Further development of the case-only design for assessing gene–environment interaction: evaluation of and adjustment for bias

Nicolle M Gatto,1† Ulka B Campbell,1† Andrew G Rundle1,2 and Habibul Ahsan1,2

Accepted 13 July 2004

Background The case-only study for investigating gene–environment interactions provides increased statistical efficiency over case-control analyses. This design has been criticized for being susceptible to bias arising from non-independence between the genetic and environmental factors in the population. Given that independence is critical to the validity of case-only estimates of interaction, researchers frequently use controls to evaluate whether the independence assumption is tenable, as advised in the literature. Our work investigates to what extent this approach is appropriate and how non-independence can be accounted for in case-only analyses.

Methods We provide a formula in epidemiological terms that illustrates the relationship between the gene–environment association measured among controls and the gene–environment association in the source population. Using this formula, we conducted sensitivity analyses to describe the circumstances in which controls can be used as proxy for the source population when evaluating gene–environment independence. Lastly, we generated hypothetical cohort data to examine whether multivariable modelling approaches can be used to control for non-independence.

Results Our sensitivity analyses show that controls should not be used to evaluate gene–environment independence in the population, even when the baseline risk of disease is low (i.e. 1%), and the interaction and independent effects are moderate (i.e. risk ratio = 2). When the factors are associated, it is possible to remove bias arising from non-independence using standard statistical multivariable techniques in case-only analyses.

Conclusions Even when the disease risk is low, evaluation of gene–environment independence in controls does not provide a consistent test for bias in the case-only study. Given that control for non-independence is possible when the source of the non-independence can be conceptualized, the case-only design may still be a useful epidemiological tool for examining gene–environment interactions.

Keywords Bias (epidemiology), case-only design, environment, epidemiological methods, genes, interaction, research design

As the name implies, the entire sample used for a case-only analysis of gene–environment interaction consists of people with the disease (cases). Each person with disease is coded as positive or negative for a genetic factor (G) and an environmental factor (E). The case-only odds ratio (OR) is derived from the cross-categorization of the study sample on G and E status. This OR is the odds of E given the presence of G divided by the odds of E given the absence of G. As always, the
OR is reversible,\(^1\) such that the case-only OR can also be interpreted as the odds of \(G\) given the presence of \(E\) divided by the odds of \(G\) given the absence of \(E\).\(^2\) The premise behind the case-only study is that this OR can be interpreted as the multiplicative interaction between \(G\) and \(E\) in causing disease.

Epidemiologists were initially enthusiastic about the case-only study to detect gene–environment interactions. This innovative design offered an opportunity to study interaction that was more statistically efficient than the analogous case-control study and was not subject to common biases arising from control selection.\(^3\) However, the validity of the case-only study hinges on one assumption—that the genetic and environmental factors of interest are independent of one another. More recently, concerns over violations of this assumption have caused the initial enthusiasm to wane.\(^4\)–\(^6\)

It has been shown that the validity of case-only estimates of gene–environment interaction is highly susceptible to bias arising from non-independence between \(G\) and \(E\).\(^4\) Many researchers advocate the use of non-diseased controls selected from case-control studies to verify the independence assumption.\(^2\)–\(^7\)–\(^10\) This examination of independence assumes that the controls provide an appropriate proxy for the gene–environment association in the population that gave rise to the cases. In practice, some authors have performed case-only analyses after observing \(G\)–\(E\) independence in controls,\(^11\),\(^12\) while others have rejected case-only analyses upon finding \(G\)–\(E\) associations in controls.\(^13\)–\(^15\) We will show that even when the disease risk is relatively low, controls may not provide a good approximation of the gene–environment association in the underlying population. Given the susceptibility to bias and that evaluation of independence may be more problematic than previously thought, it may seem that the case-only study is of limited utility. However, evaluation of and adjustment for \(G\)–\(E\) non-independence warrants further consideration.

