American Journal of Epidemiology
Vol. 168, No. 2
DOI: 10.1093/aje/kwn130
Advance Access publication May 22, 2008

Practice of Epidemiology

Population Stratification Bias in the Case-Only Study for Gene-Environment Interactions

Liang-Yi Wang and Wen-Chung Lee

From the Research Center for Genes, Environment, and Human Health and the Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Taipei, Taiwan, Republic of China.

Received for publication October 15, 2007; accepted for publication April 15, 2008.

The case-only study is a convenient approach and provides increased statistical efficiency in detecting gene-environment interactions. The validity of a case-only study hinges on one well-recognized assumption: The susceptibility genotypes and the environmental exposures of interest are independent in the population. Otherwise, the study will be biased. The authors show that hidden stratification in the study population could also ruin a case-only study. They derive the formulas for population stratification bias. The bias involves three terms: 1) the coefficient of variation of the exposure prevalence odds, 2) the coefficient of variation of the genotype frequency odds, and 3) the correlation coefficient between the exposure prevalence odds and the genotype frequency odds. The authors perform simulation to investigate the magnitude of bias over a wide range of realistic scenarios. It is found that the estimated interaction effect is frequently biased by more than 5%. For a rarer gene and a rarer exposure, the bias becomes even larger (>30%). Because of the potentially large bias, researchers conducting case-only studies should use the boundary formula presented in this paper to make more prudent interpretations of their results, or they should use stratified analysis or a modeling approach to adjust for population stratification bias in their studies.

bias (epidemiology); data interpretation, statistical; environment; epidemiologic methods; genetics

Abbreviations: CIR, confounding interaction ratio; CV, coefficient of variation; RR, relative risk.

The interplay between susceptibility genotypes and environmental exposures on the risks of human diseases has important public health implications (1–3). The traditional epidemiologic designs, for example, the case-control study and the cohort study, can be applied to detect gene-environment interactions, but they require large sample sizes to reach enough power (4, 5). The ''control-free'' case-only study, on the other hand, is a convenient approach and provides increased statistical efficiency in detecting interactions (6–9).

The validity of a case-only study hinges on one well-recognized assumption: The susceptibility genotypes and the environmental exposures of interest are independent in the population (6, 10, 11). The assumption is reasonable, because genes are constitutional and often will not predict a subject's exposure profile that is acquired environmentally through life. On the other hand, counterexamples are not difficult to find. For example, a person with a strong family history of a particular disease is more likely to carry the risk genotype and at the same time may be more likely (because of the awareness of the disease) to avoid risk exposure, thus violating the gene-environment independence assumption.

However, the nonindependence of the susceptibility genotype that a person inherits and the environmental exposure that he or she will have is not the only condition that can invalidate a case-only study. In this paper, we show that hidden stratification in the study population could easily ruin a case-only study as well, even though within each and every one of the population strata, there is perfect gene-environment independence.

Correspondence to Dr. Wen-Chung Lee, Room 536, 17 Xuzhou Road, Taipei 100, Taiwan, Republic of China (e-mail: wenchung@ntu.edu.tw).

POPULATION STRATIFICATION BIAS IN THE CASE-ONLY STUDY

Let E($\bar{\text{E}}$) indicate the presence(absence) of the environmental exposure of interest, and G($\bar{\text{G}}$), the presence(absence) of the genotype of interest. In a population composed of J strata ($j = 1, \ldots, J$), we define p_j to be the prevalence of E ($e_j = \frac{p_j}{1-p_j}$, the exposure prevalence odds); q_j as the frequency of G ($g_j = \frac{q_j}{1-q_j}$, the genotype frequency odds); b_j as the background disease risk; and m_j as the total number of subjects in the jth stratum. In a comparison of risks with the ($\bar{\text{E}}$, $\bar{\text{G}}$)-subjects, we assume that the relative risks (RR) of disease are RR_{EG} for the (E, G)-subjects, RR_{E} for the (E, $\bar{\text{G}}$)-subjects, and RR_{G} for the ($\bar{\text{E}}$, G)-subjects. Note that we assume the RR_{EG}, RR_{E}, and RR_{G} to be constant across the strata. In other words, the stratification in the population only segregates people. If sufficiently long, the segregation could result in disparate exposure prevalences, genotype frequencies, and background disease risks in various population strata. However, it could not by itself modify the magnitudes of effects of exposures and/or genes in causing human diseases.

