74	0.20	0.52	0.30	0.69	0.35	0.83	0.44	1.08	0.54	1.36	0.73	1.21
99th	0.84	2.16	0.53	1.49	0.74	1.97	0.89	2.35	1.12	2.91	1.37	3.72	2.03	3.66

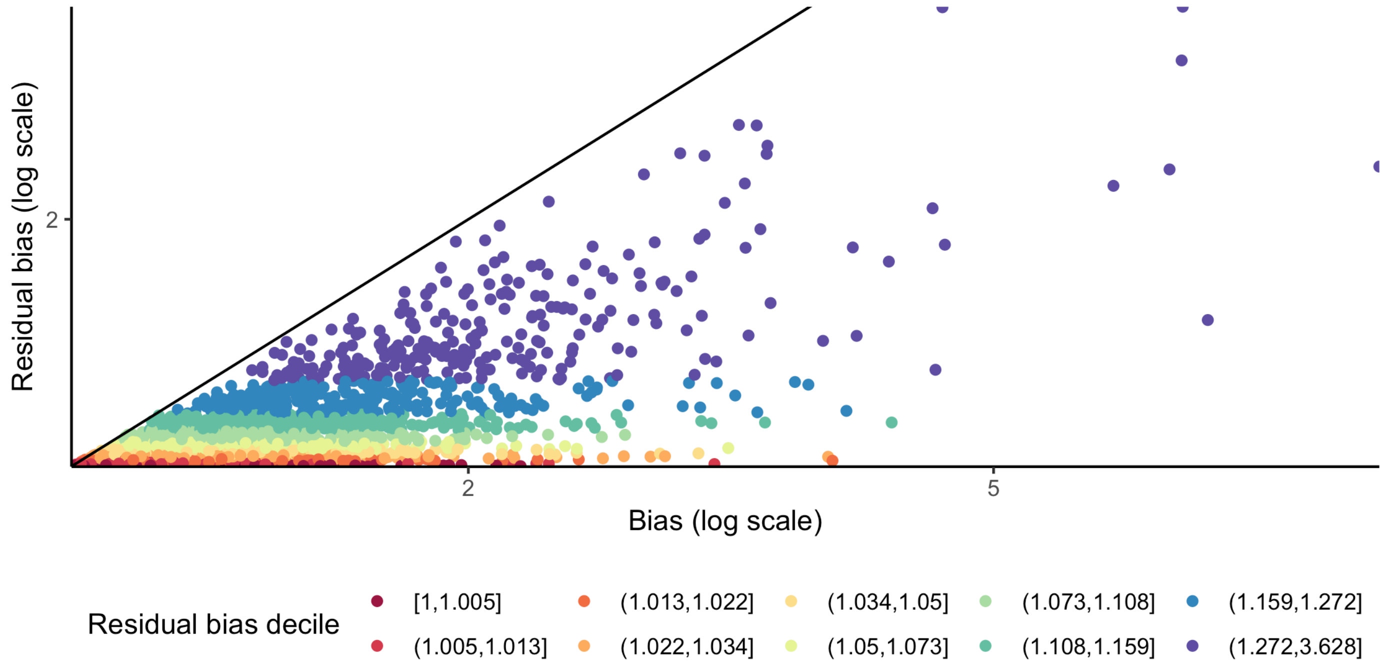
eTable 3D. Log-linear model, restricted to no qualitative interaction (RD scale)

	Overall		[0, 0.005)		[0.005, 0.01)		[0.01, 0.025)		[0.025, 0.05)		[0.05, 0.1)		[0.1, ∞)	
%ile	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.	True	Alt.
1st	0.00	-0.02	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.02	0.00	-0.05	0.00	-0.21
5th	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	-0.02	0.01	-0.11
10th	0.00	0.01	0.00	0.01	0.00	0.02	0.01	0.03	0.01	0.03	0.02	0.00	0.02	-0.07
25th	0.01	0.04	0.00	0.02	0.01	0.05	0.02	0.07	0.03	0.08	0.04	0.07	0.04	0.01
50th	0.03	0.10	0.01	0.06	0.03	0.11	0.05	0.14	0.07	0.18	0.08	0.18	0.09	0.12
75th	0.08	0.22	0.04	0.14	0.08	0.22	0.11	0.27	0.16	0.33	0.19	0.36	0.23	0.30
90th	0.19	0.42	0.10	0.30	0.17	0.40	0.22	0.48	0.29	0.60	0.37	0.72	0.52	0.63
95th	0.29	0.64	0.16	0.46	0.26	0.60	0.32	0.74	0.42	0.95	0.55	1.22	0.87	1.15
99th	0.71	1.84	0.39	1.27	0.57	1.65	0.73	2.03	1.05	2.52	1.50	3.24	2.42	3.43