For many metabolic polymorphisms, associations between genetic and environmental factors seem unlikely since these genes rarely produce symptoms and individuals have little opportunity to learn their genetic status. Thus, behavioural modification of environmental factors based on genetic status is unlikely for these types of genes. In fact, examples of independence between genetic and environmental factors appear in the recent literature on Mendelian randomization.\(^16\)

For circumstances when gene–environment associations are likely, it has been mentioned that the bias in case-only estimates caused by gene–environment associations can be controlled using stratification.\(^17\) Several studies have controlled for third variables in case-only analyses without any explanation.\(^11\),\(^18\)–\(^20\) It is not clear whether the investigators were attempting to remove bias due to gene–environment associations, or simply including variables in multivariable models because they are potential confounders of main effects. Thus far, the concept and procedures to control for non-independence between genetic and environmental factors are undeveloped, and there are no methodological guidelines for what types of third variables should be included in case-only analyses.

Despite criticisms of the case-only study, several applications of the case-only design can be found in recent publications.\(^11\),\(^12\),\(^14\)–\(^15\),\(^18\)–\(^22\) For this reason, and because of the potential advantages offered by the design, further development of the case-only design is warranted. Here, we show that the case-only study may still prove useful in some circumstances, and may in fact, lead to more valid estimation of gene–environment interaction than the analogous case-control study.

In this paper, we: (1) provide the conceptual basis for the case-only OR and its relationship to analyses of cohort data using simple epidemiological terms; (2) demonstrate the extent to which non-diseased controls can be used to evaluate gene–environment independence in the population from which the cases arose; and (3) show that control for non-independence between the genetic and environmental factors is possible using standard modelling techniques. In short, we provide a clearer conceptual basis for understanding how bias arises in the case-only study and demonstrate how such bias might be removed.

### The case-only study design

The method of identifying risk factors using a series of cases was first described by Aalen et al.\(^23\) and further developed by Prentice et al.\(^24\) The case-only design to assess gene–environment interaction was presented by Piegorsch et al.\(^25\) as derived from a traditional case-control study, in which controls are sampled from the non-diseased at the end of the study period.\(^25\) If \(G\) and \(E\) are independent among the controls, the case-only analysis yields an estimate of interaction that is based on cumulative-incidence ORs.\(^25\) The authors noted that this estimate of interaction will approximate an interaction estimate based on risk ratios (RRs) when the disease is rare. The interaction estimate based on RRs can be conceptualized as the estimate that would be obtained from a hypothetical cohort study of the source population. In two subsequent papers, Yang et al.\(^26\) and Schmidt and Schaid\(^2\) demonstrated that the approximation of the case-only OR to the interaction estimate based on RRs does not require a rare disease assumption. In fact, they noted that when \(G\) and \(E\) are independent in the source population, the case-only OR is equivalent to the interaction estimate based on RRs, regardless of disease risk.

To further the development of the case-only design, we provide a conceptual description of the mathematical relationship between the case-only OR and the estimate of multiplicative interaction based on RRs. Our approach, which begins with a description of the underlying cohort and ends with the case-only estimate of multiplicative interaction measured with RRs, incorporates the prior work of Yang et al.\(^26\) and Schmidt and Schaid.\(^2\) Understanding the relationship between the case-only OR and the measure of interaction based on RRs is essential for understanding the nature of bias in the case-only study.

Conceptually, the interaction between \(G\) and \(E\) refers to the extent to which the joint effect of the two factors on disease (\(D\)) differs from the independent effects for both factors. The joint effect of \(G\) and \(E\) is the effect on \(D\) due to the presence of both factors. The independent effect of each of the factors is its effect on \(D\) in the absence of the other. Multiplicative interaction is assessed by comparing the joint effect with the product of the independent effects. For instance, if the independent effect of \(G\) equals 3 and the independent effect of \(E\) equals 2, then we would expect the joint effect of \(G\) and \(E\) to be 6 if there is no multiplicative interaction. If \(G\) and \(E\) do interact to cause \(D\), then we would expect the joint effect to be something other than 6, the product of their independent effects. This measure of multiplicative interaction is equivalent to stratified analyses (i.e. the \(E\)–\(D\) association in \(G+\) divided by the \(E\)–\(D\) association in \(G−\)).
Population N is a hypothetical population that serves as the basis for a cohort study. The members of this population are categorized according to their G and E status and followed for incident disease. Assuming these two exposures are dichotomous, individuals are classified according to four exposure categories. This cohort study is illustrated in Figure 1a. Those positive for both G and E are represented by N11. Likewise, N10 represents the sub-population that is positive for G and negative for E. N01 corresponds to the sub-population that is negative for G and positive for E, and N00 corresponds to the sub-population that is negative for both. As seen in Figure 1a, these individuals can be further classified according to their D status. The cells labelled a1 to d1 represent those with G and the cells labelled a2 to d2 represent those without G.