The gene-environment interaction effect (RR_{INT}) on a multiplicative scale is $\text{RR}_{\text{INT}} = \frac{\text{RR}_{\text{EG}}}{\text{RR}_{\text{E}} \times \text{RR}_{\text{G}}}$. There is no gene-environment interaction if $\text{RR}_{\text{INT}} = 1$, and interaction exists when $\text{RR}_{\text{INT}} > 1$ (synergistic interaction) or $\text{RR}_{\text{INT}} < 1$ (antagonistic interaction). The RR_{INT} is estimable by use of a case-only study. If a study collects each and every diseased subject (the cases) in the population, the case numbers would be $n_{\text{EG}} = \sum_{j=1}^{J} m_j p_j q_j b_j \text{RR}_{\text{EG}}$ for the (E, G)-subjects, $n_{\text{E}\bar{\text{G}}} = \sum_{j=1}^{J} m_j p_j (1-q_j) b_j \text{RR}_{\text{E}}$ for the (E, $\bar{\text{G}}$)-subjects, $n_{\bar{\text{E}}\text{G}} = \sum_{j=1}^{J} m_j (1-p_j) q_j b_j \text{RR}_{\text{G}}$ for the ($\bar{\text{E}}$, G)-subjects, and $n_{\bar{\text{E}}\bar{\text{G}}} = \sum_{j=1}^{J} m_j (1-p_j)(1-q_j) b_j$ for the ($\bar{\text{E}}$, $\bar{\text{G}}$)-subjects (the environmental exposure and genotype are assumed to be independent within each and every stratum). Thus, the gene-environment interaction effect can be estimated as

$$\text{RR}_{\text{INT}}^{\text{c}} = \frac{n_{\text{EG}} \times n_{\bar{\text{E}}\bar{\text{G}}}}{n_{\text{E}\bar{\text{G}}} \times n_{\bar{\text{E}}\text{G}}} = \frac{\sum_{j=1}^{J} m_j p_j q_j b_j \text{RR}_{\text{EG}} \times \sum_{j=1}^{J} m_j (1-p_j)(1-q_j) b_j}{\sum_{j=1}^{J} m_j p_j (1-q_j) b_j \text{RR}_{\text{E}} \times \sum_{j=1}^{J} m_j (1-p_j) q_j b_j \text{RR}_{\text{G}}}.$$

Note that we put a superscript "c" in RR_{INT} to indicate that the estimated interaction effect in the above case-only study is a "crude" measure without respect to population stratification.

We use the confounding interaction ratio (CIR) to quantify the degree of population stratification bias of $\text{RR}_{\text{INT}}^{\text{c}}$ in the case-only study. The CIR is defined as the ratio of the crude interaction effect and the true (or adjusted) interaction effect: $\text{CIR} = \frac{\text{RR}_{\text{INT}}^{\text{c}}}{\text{RR}_{\text{INT}}}$. We found out, interestingly, that the mathematical formula for the CIR for the interaction effect is very similar to the confounding risk ratio (CRR) for the main genetic effect in our previous study (12).

$$\begin{aligned}\text{CIR} &= \frac{\sum_{j=1}^{J} w_j e_j g_j}{\bar{\varphi}_{\text{E}} \bar{\varphi}_{\text{G}}} \\ &= \frac{\sum_{j=1}^{J} w_j (e_j - \bar{\varphi}_{\text{E}})(g_j - \bar{\varphi}_{\text{G}})}{\text{SD}(\varphi_{\text{E}}) \times \text{SD}(\varphi_{\text{G}})} \times \frac{\text{SD}(\varphi_{\text{E}})}{\bar{\varphi}_{\text{E}}} \times \frac{\text{SD}(\varphi_{\text{G}})}{\bar{\varphi}_{\text{G}}} + 1 \\ &= r_{\text{EG}} \times \text{CV}_{\text{E}} \times \text{CV}_{\text{G}} + 1,\end{aligned}$$

