Potential Misinterpretation of the Case-Only Study to Assess Gene-Environment Interaction

Silke Schmidt and Daniel J. Schaid

Novel epidemiologic study designs are often required to assess gene-environment interaction. A design using only cases, without controls, is one of several approaches that have been proposed as more efficient alternatives to the typical random sampling of cases and controls. However, it has not been pointed out that a case-only analysis estimates a different interaction parameter than does a traditional case-control analysis: The latter typically estimates departure from multiplicative population odds or rate ratios, depending on the method of control selection, while the former estimates departure from multiplicative risk ratios if genotype and environmental exposure are not associated in the population. These parameters are approximately equal if the disease risk is small at all levels of the study variables. The authors quantify the impact of allowing for higher disease risk among gene carriers, a relevant situation when the gene under study is highly penetrant. Their findings show that the cross-product ratio computed from case-only data may be substantially smaller than the odds ratio computed from case-control data and may therefore underestimate either the population odds or the rate ratio. Thus, to avoid misinterpretation of interaction parameters estimated from case-only data, the definition of multiplicative interaction should be made explicit. Am J Epidemiol 1999; 150:878-85.

Interactions between susceptible genotypes and environmental risk factors have important scientific and public health implications. First, the estimated genetic effects depend on the environmental conditions that exist for any particular study, limiting inference of genetic effects to only the range of environments studied and vice versa. Second, interactions can imply dramatically increased (or decreased) risks to individuals, depending on their genotype and environment. The definition of gene-environment interaction is critical. In a traditional epidemiologic approach, case-control studies are used to estimate the main and interaction effects of interest. Here, multiplicative odds ratio models are most frequently used, and interaction in this framework means departure of joint effects from multiplicative odds ratios. Whether the odds ratio calculated from the actually studied case and control subjects estimates the population risk, or odds ratio depends on the method of control selection (1). If controls are selected from person-time at risk (density sampling), the case-control odds ratio estimates the population rate ratio; if they are selected from the base population, i.e., those at risk at the beginning of the study, the case-control odds ratio estimates the population risk ratio; and if they are selected from the survivors at the end of the study time during which cases are ascertained (cumulative sampling), the case-control odds ratio estimates the population odds ratio. Sampling controls from the base population is more typically done in the context of a case-cohort study rather than in a traditional case-control study.

It has been shown that the random sampling of cases and controls may require prohibitively large sample sizes to reach adequate power to detect gene-environment interaction, especially when the genetic factor is rare (2). Therefore, it may be useful to consider alternative sampling or analysis strategies when the main goal of a study is to estimate gene-environment interaction (3–5). One of the proposed “nontraditional” approaches is a case-only analysis, which does not use the controls to assess gene-environment interaction. With this analysis, main effects of the susceptible genotype (G) and environmental exposure (E) cannot be estimated, but a measure of association between the two factors among cases is easily computed as a cross-product ratio. This cross-product...
Potential Misinterpretation of the Case-Only Study

Table 1. Gene-environment interaction analysis in a case-only study

<table>
<thead>
<tr>
<th>Environmental exposure $E$</th>
<th>Susceptibility genotype $G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$a$</td>
</tr>
<tr>
<td>1</td>
<td>$b$</td>
</tr>
<tr>
<td>0</td>
<td>$c$</td>
</tr>
<tr>
<td>1</td>
<td>$d$</td>
</tr>
</tbody>
</table>

Ratio depends on both the interaction and the population association of $G$ and $E$. If genetic and environmental risk factors are independent in the population, the cross-product ratio is equal to a risk ratio. If, additionally, the risk of disease is small at all levels of the study variables, it is approximately equal to the odds ratio for gene-environment interaction that is computed from case-control data.

While the independence assumption and some other limitations of the case-only analysis have been discussed in the literature (4), it has not been pointed out that the case-only design and the case-control design in fact estimate different interaction parameters. In particular, it has not been demonstrated how much these parameters can differ in some realistic scenarios, e.g., when studying gene-environment interaction for highly penetrant genes in which the assumption of small disease risk for gene carriers is clearly violated. Several such genes have recently been identified, e.g., BRCA1 and BRCA2 for breast and ovarian cancer, which carry a lifetime risk on the order of 56–90 percent for breast cancer and 30–60 percent for ovarian cancer (6–8).