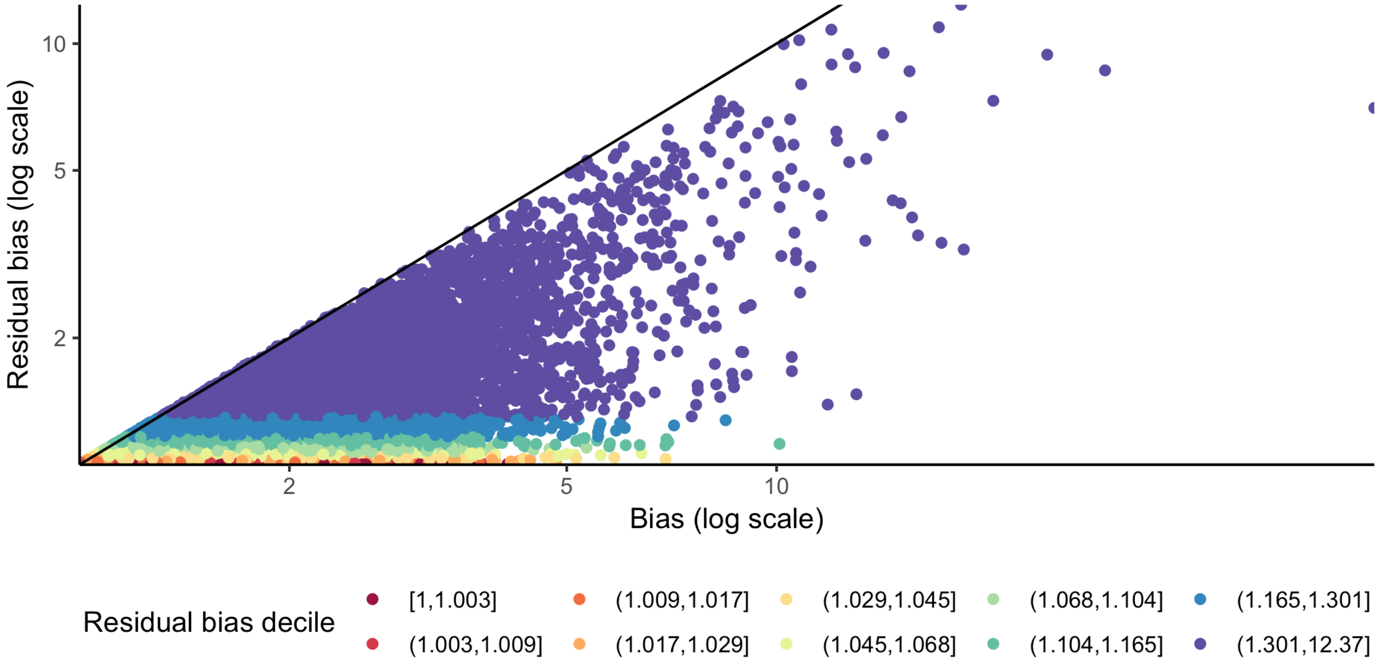
eFigure 3: Residual bias when $\mathrm{BF}_{UM=1|A}$ fails. Values near the diagonal indicate that almost none of the bias has been removed after correcting the effects with $\mathrm{BF}_{UM=1|A}$. However, the vast majority have little bias remaining after correction, as illustrated by the values of the deciles of residual bias.



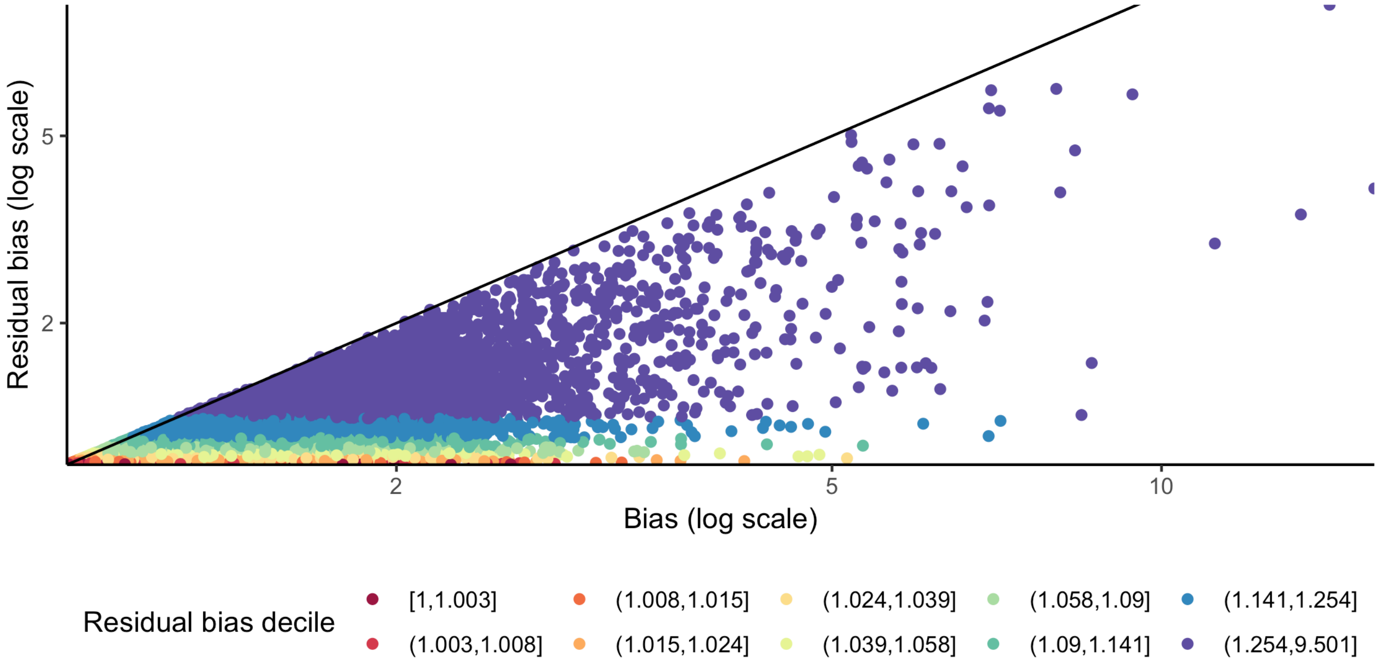
eFigure 4A. Probabilities drawn from uniform distribution