where $\bar{\varphi}_{\text{E}} = \sum_{j=1}^{J} w_j e_j$ and $\bar{\varphi}_{\text{G}} = \sum_{j=1}^{J} w_j g_j$ denote the means; $\text{SD}(\varphi_{\text{E}}) = \sqrt{\sum_{j=1}^{J} w_j (e_j - \bar{\varphi}_{\text{E}})^2}$ and $\text{SD}(\varphi_{\text{G}}) = \sqrt{\sum_{j=1}^{J} w_j (g_j - \bar{\varphi}_{\text{G}})^2}$, the standard deviations; CV_{E} and CV_{G}, the coefficients of variation of the exposure prevalence odds and the genotype frequency odds, respectively; and r_{EG} is the correlation coefficient between the exposure prevalence odds and the genotype frequency odds. Note that $w_j = \frac{m_j (1-p_j)(1-q_j) b_j}{\sum_{k=1}^{J} m_k (1-p_k)(1-q_k) b_k}$ is the "weight" used in the above summation.

It is clear from the formula above for CIR that there would be no population stratification bias ($\text{CIR} = 1$) if, across the strata, 1) the exposure prevalence odds and the genotype frequency odds are uncorrelated, 2) there is no variation in the exposure prevalence odds, or 3) there is no variation in the genotype frequency odds. Also, one notices that overestimation ($\text{CIR} > 1$) would occur when the prevalence odds of exposure and the frequency odds of genotype are positively correlated, and underestimation ($\text{CIR} < 1$) would occur when these are negatively correlated. Higher relative bias (CIR away from 1) exists with higher variation of the prevalence odds of exposure or the frequency odds of genotype.

Researchers often do not have such detailed knowledge of the study population, however. They may instead have an educated guess as to the magnitudes of the "variations." Let υ_{G} ($\upsilon_{\text{G}} \geq 1$) denote the ratio of the largest and the smallest genotype frequency odds, and υ_{E} ($\upsilon_{\text{E}} \geq 1$), the ratio of the largest and the smallest exposure prevalence odds, among all the strata in the population. Assuming that υ_{G} and υ_{E} are known to researchers, it is easy to show (similar to the derivation of the bounds for the confounding risk ratio) (13) that such limited information is enough to set an upper (U) and a lower (L) bounds of the CIR:

$$U = \frac{\sqrt{\upsilon_{\text{G}} \cdot \upsilon_{\text{E}}} \times (\sqrt{\upsilon_{\text{G}} \cdot \upsilon_{\text{E}}} + 1)^2}{(\sqrt{\upsilon_{\text{G}} \cdot \upsilon_{\text{E}}} + \upsilon_{\text{G}}) \times (\sqrt{\upsilon_{\text{G}} \cdot \upsilon_{\text{E}}} + \upsilon_{\text{E}})} \geq 1$$

and $L = \frac{1}{U} \leq 1$. Note, again, that there is no bias ($U = L = 1$) when there is either no variation in the genotype frequency odds ($\upsilon_{\text{G}} = 1$) or no variation in the exposure prevalence odds ($\upsilon_{\text{E}} = 1$).

MAGNITUDE OF BIAS

We investigate the CIR in a case-only study over some realistic ranges of genotype frequency, exposure prevalence, and background disease risk. We follow the procedure of Wacholder et al. (14) (with some modifications) by assuming that, in a population, there are eight strata of equal size but possibly with different values of genotype frequency,

TABLE 1. The theoretical upper and lower bounds, as well as the percentiles from simulated data, of the confounding interaction ratio under various situations