Here, we investigate the impact of allowing for at least moderate disease risk among carriers of susceptibility genes on the difference between the cross-product ratio computed from a case-only study and the interaction odds ratio computed from case-control data. Data are assumed to arise from a prospective logistic model that includes both main effects for $G$ and $E$ and their interactive effect. The expected value of the cross-product ratio computed from case-only data can then be expressed as a function of the logistic model parameters, which, in turn, depend on the assumed population frequencies of disease, genetic susceptibility, and exposure to the environmental risk factor. Relevant conclusions for the use of case-only studies to assess gene-environment interaction are discussed.

**MATERIALS AND METHODS**

In a case-only design, we obtain a cross-classification of cases with respect to $G$ and $E$ as in table 1. Let $G = 1$ for individuals who possess the high-risk genotype and $G = 0$ for those with a low-risk genotype; let $E = 1$ for those who are exposed to the environmental risk factor, and $E = 0$ for those who are unexposed; and let $P_{ij} = P(G = i, E = j | D = 1)$, where $D = 1$ denotes a diseased case. Then the association of $G$ with $E$ can be measured by a cross-product ratio $ad/bc$. Applying Bayes' theorem, one can show that the asymptotic expected value of this cross-product ratio ($CPR_c$) results from the multiplication of the interaction risk ratio, $RR_i$, and the population association of $G$ and $E$, $OR_{GE}$:

$$E(CPR_c) = E\left(\frac{ad}{bc}\right) \approx \frac{p_{11}/p_{01}}{p_{10}/p_{00}}$$

$$= \left\{ \frac{P(D = 1 | G = 1, E = 1)}{P(D = 1 | G = 0, E = 1)} \right\} \left\{ \frac{P(D = 1 | G = 1, E = 0)}{P(D = 1 | G = 0, E = 0)} \right\}$$

$$\times \left\{ \frac{P(G = 1, E = 1)P(G = 0, E = 0)}{P(G = 0, E = 1)P(G = 1, E = 0)} \right\}$$

$$= RR_i \times OR_{GE}.$$  

The second factor, $OR_{GE}$, can be interpreted as the odds of $G$ for $E = 1$ versus $E = 0$ or the odds of $E$ for $G = 1$ versus $G = 0$. If $G$ and $E$ are independent in the source population, $OR_{GE} = 1$, indicating no association of $G$ and $E$. In this situation, $CPR_c$ estimates the interaction risk ratio $RR_i$.

For a traditional case-control study, multiplicative interaction is measured as the ratio of two odds ratios: the odds ratio of carrier status among those exposed to the environmental factor divided by a similar odds ratio for
those not exposed to the environmental factor. The interaction parameter, denoted by $OR_i$, can be written as follows by use of Bayes’ theorem:

$$OR_i = \frac{RR_i \times \frac{P(D = 0|G = 1, E = 0)}{P(D = 0|G = 0, E = 0)} / P(D = 0|G = 0, E = 1)}{P(D = 0|G = 1, E = 0)} / P(D = 0|G = 0, E = 1)},$$

where the second term on the right side of equation 1 is an expression involving only controls.

This second term is approximately one if the disease risk is small at all levels of both study variables, which requires a low risk for carriers of the susceptibility gene.

To examine the impact of violating the assumption of small disease risk for the at-risk genotype in more detail, we consider the expected value of $CPR_e$ by appealing to the logistic model. Conditional on the disease status of cases and assuming independence of $G$ and $E$ in the population, the probability of observing $G = i, E = j$ is given by

$$p_{ij} = P(G = i, E = j|D = 1) = \frac{P(D = 1|G = i, E = j)P(G = i)P(E = j)}{P(D = 1)}.$$

The first term in the numerator of equation 2, $P(D = 1|G = i, E = j)$, is given by the logistic model:

$$\log \left( \frac{P(D = 1|G = i, E = j)}{1 - P(D = 1|G = i, E = j)} \right) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 (GE).$$