To demonstrate the calculation of the interaction between G and E using RRs, we have organized the cells a1 to d2 into the tables presented in Figure 1b. The joint and independent effects of G and E, and the calculation of the gene–environment interaction are also presented. The baseline risk of disease, or the risk of disease attributable to factors other than G and/or E, is represented by \( \frac{c_2}{N_{00}} \). The RR for the joint effect of G and E (RR_{GE}) is calculated by comparing the risk of disease among those who are G + E+ to this baseline risk. Similarly, the RR for the independent effect of E (RR_E) compares the risk of disease among those who are G – E+ to the baseline risk. The RR for the independent effect of G (RR_G) compares the risk of disease among those who are G+ E – to the baseline risk. The RR for the joint effect of G and E (RR_{GE}) is calculated by dividing the product of the independent effects of G, RR_G, and E, RR_E on disease.

The joint effect of G and E

<table>
<thead>
<tr>
<th>G+E+</th>
<th>G+E−</th>
<th>G−E+</th>
<th>G−E−</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>b1</td>
<td>c2</td>
<td>d2</td>
</tr>
</tbody>
</table>

\[ \text{RR}_{GE} = \frac{a_1/N_{11}}{c_2/N_{00}} \]

The independent effect of G (among E-)

<table>
<thead>
<tr>
<th>D+</th>
<th>D−</th>
</tr>
</thead>
<tbody>
<tr>
<td>G+</td>
<td>c1</td>
</tr>
<tr>
<td>G−</td>
<td>c2</td>
</tr>
</tbody>
</table>

\[ \text{RR}_G = \frac{c_1/N_{10}}{c_2/N_{00}} \] \[ \text{RR}_E = \frac{a_2/N_{01}}{c_2/N_{00}} \]

The joint and independent effects of G and E measured by ORs can also be derived. Just as the joint and independent effects measured by RRs are used to calculate the G × E_{RR}, the joint and independent effects measured by ORs are used to calculate the interaction estimate based on ORs (G × E_{OR}). These ORs can be calculated from a cumulative incidence case-control study (i.e. ‘traditional’ case-control study), in which all of the diseased people from the underlying cohort comprise the cases and a random sample of the non-diseased people taken at the end of the follow-up period comprise the controls. The G × E_{OR} derived from this case-control study is the same G × E_{OR} that would have been obtained had all of the non-diseased been used.

\[ \text{GxE}_{RR} = \frac{\text{RR}_{GE}}{\text{RR}_G \times \text{RR}_E} = \frac{a_1/N_{11}}{c_2/N_{00}} \times \frac{c_1/N_{10}}{c_2/N_{00}} \]

\[ \frac{a_2/N_{01}}{c_2/N_{00}} \]

Figure 1 Calculation of gene–environment interaction (G × E_{RR}) using risk ratios (RR). The G × E_{RR} is the joint effect of the gene (G) and environmental factor (E) on disease (D), \( \text{RR}_{GE} \), divided by the product of the independent effects of G, \( \text{RR}_G \), and E, \( \text{RR}_E \) on disease.
The case-only study can be conceptualized as an analysis of all individuals with disease illustrated in Figure 1. In Figure 2a, we present the calculation of the case-only OR using the notation from Figure 1. Figure 2 demonstrates how the case-only OR is equivalent to the G × E RR when term II equals one, i.e. when there is no association between G and E in the population. The components of term II represent the number of people in each of the four exposure categories (G + E+, G + E−, G − E+, G − E−). Using these components, a 2 × 2 table that relates G and E in the total population can be constructed (Figure 2b). When G and E are independent in the source population, the OR relating G and E (G–E OR) calculated from this table is equal to one.† Therefore, the case-only OR ([a1c2/ a2c1]) is equivalent to the G × E RR when term II equals one. Consequently, independence between G and E in the total population is required for equivalence between the case-only OR and the G × E RR (Figure 2c). When this requirement is met, the case-only OR should be interpreted as the multiplicative interaction between G and E in causing D (i.e. RRGE / (RRG × RRERR)).

