

## Counter-Matching in Studies of Gene-Environment Interaction: Efficiency and Feasibility

N. Andrieu,<sup>1</sup> A. M. Goldstein,<sup>2</sup> D. C. Thomas,<sup>3</sup> and B. Langholz<sup>3</sup>

The interest in studying gene-environment interaction is increasing for complex diseases. However, most methods of detecting gene-environment interactions may not be appropriate for the study of interactions involving rare genes ( $G$ ) or uncommon environmental exposures ( $E$ ), because of poor statistical power. To increase this power, the authors propose the counter-matching design. This design increases the number of subjects with the rare factor without increasing the number of measurements that must be performed. In this paper, the efficiency and feasibility (required sample sizes) of counter-matching designs are evaluated and discussed. Counter-matching on both  $G$  and  $E$  appears to be the most efficient design for detecting gene-environment interaction. The sensitivity and specificity of the surrogate measures, the frequencies of  $G$  and  $E$ , and, to a lesser extent, the value of the interaction effect are the most important parameters for determining efficiency. Feasibility is also more dependent on the exposure frequencies and the interaction effect than on the main effects of  $G$  and  $E$ . Although the efficiency of counter-matching is greatest when the risk factors are very rare, the study of such rare factors is not realistic unless one is interested in very strong interaction effects. Nevertheless, counter-matching appears to be more appropriate than most traditional epidemiologic methods for the study of interactions involving rare factors. *Am J Epidemiol* 2001;153:265–74.

case-control studies; cohort studies; epidemiologic methods; interaction; matched-pair analysis; research design; statistics

The interest in studying gene-environment interaction is increasing for most chronic and complex diseases, mostly because of considerable advances in molecular genetic techniques. However, investigation of interactions requires sample sizes much larger than those needed to investigate main effects. Smith and Day (1) showed that detecting interactions of the same magnitude as postulated main effects in 1:1 unmatched case-control studies always requires increases in study size of at least a factor of 4 (and, in some circumstances, considerably more). Methods of detecting gene-environment interactions have been reviewed (2–4). Most methods may not be appropriate for the study of interactions involving rare genes or uncommon environmental exposures, particularly for moderate values of the interaction effect. Required sample sizes are often unattainable, ranging from tens of thousands to hundreds of thousands (or more) of study subjects.

To increase a study's power to detect a gene-environment interaction when one of the factors under study is rare, one possible alternative is the counter-matching design. Counter-matching was introduced by Langholz and Clayton (5) as a method of sampling controls from a cohort, or more generally from an at-risk population, for nested case-control studies. One purpose of the design is to increase the numbers of cases and controls with the rare factor of interest without prohibitively increasing the number of measurements that must be performed. The goal of counter-matching is to maximize the number of discordant case-control pairs, from which information comes in a matched case-control study. The efficiency of this method in assessing main effects of uncommon factors has already been evaluated. Counter-matching has been shown to increase the efficiency of main effect estimation by approximately 25 percent in comparison with classical random sampling (6). In recent work, Cologne and Langholz found that counter-matching was advantageous in assessing interaction between two factors for which data on one were available for the entire cohort and data on the other were to be obtained from the sample (J. B. Cologne, Radiation Effects Research Foundation (Hiroshima, Japan) and B. Langholz, University of Southern California (Los Angeles, California), personal communication, 1999). The design they explored is quite different from that considered here, in that information on one of the exposures (used for the counter-matching) is known for the entire cohort and information on the other is collected in the sample. Furthermore, Cologne and Langholz were interested only in

Received for publication June 16, 1999, and accepted for publication February 14, 2000.

Abbreviations: ARE, asymptotic relative efficiency; RR, rate ratio.

<sup>1</sup>Unité de Recherche en Épidémiologie des Cancers, Institut de la Santé et de la Recherche Médicale (INSERM) U521, Institut Gustave-Roussy, 94805 Villejuif, France.

<sup>2</sup>Genetic Epidemiology Branch, National Cancer Institute, Bethesda, MD.

<sup>3</sup>Department of Preventive Medicine, School of Medicine, University of Southern California, Los Angeles, CA.

Reprint requests to Dr. Nadine Andrieu, INSERM U521, Institut Gustave-Roussy, 94805 Villejuif, France (e-mail: nandrieu@igr.fr).

the interaction effect of the exposures, not in the main effects. Assessment of effects of gene-environment interaction without knowledge about both the genetic and the environmental main effects might be of little use for public health or individual risk assessment. In this paper, we propose counter-matching designs that allow for estimation of the gene-environment interactive effect as well as both genetic and environmental main effects. The efficiency and feasibility of this counter-matching design are evaluated and discussed for different interaction scenarios.

## MATERIALS AND METHODS

### Models for interaction between a gene and an environmental exposure

We used the following parameters for modeling an interaction between a genetic factor ( $G$ ) and an environmental exposure ( $E$ ).  $E$  and  $G$  are assumed to be dichotomous, with  $E = 1$  indicating the exposed-to- $E$  status and  $G = 1$  indicating the susceptible genotype; exposure,  $E$ , and genotype,  $G$ , are assumed to occur independently. Let  $P(G)$  equal the frequency of the susceptible genotype in the cohort/population at risk and  $P(E)$  the frequency of the environmental factor in the cohort/population at risk.  $RR_{eg}$  is the rate ratio (RR) for environmental exposure and the genetic factor, with  $e = 1$  denoting exposure to  $E$  and  $e = 0$  nonexposure to  $E$ , and with  $g = 1$  denoting the presence of  $G$  and  $g = 0$  the absence of  $G$ . Thus, the rate ratio when  $e = 1$  and  $g = 0$  is  $RR_{10}$  and is denoted  $RR_E$ ; the rate ratio when  $e = 0$  and  $g = 1$  is  $RR_{01}$  and is denoted  $RR_G$ ; and the rate ratio when  $e = 1$  and  $g = 1$  is  $RR_{11}$  and is denoted  $RR_{EG}$ .

The classical definition of interaction was used for this analysis and is as follows. Gene-environment interaction exists if the joint effect of the genetic factor and the environmental exposure differs from the product of the risks for the individual factors on a multiplicative scale ( $RR_{int} = RR_{EG}/(RR_E RR_G)$ ). An interaction effect of more than 1 indicates a greater than multiplicative effect between  $E$  and  $G$ , while an interaction effect less than 1 indicates a less than multiplicative effect. An additive effect may also be considered when the joint effect of the genetic factor and the environmental exposure differs from the sum of the background disease rate and the excess rates for the environmental exposure and the genetic factor ( $RR_{int} = RR_{EG}/(RR_E + RR_G - 1)$ ). However, this exercise focuses on the multiplicative model, the model most commonly used in chronic disease epidemiology.

