**Appendix 2. Confounding**

|  |  |  |  |
| --- | --- | --- | --- |
| **DAG** | **SCC model equivalent of DAG** | **Parameter inputs from simulation** | **True**  **effects** |
|  |  | Prevalence of exogenous variables in simulation:  pQ1=0.8  pQ2=0.2  pZ1=0  pZ3=0.1  pZ4=0.15  pLowSES=0.7 | RR = 1.0  RD = 0.0 |

**I. Estimate crude measures**

We first compute the crude risk ratio and risk difference as follows:

)

The numerator and denominator quantities can be derived from the relevant component causes in Figure 4B and their parameter values specified in the simulation.

=

=

,

=

=

= ,

Because confounding creates a dependency between Low SES (confounder) and Low Education (exposure) this must be accounted for when computing the crude risk ratio and risk difference for dementia (in contrast to the causation example in Appendix 1 where exposure was exogenous and calculation of the crude measures followed directly from the simulation parameter inputs with no additional computation required).

We can compute probabilities for the dependency between Low SES and Low Education using Bayes Theorem where quantities in the numerator and denominator are derived from the relevant component causes in Figure 4B:

*Where*:

, because

= 0.8 + 0.2 – 0.8\*0.2

= 0.84

= 0.16

= (0.7\*0.8) + (0.2) – (0.7\*0.8\*0.2)

= 0.648

= 1 – 0.648

= 0.352

= 0.7

*Thus*:

= 0.9074

= 0.3182

We now have all of the necessary probabilities to compute the crude measures in the sample. In the numerator we include all ways in which someone with low education could get dementia (making sure to account for the dependency between low SES and low education in the formula). The denominator includes all ways in which someone without low education could get dementia. Note that because the prevalence of Z1 was set to 0, we can eliminate these terms, where relevant, from the equations below.

= 0.10 + 0.9074\*0.15 – 0.10\*0.9074\*0.15

= 0.222499

=

= 0.10 + 0.3182\*0.15 – 0.10\*0.3182\*0.15

= 0.142957

*Thus:*

= 0.222499/0.142957

= 1.5564

= 0.222499 – 0.142957

= 0.07954

Note that the 22.2% prevalence of dementia among those with low education and the 14.3% prevalence of dementia among those without low education matches the distribution of the doomed response types within exposure groups in Table 3. This makes sense: because we simulated under the sharp null, the only response types contributing to someone getting the outcome in our sample are the doomed. This imbalance of doomed response types between exposure shows there is non-exchangeability in the sample (due to confounding), which creates a crude risk ratio and risk difference in the sample that do not match the truth (no effect).

**II. Correcting for confounding bias through standardization**

To correct for confounding, we need to standardize by the distribution of the confounder, low SES, which requires the calculation of a few more probabilities. We will use standardization to compute the causal risk ratio. The formula for the standardized risk ratio is 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):

=

= *0.10 + 0.15 – 0.10\*0.15*

= *0.235*

= *,*

=

= *0.10*

= *,*

= *0.10 + 0.15 – 0.10\*0.15*

= *0.235*

= *,*

= *0.10*

*Thus:*

After standardization, the RRadj and RDadj are equal to the true value of 1.0 and 0.0, respectively.