Supplementary information

Let us consider an initial cohort of 30,000 participants without rheumatoid arthritis (RA) at baseline and assume that there are four risk factors involved in the aetiology of RA incidence (E1) or progression (E2), namely the risk factor of interest (R) and three other unmeasured risk factors (U1, U2, and U3). We set the marginal frequencies of exposure R in the 30,000 participants to 0.33, and ensure that the occurrence of each unknown factor is independent of the risk factor of interest (Table 1). For instance, the likelihood that U1=1 is 0.02, whether R=1 or R=0.

Table 1 | Joint distribution of risk factors R, U1, U2, and U3 in 30,000 individuals without RA at baseline

Risk factors	U2=1 U1=1 U0=1	U2=1 U1=1 U0=0	U2=1 U1=0 U0=1	U2=0 U1=1 U0=1	U2=1 U1=0 U0=0	U2=0 U1=1 U0=0	U2=0 U1=0 U0=1	U2=0 U1=0 U0=0	Total
R=1	100	20	20	20	80	80	80	9,600	10000
R=0	200	40	40	40	160	160	160	19,200	20000
Total	300	60	60	60	240	240	240	28,800	30000

Analogous to Smits' article[1] (following contemporary disease causation theory), we assume that the development of disease is the result of the combined action of multiple components. We stipulate that causation of E1 requires the presence of at least two risk factors (for example, R=1, U1=1, U2=0), whereas E2 (that is, a sequelae event of E1) requires at least three risk factors (for example, R=1, U1=1, U2=1). We opted for the latter condition because more causal factors may be involved in RA progression than incidence, but requiring the same number of factors for both incidence and progression would lead to the same conclusion as did Smits' article.[1] Table 2 lists all combinations of values of risk factors sufficient for E1 and E2 to develop.

Table 2 | Combinations of risk factors sufficient to cause E1 (RA incidence) and E2 (RA progression)

R	U0	U1	U2	E1	E2
1	1	1	1	Y	Y
1	0	1	1	Y	Y
1	1	0	1	Y	Y
1	1	1	0	Y	Y
1	0	0	1	Y	N
1	0	1	0	Y	N
1	1	0	0	Y	N
1	0	0	0	N	N
0	1	1	1	Y	Y
0	0	1	1	Y	N
0	1	0	1	Y	N
0	1	1	0	Y	N
0	0	0	1	N	N
0	0	1	0	N	N

0	1	0	0	N	N
0	0	0	0	N	N

Application of this scheme to our study population means that 720 individuals (2.4%) develop incident RA (Table 1, italicized numbers). Risks of developing incident RA among individuals with R=1 and R=0 are 0.04 (400/10000) and 0.016 (320/20000), respectively, and the RR is 2.5 (0.04/ 0.016).

Table 3 | Joint distribution of risk factors R, U1, U2, and U3 among who developed incident RA (E1)

Risk factors	U2=1 U1=1 U0=1	U2=1 U1=1 U0=0	U2=1 U1=0 U0=1	U2=0 U1=1 U0=1	U2=1 U1=0 U0=0	U2=0 U1=1 U0=0	U2=0 U1=0 U0=1	U2=0 U1=0 U0=0	Total
R=1	100	20	20	20	80	80	80	0	400
R=0	200	40	40	40	0	0	0	0	320
Total	300	60	60	60	80	80	80	0	720

In Table 3, frequencies of combinations of R, U1, U2, and U3 are displayed for RA progression among the 720 individuals who developed RA. Now, the risk factor of interest (R) has become inversely associated with each unknown risk factor (U1, U2, and U3). For instance, if R=1, the likelihood that U1=1 is 0.35, whereas it is 0.88 if R=0. Because E2 requires at least three risk factors (Table 3, italicized numbers), the crude RR for R=1 vs. R=0 is 0.64 ((160/400)/(200/320)). This inverse crude RR is the result of extreme bias caused by the introduction of a negative association between R and the other risk factors. The real (causal) RR can be calculated by means of a counterfactual approach.[2] That is, instead of comparing the observed risks of RA patients with R=1 and R=0, the observed risk among RA patients with R=1 is compared with the hypothetical risk that would apply if the RA patient would have R=0 instead of R=1. Computed in this way, the RR amounts to 1.6 because its denominator is 0.25 (100/400, Table 4) instead of 0.63 (200/ 320, Table 3).

Table 4 | Counterfactual condition among who developed incident RA (E1) and R=1

Risk factors	U2=1 U1=1 U0=1	U2=1 U1=1 U0=0	U2=1 U1=0 U0=1	U2=0 U1=1 U0=1	U2=1 U1=0 U0=0	U2=0 U1=1 U0=0	U2=0 U1=0 U0=1	U2=0 U1=0 U0=0	Total
R=1	100	20	20	20	80	80	80	0	400
Counterfactual R=0	100	20	20	20	80	80	80	0	400

Supplementary References:

1. Smits, L. J. et al. Index event bias-a numerical example. J Clin Epidemiol **66**, 192-6 (2013).
2. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol **31**, 422e9 (2002).