**Text S1. Correction of coefficients *a1* and *a2* for unmatched case-control study design**

**Correction of coefficient *a1***.

The regression coefficient *a1*, of *M1*-*X* association while regressing *M1* on *X*, could be biased. Here we provide the non-linear estimating equation approach to correct the bias.

The OR for the association between the SNP *X* and the mediator *M1* (exp(*a1*)) can be expressed as below

. (7)

In this scenario, *Eji* is the expected number of individuals in the sample with *M1*=*j* and *X*=*i* and is given as

, where *i =* 0, 1, and 2, and *j*, *r* = 0, 1. The conditional probability *pji|r* is written as

;

for *i =* 0, 1, and 2, and *j*, *r* = 0, 1.

The probabilities *pi*, *i* = 0, 1, and 2 represent the genotype frequencies, and *q1* represents the prevalence of the disease. The conditional probabilities *pr|ji* and *pj|i* are given as functions of regression coefficients:

Pr(*M1*=*j*|*X*=*i*) =  and Pr(*Y*=*r*|*X*=*i*, *M1*=*j*) = , (8)

where *a0*, *c0*, and *a1* are unknown coefficients of interest. Based on the conditional probabilities given above, we can write similar equations for the estimated prevalences of the disease and the mediator *M1* as follows:

, (9)

. (10)

Note that the bias in estimating *a1* was not only caused by the disproportionate number of individuals with mediator *M1* but was also influenced by the disproportionate number of individuals with mediator *M2* in the primary disease case-control study. We, therefore, evaluate the regression coefficients *b1* and *c'* using logistic regressions in Equation (8) in order to reduce the bias due to the mediator *M2*. Given the biased coefficient *a1* estimated based on Equation (1), the solution to Equations (7), (9) and (10) will give us the corrected estimate for coefficient *a1* for the association between the SNP and the mediator *M1*, which is denoted as . We also have tried to use an alternative approach to correcting the coefficient *a1* based on the estimates of *d1* and *a2*, in which the mediator *M2* is treated as the primary disease, as described in our previous study of secondary phenotype [[1](#_ENREF_1)]. However, we found that the alternative approach was less efficient compared with the one we have described above, based on the simulation studies. Therefore, in the current article, we only reported results based on the approach that uses the coefficients *b1* and *c'*.

**Correction of coefficient *a2***.

The regression coefficient *a2*, of *M2*-*X* association while regressing *M2* on *M1* and *X*, could also be biased. A similar non-linear estimating equation approach can be applied to correct the bias. The OR for the association between the SNP *X* and the mediator *M2* (exp(*a2*)) can be expressed as below

, (11)

where *Eki* is the expected number of individuals in the sample with *M2*=*k* and *X*=*i* and is given as

,

where *i =* 0, 1, and 2, and *k*, *r* = 0, 1. The conditional probability *pki|r* is written as

;

for *i =* 0, 1, and 2, and *k*, *r* = 0, 1.

The probabilities *pi*, *i* = 0, 1, and 2 represent the genotype frequencies, and *q1* represents the prevalence of the disease. The conditional probabilities *pr|ki* and *pk|i* are given as functions of regression coefficients:

Pr(*M2*=*k*|*X*=*i*) =  and Pr(*Y*=*r*|*X*=*i*, *M2*=*k*) = ,

where *b0*, *c0*, and *a2* are the unknown coefficients of interest. Based on the conditional probabilities given above, we can write similar equations for the estimated prevalences of the disease and the mediator *M2* as follows:

, (12)

. (13)

Given the biased coefficients *b2*, *c'*, and *a2* estimated based on Equations (1)~(3), the solution to Equations (11), (12), and (13) will give us the corrected estimate for coefficient *a2* for the association between the SNP and the mediator *M2*, which is denoted as .

Reference List

1. Wang J, Shete S (2011) Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. Genet Epidemiol 35: 190-200.