Background risk ratio (range)	Genotype frequency (range)	Exposure prevalence (range)	Theoretical lower bound of CIR*	CIRs among 100,000 simulated sets of eight strata of equal size			Theoretical upper bound of CIR
				25th percentile	50th percentile	75th percentile	
1.0–1.5	0.01–0.30	0.01–0.30	0.09	0.69	0.95	1.31	11.11
1.0–1.5	0.01–0.30	0.10–0.40	0.28	0.84	0.99	1.17	3.58
1.0–1.5	0.01–0.30	0.30–0.60	0.40	0.89	1.00	1.13	2.47
1.0–1.5	0.10–0.40	0.01–0.30	0.28	0.84	0.99	1.17	3.58
1.0–1.5	0.10–0.40	0.10–0.40	0.49	0.91	1.00	1.09	2.04
1.0–1.5	0.10–0.40	0.30–0.60	0.60	0.94	1.00	1.06	1.67
1.0–1.5	0.30–0.60	0.01–0.30	0.40	0.88	1.00	1.13	2.47
1.0–1.5	0.30–0.60	0.10–0.40	0.60	0.94	1.00	1.06	1.67
1.0–1.5	0.30–0.60	0.30–0.60	0.69	0.96	1.00	1.05	1.45
1.0–3.0	0.01–0.30	0.01–0.30	0.09	0.68	0.95	1.32	11.11
1.0–3.0	0.01–0.30	0.10–0.40	0.28	0.83	0.99	1.17	3.58
1.0–3.0	0.01–0.30	0.30–0.60	0.40	0.88	1.00	1.13	2.47
1.0–3.0	0.10–0.40	0.01–0.30	0.28	0.83	0.99	1.17	3.58
1.0–3.0	0.10–0.40	0.10–0.40	0.49	0.91	1.00	1.09	2.04
1.0–3.0	0.10–0.40	0.30–0.60	0.60	0.94	1.00	1.07	1.67
1.0–3.0	0.30–0.60	0.01–0.30	0.40	0.88	1.00	1.13	2.47
1.0–3.0	0.30–0.60	0.10–0.40	0.60	0.94	1.00	1.07	1.67
1.0–3.0	0.30–0.60	0.30–0.60	0.69	0.96	1.00	1.05	1.45

* CIR, confounding interaction ratio.

exposure prevalence, and background disease risk. The ranges of the genotype frequencies and the exposure prevalence considered are 0.01–0.30, 0.10–0.40, or 0.30–0.60. Note that the variations in terms of the difference between the largest and the smallest values are ~0.3 for all the situations considered, but that the variations in terms of odds ratios of the largest and the smallest values are extremely disparate: 42.43 (for 0.01–0.30), 6.00 (for 0.10–0.40), and 3.50 (for 0.30–0.60). The risk ratio ranges of the background disease considered are 1.0–1.5 or 1.0–3.0.

Within the indicated ranges above, the genotype frequency and the exposure prevalence are spaced to be equidistant on a logistic scale (for eight values), while the background disease risk ratio is spaced to be equidistant on a logarithmic scale (also for eight values). In the simulation, we fix the background disease risk ratio for the eight strata, and in each round of the simulation, we randomly permute the eight values for the genotype frequencies and the eight values for the exposure prevalences to the eight strata. The simulation was done for a total of 100,000 times. Using the equations in the previous section, we calculate the CIRs for all the simulations and show the 25th, 50th, and 75th percentiles in table 1. The same table also presents the upper and the lower bounds of CIR over these ranges.

From table 1, we see that the estimated interaction effect in a case-only study is frequently biased by more than 5 percent (CIR $\geq$1.05 or $\leq$ 0.95), a nonnegligible amount of bias. For a rarer gene and/or a rarer exposure, the bias becomes larger. For example, with a genotype frequency range of 0.01–0.30 and an exposure prevalence range of 0.01–0.30, the bias is usually over 30 percent. The theoretical analysis of the bounds of CIR reveals that up to 1,000 percent bias is possible. On the other hand, the magnitude in the variation in background disease risks has no influence on the magnitude of bias.

DISCUSSION

The impacts of population stratification bias on the ''main genetic effect'' in a case-control study have been extensively studied in the literature, both theoretically and empirically (12–19). It is of interest to compare the results of the ''main-effect/case-control'' in those studies and the ''interaction/case-only'' in the present study. On the surface, the biases in both situations appear similar, viewed through their respective mathematical formulas (13). The variation in the genotype frequency odds is involved in both situations. The variation in the exposure prevalence odds is involved in the present interaction/case-only situation but not in the main-effect/case-control situation, and the variation in the background disease risks is to the contrary. However, the magnitudes of the biases in the two situations are drastically different: Wacholder et al. (14) reported that the main-effect/

case-control bias is generally no more than 1 percent, whereas our study showed that the interaction/case-only bias is frequently over 5 percent and could be alarmingly larger. The reason for this disparity lies in that ''minor'' difference in formulas: the variation in ''background disease risks'' (main effect/case-control) versus the variation in ''exposure prevalence odds'' (interaction/case-only). It turns out that the measurement scale of an ''odds'' acts as a bias magnifier: What appears to be a nothing-out-of-the-ordinary variation in exposure prevalence from 0.1 to 0.4 can easily translate into a sixfold variation when measured in prevalence odds.