### Counter-matching for gene-environment interaction

The design and analysis of counter-matched studies are presented in detail elsewhere (6–8) and thus are only briefly described below. In counter-matching, controls are selected to increase the variation in factors of interest in a case-control set relative to random sampling. The goal is thus the opposite of that of matching, where one selects factors for controls that are similar to the cases' factors. A partial likelihood method has been developed for estimating different exposure effects in counter-matching (9) using weighting

that takes into account the probabilities that subjects were selected from specific strata.

In general, the number of case-control subjects from each factor-of-interest status group is fixed by the design. Here, three different variants of counter-matching for assessment of gene-environment interaction are proposed. In order to make assessment of main effects possible, we suppose that surrogates for  $G$  and  $E$  are available for the entire cohort/population at risk in which the case-control study is nested. Thus, counter-matching is performed either on the genetic factor or on the environmental factor, or on both the genetic and the environmental factors. Each case's risk set would be stratified by either a surrogate of  $G$  or a surrogate of  $E$ , or by surrogates of both  $G$  and  $E$ , and controls for that risk set would be selected from the strata other than the case's stratum.

We compared the following population-based designs to evaluate the efficiency of counter-matching in a study examining gene-environment interaction: 1) a full cohort study with no matching and with infinite numbers of controls for each case; 2) a standard nested case-control study with three controls per case; 3) a 2-2 case-control design with counter-matching on a surrogate of  $E$ ; 4) a 2-2 case-control design with counter-matching on a surrogate of  $G$ ; and 5) a 1-1-1-1 case-control design with counter-matching on surrogates of both  $E$  and  $G$ .

The third and fourth designs have two individuals exposed and two unexposed for either the  $G$  surrogate ( $G_{sur}$ ) or the  $E$  surrogate ( $E_{sur}$ ), respectively. For example, if a case is exposed for a given surrogate, then one exposed control and two unexposed controls for the given surrogate are drawn. If a case is unexposed for a given surrogate, then one unexposed control and two exposed controls for the given surrogate are drawn. Design 5 includes one individual who is unexposed for both surrogates, one who is exposed for both surrogates, one who is exposed for  $E_{sur}$  and unexposed for  $G_{sur}$ , and one who is exposed for  $G_{sur}$  and unexposed for  $E_{sur}$ . Thus, for designs 2–5, each sampled risk set includes one case and three controls, with controls sampled according to  $E_{sur}$  or/and  $G_{sur}$  status, depending on the design.

### Calculating asymptotic relative efficiency

To evaluate efficiency, we calculate the asymptotic relative efficiency (ARE). For assessment of gene-environment interaction, ARE is defined as the ratio of the gene-environment interaction variance for each counter-matched case-control design to the variance for either the classical 1:3 nested case-control study or the full cohort. The ratio indicates proportionally how many more (or fewer) observations (in large samples) are needed by the counter-matched design to achieve the same precision as the reference design (10). Asymptotic variances are calculated as described by Langholz and Borgan (9), using a FORTRAN program developed by Langholz (11).

### Calculations of sample size

To calculate sample sizes for a gene-environment interaction study using counter-matching, we use the following

classical formulation as described by Breslow and Day (12):

$$n = \frac{(z_{\alpha}\sigma_{\text{intH}_0} + z_{1-\beta}\sigma_{\text{intH}_1})^2}{(\log \text{RR}_{\text{int}})^2},$$

where  $\text{RR}_{\text{int}}$  is the gene-environment interaction effect and  $n$  is the number of sets. The asymptotic variance of  $\log(\text{RR}_{\text{int}})$  under the null hypothesis ( $\sigma_{\text{intH}_0}^2$ ) is calculated with  $\text{RR}_{\text{int}}$  set to 1. The asymptotic variance of  $\log(\text{RR}_{\text{int}})$  under the alternative hypothesis ( $\sigma_{\text{intH}_1}^2$ ) is calculated with  $\text{RR}_{\text{int}}$  set to the alternative value. The numbers of counter-matching sets required to detect a given  $\text{RR}_{\text{int}}$  interaction value are calculated for 80 percent and 90 percent power ( $1 - \beta$ ) using a two-sided test at the 5 percent ( $\alpha$ ) level.

## RESULTS

### Efficiency of counter-matching

*Efficiency of counter-matching according to study design.* Table 1 presents the AREs for different values of the main effect of  $E$  ( $\text{RR}_E = 1, 2$ ),  $G$  ( $\text{RR}_G = 2, 3$ ), and gene-environment ( $G \times E$ ) interaction ( $\text{RR}_{\text{int}} = 2, 5, 10$ ). The RRs are rate ratios for  $E$  ( $\text{RR}_E$ ),  $G$  ( $\text{RR}_G$ ), and the  $G \times E$  interaction term ( $\text{RR}_{\text{int}}$ ). The sensitivity and specificity of the  $G$  and  $E$  surrogates are both fixed at 0.8. Counter-matching on both  $G$  and  $E$  appears to be the most efficient of the four case-control designs, whatever the main and interactive effect values (table 1) and whatever the frequencies of  $E$  and  $G$  (data not shown). For example, when  $\text{RR}_E = 2$ ,  $\text{RR}_G = 3$ , and  $\text{RR}_{\text{int}} = 5$ ,  $\text{ARE} = 1.44$  when counter-matching on  $E$ ,  $\text{ARE} = 1.81$  when counter-matching on  $G$ , and  $\text{ARE} = 2.31$  when counter-matching on both  $E$  and  $G$ , relative to the standard 1:3 nested case-control study. In addition, when counter-matching is performed on only one surrogate, counter-matching on the rarer factor (e.g.,  $G$  in table 1) appears to be more efficient (table 1; data not shown).

Counter-matching on both  $G$  and  $E$  also appears to be the most efficient design for simultaneous detection of main effects. Indeed, although the most efficient design for detecting a given main effect is a design with counter-matching on only the main effect of interest, counter-matching on both  $G$  and  $E$  remains the most efficient when one is interested in detecting the main effects of both  $G$  and  $E$ . For example, if one is interested in detecting the main effect of  $G$  in a situation where  $\text{RR}_E = 2$ ,  $\text{RR}_G = 2$ , and  $\text{RR}_{\text{int}} = 2$ , then  $\text{ARE} = 0.89$  when counter-matching on  $E$ , 1.21 when counter-matching on  $G$ , and 1.14 when counter-matching on both  $E$  and  $G$ , relative to the standard 1:3 nested case-control study. Similarly, when one wants to detect the main effect of  $E$ ,  $\text{ARE} = 1.20$  when counter-matching on  $E$ , 0.86 when counter-matching on  $G$ , and 1.10 when counter-matching on both  $E$  and  $G$  (table 1).