Problems with using controls to assess non-independence

In practice, case-only analyses have been most commonly performed using the case series of an existing case-control study. This approach takes advantage of the statistical efficiency offered by case-only analyses and allows evaluation of independence between G and E in the controls, as is advised in the literature.2,7–10

† When the G–E RR in the population is equal to one, the G–E OR will also be equal to one. The G–E OR is used to test for non-independence in the population because it is this ratio that must equal one for the case-only OR to be mathematically equivalent to the G × E RR.
In the next example (Figure 4), the G–E OR in the total cohort is 2.0, a violation of the independence requirement of the case-only design. The baseline risk of disease is 4% and the overall risk of disease is 8%. The RR for the independent effect of G is 6 [RRG = (132/548)/(538/13 452)] and the RR for the independent effect of E is 2.7 [RRE = (599/5548)/(538/13 452)]. The interaction estimate based on RRs from the underlying cohort study is 1.

Again, the non-diseased provide a poor approximation to the G–E independence in the population. Here, an investigator using the controls to approximate the G–E association in the population would erroneously presume that the case-only analysis was valid, concluding that there is a G × E interaction when none exists among the corresponding RRs.

These examples show that the use of controls to evaluate G–E interaction in the population is problematic. Even when the disease risk is low, the G–E OR in the controls may not accurately reflect the G–E OR in the source population. Therefore, the observation of independence or non-independence between G and E among controls does not provide a consistent test for bias in case-only analyses.

Below, we provide a formula to describe the circumstances in which the OR for the gene–environment association derived from controls can be used to approximate the gene–environment OR in the source population (Equation 1).

\[
G-E\text{OR}_{\text{at}\ D-} = G-E\text{OR}_{\text{at}\ pop} \times \left( \frac{1}{p(D|G-E)} - 1 \right) \times \left( \frac{1}{p(D|G-E)} - RR_{GE} \right)
\]

Figure 3 Numeric example in which the case-only odds ratio (OR) is equal to the gene–environment interaction based on risk ratios (G × E) since the OR relating the gene (G) and the environmental factor (E) in the population (G–E OR) is 1. The G–E OR in the controls is a poor proxy for the G–E OR in the population.

Figure 4 Numeric example in which the case-only odds ratio (OR) is not equal to the gene–environment interaction based on risk ratios (G × E) since the OR relating the gene (G) and the environmental factor (E) in the population (G–E OR) is 2. The G–E OR in the controls is a poor proxy for the G–E OR in the population.

All of the factors that determine the relationship between the G–E OR in the controls and the G–E OR in the source population are visible in Equation 1. They are: the baseline risk of disease \([p(DG - E -)]\), the RRs representing the independent effects of the genetic and environmental factors \([RR_G\text{ and } RR_E\text{, respectively}])\), and their joint effect on disease \([RR_{GE}]\). Alternatively, the relationship between the G–E OR in controls and the G–E OR in the population can be expressed in terms of the disease risk in each of the four exposure categories \((G + E+, G + E-, G - E+, G - E-)\) by factoring out the baseline risk. Equation 1 was adapted from previous research, in which two of the authors of this report (UBC and NMG) derived a formula that describes the mathematical relationship between interaction estimates based on ORs and RRs. In fact, the same factors that cause divergence between the G–E OR in the controls and the G–E OR in the underlying population are the same factors that cause the divergence between the G × EOR and the G × ERR. Therefore, the circumstances in which ORs and RRs yield materially different estimates of interaction are the same circumstances in which controls cannot be used to estimate the G–E OR in the population.

To illuminate some situations in which valid case-only results would be rejected based on an observed G–E association in controls, we used Equation 1 to perform sensitivity analyses. Specifically, we assessed the impact of the baseline risk of disease and the independent effect of G on the G–E OR in non-diseased controls under G–E independence in the source population. For these analyses, the independent effect of E and the G × ERR were each equal to 2.

Figure 5 highlights situations in which the baseline risk of disease ranges from 0.1% to 6%. As illustrated in the figure, the
G–E OR in the controls is a good approximation to the G–E OR of 1 in the population when either the baseline risk of disease is low (baseline risk is 0.1%) or when the baseline disease risk is close to 1% and the independent effect of G is moderate (RR_G < 2.5). However, as the baseline risk of disease approaches 3%, the G–E OR in the controls begins to appreciably diverge from the G–E OR in the population. For example, when the independent effect of G is 2.3 and the baseline risk is 3%, the G–E OR in the controls is 0.8. This finding among controls can have important implications for interpretation. A researcher may conclude that the assumption of independence required for a valid case-only analysis is not met and reject the case-only design, despite G–E independence in the population. For minimal increases in the baseline risk of disease, the approximation becomes increasingly worse.

