



Selection Bias in the Assessment of Gene-Environment Interaction in Case-Control Studies

Libby M. Morimoto^{1,2}, Emily White^{1,2}, and Polly A. Newcomb^{1,2,3}

¹ Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA.

² Department of Epidemiology, School of Public Health and Community Medicine, University of Washington, Seattle, WA.

³ University of Wisconsin Comprehensive Cancer Center, Madison, WI.

Received for publication October 16, 2002; accepted for publication February 25, 2003.

Selection bias is a common concern in epidemiologic studies, particularly case-control studies. Selection bias in the odds ratio occurs when participation depends jointly on exposure and disease status. General results for understanding when selection bias may affect studies involving gene-environment interactions have not yet been developed. In this paper, the authors show that the assessment of gene-environment interactions will not be subject to selection bias under the assumption that genotype does not influence participation conditional on exposure and disease status. This is true even when selection, including self-selection of subjects, is jointly influenced by exposure and disease and regardless of whether the genotype is related to exposure, disease, or both. The authors present an example to illustrate this concept.

bias (epidemiology); case-control studies; environment; epidemiologic methods; genotype; polymorphism (genetics); risk factors; selection bias

In the search for causative agents of human diseases, both environmental and genetic risk factors have been identified. The relative influence of the two is highly variable, yet for most diseases, it is unlikely that purely environmental or purely genetic etiologies will sufficiently explain the observed variability in disease occurrence. One challenge in epidemiologic studies is determining the nature and extent to which environmental agents and genetic factors influence disease risk, both as independent factors and as modifiers of each other. Such studies have led investigators to examine the role of genetic susceptibility in exposure-disease relations. Studies examining the modifying effect of genes on environmental exposures, often referred to as gene-environment interactions, are increasingly common. Such analyses can identify genetic subpopulations of persons for whom risk factors are most relevant, as well as clarify the biologic mechanisms of exposure-disease relations.

Selection bias occurs in epidemiologic studies when there are systematic differences in characteristics between persons who are selected for study and those who are not (1). Such differences can arise from the procedures used to select subjects and/or from factors that influence study participation (self-selection), the end result being that the relation between exposure and disease is different for persons who

participate and persons who should theoretically be eligible for study (2). It is well known that selection bias occurs when response proportions are jointly dependent on exposure and disease status (3, 4). Selection bias is a particular problem in case-control studies, since exposure and disease have both occurred at the time study subject selection is made (5). While selection bias is a possibility even when response rates for recruitment are high for both cases and controls, it is a particular concern when response rates are low. Studies analyzing participation rates have found that responders and nonresponders differ with regard to various characteristics, such as age, employment status, and race (6). People who have the disease under study are more likely to participate in a study than nondiseased controls (7). Participation may also depend on exposures—for example, persons with exposures perceived to be socially unacceptable, such as consumption of alcoholic beverages or smoking, may be less likely to participate (8).

While selection bias is a major concern in exposure-disease associations, it can also be a concern in gene-disease associations, because genotype could be associated with exposures that influence participation. However, we show here that the assessment of gene-environment interactions will not be subject to selection bias under the likely scenario

Target population		
	Cases	Controls
Exposed	A	B
Unexposed	C	D
$OR_T = AD/BC$		

Selection probabilities		
	Cases	Controls
Exposed	α	β
Unexposed	γ	δ

Epidemiologic study population		
	Cases	Controls
Exposed	a	b
Unexposed	c	d
$OR_O = ad/bc$		

FIGURE 1. Exposure-disease 2×2 tables for target and study populations. OR_T , target population odds ratio; OR_O , observed epidemiologic study population odds ratio.

that genotype does not influence participation conditional on the exposure and disease status. This is true even when selection, including self-selection of subjects, is jointly influenced by exposure and disease and regardless of whether the genotype is related to exposure, disease, or both.

METHODS AND RESULTS

Selection bias in the exposure-by-disease 2×2 table

The degree of selection bias in the estimate of exposure-disease relations for a dichotomous exposure and a dichotomous disease can be expressed in terms of selection proportions. Suppose the number of people in the exposure-by-disease 2×2 table representing the target population of the epidemiologic study is as shown in figure 1. The target (true) exposure-disease odds ratio is $OR_T = AD/BC$. In the observed case-control study, which is subject to selection bias due to sample selection and self-selection, the number of subjects by exposure and disease status may be expressed as in figure 1. In this case, the observed exposure-disease relation is $OR_O = ad/bc$.