Since counter-matching on both  $G$  and  $E$  appears to be the most efficient case-control design in the detection of gene-environment interaction, we use this design to examine the efficiency when parameters such as frequencies of  $G$  and  $E$ , main effects of  $G$  and  $E$ , or  $G \times E$  interaction effects are varied.

*Efficiency of counter-matching according to the sensitivity and specificity of surrogates.* Figure 1 shows the effects of the sensitivity (proportion of truly exposed (to either  $E$  or  $G$ ) subjects who are so identified by the surrogate ( $E_{\text{sur}}$  or  $G_{\text{sur}}$ , respectively)) and specificity (proportion of truly nonexposed subjects who are so identified by the nonexposed surrogate) of  $G_{\text{sur}}$  on ARE for different frequencies of  $G$  ( $P(G) = 0.01, 0.1, 0.2$ ). The AREs are calculated comparing the design that counter-matches on both  $G$  and  $E$  with the standard 1:3 case-control study. All other parameters are fixed with  $\text{RR}_E = \text{RR}_G = \text{RR}_{\text{int}} = 2$  and  $P(E) = 0.1$ , and the sensitivity and specificity of  $E$  are fixed at 0.8. AREs increase as the sensitivity of  $G_{\text{sur}}$  increases (specificity of  $G_{\text{sur}}$  fixed at 0.8 (figure 1, lines with circles)) or as the specificity of  $G_{\text{sur}}$  increases (sensitivity of  $G_{\text{sur}}$  fixed at 0.8 (figure 1, lines with asterisks)) for different frequencies of  $G$ . In this scenario, the counter-matched design is more efficient than or at least as efficient as the 1:3 case-control study regardless of the specificity, and it becomes more efficient than the 1:3 case-control study when the sensitivity of  $G$  is greater than 0.1. Actually, the threshold of specificity and sensitivity of  $G$  for obtaining a gain in efficiency depends mainly on the fixed values of the sensitivity and specificity of  $E$  and vice versa (data not shown). For example, when the sensitivity and specificity of  $E$  are equal to 0.5, the threshold ( $\text{ARE} = 1$ ) for the sensitivity of  $G$  is equal to 0.3 and the specificity of  $G$  is equal to 0.5. In other words, if one of the two surrogates has low sensitivity and specificity, the other factor must be highly sensitive and specific to produce a gain in efficiency for a gene-environment interaction study. This gain increases as the specificity or sensitivity increases. Thus, high sensitivity and specificity for the surrogates of  $G$  and  $E$  are important for making counter-matching on  $G$  and  $E$  an efficient study design.

Since family history may often be an easily, inexpensively, and efficiently measured surrogate for many genetic factors of interest, we undertook some calculations to determine values that might reasonably be expected for the sensitivity and specificity of family history as a surrogate for genotype. We performed calculations for a single gene under a variety of assumptions about penetrance, dominance, allele frequency, gene-environment interaction, and within-family concordance in exposure. We computed  $P(\text{FH}|G1)$  for nuclear families with two siblings of the case or the control (case/control) under the models described by Witte et al. (13), where  $G1$  denotes the genotype of the case/control and FH (family history) is 1 if either parent or either sibling is affected and 0 otherwise, averaging over the joint distribution of exposures in the family and over the disease status of the case/control.

The results of these calculations are summarized in table 2. Over a wide range of genetic parameters, the specificities  $P(\text{FH} = 0|G1 = aa)$  were consistently high, ranging from approximately 68 percent to 84 percent. Under a recessive model, the specificities for heterozygous ( $Aa$ ) cases/controls  $P(\text{FH} = 0|G1 = Aa)$  were also fairly high, generally about 65–75 percent. Sensitivities were more variable from one set of model parameters to another, but they generally ranged from 40 percent to 90 percent for homozygous ( $AA$ ) cases/controls. For heterozygous cases/controls, the sensi-

**TABLE 1. Asymptotic relative efficiency (ARE) of three counter-matching designs for studying gene (G)-environment (E) interaction in comparison with either a standard nested case-control study design or a full cohort study design\***

	Main effect		Interaction effect		ARE for detecting interaction effect		ARE for detecting main effect (nested case-control study used as referent)	
	RR <sub>E</sub> †	RR <sub>G</sub>	RR <sub>int</sub>	Variance for one set‡	Nested case-control study used as referent	Full cohort study used as referent	Main effect of E	Main effect of G
Full cohort study	1	2	2	320.6		1.00		
1:3 nested case-control study	1	2	2	695.6	1.00	0.46	1.00	1.00
2-2 counter-matching on E	1	2	2	571.0	1.22	0.56	1.12	0.87
2-2 counter-matching on G	1	2	2	511.4	1.36	0.63	0.88	1.23
1-1-1-1 counter-matching on E and G	1	2	2	457.5	1.52	0.70	1.05	1.14
Full cohort study	1	2	5	169.8		1.00		
1:3 nested case-control study	1	2	5	546.5	1.00	0.31	1.00	1.00
2-2 counter-matching on E	1	2	5	410.7	1.33	0.41	1.12	0.87
2-2 counter-matching on G	1	2	5	347.7	1.57	0.49	0.88	1.23
1-1-1-1 counter-matching on E and G	1	2	5	294.5	1.86	0.58	1.05	1.14
Full cohort study	1	2	10	120.0		1.00		
1:3 nested case-control study	1	2	10	500.1	1.00	0.24	1.00	1.00
2-2 counter-matching on E	1	2	10	358.2	1.40	0.34	1.12	0.87
2-2 counter-matching on G	1	2	10	293.4	1.70	0.41	0.88	1.23
1-1-1-1 counter-matching on E and G	1	2	10	238.7	2.10	0.50	1.05	1.14
Full cohort study	2	2	2	208.2		1.00		
1:3 nested case-control study	2	2	2	586.6	1.00	0.35	1.00	1.00
2-2 counter-matching on E	2	2	2	455.4	1.29	0.46	1.20	0.89
2-2 counter-matching on G	2	2	2	391.7	1.50	0.53	0.86	1.21
1-1-1-1 counter-matching on E and G	2	2	2	338.4	1.73	0.62	1.10	1.14
Full cohort study	2	2	5	125.9		1.00		
1:3 nested case-control study	2	2	5	506.7	1.00	0.25	1.00	1.00
2-2 counter-matching on E	2	2	5	367.9	1.38	0.34	1.20	0.89
2-2 case-m on G	2	2	5	300.7	1.70	0.42	0.86	1.21
1-1-1-1 counter-matching on E and G	2	2	5	246.5	2.06	0.51	1.10	1.14
Full cohort study	2	2	10	99.5		1.00		
1:3 nested case-control study	2	2	10	486.3	1.00	0.20	1.00	1.00
2-2 counter-matching on E	2	2	10	342.4	1.42	0.29	1.20	0.89
2-2 counter-matching on G	2	2	10	272.8	1.78	0.36	0.86	1.21
1-1-1-1 counter-matching on E and G	2	2	10	216.6	2.25	0.46	1.10	1.14
Full cohort study	2	3	2	142.7		1.00		
1:3 nested case-control study	2	3	2	521.9	1.00	0.27	1.00	1.00
2-2 counter-matching on E	2	3	2	385.9	1.35	0.37	1.20	0.87
2-2 counter-matching on G	2	3	2	321.0	1.63	0.44	0.86	1.30
1-1-1-1 counter-matching on E and G	2	3	2	266.7	1.96	0.54	1.10	1.20
Full cohort study	2	3	5	87.7		1.00		
1:3 nested case-control study	2	3	5	471.3	1.00	0.19	1.00	1.00
2-2 counter-matching on E	2	3	5	328.3	1.44	0.27	1.20	0.87
2-2 counter-matching on G	2	3	5	260.5	1.81	0.34	0.86	1.30
1-1-1-1 counter-matching on E and G	2	3	5	204.3	2.31	0.43	1.10	1.20
Full cohort study	2	3	10	70.4		1.00		
1:3 nested case-control study	2	3	10	463.3	1.00	0.15	1.00	1.00
2-2 counter-matching on E	2	3	10	314.6	1.47	0.22	1.20	0.87
2-2 counter-matching on G	2	3	10	244.3	1.90	0.29	0.86	1.30
1-1-1-1 counter-matching on E and G	2	3	10	185.7	2.50	0.38	1.10	1.20