Conventional epidemiological wisdom suggests that as long as the disease is relatively rare (i.e. overall disease risk is at or below 5%), the OR is a good approximation for the RR. However, this rule of thumb does not necessarily apply when using controls to assess independence in the population. As our numerical example illustrates, controls may not provide a good approximation for the G–E association in the population, even when the disease risk is 5% (Figure 3). Furthermore, Equation 1 shows that the risk of disease is only one of several factors that influence whether controls can be used to evaluate G–E independence in the population.

Although the case-only design is not specific to cancer epidemiology, it has been most often used in this context. The baseline risks of disease used in our analyses are higher than is typical for cancers, and indeed our results show that evaluation of G–E associations in controls may be satisfactory when the disease risk is less than 0.1%. However, the prevalences used in our analyses are applicable to studies of precursor lesions such as colorectal adenomas, Barrett’s oesophagus, and benign breast disease. In addition, researchers are increasingly studying intermediate biomarkers as endpoints, where the risk of the endpoint may be close to 50%.32

Although there are some situations in which the assessment of G–E independence in the controls will be a good proxy for assessment in the underlying population, recognizing these situations in practice requires assurance of many unknown quantities (e.g. the risk of disease among those without the genetic or the environmental factor, the underlying magnitude of interaction, etc.). Even when it is possible to estimate these quantities, minimal errors at critical thresholds can cause a researcher to come to the incorrect conclusion about whether controls may be safely used to evaluate independence. Consequently, interpretation of the G–E OR in the non-diseased controls as the G–E OR in the population should be made cautiously.

**Sources of non-independence**

Clearly, independence between G and E is central to valid interpretation of a case-only study. Associations between G and E seem most likely when an individual is driven to alter his/her exposure according to gene status. Gene-dependent behaviour modification may occur when an individual knows or can guess at his/her gene status, or when the gene status is symptomatic.

Here, we posit two examples of potential G–E relationships as a matter of illustration. For instance, in a study assessing interaction between BRCA1 status and use of post-menopausal hormones in causing breast cancer, a strong family history of breast cancer may cause non-independence between BRCA1 status and hormone use. Specifically, a woman with a strong family history is more likely to carry the BRCA1 mutation and knowing her family history, may be more likely to avoid post-menopausal hormones. In this scenario, there is a positive association between family history and BRCA1 mutation and a negative association between family history and hormone use.
resulting in a negative association between BRCA1 and hormone use. Another example might be a study assessing interaction between alcohol intake and the alcohol dehydrogenase (ADH) polymorphism in causing liver cancer. The adverse reaction to alcohol common in those with the ADH polymorphism may cause non-independence between the polymorphism and alcohol intake. Specifically, those with the ADH polymorphism may alter their drinking behaviour to avoid the resulting adverse reaction. There is a positive association between ADH polymorphism and alcohol-induced adverse reaction and a negative association between alcohol-induced adverse reaction and alcohol use, resulting in a negative association between ADH polymorphism and alcohol use. Here, the third variable, adverse reaction to alcohol, is a step in the causal pathway and accounts for the association between the polymorphism and alcohol intake.

In both examples, the covariate (family history, adverse reaction to alcohol) is responsible for a negative association between the gene and the environmental factor. In a case-only study of a positive multiplicative interaction, this negative association between G and E in the underlying population will cause the case-only OR to be biased toward the null.

In the worst-case scenario, G and E are plausibly related as in the examples discussed above and controls are deemed inappropriate to evaluate independence because of a high baseline risk or strong independent effects. However, there may be alternatives for assessing independence in this situation. For instance, Sturmer et al. measured the association between alcohol consumption (the exposure of interest) and the dehydrogenase II gene (the gene of interest) in a small random sample of the general population.22

Control for violations of independence

In their 1997 paper, Umbach and Weinberg stated that while gene–environment independence might not hold overall, independence may be tenable within certain strata of the population.17 This implies that stratification may improve the validity of case-only interaction estimates. Below, we show that adjustment for non-independence using multivariable modelling is possible.