Wang et al. (20) studied the impacts of population stratification bias on gene-environment interaction in a case-control study, through computer simulations. They concluded that the interaction/case-control bias is small to nonexistent. This is in sharp contrast to our interaction/case-only situation, where the bias is real and isn't to be taken lightly. A possible explanation for this disparity is that the exposed and the unexposed subgroups of a case-control study suffer from a similar population stratification bias. Therefore, the bias in a gene-environment interaction and the ratio of the main genetic effects between the exposed and the unexposed cancel each other out. By contrast, our interaction/case-only situation does not have two subgroups to cancel each other out.

In this paper, we have shown that, in addition to the bias due to gene-environment nonindependence, the bias due to population stratification is also a major threat to the validity of a case-only study. One may argue that the latter is but a special case for the former, because population stratification by itself also creates gene-environment nonindependence at the population level. However, we believe that distinguishing these two types of biases is important. For the ''ordinary'' gene-environment nonindependence bias, the researchers should try to evaluate on the basis of their subject-matter knowledge whether the gene under study will influence the exposure status, from the point of view of an individual subject in the population. For the population stratification bias, the attention should, instead, be paid to the study population itself. If the researchers have some rough guess about the ratio of the largest and the smallest genotype frequency odds (υ_G) and that of the exposure prevalence odds (υ_E), they can use the boundary formula in this paper to set the limits of the population stratification bias. This would help for a more prudent interpretation of the results. To really achieve control of the two types of biases in a case-only study, a stratified analysis (11) or a modeling approach (21, 22) should be used, adjusting for any confounding factor causing the gene-environment nonindependence in the population, as well as any variable, such as race, ethnicity, nationality, ancestry, and birthplace, that helps to delineate population strata.

ACKNOWLEDGMENTS

This paper is partly supported by funding from the National Science Council, Taiwan, Republic of China (grants NSC 95-2314-B-002-242, NSC 95-3114-P-002-005-Y, and NSC 96-2314-B-002-143).

Conflict of interest: none declared.

REFERENCES

1. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet 2005;6:287–98.
2. Olden K. Commentary: From phenotype, to genotype, to gene-environment interaction and risk for complex diseases. Int J Epidemiol 2007;36:18–20.
3. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 2007;63:111–19.
4. Hwang SJ, Beaty TH, Liang KY, et al. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 1994;140:1029–37.
5. Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. Am J Epidemiol 1997;146:596–604.
6. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994;13:153–62.
7. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996;144:207–13.
8. Yang Q, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 1997;146:713–20.
9. Hamajima N, Yuasa H, Matsuo K, et al. Detection of gene-environment interaction by case-only studies. Jpn J Clin Oncol 1999;29:490–3.
10. Albert PS, Ratnasinghe D, Tangrea J, et al. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 2001;154:687–93.
11. Gatto NM, Campbell UB, Rundle AG, et al. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int J Epidemiol 2004;33:1014–24.
12. Lee WC, Wang LY. Reducing population stratification bias: stratification matching is better than exposure matching. J Clin Epidemiol (in press).
13. Lee WC, Wang LY. Simple formulas for gauging the potential impacts of population stratification bias. Am J Epidemiol 2008;167:86–9.
14. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000;92:1151–8.
15. Millikan RC. Re: Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. (Letter). J Natl Cancer Inst 2001;93:156–8.
16. Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. Am J Hum Genet 2002;71:304–11.
17. Pankow JS, Province MA, Hunt SC, et al. Regarding ''Testing for population subdivision and association in four case-control studies.'' (Letter). Am J Hum Genet 2002;71:1478–80.

18. Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in case-control association studies of admixed populations. Genet Epidemiol 2004;27:14–20.
19. Helgason A, Yngvadottir B, Hrafnkelsson B, et al. An Icelandic example of the impact of population structure on association studies. Nat Genet 2005;37:90–5.
20. Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions. Cancer Epidemiol Biomarkers Prev 2006;15:124–32.
21. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 2005;92:399–418.
22. Cheng KF. A maximum likelihood method for studying gene-environment interactions under conditional independence of genotype and exposure. Stat Med 2006;25:3093–109.