**Appendix 3. Collider bias**

|  |  |  |  |
| --- | --- | --- | --- |
| **DAG** | **SCC model equivalent of DAG** | **Parameter inputs from simulation** | **True**  **effects** |
|  |  | Prevalence of exogenous variables in simulation:  pX1=0.5  pX2=0.1  pX3=0.5  pZ1=0  pZ5=0.1  pZ6=0.4  pAPOE4=0.25  pLowEd=0.7 | RR = 1.0  RD = 0.0 |

**I. Estimates in the full simulated sample**

We first compute the crude risk ratio and risk difference in the full simulated sample (“full sample”) as follows:

The numerator and denominator quantities for this formula can be derived from the relevant causal components in Figure 5B and their parameter values specified in the simulation:

`

= P(

= P(,

= *#P(Z1) is eliminated from formula because its prevalence is 0*

= 0.1 + (0.25\*0.40) – (0.1\*0.25\*0.40)

= 0.19

= P((

= P(,

= 0.1 + (0.25\*0.40) – (0.1\*0.25\*0.40)

= 0.19

Thus:

In Table 4 we see that in our simulations, the prevalence of doomed response types in the full simulated sample is also 0.19 among individuals exposed and unexposed to low education. Again, for ease of demonstration, we simulate under the sharp null; thus, the prevalence of the outcome in each exposure group is entirely determined by doomed response types i.e., P(Dementia|Low Education) = 0.19 and analogously the prevalence of doomed risk type in Table 4 is 0.19. Thus, in the full simulated sample there is exchangeability, and the crude RR and RD equal the true values dictated by the data generating mechanism.

**II. Estimates in the analytic sample**

The crude risk ratio and risk difference in the analytic sample (the target population for our estimand) are computed as follows, where S indicates being in the analytic sample and indicates not:

-

Because High Education (or, equivalently, being *unexposed* to low education) and APOE4 are both causes of study participation, restricting to study participation creates a dependency between Low Education and APOE4 *in the analytic sample*. When we compute crude measures in the analytic sample, we must consider this dependency. These quantities can be derived from the relevant component causes in Figure 5B:

=

=

,

=

= ,

Thus, we must also compute the following probabilities:

*i.e., the probability you have APOE4 given you are in the study sample and have low education*

*i.e., the probability you have APOE4 given you are in the study sample and have high education*

We can compute these probabilities using Bayes Theorem where quantities in the numerator and denominator are derived from derived from the relevant component causes in Figure 5B:

*Where*:

= *0.1 + 0.5 – 0.1\*0.5*

= *0.55*

=

*#Because low/high education and APOE4 are independent in the full simulated sample (i.e., before selection)*

= *0.1 + (0.25\*0.5) – 0.1\*0.25\*0.5*

= *0.2125*

*,*

*Thus*:

Intuitively, this probability of APOE4 within the analytic sample makes sense. Having high education or APOE4 are two possible ways that an individual makes it into the analytic sample. If you are in the analytic sample and you do *not* have high education, this should mean that, within the analytic sample, your probability of having APOE4 is higher than 0.25.

We can likewise perform these calculations for those with high education (i.e., not low education):

*Where*:

= )

*= 0.5 + 0.1 + 0.5 – (0.1\*0.5) – (0.1\*0.5) – (0.5\*0.5) + (0.5\*0.5\*0.1)*

*= 0.775*

=

*#Because low/high education and APOE4 are independent in the full simulated sample (i.e., before selection)*

= *0.5+0.1+(0.25\*0.5)–(0.5\*0.1)–(0.1\*0.125)–(0.5\*0.125)+(0.5\*0.1\*0.125)*

= *0.60625*

*Therefore*:

We now have what we need to compute the risk ratio in the analytic sample:

– (0.10 \* 0.6471 \* 0.4)

**= 0.33296**

=

– (0.10 \* \* 0.4)

**= 0.2151**

-

= 0.333 – 0.215 = 0.118

**= 0.12**

Again, because we simulate under the sharp null, these outcome probabilities are entirely determined by the doomed response types. We see in Table 4 that, indeed, within the analytic sample, the prevalence of dementia (i.e., of doomed response types) among those with low education is 33.3% and among those with high education (i.e., without low education) is 21.6%. Thus, there is non-exchangeability due to collider bias within the analytic sample, and this bias is reflected in the crude estimates.

**III. Correcting for collider bias in study sample estimates through standardization**

In this example, if we knew both high education and APOE4 were the causes of study participation (the collider), we could compute a risk ratio for the effect of education on dementia that corrected for collider bias by standardizing to the distribution of the collider-inducing variable, APOE4, in the analytic sample and obtain the analytic sample adjusted (i.e., standardized) risk ratio and risk difference as follows:

*=*

*=*

Again, we can derive the probabilities of interest from the component causes in the figure as follows (and because the prevalence of Z1 was set to 0, we can eliminate these terms, where relevant, from the equations below):

*Where*:

=

= *0.10 + 0.40 – 0.10\*0.40*

= *0.46*

=

= *0.10*

= *0.10 + 0.40 – 0.10\*0.40*

= *0.46*

= *0.10*

*And:*

*Where:*

*,*

*#Because low/high education and APOE4 are independent in the full simulated sample (i.e., before selection)*

= (0.3\*0.5) + 0.1 + 0.5 – (0.3\*0.5\*0.1) – (0.1\*0.5) – (0.3\*0.5\*0.5) + (0.3\*0.5\*0.1\*0.5)

= 0.6175

*=*

= (0.3\*0.5)+0.1+(0.25\*0.5)–(0.3\*0.5\*0.1)–(0.1\*0.25\*0.5)–(0.3\*0.5\*0.25\*0.5)+(0.3\*0.5\*0.1\*0.25\*0.5)

=

0.25

*Thus:*

*and*

We have now computed all quantities required to calculate the standardized risk ratio and risk difference in the analytic sample:

After standardization, the RR and RD in the analytic sample now equal the RR and RD in the full simulated sample and, in both, the RR and RD are equal to the true value of 1.0 and 0.0, respectively.