Aspects of controlling for non-independence in case-only studies examining interaction are similar to controlling for confounding in studies of main effects. When main effects are of interest, researchers are concerned with controlling for covariates that are risk factors for the disease and related to the exposure to ensure that any observed exposure–disease association reflects a causal effect. A covariate with these characteristics may represent a common cause of the exposure and disease or an intermediate variable that participates in a causal pathway between the exposure and disease. If such a covariate is indeed a confounder, adjustment for the covariate will remove its influence on the exposure–disease association and improve the validity of the effect estimate.

In case-only studies of interaction, researchers should be concerned with controlling for covariates that cause non-independence between G and E to ensure that any G–E association among cases reflects their interaction in causing disease. In this context, a covariate of interest (C) may represent a common cause of G and E (e.g. family history of breast cancer) or an intermediate variable that participates in the causal pathway between them (e.g. adverse reaction to alcohol). If such a covariate is indeed the source of non-independence between G and E, adjustment for the covariate will remove the G–E association and yield a valid estimate of interaction among cases.

It is important to note that the purpose of controlling for non-independence in a case-only study is to remove the association between G and E entirely, regardless if the association is causal or non-causal. This is in contrast to the purpose of controlling for confounding in a study of a main effect, which is to separate the causal effect of interest from the non-causal association. However, we find the analogy to confounding useful, especially when thinking about whether G and E are related in the population. Just as an investigator would not worry about covariates that are not plausible potential confounders, an investigator using the case-only design need not worry about third variables that are not plausibly associated with both G and E.

The diagrams shown in Figure 6 depict the scenario in which the third variable, C, is related to both G and E and the scenario in which C is in the causal pathway between G and E. The diagrams have been labelled with the factors involved in the two examples of non-independence described above. The mechanisms are simplified to include only the relevant pathways (e.g. other causes of reduced alcohol intake and adverse reaction to alcohol that do not result from the polymorphism are omitted).

To illustrate how to use multivariable approaches to control for non-independence, we used the cases from the hypothetical populations to conduct case-only analyses with and without control for C. Data from 12 cohort studies were generated, representing 3 dichotomous exposures to G, E, and a third variable, C. In each cohort, G and E interact to cause D. In addition, G and E are associated with one another due to their mutual associations with C. That is, C represents the variable(s) that cause a G–E association. C may be an antecedent that causes both G and E, leading to their association. Alternatively, C may be in the causal pathway between G and E, accounting for their association.

Twelve hypothetical populations of 100 000 people were generated using SAS version 8.12.33 Each population can be

![Figure 6](https://academic.oup.com/ije/article-abstract/33/5/1014/624059/fig6)

Figure 6 Two hypothetical scenarios in which a third variable, C, may be the source of an association between a genetic (G) and an environmental (E) factor in the population.
CASE-ONLY DESIGN FOR ASSESSING GENE–ENVIRONMENT INTERACTION

conceptualized as a study cohort, similar to Figure 1a, but further stratified by C+ and C−, in which a1 to d2 represent those exposed to C and a3 to d4 represent those unexposed to C (Fig. 7). For each population, the SAS MODEL procedure was used to simultaneously solve for the desired relationships among the variables C, G, E, and D. Each solution was based on set parameter values. For details regarding these parameters, see Appendix 1. For each of the 12 examples, the cell counts, a1 to d4, derived from the SAS MODEL procedure were used to generate a dataset of 100 000 observations to which logistic regression modelling was applied.

The relationship between any two variables was determined by setting the RR for that relationship. For example, if no association between two variables was desired, the RR was set at 1. The strength of the G–E association was varied from 0.38 to 2.63, by varying the C–E RR from 0.33 to 6 and C–G RR from 1 to 6. For every cohort, the G × E RR was 4 and the RR for the G–E relationship was 1 within each stratum of C. In addition, the overall prevalence of C was set at 0.40 and the C–D RR was 1.