eFigure 4B. Uniform probabilities, restricted to no qualitative interaction eFigure 4C. Log-linear model for M



eFigure 4D. Log-linear model, restricted to no qualitative interaction



eTable 4: Distribution of the residual bias of on the risk difference scale after correction by $\text{BF}_{UM=1|A}$, among those cases in which the bound failed. That is, the first value in the first column shows that at least 10% of the time that the alternative bound failed, the corrected effect was less than 0.002 from the true effect.

	Uniform probabilities		Log-linear model	
%ile	No restriction	No qualitative interaction	No restriction	No qualitative interaction
10th	0.002	0.002	0.001	0.001
20th	0.005	0.004	0.003	0.003
30th	0.009	0.008	0.007	0.005
40th	0.014	0.012	0.011	0.009
50th	0.020	0.017	0.017	0.014
60th	0.028	0.025	0.025	0.021
70th	0.040	0.035	0.038	0.031
80th	0.059	0.051	0.057	0.048
90th	0.096	0.079	0.096	0.079
Maximum	0.537	0.449	0.656	0.542

6 COMPUTING CODE

The results can be replicated using the R code in this document. The code is also available at https://github.com/louisahsmith/mediational-e-values.

The size of the simulation can be controlled by the object `N`. Note that large `N` (e.g., greater than the 5,000,000 to produce the results in this paper) requires several hours of computing time and may overload a personal computer.

In the code, the variable names correspond to the following probabilities:

$$\texttt{a} = \mathrm{P}(U = 1)$$

$$\texttt{b} = \mathrm{P}(M = 1|A = 1, U = 1)$$

$$\texttt{c} = \mathrm{P}(M = 1|A = 1, U = 0)$$

$$\texttt{d} = \mathrm{P}(M = 1|A = 0, U = 1)$$

$$\texttt{e} = \mathrm{P}(M = 1|A = 0, U = 0)$$

$$\texttt{f} = \mathrm{P}(Y = 1|A = 1, U = 1, M = 1)$$

$$\texttt{g} = \mathrm{P}(Y = 1|A = 1, U = 1, M = 0)$$

$$\texttt{h} = \mathrm{P}(Y = 1|A = 1, U = 0, M = 1)$$

$$\texttt{i} = \mathrm{P}(Y = 1|A = 1, U = 0, M = 0)$$

$$\texttt{j} = \mathrm{P}(Y = 1|A = 0, U = 1, M = 1)$$

$$\texttt{k} = \mathrm{P}(Y = 1|A = 0, U = 1, M = 0)$$

$$\texttt{l} = \mathrm{P}(Y = 1|A = 0, U = 0, M = 1)$$

$$\texttt{m} = \mathrm{P}(Y = 1|A = 0, U = 0, M = 0)$$

$$\texttt{n} = \mathrm{P}(A = 1)$$

(Strictly speaking, the bias we are interested in does not depend on $\mathrm{P}(Y = 1 \mid A = 0, M = m, U = u)$ for any m or u, but we included those values in order to calculate the true and observed NDEs and not just their ratio. The marginal probability of the exposure, $\mathrm{P}(A = 1)$, is only useful for the calculation of one of the possible RR alternatives and also does not directly factor into the bias or the effect sizes.)

For simplicity, we also computed the probabilities of each of A, M, U, and Y taking on values of 0. Since we assumed that each is binary, $\mathrm{P}(U = 0) = 1 - \mathrm{P}(U = 1)$, $\mathrm{P}(M = 0 \mid A = 1, U = 1) = 1 - \mathrm{P}(M = 1|A = 1, U = 1)$, etc. In the code, those values are referred to as `a_comp`, `b_comp`, etc., for each of the above letters.

The alternative RRs are labeled as follows:

$$\texttt{RRau} = \text{RR}_{AU|M}$$

$$\texttt{RRuy} = \text{RR}_{UY|(A=1,M)}$$

$$\texttt{RRam} = \text{RR}_{AM}$$

$$\texttt{RRam\_no0} = \text{RR}_{AM=1}$$

$$\texttt{RRam\_strat} = \text{RR}_{AM|U}$$

$$\texttt{RRam\_strat\_no0} = \text{RR}_{AM=1|U}$$

$$\texttt{RRum} = \text{RR}_{UM}$$

$$\texttt{RRum\_no0} = \text{RR}_{UM=1}$$

$$\texttt{RRum\_strat} = \text{RR}_{UM|A}$$

$$\texttt{RRum\_strat\_no0} = \text{RR}_{UM=1|A}$$

The following code was used to run the simulation and produce the tables and figures in the eAppendix.