\* In this example, P(E) = 0.1, P(G) = 0.01, and sensitivity and specificity are 80% for both the E surrogate and the G surrogate.

† RR, rate ratio.

‡ The expected variance from a study with n sets is the variance per set divided by n.

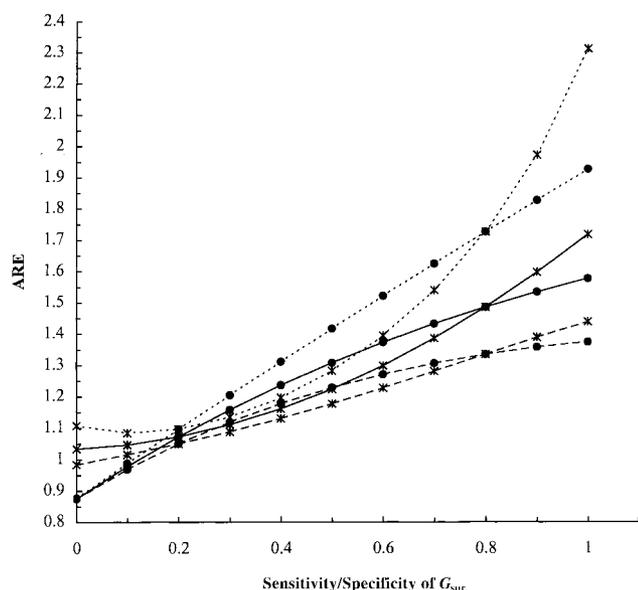
tivities ranged from about 45 percent to 75 percent under a dominant model.

In general, both the sensitivity and the specificity of surrogate family history tended to be somewhat higher for rare major susceptibility genes than for common low penetrance genes, and were barely affected by the degree of environmental concordance. Family history appears to be quite appropriate as a surrogate for genotype in the counter-

matched design, with 80 percent sensitivity and specificity being not unrealistic values for some genetic models.

Thus, we fixed sensitivity and specificity at 80 percent to examine the efficiency when frequencies of G and E, main effects of G and E, or G × E interaction effects are varied.

*Efficiency of counter-matching according to the main effects of G and E.* In parts a and b of figure 2, we examine the AREs for different RR<sub>E</sub> and RR<sub>G</sub> values with sensi-



**FIGURE 1.** Effect of the sensitivity and specificity of a surrogate for genetic exposure ( $G_{sur}$ ) on asymptotic relative efficiency (ARE) for different values of the frequency of  $G$  ( $P(G) = 0.01, 0.1, 0.2$ ), with  $RR_E = RR_G = RR_{int} = 2$ ,  $P(E) = 0.1$ , and the sensitivity and specificity of the surrogate for environmental exposure ( $E$ ) fixed at 0.8. Lines with circles denote the effect of the sensitivity of  $G_{sur}$  on ARE; lines with asterisks denote the effect of the specificity of  $G_{sur}$  on ARE. Key:  $\cdots$ ,  $P(G) = 0.01$ ; —,  $P(G) = 0.1$ ; - - -,  $P(G) = 0.2$ .

tivity and specificity set at 0.8 for both  $G$  and  $E$ ,  $P(E)$  equal to 0.2, and  $P(G)$  equal to 0.01. AREs are calculated for two different values of  $G \times E$  interaction ( $RR_{int} = 3, 10$ ). The results show an increase in ARE as  $RR_G$  increases. For example, when  $RR_{int} = 3$  and  $RR_E = 3$  (figure 2, part *a*,

dashed line with asterisks), the AREs increase from 1.35 when  $RR_G = 1$  to 2.30 when  $RR_G = 10$ . The slopes of the AREs barely change across the range of  $RR_{int}$ . For example, when  $RR_E = 3$ , comparison of AREs for  $RR_G = 10$  versus  $RR_G = 1$  shows increases of 0.93 when  $RR_{int} = 3$  (part *a*, dashed line), 0.95 when  $RR_{int} = 5$  (data not shown), and 0.96 when  $RR_{int} = 10$  (part *b*, dashed line).

Parts *a* and *b* of figure 2 also show a somewhat complicated ARE pattern for given  $RR_G$ 's and  $RR_E$ 's. For example, when  $RR_{int} = 3$ , the ARE is highest for  $RR_E = 3$  and lowest for  $RR_E = 10$ . When  $RR_{int} = 10$ , the AREs decrease as  $RR_E$  increases from 1 to 10. Parts *c* and *d* of figure 2 better illustrate this pattern. For example, when  $P(G) = 0.01$ ,  $RR_{int} = 3$ , and  $RR_G = 3$ , the AREs increase from 1.68 when  $RR_E = 1$  to 1.76 when  $RR_E = 2$  and then decrease to 1.62 when  $RR_E = 10$  (part *c*, line with diamonds). This decrease in ARE for detecting a gene-environment interaction is a consequence of the relative frequencies of the two factors under study and is always observed as the main effect of the more common factor increases (data not shown). That is, as the absolute difference in frequency between  $G$  and  $E$  increases, the AREs will generally decrease as the rate ratio of the more common factor increases.

These results show that as the main effect of the rarer factor increases, AREs for detecting gene-environment interaction increase. In contrast, as the main effect of the more common factor increases, AREs for detecting gene-environment interaction may increase slightly but usually decrease. When the frequencies of  $G$  and  $E$  are equal, AREs increase as the main effect of either  $E$  or  $G$  increases (data not shown).