All of the cases from each hypothetical cohort were used in the corresponding case-only analysis. All crude and adjusted case-only ORs were generated using the LOGISTIC procedure in SAS.33 In general, modelling any variable as a function of another in a case-only analysis gives the estimate of interaction between these two variables. The crude case-only OR was calculated by modelling G as a function of E (Equation 2). C was added to this model as a covariate to obtain the adjusted case-only OR (Equation 3). The exponentiations of the ORs represented in equations 2 and 3 are the unadjusted and adjusted case-only ORs respectively.

log odds (G) = β0 (Equation 2)

log odds (G) = β0 + β1 (E) + β2 (C) (Equation 3)

Table 1 presents the results from the case-only analyses. Each row of the table represents a hypothetical cohort of 100 000 people in which one of the relationships involving C varies across rows while the other relationships are held constant. All relationships among C, E, G, and D pertain to the underlying cohort. Each case-only analysis is an attempt to estimate a G × E RR of 4.

As expected, the case-only OR is biased when the G–E OR in the population is not 1. Furthermore, the extent of bias is related to the strength and direction of the G–E association. Adjustment for C in the case-only analysis (by including C in the logistic model) completely removes this bias.

Figure 7 Two-by-two tables underlying each hypothetical cohort, where E represents the environmental factor, G is the gene, D is the disease and C is the covariate

Table 1 Effect of controlling for covariate C, the source of non-independence between the gene (G) and environmental factor (E), on the case-only odds ratio (OR)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>3.00</td>
<td>1.00</td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td>2.00</td>
<td>3.00</td>
<td>1.67</td>
<td>4.00</td>
<td>6.67</td>
</tr>
<tr>
<td>3.00</td>
<td>3.00</td>
<td>2.04</td>
<td>4.00</td>
<td>8.17</td>
</tr>
<tr>
<td>4.00</td>
<td>3.00</td>
<td>2.30</td>
<td>4.00</td>
<td>9.19</td>
</tr>
<tr>
<td>5.00</td>
<td>3.00</td>
<td>2.49</td>
<td>4.00</td>
<td>9.90</td>
</tr>
<tr>
<td>6.00</td>
<td>3.00</td>
<td>2.63</td>
<td>4.00</td>
<td>10.54</td>
</tr>
<tr>
<td>0.33</td>
<td>1.00</td>
<td>1.00</td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td>0.33</td>
<td>2.00</td>
<td>0.62</td>
<td>4.00</td>
<td>2.46</td>
</tr>
<tr>
<td>0.33</td>
<td>3.00</td>
<td>0.50</td>
<td>4.00</td>
<td>2.00</td>
</tr>
<tr>
<td>0.33</td>
<td>4.00</td>
<td>0.44</td>
<td>4.00</td>
<td>1.76</td>
</tr>
<tr>
<td>0.33</td>
<td>5.00</td>
<td>0.40</td>
<td>4.00</td>
<td>1.60</td>
</tr>
<tr>
<td>0.33</td>
<td>6.00</td>
<td>0.38</td>
<td>4.00</td>
<td>1.50</td>
</tr>
</tbody>
</table>

a Crude model included the following covariate: E (Equation 1).
b Adjusted model included the following covariates: E and C (Equation 2).
Discussion

The case-only design to study gene–environment interaction has been criticized for its susceptibility to bias caused by non-independence between genetic and environmental factors. As advised in the literature, researchers frequently use controls to evaluate whether the independence assumption is tenable. In this paper, we demonstrated that the OR for the gene–environment association measured in controls may not accurately reflect the gene–environment association in the underlying population, even when the baseline risk of disease is low. Using controls to assess gene–environment independence in the source population can lead an investigator to reject valid case-only data, overestimate or underestimate the magnitude of the interaction, or observe an interaction when none exists among the corresponding risk ratios. Our formulation of the relationship between the ORs relating the genetic and environmental factors in the controls and in the underlying population provides some guidelines for deciding when assessment in controls may be safe. We found that as long as the baseline risk of disease is small (<1%) and the interaction and independent effects are moderate (<2), controls provide a reasonable approximation of the gene–environment association in the population. Alternatively, controls can be used if the disease risk is low (e.g. <5%) in all strata of G and E, as previously suggested by Schmidt and Schaid. However, as our equation implies, the approximation becomes progressively worse with minimal increases in the baseline risk of disease, interaction, or independent effects. Previous studies which used controls to determine whether a case-only study was valid may need re-evaluation.

Furthermore, we discussed how non-independence seems most likely when an individual's genetic status is knowable or symptomatic, enabling modification of exposure to an environmental factor dependent on gene status. For most metabolic polymorphisms, it seems that obvious symptoms or knowledge of gene status would be uncommon. In studies of these genetic factors, gene–environment associations are less likely. Although the data are limited, some studies have shown a lack of association between metabolic polymorphisms and behavioural risk factors.