File: `functions.R`

```
if (!require("tidyverse")) {install.packages("tidyverse"); library(tidyverse)}
if (!require("xtable")) {install.packages("xtable"); library(xtable)}


# function for bounding factor from any two RRs
BF_func <- function(RRa, RRb) {
  RRa * RRb / (RRa + RRb - 1)
}

# function to create dataset with NIE, NDE (true and observed), bias, etc.
# N: number of samples to draw
# loglin: whether M should be drawn according to loglinear model
make_data <- function(N, loglin = FALSE) {

  # draw N sets of 14 probabilities from uniform(0, 1)
  dat <- data.frame(matrix(runif(N * 14), ncol = 14))

  # letter names correspond to probabilities
  names(dat) <- letters[1:14]

  dat$loglin <- "unrestricted"

  # force to follow a loglinear model
  if (loglin) {
    dat <- dat %>%
      mutate(
        # randomly choose one probability to throw out
        keep = runif(N),
        b = case_when(keep < .25 ~ d * c / e, TRUE ~ b),
        c = case_when(between(keep, .25, .5) ~ b * e / d, TRUE ~ c),
```

```
      d = case_when(between(keep, .5, .75) ~ b * e / c, TRUE ~ d),
      e = case_when(keep > .75 ~ d * c / b, TRUE ~ e),
      loglin = "loglinear"
    ) %>%
    select(-keep) %>%
    # remove any impossible probabilities
    filter(b < 1 & c < 1 & d < 1 & e < 1)
}

# make new variables
dat %>%

  # 1 - prob for P(var = 0)
  mutate_if(is.numeric, funs(comp = 1 - .)) %>%

  mutate(

    # ratio of obs/true NDE
    obs_true = ((1 / (b * a + c * a_comp)) * (f * b * a + h * c * a_comp) *
      (d * a + e * a_comp) +
      (1 / (b_comp * a + c_comp * a_comp)) *
        (g * b_comp * a + i * c_comp * a_comp) * (d_comp * a + e_comp * a_comp)) /
      (f * d * a + g * d_comp * a + h * e * a_comp + i * e_comp * a_comp),

    # values of true and observed effects
    NIE_true = (f * b * a + g * b_comp * a + h * c * a_comp + i * c_comp * a_comp) /
      (f * d * a + g * d_comp * a + h * e * a_comp + i * e_comp * a_comp),
    NIE_obs = NIE_true / obs_true,
    NDE_true = (f * d * a + g * d_comp * a + h * e * a_comp + i * e_comp * a_comp) /
      (j * d * a + k * d_comp * a + l * e * a_comp + m * e_comp * a_comp),
    NDE_obs = NDE_true * obs_true,

    # make sure the bias > 1 so bound is sensical
    bias = (obs_true > 1) * obs_true + (obs_true < 1) / obs_true,

    # make loglin into a factor for later grouping
    loglin = fct_recode(loglin,
      "Unrestricted probabilities" = "unrestricted",
      "Log-linear model" = "loglinear"
    ),

    # create possible RRau values
    RRau_m1u1 = b * (d * a + e * a_comp) / (d * (b * a + c * a_comp)),
    RRau_m1u0 = c * (d * a + e * a_comp) / (e * (b * a + c * a_comp)),
    RRau_m0u1 = b_comp * (d_comp * a + e_comp * a_comp) /
      (d_comp * (b_comp * a + c_comp * a_comp)),
    RRau_m0u0 = c_comp * (d_comp * a + e_comp * a_comp) /
      (e_comp * (b_comp * a + c_comp * a_comp)),

    # create possible RRuy values
    RRuy_m1u1 = f / h,
    RRuy_m1u0 = 1 / RRuy_m1u1,
    RRuy_m0u1 = g / i,
    RRuy_m0u0 = 1 / RRuy_m0u1,

    # create possible RRam values
```

```
    RRamm_a1m1 = (b * a + c * a_comp) / (d * a + e * a_comp),
    RRamm_a0m1 = 1 / RRamm_a1m1,
    RRamm_a1m0 = (b_comp * a + c_comp * a_comp) / (d_comp * a + e_comp * a_comp),
    RRamm_a0m0 = 1 / RRamm_a1m0,

    # create possible RRam|u values
    RRamc_a1m1u1 = b / d,
    RRamc_a0m1u1 = 1 / RRamc_a1m1u1,
    RRamc_a1m0u1 = b_comp / d_comp,
    RRamc_a0m0u1 = 1 / RRamc_a1m0u1,
    RRamc_a1m1u0 = c / e,
    RRamc_a0m1u0 = 1 / RRamc_a1m1u0,
    RRamc_a1m0u0 = c_comp / e_comp,
    RRamc_a0m0u0 = 1 / RRamc_a1m0u0,

    # create possible RRum values
    RRumm_u1m1 = (b * n + d * n_comp) / (c * n + e * n_comp),
    RRumm_u0m1 = 1 / RRumm_u1m1,
    RRumm_u1m0 = (b_comp * n + d_comp * n_comp) / (c_comp * n + e_comp * n_comp),
    RRumm_u0m0 = 1 / RRumm_u1m0,

    # create possible RRum|a values
    RRumc_a1m1u1 = b / c,
    RRumc_a1m1u0 = 1 / RRumc_a1m1u1,
    RRumc_a1m0u1 = b_comp / c_comp,
    RRumc_a1m0u0 = 1 / RRumc_a1m0u1,
    RRumc_a0m1u1 = d / e,
    RRumc_a0m1u0 = 1 / RRumc_a0m1u1,
    RRumc_a0m0u1 = d_comp / e_comp,
    RRumc_a0m0u0 = 1 / RRumc_a0m0u1,

    # look at interaction direction
    AMdirection = case_when(
      c / e > 1 & b / d > 1 ~ "A-M positive",
      c / e < 1 & b / d < 1 ~ "A-M negative",
      c / e > 1 & b / d < 1 ~ "A-M positive U0, negative U1",
      c / e < 1 & b / d > 1 ~ "A-M negative U0, positive U1"
    ),
    UMdirection = case_when(
      d / e > 1 & b / c > 1 ~ "U-M positive",
      d / e < 1 & b / c < 1 ~ "U-M negative",
      d / e > 1 & b / c < 1 ~ "U-M positive A0, negative A1",
      d / e < 1 & b / c > 1 ~ "U-M negative A0, positive A1"
    ),
    MYdirection = case_when(
      h / i > 1 & f / g > 1 ~ "M-Y positive",
      h / i < 1 & f / g < 1 ~ "M-Y negative",
      h / i > 1 & f / g < 1 ~ "M-Y positive U0, negative U1",
      h / i < 1 & f / g > 1 ~ "M-Y negative U0, positive U1"
    ),
    UYdirection = case_when(
      g / i > 1 & f / h > 1 ~ "U-Y positive",
      g / i < 1 & f / h < 1 ~ "U-Y negative",
      g / i > 1 & f / h < 1 ~ "U-Y positive M0, negative M1",
      g / i < 1 & f / h > 1 ~ "U-Y negative M0, positive M1"
    ),
```

```
    M1int = case_when(
      d / e > b / c ~ "A0 > A1",
      d / e < b / c ~ "A0 < A1"
    ),
    M0int = case_when(
      d_comp / e_comp > b_comp / c_comp ~ "A0 > A1",
      d_comp / e_comp < b_comp / c_comp ~ "A0 < A1"