*Efficiency of counter-matching according to the frequencies of  $G$  and  $E$ .* Figure 3 presents AREs from counter-matching on both  $G$  and  $E$  for different frequencies of  $G$  and  $E$ , with the sensitivity and specificity of both  $G$  and  $E$  fixed

**TABLE 2.** Calculation of the sensitivity and specificity of family history (FH) as a surrogate for a genetic exposure ( $RR_{int} = RR_E = 2$ ,  $P(E) = 0.25$ )

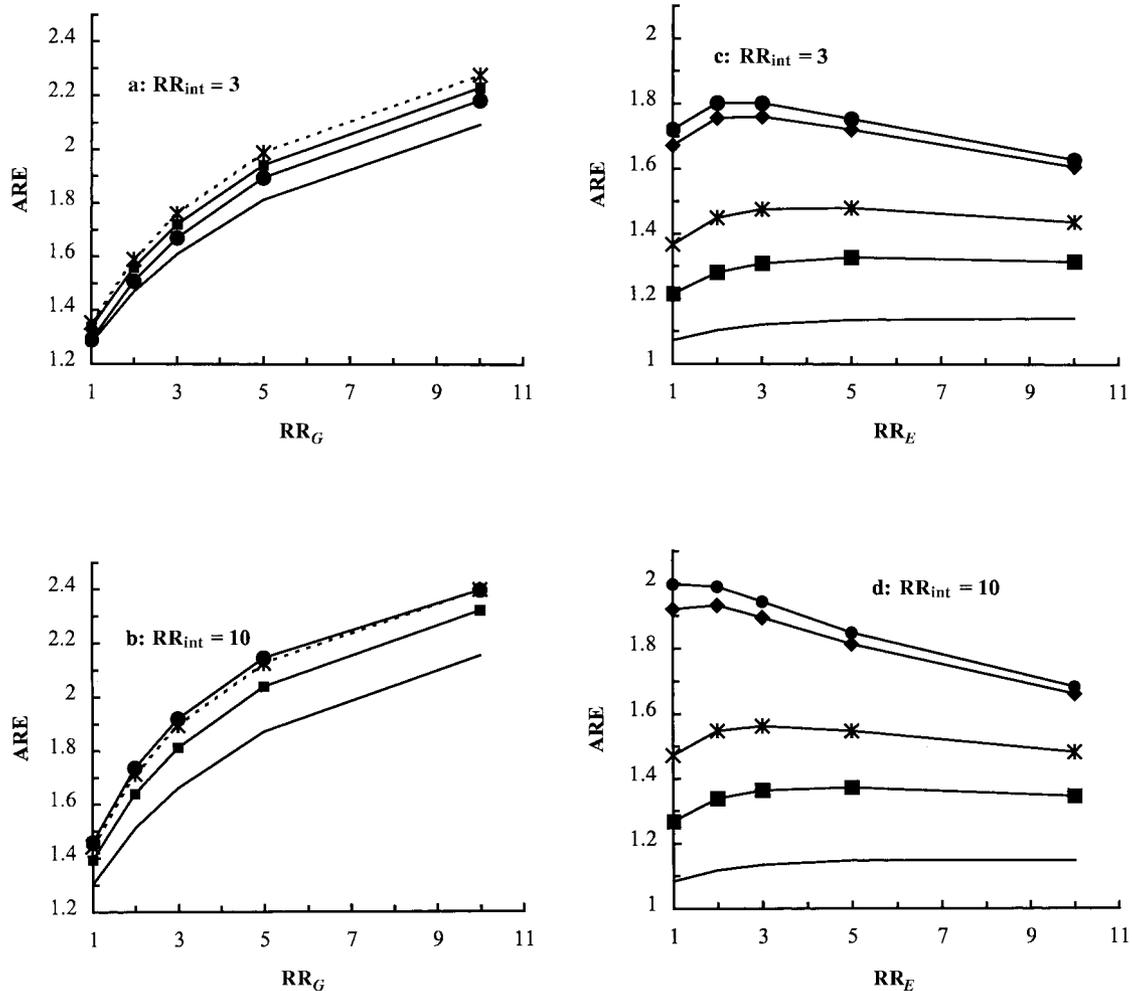
Gene dominance	Allele frequency	$RR_G^*$	Disease cumulative risk	Exposure concordance (odds ratio) <sup>†</sup>	Sensitivity ( $P(FH = 1   G1 = AA)\ddagger$ )	Specificity ( $P(FH = 0   G1 = aa)$ )	$P(FH = 1   G1 = Aa)\S$
Dominant	0.01	20	0.05	2	89.6	83.3	73.8
	0.01	20	0.05	1	90.0	83.4	73.9
	0.10	2	0.10	2	53.0	68.4	45.0
	0.10	2	0.10	1	53.2	68.6	44.9
Recessive	0.14	20	0.05	2	44.9	84.0	23.7
	0.14	20	0.05	1	44.8	84.1	23.6
	0.44	2	0.10	2	42.3	70.7	34.3
	0.44	2	0.10	1	42.2	70.8	34.2
Codominant	0.02	20	0.05	2	61.8	82.4	
	0.02	20	0.05	1	61.8	82.5	
	0.19	2	0.10	2	45.1	68.2	
	0.19	2	0.10	1	45.1	68.3	

\*  $RR_G$ , rate ratio for the genetic factor ( $G$ ).

<sup>†</sup> Defined as  $[P_R(E) \times (1 - P(E))]/[(1 - P_R(E)) \times P(E)]$ , with  $P_R(E)$  being the frequency of the environmental exposure ( $E$ ) among relatives.

<sup>‡</sup> "A" represents the deleterious allele.

<sup>§</sup> Under a dominant model,  $P(FH = 1 | G1 = Aa)$  is also interpretable as sensitivity; under a recessive model,  $P(FH = 1 | G1 = Aa)$  is interpretable as  $(1 - \text{specificity})$ ; under a codominant model,  $P(FH = 1 | G1 = Aa)$  is uninterpretable.