Selection proportions are the proportions of persons in the target population who participate in the epidemiologic study, and they are influenced by sampling methods and self-selection by the study subjects. For each of the disease \times exposure groups, the selection probabilities are as follows: α = selection of persons with the exposure and the disease (a/A); β = selection of persons with the exposure but not the disease (b/B); γ = selection of persons without the exposure and with the disease (c/C); and δ = selection of persons without the exposure or the disease (d/D). The observed exposure-disease odds ratio, relative to the true odds ratio, may be expressed as $OR_O = ad/bc = \alpha A \delta D / \beta B \gamma C = (\alpha \delta / \beta \gamma) \times OR_T$.

If the cross-product $\alpha \delta / \beta \gamma = 1$, then selection bias does not affect estimates. For example, there is no selection bias when disease influences participation, but within disease groups, the response proportions for those exposed and not exposed

are the same (i.e., $\alpha = \gamma$ and $\beta = \delta$). Similarly, there is no selection bias when only exposure is related to participation rates (i.e., $\alpha = \beta$ and $\gamma = \delta$) or if these two effects are independent. If disease and exposure status jointly affect participation rates, that is, if $\alpha \delta / \beta \gamma \neq 1$, then the observed odds ratio is biased (4).

Selection bias in gene-environment interaction

In a gene-environment interaction, one stratifies subjects on genotype status to determine whether the exposure-disease relation differs according to genotype. Then the distributions of exposure and disease in the target population and the exposure-disease relation, within genotype, are as shown in figure 2. The true interaction odds ratio (OR_{INT-T}), defined (in this case) as the multiplicative factor by which the mutant exposure-disease odds ratio differs from the wild-type exposure-disease odds ratio, is $OR_{INT-T} = (A_2 D_2 / B_2 C_2) / (A_1 D_1 / B_1 C_1)$. (OR_{INT} is analogous to e^c , where c is the coefficient of the interaction term in a logistic model with terms for exposure (1 = exposed, 0 = not exposed), genotype (1 = mutant, 0 = wild-type), and their interaction.)

Similarly, in an epidemiologic study, the distributions of exposure and disease and the exposure-disease relation by genotype may be represented as in figure 2. The observed interaction odds ratio, OR_{INT-O} , is $OR_{INT-O} = (a_2 d_2 / b_2 c_2) / (a_1 d_1 / b_1 c_1)$. The selection proportions among persons with the wild-type genotype are denoted by α_1 , β_1 , γ_1 , and δ_1 , and the selection proportions among persons with mutant genotypes are denoted by α_2 , β_2 , γ_2 , and δ_2 . Therefore, the observed interaction odds ratio can be represented as

$$OR_{INT-O} = (\alpha_2 A_2 \delta_2 D_2 / \beta_2 B_2 \gamma_2 C_2) / (\alpha_1 A_1 \delta_1 D_1 / \beta_1 B_1 \gamma_1 C_1) = [(\alpha_2 \delta_2 / \beta_2 \gamma_2) / (\alpha_1 \delta_1 / \beta_1 \gamma_1)] \times OR_{INT-T}$$

Under the circumstances described in the section above, each of the stratum- (genotype-) specific odds ratios would be affected by selection bias. However, if genotype does not influence response proportions conditional on exposure and disease status, that is, if $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$, and $\delta_1 = \delta_2$, then the observed interaction odds ratio is equal to the target interaction odds ratio: $OR_{INT-O} = OR_{INT-T}$. Thus, under these conditions, the interaction odds ratio is *not* affected by selection bias even though the stratum-specific odds ratios are biased.