The asymptotic expected value of the cross-product ratio can now be expressed as a function of $\beta_0, \ldots, \beta_3$ as follows:

$$E(CPR_e) \approx \exp(\beta_3) \frac{(1 + \exp(\beta_0 + \beta_1))(1 + \exp(\beta_0 + \beta_2))}{(1 + \exp(\beta_0))(1 + \exp(\beta_0 + \beta_1 + \beta_2 + \beta_3))}.$$

Thus, the expected value of $CPR_e$ is approximately equal to the odds ratio, $\exp(\beta_3)$, times a factor that is very close to one if $\beta_0 \ll 0$. If the population disease risk is larger, $\beta_0$ is less negative, and the multiplicative factor in equation 3 is smaller than one when both the genetic and environmental factors increase disease risk (i.e., $\beta_1 > 0, \beta_2 > 0$).

If $\beta_1, \beta_2, \text{ and } \beta_3$ are chosen to reflect plausible effect sizes (e.g., $\beta_1$ might be assumed to be fairly large to model a highly penetrant gene), the intercept parameter $\beta_0$ can be determined from the overall population disease risk $P(D = 1)$ by solving the following equation for $\beta_0$:

$$P(D = 1) = \sum_{i=0}^{1} \sum_{j=0}^{1} P(D = 1|G = i, E = j)P(G = i)P(E = j).$$

We used an iterative bisection algorithm to compute $\beta_0$ for a given population disease risk.

RESULTS

To illustrate the extent to which the two parameters measuring multiplicative interaction can differ under reasonable study scenarios, the expected $CPR_e$ for gene-environment interaction (equation 3) is plotted against the assumed odds ratio, $\exp(\beta_3)$, in figures 1–3 for different values of population disease risk ($P(D)$), frequency of susceptible genotype ($P(G)$), and main effect of the genetic factor ($\beta_1$). Also displayed is a solid line representing equality of the expected cross-product ratio and assumed odds ratio. For the parameters relating to the environmental risk factor, a common exposure ($P(E) = 0.3$) with a small odds ratio (1.5) was assumed. For the genetic risk factor, we compared the situation of a rare susceptible genotype ($P(G) = 0.05$) with that of a more frequent one ($P(G) = 0.2$). The parameter that has the strongest impact on the intercept $\beta_0$ when solving equation 4 is the assumed population disease risk $P(D)$, which was fixed at 0.01 in figure 1, at 0.05 in figure 2, and at 0.10 in figure 3. In each figure, we
compare the extent of underestimation of the assumed interaction odds ratio when the genetic main effect odds ratio is moderate ($OR(G) = 2$), higher ($OR(G) = 6$), or very high ($OR(G) = 10$).

Figure 1 ($P(D) = 0.01$) illustrates that when the disease is rare, the risk ratio estimated from a case-only analysis approximates the assumed odds ratio for gene-environment interaction very closely, both when the
susceptible genotype is rare (figure 1A, $P(G) = 0.05$) and when it is common (figure 1B, $P(G) = 0.20$). This is true even for genes as highly penetrant as those characterized by an odds ratio of 10. The approximation is considerably worse when the disease is more common, i.e., with $P(D) = 0.05$ (figure 2). The difference
FIGURE 3. Expected cross-product ratio (CPR) for gene-environment interaction from case-only study versus assumed odds ratio (OR) from case-control study. We assume a disease risk of $P(D) = 0.10$ and either a rare susceptible genotype ($A, P(G) = 0.05$) or a common susceptible genotype ($B, P(G) = 0.20$). The odds ratio for the susceptible genotype in the absence of the environmental exposure is $OR(G) = 2, 6, \text{or } 10$.

between the two parameters is quite striking when the odds ratio for the main genetic effect is on the order of 6 or more, with underestimation of the odds ratio more extreme when the susceptible genotype is more rare (figure 2A, $P(G) = 0.05$, versus figure 2B, $P(G) = 0.20$). Finally, with a population disease frequency of 0.10
(figure 3), a realistic value for some common cancers (e.g., breast and colorectal cancer), the extent of underestimation of the odds ratio can be dramatic, even if the susceptible genotype is as common as $P(G) = 0.2$ (figure 3B).