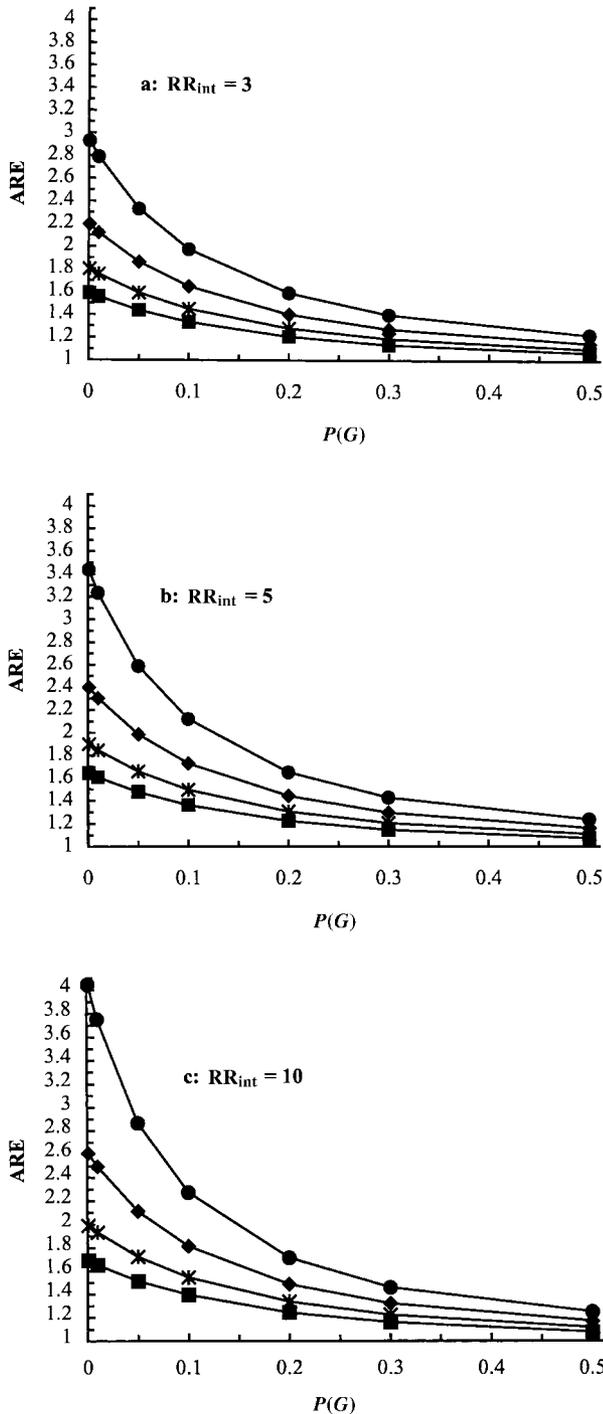
**FIGURE 2.** Asymptotic relative efficiency (ARE) according to (a and b)  $RR_G$  (main effect of  $G$ ) and  $RR_E$  (main effect of  $E$ ) and according to (c and d)  $RR_E$  and  $P(G)$  for different gene ( $G$ )-environment ( $E$ ) interaction values ( $RR_{int} = 3, 10$ ). In parts a and b, lines with circles denote the effect of  $RR_G$  on ARE for  $RR_E = 1$ ; dotted lines with asterisks denote the effect for  $RR_E = 3$ ; lines with squares denote the effect for  $RR_E = 5$ ; and plain lines denote the effect for  $RR_E = 10$ . In parts c and d, lines with circles denote the effect of  $RR_E$  on ARE for  $P(G) = 0.001$ ; lines with diamonds, for  $P(G) = 0.01$ ; lines with asterisks, for  $P(G) = 0.1$ ; lines with squares, for  $P(G) = 0.2$ ; and plain lines, for  $P(G) = 0.5$ .

at 0.8,  $RR_E$  equal to 2, and  $RR_G$  equal to 3. AREs are calculated for three different values of  $G \times E$  interaction ( $RR_{int} = 3, 5, 10$ ). The results show a decrease in the ARE as the frequency of  $G$  increases. For example, when  $RR_{int} = 5$  and  $P(E) = 0.1$  (figure 3, part b, line with diamonds), AREs decrease from 2.40 when  $P(G) = 0.001$  to 1.17 when  $P(G) = 0.5$ . At a given frequency of  $E$ , the slope of ARE increases as  $RR_{int}$  increases. For example, when  $P(E) = 0.1$ , comparison of AREs for  $P(G) = 0.001$  versus  $P(G) = 0.5$  shows a decrease in ARE of 1.04 when  $RR_{int} = 3$  (part a, line with diamonds), 1.23 when  $RR_{int} = 5$  (part b, line with diamonds), and 1.44 when  $RR_{int} = 10$  (part c, line with diamonds).

Figure 3 also shows that at a given frequency of  $G$ , the AREs decrease as the frequency of  $E$  increases. For example, when  $RR_{int} = 5$  and  $P(G) = 0.05$ , the ARE decreases from 2.59 when  $P(E) = 0.01$  to 1.48 when  $P(E) = 0.3$ . In addition, AREs increase as the frequency of  $G$  decreases (see figure 1). These results show increases in AREs for

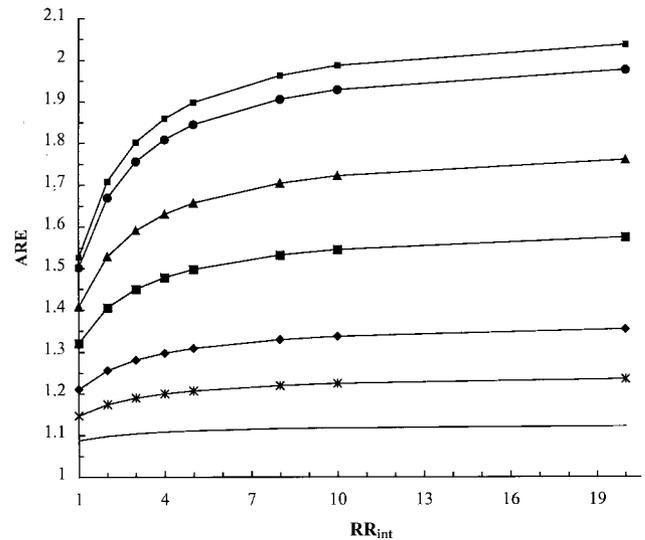
detecting gene-environment interaction when the frequencies of the factors under study decrease. Thus, as  $G$  and  $E$  become rarer, counter-matching on both  $G$  and  $E$  becomes more efficient.