Finally, we demonstrated that control of non-independence in the case-only analysis can yield a valid interaction estimate. If the factor that causes a gene–environment association is measured, multivariable logistic regression modelling can be used to remove the bias. If one can posit a mechanism by which G and E are related, then the source of this relationship should be measured and included in the case-only analysis. The extent to which complete control is possible depends on the same challenges that face epidemiological studies of main effects. These challenges include articulation of explicit hypotheses, and attention to the construct and measurement of all variables.

In practice, control for non-independence will not always be straightforward. For instance, control for non-independence requires that the source(s) of non-independence can be conceptualized and measured. However, this may be difficult or impossible in some situations. For example, when gene–gene interaction of interest, linkage disequilibrium may cause non-independence between the genes. In this circumstance, the source of non-independence may not easily be attributable to a third variable, making adjustment impossible. Furthermore, in studies of main effects, adjustment for variables that does not change the validity of the effect estimate costs degrees of freedom and reduces the precision of the estimate. Similarly, control for variables that do not generate a G–E association in the underlying population may have the same costs in case-only studies. Finally, given that it is difficult to control for sources of bias in cohort and case-control studies of main effects and/or interaction (e.g. misclassification, loss to follow-up, etc.) it may be difficult to control for these sources of bias in case-only studies. However, sensitivity analysis methods that prove useful in cohort and case-control studies may also apply to case-only studies.

In conclusion, recent criticisms of the case-only study may be overstated. Since non-independence can be accounted for in the analysis, the case-only study may still be a valuable tool for investigations of gene–environment interaction.

Acknowledgements

This work was supported by grants from the National Institute of Health (T32-CA09529, KO7-CA92348 & P42ES10349-01), Department of Defense (DAMD 17-02-1-0354 & DAMD 17-00-1-0213) and the New York City Council Speaker's Fund for Public Health Research. We are extremely grateful to Sharon Schwartz and the referees for their exceptionally detailed comments that led to substantial improvement of the manuscript.

KEY MESSAGES

- The validity of the case-only study for assessing gene–environment interaction hinges upon independence between the genetic and environmental factors in the population.
- Using controls to assess gene–environment independence in the population can lead an investigator to reject valid case-only data, or over/underestimate the magnitude of the interaction or observe an interaction when none exists among the corresponding risk ratios in the underlying cohort.
- As long as the baseline disease risk is less than 1% and the interaction and independent effects are less than 2, controls provide a reasonable approximation of the gene–environment association in the underlying population.
- If the source(s) of non-independence can be conceptualized and measured, adjustment for non-independence using multivariable models can be used to remove the bias in case-only analyses of interaction.
References


Appendix 1

To generate hypothetical cohort data, we developed a program in SAS version 8.12,33 in which the following parameters were controlled:

1) The overall prevalence of C+,
2) The overall prevalence of C−,
3) The prevalence of G among C+,
4) The prevalence of G among C−,
5) The prevalence of E among C+,
6) The prevalence of E among C−,
7) The risk of D among C+, E− and G−,
8) The risk of D among C−, E− and G−,
11) The G × FRR among C+, G × FRR among C−,
12) The E–D RR among C+, C–E RR among C−,
13) The E–D RR among C+, C–E RR among C−,
14) The E–D RR among C−, C–E RR among C+, and,
15) The G–D RR among C+, E− and,
The C–G RR is varied by manipulating parameters (3) and (4). Likewise, the C–E RR is varied by manipulating parameters (5) and (6). The independent effect of C (among E−, G−) on D is varied by manipulating parameters (7) and (8).

The 16 parameters described above were translated into 16 equations relating 16 cells. These 16 cells (labelled a₁ to d₄) represent the cell counts for four 2 × 2 tables (Figure 7).

For each population, the SAS MODEL procedure was used to simultaneously solve for the 16 variables (a₁ to d₄), based on the parameter values entered. The solution values were used to generate a dataset of 100,000 observations in order to perform regression modelling. For example, if a₁ equalled 5000, the dataset contained 5000 observations of people who were labelled as ‘exposed’ to variables E, D, G, and C.

For additional details regarding this program, contact Ulka Campbell at uvb2@columbia.edu, or Nicolle Gatto at nmg22@columbia.edu.