    ),
    Mint_directions = case_when(
      M1int == "A0 > A1" & M0int == "A0 < A1" ~
        "Not consistent",
      M1int == "A0 < A1" & M0int == "A0 > A1" ~
        "Not consistent",
      TRUE ~ "Consistent interaction"
    ),
    # summarize interaction directions
    qual_int = case_when(
      AMdirection %in% c("A-M positive", "A-M negative") &
        UMdirection %in% c("U-M positive", "U-M negative") &
        MYdirection %in% c("M-Y positive", "M-Y negative") &
        UYdirection %in% c("U-Y positive", "U-Y negative") ~ "no qual",
      TRUE ~ "qual"
    )
  ) %>%

  select(-(a:n)) %>%
  select(-(a_comp:n_comp)) %>%

  mutate(

    # choose RRs that match criteria
    RRa1u = pmap_dbl(select(., starts_with("RRau")), ~max(c(...))),
    RRa0u = pmap_dbl(select(., starts_with("RRau")), ~1 / min(c(...))),
    RRau = (obs_true > 1) * RRa1u + (obs_true < 1) * RRa0u,

    RRuy = pmap_dbl(select(., starts_with("RRuy")), ~max(c(...))),

    RRam = pmap_dbl(select(., starts_with("RRamm")), ~max(c(...))),
    RRam_no0 = pmap_dbl(select(., matches("RRamm_a\\dm1")), ~max(c(...))),

    RRam_strat = pmap_dbl(select(., starts_with("RRamc")), ~max(c(...))),
    RRam_strat_no0 = pmap_dbl(select(., matches("RRamc_a\\dm1u\\d")), ~max(c(...))),

    RRum = pmap_dbl(select(., starts_with("RRumm")), ~max(c(...))),
    RRum_no0 = pmap_dbl(select(., matches("RRumm_u\\dm1")), ~max(c(...))),

    RRum_strat = pmap_dbl(select(., starts_with("RRumc")), ~max(c(...))),
    RRum_strat_no0 = pmap_dbl(select(., matches("RRumc_a\\dm1u\\d")), ~max(c(...))),

    # Create bounding factors with new RRs
    BF_true = BF_func(RRau, RRuy),
    BF_am = BF_func(RRam, RRuy),
    BF_am_no0 = BF_func(RRam_no0, RRuy),
    BF_am_strat = BF_func(RRam_strat, RRuy),
    BF_am_strat_no0 = BF_func(RRam_strat_no0, RRuy),
    BF_um = BF_func(RRum, RRuy),
```

```
      BF_um_no0 = BF_func(RRum_no0, RRuy),
      BF_um_strat = BF_func(RRum_strat, RRuy),
      BF_um_strat_no0 = BF_func(RRum_strat_no0, RRuy)
    ) %>%

    select(-(RRau_m1u1:Mint_directions)) %>%

    mutate_at(
      # for all the bounding factors, create indicators if they succeed/fail
      # to bound bias, as well as correct NIE and NDE by each
      vars(starts_with("BF")),
      funs(
        fail = bias > .,
        GT = . > BF_true,
        NIE_correct = ifelse(obs_true > 1, NIE_obs * ., NIE_obs / .),
        NDE_correct = ifelse(obs_true < 1, NDE_obs * ., NDE_obs / .)
      )
    ) %>%

    select(-(RRa1u:BF_um_strat_no0)) %>%

    mutate(

      # create some variables for visualization/tables
      NIE_direction = ifelse(obs_true > 1, "NIE underestimated", "NIE overestimated"),
      NDE_direction = ifelse(obs_true < 1, "NDE underestimated", "NDE overestimated"),
      Bias = cut(bias, c(1, 1.05, 1.15, 1.3, 1.5, 2, Inf), right = F),
      qual_int = fct_recode(qual_int,
        "No interaction restriction" = "qual",
        "No qualitative interaction" = "no qual"
      )
    )
}

# function to get frequencies of failure
sum_data <- function(dat) {

  full_res <- dat %>%
    group_by(loglin) %>%
    summarise_at(vars(ends_with("GT"), ends_with("fail")), mean) %>%
    mutate(int = "any_int")

  no_int_res <- dat %>%
    filter(qual_int == "No qualitative interaction") %>%
    group_by(loglin) %>%
    summarise_at(vars(ends_with("GT"), ends_with("fail")), mean) %>%
    mutate(int = "no_int")

  rbind(full_res, no_int_res)
}

# function to make table 1
overall_res <- function(res) {
  GT <- res %>%
    arrange(loglin, int) %>%
    select(ends_with("GT")) %>%
```

```
    t()

  fail <- res %>%
    arrange(loglin, int) %>%
    select(ends_with("fail")) %>%
    t()

  tab <- cbind(GT, fail)
  tab[1, 1:4] <- NA

  addtorow <- list()
  addtorow$pos <- list(0, 0, 0)
  addtorow$command <- c(
    "& \\multicolumn{4}{c}{Alternative $\\text{RR }> \\RRau$} & \\multicolumn{4}{c}{Bias \\textgre
    "& \\multicolumn{2}{c}{$\\pr(M | A, U)$ uniform} & \\multicolumn{2}{c}{$\\pr(M | A, U)$ log-li
    "Alternative & Full & Restricted & Full & Restricted & Full & Restricted & Full & Restricted \
  )
  rownames(tab) <- c(
    "$\\RRau$", "$\\text{RR}_{AM}$", "$\\text{RR}_{AM=1}$",
    "$\\text{RR}_{AM|U}$",
    "$\\text{RR}_{AM=1|U}$", "$\\text{RR}_{UM}$",
    "$\\text{RR}_{UM=1}$", "$\\text{RR}_{UM|A}$",
    "$\\text{RR}_{UM=1|A}$"
  )

  tab <- xtable(tab,
    digits = 6,
    align = c("r", rep("c", 8))
  )

  list(
    x = tab,
    comment = FALSE,
    booktabs = TRUE,
    include.rownames = TRUE,
    add.to.row = addtorow,
    include.colnames = FALSE,
    table.placement = "H",
    sanitize.text.function = function(x) x
  )
}

# function to make figure 2
ratio_true_plot <- function(dat, loglin_val, qualint_val, tag) {
  dat %>%
    gather(key = BF, value = corrected, BF_um_strat_no0_NIE_correct, BF_true_NIE_correct) %>%
    mutate(
      ratio = corrected / NIE_true,
      BF = fct_recode(BF,
        "True" = "BF_true_NIE_correct",
        "Alternative" = "BF_um_strat_no0_NIE_correct"
      )
    ) %>%
    filter(
      loglin == loglin_val,
      qual_int %in% qualint_val
```

```
    ) %>%
    ggplot() +
    geom_density(aes(ratio, col = BF)) +
    scale_x_log10(limits = c(.1, 10)) +
    scale_y_continuous(expand = c(0, 0)) +
    facet_grid(NIE_direction ~ Bias,
      labeller = labeller(.rows = label_value, .cols = label_both)
    ) +
    theme_classic() +