*Efficiency of counter-matching according to the  $G \times E$  interaction value.* The ARE of counter-matching on both  $G$  and  $E$  is calculated according to different values of  $G \times E$  interaction for various frequencies of  $G$  (figure 4). As before, the sensitivity and specificity of both  $G$  and  $E$  are fixed at 0.8;  $RR_E = 2$ ,  $RR_G = 3$ , and  $P(E) = 0.2$ . Figure 4 shows a slight increase in the ARE as  $RR_{int}$  increases. The increase becomes greater as the frequency of  $G$  becomes smaller. For example, when  $P(G) = 0.01$  (figure 4, line with circles), AREs increase from 1.50 when  $RR_{int} = 1$  to 1.98 when  $RR_{int} = 20$ . The slopes of the AREs increase as the frequency of  $G$  decreases. For example, comparing  $RR_{int} = 1$  with  $RR_{int} = 20$  produces an increase in ARE of 0.09 when  $P(G) = 0.3$ , an increase of 0.26 when  $P(G) = 0.1$ , and an increase of 0.52 when  $P(G) = 0.001$ . These results show



**FIGURE 3.** Asymptotic relative efficiency (ARE) according to the frequencies of  $G$  ( $P(G)$ ) and  $E$  ( $P(E)$ ) for different gene ( $G$ )-environment ( $E$ ) interaction values ( $RR_{int} = 3, 5, 10$ ), with  $RR_E = 2$  and  $RR_G = 3$ . Lines with circles denote the effect of  $P(G)$  on ARE for  $P(E) = 0.01$ ; lines with diamonds, for  $P(E) = 0.1$ ; lines with squares, for  $P(E) = 0.2$ ; and lines with asterisks, for  $P(E) = 0.3$ .

increases in AREs for detecting gene-environment interaction as  $RR_{int}$  increases. The greater the interaction effect, the greater the efficiency of counter-matching on both  $G$  and  $E$ ,



**FIGURE 4.** Asymptotic relative efficiency (ARE) according to the gene ( $G$ )-environment ( $E$ ) interaction effect ( $RR_{int}$ ) for various frequencies of  $G$ , with  $P(E) = 0.2$ ,  $RR_E = 2$ , and  $RR_G = 3$ . Line with small squares denotes the effect of  $RR_{int}$  on ARE for  $P(G) = 0.001$ ; line with circles, for  $P(G) = 0.01$ ; line with triangles, for  $P(G) = 0.05$ ; line with large squares, for  $P(G) = 0.1$ ; line with diamonds, for  $P(G) = 0.2$ ; line with asterisks, for  $P(G) = 0.3$ ; and plain line, for  $P(G) = 0.5$ .

particularly when the frequency of  $G$  is small ( $P(G) < 0.1$ ). However, the AREs are much more strongly influenced by  $P(G)$  than by  $RR_{int}$ .

### Feasibility of counter-matching

To evaluate the feasibility of counter-matching in the assessment of gene-environment interaction, required sample sizes are calculated. The necessary number of counter-matching sets is calculated using a study design that counter-matches on surrogates of both  $G$  and  $E$  with sensitivity and specificity set equal to 0.8. We present sample sizes for two different values of power (80 percent and 90 percent) using a two-sided test at the 5 percent level for different frequencies of  $G$  and  $E$  and the  $G \times E$  interaction effect. Tables 3 and 4 show the number of counter-matching sets required when  $RR_E = RR_G = 2$  (table 3) and when  $RR_E = 2$  and  $RR_G = 10$  (table 4).

In table 3, because  $RR_G = RR_E$ , there is a symmetry in the sample sizes. That is, the required sample sizes for ( $P(G) = x$ ,  $P(E) = y$ ) are equal to the sample sizes for ( $P(G) = y$ ,  $P(E) = x$ ). The change in  $RR_G$  from 2 in table 3 to 10 in table 4 has a large effect on the required sample sizes when  $G$  or  $E$  is rare (i.e.,  $< 0.1$ ) and has little effect when  $G$  or  $E$  is common (i.e.,  $\geq 0.1$ ). For example, when  $P(E) = 0.1$  and  $P(G) = 0.01$ , the sample size required to detect an  $RR_{int}$  of 3 (80 percent power) is 2,738 counter-matched sets when  $RR_G = 2$  (table 3) and 1,207 sets when  $RR_G = 10$  (table 4). When  $P(E) = 0.1$  and  $P(G) = 0.1$ , the sample size required to detect an  $RR_{int}$  of 3 (80 percent power) is 379 counter-matched sets when  $RR_G = 2$  (table 3) and 308 sets when  $RR_G = 10$  (table 4).

**TABLE 3. Numbers of matching sets (sample sizes) required to have 80% and 90% power to detect a gene-environment interaction, using a two-sided test at the 5% level, for different frequencies of *G* and *E* and different interaction effects ( $RR_{int}$ ) when  $RR_E = RR_G = 2^*$**

<i>P</i> ( <i>E</i> )	$RR_{int}$	<i>P</i> ( <i>G</i> ) = 0.001		<i>P</i> ( <i>G</i> ) = 0.01		<i>P</i> ( <i>G</i> ) = 0.1		<i>P</i> ( <i>G</i> ) = 0.2		<i>P</i> ( <i>G</i> ) = 0.3	
		(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9
0.01	2	503,984	645,452	52,166	66,891	7,170	9,284	4,919	6,416	4,413	5,790
	3	189,164	237,277	19,620	24,659	2,738	3,495	1,899	2,452	1,717	2,235
	5	83,287	102,318	8,659	10,669	1,230	1,549	863	1,104	787	1,018
	10	38,645	46,550	4,029	4,874	584	727	415	528	382	493
0.1	2	69,366	89,718	7,170	9,284	976	1,276	666	877	596	790
	3	26,439	33,688	2,738	3,495	379	491	262	344	237	314
	5	11,841	14,877	1,230	1,549	174	224	123	161	113	150
	10	5,593	6,942	584	727	86	111	63	83	59	80
0.2	2	47,619	62,051	4,919	6,416	666	877	453	602	406	542
	3	18,344	23,640	1,899	2,452	262	344	181	241	165	221
	5	8,304	10,595	863	1,104	123	161	87	116	80	109
	10	3,964	5,016	415	528	63	83	46	63	44	62
0.3	2	42,749	56,011	4,413	5,790	596	790	406	542	364	489
	3	16,590	21,552	1,717	2,235	237	314	165	221	150	203
	5	7,564	9,754	787	1,018	113	150	80	109	75	103
	10	3,634	4,659	382	493	59	80	44	62	42	61

\* *G*, genetic exposure; *E*, environmental exposure; *RR*, rate ratio;  $RR_{int}$ , rate ratio for interaction between *G* and *E*.

4). The frequencies of *G* and *E* and  $RR_{int}$  are the parameters that have the largest effect on the sample sizes required to detect gene-environment interaction. Sample sizes needed to detect gene-environment interaction increase as the frequen-

cies of *G* and *E* decrease. In addition, sample sizes decrease as the *G* × *E* interaction values increase.