Example: alcohol consumption, aldehyde dehydrogenase 2, and esophageal cancer risk

To illustrate this result, let us consider a hypothetical study evaluating the relation of alcohol consumption, aldehyde dehydrogenase 2 (one of the key enzymes in the oxidation of alcohol to acetate), and esophageal cancer risk. A polymorphism in the aldehyde dehydrogenase 2 gene *ALDH2* caused by a structural point mutation that results in a Glu \rightarrow Lys substitution is associated with phenotypic loss of enzymatic activity (9). Persons with this allele, *ALDH2**2, are deficient in aldehyde dehydrogenase 2 activity and tend to refrain from excessive alcohol drinking because of their adverse

Target population

Genotype wild-type

	Cases	Controls
Exposed	A ₁	B ₁
Unexposed	C ₁	D ₁

$OR_{T1} = A_1D_1/B_1C_1$

Genotype mutant

	Cases	Controls
Exposed	A ₂	B ₂
Unexposed	C ₂	D ₂

$OR_{T2} = A_2D_2/B_2C_2$

$$OR_{INT-T} = OR_{T2}/OR_{T1} = (A_2D_2/B_2C_2)/(A_1D_1/B_1C_1)$$

Selection probabilities

Genotype wild-type

	Cases	Controls
Exposed	α ₁	β ₁
Unexposed	γ ₁	δ ₁

Genotype mutant

	Cases	Controls
Exposed	α ₂	β ₂
Unexposed	γ ₂	δ ₂

Epidemiologic study population

Genotype wild-type

	Cases	Controls
Exposed	a ₁	b ₁
Unexposed	c ₁	d ₁

$OR_{O1} = a_1d_1/b_1c_1$

Genotype mutant

	Cases	Controls
Exposed	a ₂	b ₂
Unexposed	c ₂	d ₂

$OR_{O2} = a_2d_2/b_2c_2$

$$OR_{INT-O} = OR_{O2}/OR_{O1} = (a_2d_2/b_2c_2)/(a_1d_1/b_1c_1)$$

FIGURE 2. Exposure-disease 2 × 2 tables for target and study populations, stratified by genotype. OR_T, target population odds ratio; OR_O, observed epidemiologic study population odds ratio; OR_{INT-T}, target population interaction odds ratio; OR_{INT-O}, observed epidemiologic study population interaction odds ratio.

reaction to alcohol (10). This polymorphism is highly prevalent in Asian populations, although it is rare in other ethnic groups (10, 11). Because this gene encodes an enzyme that is critical for the elimination of acetaldehyde (a carcinogen) generated by alcohol, the *2 allele (which is inactive) is associated with several cancers, including esophageal cancer (12). While case-only analyses of gene-environment interactions require independence between the gene and the environment, this assumption is not required in order for case-control analyses to yield unbiased results (13).

Let us assume that our target population is selected to have a high prevalence of this polymorphism (i.e., an Asian population) and that there are 1,000 cases of esophageal cancer and 1,000,000 cancer-free individuals. Figure 3 shows the distributions of exposure, genotype, and disease in this hypothetical target population. Assume that among cases, 73 percent have high alcohol consumption, while among controls, 50 percent have high alcohol consumption, such that the true exposure-disease odds ratio is 2.75. In this population, the frequency of homozygosity for the *2 allele (called “mutant” in this example) is assumed to be 50 percent, while persons with the other genotypes (the homozygous wild-types and heterozygotes, called “wild-type”) comprise the remaining 50 percent of the population. Among the cases, 75 percent have the mutant genotype and 25 percent have the wild-type genotype. Thus, the true gene-

disease odds ratio in this population is 3.0. Finally, among persons with the mutant genotype, the proportion with high alcohol consumption is 40 percent, while the proportion with high alcohol consumption among those with the wild-type genotype is 60 percent. Therefore, the exposure-genotype odds ratio is 0.45. Among persons with the mutant genotype, the true exposure-disease odds ratio is 4.0, while among those with the wild-type genotype, the exposure-disease odds ratio is 2.0; this results in an interaction odds ratio of 2.0.