    theme(
      legend.position = "bottom",
      strip.background.x = element_blank(),
      strip.background.y = element_rect(fill = "lightgrey", linetype = "blank")
    ) +
    labs(
      x = "Ratio of corrected NIE to true NIE (log scale)",
      subtitle = tag
    )
}

# function to make table 2
quant_rat <- function(dat) {

  p <- c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99)

  newdat <- dat %>%
    gather(key = BF, value = corrected, BF_um_strat_no0_NIE_correct, BF_true_NIE_correct) %>%
    mutate(
      ratio = corrected / NIE_true,
      BF = fct_recode(BF,
        "True" = "BF_true_NIE_correct",
        "Alternative" = "BF_um_strat_no0_NIE_correct"
      )
    ) %>%
    mutate(ratiopos = case_when(obs_true > 1 ~ ratio, TRUE ~ 1 / ratio))

  ratio_vals <- newdat %>%
    nest(-BF) %>%
    mutate(x = map(.$data, ~tibble(
      q = paste0("q", p),
      val = quantile(.$ratiopos, p)
    ))) %>%
    select(-data) %>%
    unnest() %>%
    spread(q, val)

  ratio_vals_bias <- newdat %>%
    nest(-Bias, -BF) %>%
    mutate(x = map(.$data, ~tibble(
      q = paste0("q", p),
      val = quantile(.$ratiopos, p)
    ))) %>%
    select(-data) %>%
    unnest() %>%
    spread(q, val)
```

```
  tab <- cbind(t(ratio_vals[, -1]), t(ratio_vals_bias[, -(1:2)]))

  addtorow <- list()
  addtorow$pos <- list(0, 0)
  addtorow$command <- c(
    "& \\multicolumn{2}{c}{Overall} & \\multicolumn{2}{c}{[1, 1.05)} & \\multicolumn{2}{c}{[1.05,
    "\\%ile & True & Alt. & True & Alt. & True & Alt. & True & Alt. & True & Alt. & True & Alt. &
  )
  rownames(tab) <- c("1st", "5th", "10th", "25th", "50th", "75th", "90th", "95th", "99th")

  tab <- xtable(tab,
    digits = 2,
    align = c("r", rep("c", ncol(tab)))
  )

  list(
    x = tab,
    comment = FALSE,
    booktabs = TRUE,
    include.rownames = TRUE,
    add.to.row = addtorow,
    include.colnames = FALSE,
    table.placement = "H",
    sanitize.text.function = function(x) x
  )
}

# function to make figure 2
resid_bias <- function(dat, loglin_val, qualint_val, tag) {

  dat %>%
    filter(
      BF_um_strat_no0_fail,
      loglin == loglin_val,
      qual_int %in% qualint_val
    ) %>%
    mutate(
      ratio = BF_um_strat_no0_NIE_correct / NIE_true,
      residual_bias = ratio * (obs_true < 1) + (1 / ratio) * (obs_true > 1),
      residual_quantile = cut(residual_bias, quantile(residual_bias, p = seq(0, 1, .1)), include.l
    ) %>%
    ggplot() +
    geom_point(aes(bias, residual_bias, col = residual_quantile)) +
    scale_y_log10(breaks = c(1, 2, 5, 10), labels = c(1, 2, 5, 10), expand = c(0, 0)) +
    scale_x_log10(breaks = c(1, 2, 5, 10), labels = c(1, 2, 5, 10), expand = c(0, 0)) +
    geom_abline(slope = 1, intercept = 0) +
    scale_color_brewer(name = "Residual bias decile", palette = "Spectral") +
    theme_classic() +
    theme(legend.position = "bottom") +
    labs(
      x = "Bias (log scale)", y = "Residual bias (log scale)",
      subtitle = tag
    )
}

# function to make table 3
```

```
resid_bias_add <- function(dat){

  tab_any <- dat %>%
    filter(BF_um_strat_no0_fail) %>%
    mutate(
      diff = BF_um_strat_no0_NIE_correct - NIE_true,
      residual_bias = diff * (obs_true < 1) + (-diff) * (obs_true > 1)
    ) %>%
    nest(-loglin) %>%
    mutate(x = map(.$data, ~tibble(
      q = paste0("q", seq(.1, 1, .1)),
      val = quantile(.$residual_bias, seq(.1, 1, .1))
    ))) %>%
    select(-data) %>%
    unnest() %>%
    spread(q, val) %>%
    mutate_if(is.numeric, round, 3) %>%
    t()

  tab_no_int <- dat %>%
    filter(BF_um_strat_no0_fail, qual_int == "No qualitative interaction") %>%
    mutate(
      diff = BF_um_strat_no0_NIE_correct - NIE_true,
      residual_bias = diff * (obs_true < 1) + (-diff) * (obs_true > 1)
    ) %>%
    nest(-loglin) %>%
    mutate(x = map(.$data, ~tibble(
      q = paste0("q", seq(.1, 1, .1)),
      val = quantile(.$residual_bias, seq(.1, 1, .1))
    ))) %>%
    select(-data) %>%
    unnest() %>%
    spread(q, val) %>%
    mutate_if(is.numeric, round, 3) %>%
    t()

  tab <- cbind(tab_any[,1], tab_no_int[,1], tab_any[,2], tab_no_int[,2])[-1,]

  addtorow <- list()
  addtorow$pos <- list(0, 0)
  addtorow$command <- c(
    "& \\multicolumn{2}{c}{Uniform probabilities} & \\multicolumn{2}{c}{Log-linear model} \\\\ \\c
    "\\%ile & No restriction & No qualitative interaction & No restriction & No qualitative intera
  )
  rownames(tab) <- c("10th", "20th", "30th", "40th", "50th", "60th", "70th", "80th", "90th", "Maxi

  tab <- xtable(tab,
                digits = 2,
                align = c("r", rep("c", ncol(tab)))
  )

  list(
    x = tab,
    comment = FALSE,
    booktabs = TRUE,
    include.rownames = TRUE,
```

```
    add.to.row = addtorow,
    include.colnames = FALSE,
    table.placement = "H",
    sanitize.text.function = function(x) x
  )
}

# function to create other values for paper (how many had interaction)
vals <- function(dat) {
  dat %>%
    group_by(loglin) %>%
    summarise(
      qual_int = "No interaction restriction",
      n_noqual = n(),
      resUM = mean(BF_um_strat_no0_fail)
    ) %>% rbind(
  dat %>%
    filter(qual_int == "No qualitative interaction") %>%
    group_by(loglin) %>%
    summarise(
      qual_int = "No qualitative restriction",
      n_noqual = n(),
      resUM = mean(BF_um_strat_no0_fail)
    )
    ) %>% arrange(loglin, desc(qual_int))
}
```