**DISCUSSION**

We have demonstrated that using only cases to assess gene-environment interaction can lead to underestimation of the interaction odds ratio computed from case-control data because interaction is measured as departure from multiplicative risk ratios. As pointed out in the introduction, the population parameter that is estimated by the case-control odds ratio depends on the method of control selection. More precisely, therefore, the cross-product ratio computed from case-only data may underestimate the population odds ratio if controls are selected from the survivors at the end of the study time (cumulative sampling), or it may underestimate the population risk ratio if controls are selected longitudinally throughout the course of the study (density sampling). If controls are selected from the base population at the beginning of a study, both the case-control interaction odds ratio and the case-only cross-product ratio estimate the interaction risk ratio. However, this method of control selection is not typically used in actual case-control studies.

On the basis of our results, we conclude that the interaction parameters estimated from case-only and case-control data are of comparable magnitude if the disease is truly rare (on the order of 1 percent) or if the disease is somewhat more common (around 5 percent), but the genes under study confer only moderately increased disease risk (e.g., OR($G) < 6$). For diseases as common as breast or colorectal cancer, a case-only analysis is likely to underestimate population odds or rate ratios for multiplicative interactions greater than two, but always yields unbiased estimates of interaction risk ratios. These considerations are based on the assumption that the genetic factor is associated with an increased disease risk in the absence of the environmental factor and vice versa. If, however, the environmental exposure confers an increased disease risk only in the presence of the susceptible genotype (i.e., $\beta_2 = 0$), it can be seen from equation 3 that the amount of underestimation of the odds ratio with a case-only analysis is less dramatic, even if the disease is more common and the genetic main effect is large. An analogous argument can be made when the genetic factor increases disease risk only in the presence of the environmental factor ($\beta_1 = 0$).

If exposure affects average risk, the rate ratio is usually expected to fall between the risk ratio and odds ratio (9). It is important to bear in mind that effects on incidence rates are not the same as those on incidence proportions (average risks) and that the three effect measures should therefore be kept distinct. The common practice of referring to all three measures under the generic term "relative risk" unfortunately makes confusion among effect measures likely (1, 9). It is also important to recognize that the estimation procedures used in case-control or case-only studies do not depend on any rare disease assumption. Which of the three effect measures is estimated by the case-control odds ratio is determined by the method of control selection. The rare disease assumption is relevant only when the investigator wants to compare the actually estimated effect measure with one of the other effect measures: If the disease is rare for all exposure patterns under study during the interval of interest, the population odds ratio estimated from case-control data with cumulative sampling is approximately equal to that estimated from density-sampled case-control data and is also approximately equal to the population risk ratio estimated from case-only data under the independence assumption.

It is helpful to clarify when a case-only analysis can and cannot be recommended to practicing epidemiologists. If the independence assumption for $G$ and $E$ is met, the case-only design allows for substantial savings in sample size because it offers greater precision for estimating gene-environment interaction than does a traditional case-control study (3). For hypothesis tests, the variance of the test statistic when using only cases is smaller than the variance of the test statistic when using cases and controls, leading to better power for detecting gene-environment interaction (10). If the particular gene under study is very rare, small numbers of controls will probably carry the susceptibility gene, leading to very unstable effect estimates and making the case-only analysis a more attractive approach in this situation. On the other hand, if a model simultaneously incorporating main and interaction effects is desired, there is no alternative to the traditional case-control design.

The assumption of independence between genetic and environmental risk factors seems reasonable for a wide variety of genes and exposures. However, it may be violated in practice if, for example, both exposure pattern and allele frequency vary with confounding factors such as age or ethnic group. The assumption can be tested only with data from a random sample of controls, i.e., by cross-classifying controls according to $G$ and $E$ and using a $\chi^2$ test for association in this table. More recently, it has been demonstrated how the independence assumption, if tenable, can be explicitly imposed in log-linear models, allowing better precision for estimates of gene-environment inter-
action than is provided by logistic regression analyses (5).

In summary, the case-only analysis may be an option to assess gene-environment interaction when the main effects of both risk factors have already been well established. However, to avoid potential misinterpretation of the results, it should be made explicit that interaction is measured in terms of population risk ratios, whereas traditional case-control studies estimate population odds or rate ratios depending on the method of control selection.

ACKNOWLEDGMENTS

This research was partially supported by the German Academic Exchange Service (DAAD), Doktorandenprogramm HSP III, and United States National Institutes of Health grant GM51256.

REFERENCES