In our hypothetical observational epidemiologic study, 100 percent of eligible cases and 0.1 percent of eligible controls are selected to participate. However, because of self-selection in our case population, only 85 percent of cases with high alcohol consumption agree to participate, while 90 percent of cases with low alcohol consumption agree to participate. Among controls, only 50 percent of those with high alcohol consumption agree to participate, while 90 percent of those with low alcohol consumption agree to participate. Thus, the selection probabilities are α = 0.85, β = 0.0005, γ = 0.90, and δ = 0.0009 and are independent of genotype within each exposure/disease stratum. The numbers of participants in each disease and exposure stratum who enter the epidemiologic study are shown in figure 3. Since the cross-product of the selection probabilities is 1.7, the observed odds ratio is biased by a

Target population*Exposure-disease relation*

Alcohol consumption	Cases	Controls
High	733	500,000
Low	267	500,000

$OR_T = 2.75$

Gene-disease relation

	Cases	Controls
Mutant	750	500,000
Wild-type	250	500,000

$OR_T = 3.00$

Gene-exposure relation

	Mutant	Wild type
High	200,545	300,188
Low	300,205	200,062

$OR_T = 0.45$

Exposure-disease relation, stratified by genotype

Mutant

Alcohol consumption	Cases	Controls
High	545	200,000
Low	205	300,000

$OR_{T2} = 4.00$

Wild-type

Alcohol consumption	Cases	Controls
High	188	300,000
Low	62	200,000

$OR_{T1} = 2.00$

Interaction odds ratio

$$OR_{INT-T} = 4.00/2.00 = 2.00$$

Epidemiologic study population*Exposure-disease relation*

Alcohol consumption	Cases	Controls
High	623	250
Low	241	450

$OR_O = 4.65$

Gene-disease relation

	Cases	Controls
Mutant	648	370
Wild-type	216	330

$OR_O = 2.68$

Gene-exposure relation

	Mutant	Wild type
High	563	310
Low	455	236

$OR_O = 0.94$

Exposure-disease relation, stratified by genotype

Mutant

Alcohol consumption	Cases	Controls
High	463	100
Low	185	270

$OR_{O2} = 6.76$

Wild-type

Alcohol consumption	Cases	Controls
High	160	150
Low	56	180

$OR_{O1} = 3.43$

Interaction odds ratio

$$OR_{INT-O} = 6.76/3.43 = 2.00$$

FIGURE 3. Example of the lack of bias in the interaction odds ratio when selection probabilities do not depend on genotype conditional on exposure and disease. OR_T , target population odds ratio; OR_O , observed epidemiologic study population odds ratio; OR_{INT-T} , target population interaction odds ratio; OR_{INT-O} , observed epidemiologic study population interaction odds ratio. Under selection probabilities $\alpha = 0.85$, $\beta = 0.0005$, $\gamma = 0.90$, and $\delta = 0.0009$, $\alpha\delta/\beta\gamma = 1.70$.

factor of 1.7, resulting in an observed odds ratio of 4.65 ($2.75 \times 1.7 \approx 4.65$). The joint effect of case status and exposure status affects participation in this example; subsequently, our main exposure-disease association is biased.

Similarly, the gene-disease odds ratio and the gene-exposure odds ratio are biased, because genotype is related to exposure and disease, which influence the selection proportions (to compute this, one must use the stratum-specific numbers at the bottom of figure 3). However,

because participation is not affected by genotype itself but is only affected through its association with exposure and disease, stratification on genotype and calculation of the ratio of the exposure-disease odds ratio will not produce a biased estimate of the interaction odds ratio. In this example, we demonstrate that even in the instance where there is a gene-exposure relation and a gene-disease relation, and exposure and disease jointly affect selection probabilities, selection bias in case-control studies does not bias gene-environment interaction estimates.

DISCUSSION

The possibility of introducing selection bias when conducting epidemiologic studies, particularly case-control studies, is a major concern. Reasons for participating in a study differ between cases and controls, and the decision as to whether to participate may be influenced by a person's exposure history as well as his or her disease status. Studies examining genetic exposures may involve an additional layer of complexity, since willingness by participants to provide a biologic specimen may be affected by many of the same, or different, factors (14).

There are limited methods available to correct for the effects of selection bias on estimates of the exposure-disease relation. First, if one has information on the four (exposure-by-disease) selection proportions, one can adjust the observed odds ratio to correct for selection bias (4). However, selection proportions are rarely known, because this requires information on the exposure-by-disease distribution of the study nonrespondents. Secondly, selection bias due to selection factors that are not the main exposure of interest can be adjusted for in the analysis (as confounding factors) to correct for selection bias (2), but this technique cannot be applied when the exposure of interest is a selection factor.