File: `make.R`

```
if (!require("drake")) {install.packages("drake"); library(drake)}

pkgconfig::set_config("drake::strings_in_dots" = "literals")

source("functions.R")

set.seed(1692)

N <- 25e5

plan <- drake_plan(

  full_dat = rbind(make_data(N = N, loglin = FALSE),
                   make_data(N = round(N / .75), loglin = TRUE)),

  res = sum_data(full_dat),

  vals_4_rmd = vals(full_dat),

  table1 = overall_res(res),

  table2a = full_dat %>%
    filter(loglin == "Unrestricted probabilities") %>%
    quant_rat(),
  table2b = full_dat %>%
    filter(loglin == "Unrestricted probabilities",
           qual_int == "No qualitative interaction") %>%
```

```r
    quant_rat(),
  table2c = full_dat %>%
    filter(loglin != "Unrestricted probabilities") %>%
    quant_rat(),
  table2d = full_dat %>%
    filter(loglin != "Unrestricted probabilities",
           qual_int == "No qualitative interaction") %>%
    quant_rat(),

  table3 = resid_bias_add(full_dat),

  compare_correctedA = ratio_true_plot(
    full_dat,
    "Unrestricted probabilities",
    c("No interaction restriction", "No qualitative interaction"),
    "eFigure 1A. Probabilities drawn from uniform distribution"
  ),
  compare_correctedB = ratio_true_plot(
    full_dat,
    "Unrestricted probabilities",
    "No qualitative interaction",
    "eFigure 1A. Uniform probabilities, restricted to no qualitative interaction"
  ),
  compare_correctedC = ratio_true_plot(
    full_dat,
    "Log-linear model",
    c("No interaction restriction", "No qualitative interaction"),
    "eFigure 1C. Log-linear model for M"
  ),
  compare_correctedD = ratio_true_plot(
    full_dat,
    "Log-linear model",
    "No qualitative interaction",
    "eFigure 1D. Log-linear model, restricted to no qualitative interaction"
  ),

  resid_plotA = resid_bias(
    full_dat,
    "Unrestricted probabilities",
    c("No interaction restriction", "No qualitative interaction"),
    "eFigure 2A. Probabilities drawn from uniform distribution"
  ),
  resid_plotB = resid_bias(
    full_dat,
    "Unrestricted probabilities",
    "No qualitative interaction",
    "eFigure 2B. Uniform probabilities, restricted to no qualitative interaction"
  ),
  resid_plotC = resid_bias(
    full_dat,
    "Log-linear model",
    c("No interaction restriction", "No qualitative interaction"),
    "eFigure 2C. Log-linear model for M"
  ),
  resid_plotD = resid_bias(
    full_dat,
```

```
    "Log-linear model",
    "No qualitative interaction",
    "eFigure 2D. Log-linear model, restricted to no qualitative interaction"
  ),
  eAppendix = rmarkdown::render(
    knitr_in("eAppendix.Rmd"),
    output_file = file_out("eAppendix.pdf"),
    quiet = TRUE
  )
)

make(plan)
```