Therefore, one usually cannot correct for selection bias if the exposure of interest influences selection (jointly with disease), and this may bias the existence, strength, and direction of main associations between exposures and disease. However, we have shown that the assessment of gene-environment interaction odds ratios in epidemiologic studies is not affected by selection bias when the genotype does not influence selection conditional on exposure and disease status. Wacholder et al. (15) recently presented a related result for selection bias in gene-environment interactions in case-control studies using hospital controls. Their results apply to the more limited situation in which the only sources of selection bias are the risk factors for the control disease (i.e., the controls do not have the same gene-exposure distribution as the ideal "target" controls, and this leads to the selection bias). Wacholder et al. concluded that there is no bias in the estimation of gene-environment interactions for the disease of interest when there is no gene-environment interaction for the control disease, even when the control condition is caused by the genetic or environmental factor.

The main assumption for our results, that genotype does not influence selection conditional on exposure and disease, seems likely to be true in most situations. Specifically, it seems reasonable to assume that one's genotype cannot influence participation in a study, other than through some association of genotype with phenotype or with behavior. Genotype could, of course, influence participation through the relation of genotype with selection factors other than the exposure of interest, in which case the gene-environment interaction estimate will be biased as well. As an illustration of this, let us assume in the above example that the allele frequency of our polymorphism in *ALDH2* is more common among certain Asian ethnic groups and that persons from those groups are less likely to participate as controls than persons from other racial groups. Then the gene-environment interaction estimate will be biased. This is because the genotype affects participation rates, not directly but through its association with another

selection factor. However, as we noted above, the selection bias contributed by a selection factor (in this case, ethnic group) other than the main exposure can be controlled for by controlling for the selection factor in the statistical analysis (2). Thus, the effect of this source of selection bias can be corrected using standard adjustment methods.

In summary, studies in which high nonparticipation rates raise concerns about selection bias are still able to generate valid estimates of gene-environment interactions. Identification of such interaction can help to define and clarify biologic pathways between exposure and disease in certain subsets of people.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health training grant 5-T32-CA09168.

REFERENCES

1. Last JM, ed. A dictionary of epidemiology. New York, NY: Oxford University Press, 1995.
2. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven Publishers, 1998.
3. Greenland S. Response and follow-up bias in cohort studies. *Am J Epidemiol* 1977;106:184-7.
4. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.
5. Hennekens CH, Buring JE. Epidemiology in medicine. Boston, MA: Little, Brown and Company, 1987.
6. Armstrong BK, White E, Saracci R. Principles of exposure measurement in epidemiology. New York, NY: Oxford University Press, 1992.
7. Newcomb PA, Storer BE, Longnecker MP, et al. Pregnancy termination in relation to risk of breast cancer. *JAMA* 1996;275:283-7.
8. Schlesselman JJ. Case-control studies. New York, NY: Oxford University Press, 1982.
9. Yoshida A, Huang IY, Ikawa M. Molecular abnormality of an inactive aldehyde dehydrogenase variant commonly found in Orientals. *Proc Natl Acad Sci U S A* 1984;81:258-61.
10. Crabb DW, Edenberg HJ, Bosron WF, et al. Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity: the inactive *ALDH2*(2) allele is dominant. *J Clin Invest* 1989;83:314-16.
11. Frenzer A, Butler WJ, Norton ID, et al. Polymorphism in alcohol-metabolizing enzymes, glutathione S-transferases and apolipoprotein E and susceptibility to alcohol-induced cirrhosis and chronic pancreatitis. *J Gastroenterol Hepatol* 2002;17:177-82.
12. Yokoyama A, Muramatsu T, Ohmori T, et al. Oesophageal cancer and aldehyde dehydrogenase-2 genotypes in Japanese males. *Cancer Epidemiol Biomarkers Prev* 1996;5:99-102.
13. 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.
14. Engel LS, Rothman N, Knott C, et al. Factors associated with refusal to provide a buccal cell sample in the Agricultural Health Study. *Cancer Epidemiol Biomarkers Prev* 2002;11:493-6.
15. Wacholder S, Chatterjee N, Hartge P. Joint effect of genes and environment distorted by selection biases: implications for hospital-based case-control studies. *Cancer Epidemiol Biomarkers Prev* 2002;11:885-9.