File: `eAppendix.Rmd`

```
---
  title: "Tables and Figures for *Mediational E-values: Approximate Sensitivity Analysis for Unmea
output:
  pdf_document
---
```

```
if (!require("drake")) {install.packages("drake"); library(drake)}
if (!require("xtable")) {install.packages("xtable"); library(xtable)}
if (!require("ggplot2")) {install.packages("ggplot2"); library(ggplot2)}
if (!require("captioner")) {install.packages("captioner"); library(captioner)}
if (!require("knitr")) install.packages("knitr")

knitr::opts_chunk$set(
  message = FALSE,
  warning = FALSE,
  echo = FALSE,
  fig.align = "center"
)

loadd(vals_4_rmd)
```

```
do.call(print, readd(table1))
```

```
print(readd(compare_correctedA))
```

```
print(readd(compare_correctedB))
```

```
print(readd(compare_correctedC))
```

```
print(readd(compare_correctedD))
```

```
do.call(print, readd(table2a))
```

```
do.call(print, readd(table2b))
```

```
do.call(print, readd(table2c))
```

```
do.call(print, readd(table2d))
```

```
print(readd(resid_plotA))
```

```
print(readd(resid_plotB))
```

```
print(readd(resid_plotC))
```

```
print(readd(resid_plotD))
```

```
do.call(print, readd(table3))
```

REFERENCES

1. Ding P, VanderWeele TJ. Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika.* 2016;103:483–490.

2. Greenland S. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology.* 2003;14:300–306.

3. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: Introducing the E-value. *Ann Intern Med.* 2017;167:268–275.

4. VanderWeele TJ, Ding P, Mathur M. Technical considerations in the use of the E-value. *Journal of Causal Inference.* 2019;in press.