When *G* and *E* are rare (e.g., ≤0.01) and gene-environment interaction is moderate (e.g.,  $RR_{int} \leq 5$ ), the required

**TABLE 4. Numbers of matching sets (sample sizes) required to have 80% and 90% power to detect a gene-environment interaction, using a two-sided test at the 5% level, for different frequencies of *G* and *E* and different interaction effects ( $RR_{int}$ ) when  $RR_E = 2$  and  $RR_G = 10^*$**

<i>P</i> ( <i>E</i> )	$RR_{int}$	<i>P</i> ( <i>G</i> ) = 0.001		<i>P</i> ( <i>G</i> ) = 0.01		<i>P</i> ( <i>G</i> ) = 0.1		<i>P</i> ( <i>G</i> ) = 0.2		<i>P</i> ( <i>G</i> ) = 0.3	
		(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9	(1 - β) = 0.8	(1 - β) = 0.9
0.01	2	191,548	150,010	21,735	28,403	5,596	7,372	5,797	7,678	6,969	9,263
	3	73,952	95,532	8,404	10,875	2,188	2,865	2,284	3,014	2,759	3,660
	5	33,552	42,948	3,819	4,899	1,006	1,311	1,059	1,395	1,286	1,706
	10	16,045	20,384	1,830	2,332	490	638	521	689	637	849
0.1	2	27,352	35,934	3,086	4,061	768	1,022	787	1,053	943	1,269
	3	10,652	13,893	1,207	1,578	308	410	319	430	386	523
	5	4,874	6,318	556	725	147	198	156	213	191	263
	10	2,351	3,034	273	356	78	107	85	120	106	151
0.2	2	18,910	24,952	2,131	2,818	526	706	538	728	646	879
	3	7,407	9,722	840	1,108	215	290	224	306	272	375
	5	3,408	4,455	392	517	106	146	113	159	139	198
	10	1,654	2,156	196	261	59	84	66	96	82	123
0.3	2	16,721	22,130	1,887	2,506	471	636	485	661	586	803
	3	6,575	8,667	749	994	195	267	206	285	251	351
	5	3,037	3,991	353	469	99	138	107	153	132	191
	10	1,480	1,943	179	241	57	84	64	97	81	124

\* *G*, genetic exposure; *E*, environmental exposure; *RR*, rate ratio;  $RR_{int}$ , rate ratio for interaction between *G* and *E*.

sample size is very large (>8,000 sets), often reaching unrealistic numbers of sets. When the frequency of  $G$  is very low ( $P(G) = 0.001$ ), as might be observed for major genes such as *BRCA1/BRCA2* or *CDKN2A*, the needed sample size is only realistic when  $E$  is common ( $P(E) \geq 0.2$ ) and there is a strong gene-environment interaction effect:  $RR_{int} > 5$  when  $RR_G = 2$  and  $RR_{int} \geq 5$  when  $RR_G = 10$ .

## DISCUSSION

Counter-matching on both  $G$  and  $E$  in studies formulated to detect gene-environment interaction appears to be the most efficient of the four case-control designs considered. The parameters that are the most important for determining efficiency are the sensitivity and specificity of the surrogates, the frequencies of the risk factors of interest ( $E$  and  $G$ ), and, to a lesser extent, the value of the interaction effect. Feasibility, as measured by the required sample sizes, is also more dependent on the risk factor frequencies and the interaction effect than on the main effects (particularly for common exposures).

Since the sensitivity and specificity of the surrogates are very important for the gain in efficiency of counter-matching, the choice of highly specific and sensitive surrogates in the first stage of this method is critical. However, the requirement for highly specific and sensitive surrogates must be balanced against the need to use surrogates on which data are available or are easily measured in the cohort/population at risk. For example, at present, it would probably be too costly to genotype an entire cohort for a specific gene. As such, one might consider using a family history of the disease under study as a surrogate for the genetic factor. Family history is relatively easily measured, and the information is not too expensive to obtain. However, before designing the study, one must assess how predictive family history of disease is for the particular gene of interest. Indeed, we have shown that both the sensitivity and the specificity of surrogate family history tend to be higher for rare major susceptibility genes than for common low penetrance genes. Thus, in many complex and chronic diseases, family history may not be highly sensitive or specific for  $G$  if the disease under study is genetically heterogeneous (i.e., if more than one gene is involved, leading to low sensitivity), if most gene carriers are not affected (i.e., there is low penetrance), or if family sizes are not sufficiently large (the latter two conditions' leading to low specificities). In such scenarios, family history would be expected to be only a weak surrogate of  $G$  and thus produce only a modest or minimal gain in efficiency for the counter-matching design. When genes of interest have low penetrance, physiologic  $G$  surrogates (such as inexpensive phenotypic assays of urine, saliva, hair, etc.) may be considered and may be expected to be more sensitive and specific than family history.

The results of this analysis show that as the main effect of the rarer factor increases, the relative efficiency of counter-matching on both  $G$  and  $E$  for detecting gene-environment interaction increases. Conversely, as the main effect of the more common factor increases, the relative efficiency usually decreases. Moreover, it has been shown that the larger the gene-environment interaction and the rarer the risk fac-

tors  $G$  and  $E$ , the greater the efficiency of counter-matching. However, the gain in efficiency must be balanced by the feasibility of the study, as measured by the needed sample size. Indeed, when the two factors are rare (i.e., frequency < 0.1) and the interaction value is moderate (i.e.,  $\leq 5$ ), the relative efficiency of the counter-matching design is very high but the corresponding required sample size is very large (i.e., >8,000 sets, >8,000 cases, and >24,000 controls). Even if sample sizes for alternative designs would be even larger than those needed for counter-matching, these sample sizes are generally not realistic, and studies of this size tend to be prohibitively expensive.

When the frequency of  $G$  is very small (e.g., 0.001), as might be observed for major genes in cancer or other chronic diseases, the needed sample size might remain realistic only when factor  $E$  is common (i.e., >0.2) and when the interaction effect is high (i.e., >5).

For more common factors, the gain in efficiency derived from use of the counter-matching design, which may be minimal in some situations, must be balanced against the complexity of the design, particularly the difficulty involved in obtaining two specific and sensitive surrogates for the risk factors of interest. At present, identification of good surrogates for the factor(s) of interest and the costs associated with measuring these surrogates in large numbers of subjects (i.e., the entire cohort) may be the major determinants in deciding whether or not to conduct a counter-matched study.

This evaluation of the counter-matching design for assessment of gene-environment interaction used unrelated individuals drawn from a cohort or population at risk. The potential problem of population stratification or genetic admixture might affect the efficiency of this design in a gene-environment interaction study, although the potential loss of efficiency would be expected to be small (14). The use of related individuals as controls from a family population-based cohort may be an alternative. This design has recently been proposed for counter-matching in assessment of the effect of genetic factors and their interaction with environmental exposures (15).

Another multistage design has been proposed for the study of rare factors: the so-called "balanced design" (16). In this study design, rather than choosing a subset at random, one selects cases and controls in order to oversample for the rare factor of interest. The oversampling is taken into account in the analysis to obtain unbiased estimates of the effects of the individual factors and their interaction. One important difference between this design and counter-matching is that the balanced design is for grouped data (i.e., the case-control set must have multiple cases in each set), while the counter-matching design we have used is individually matched. In some situations, grouping may offer some logistical advantages, while in others the ability to match finely may be desirable. When the individual factor effects and their interaction are to be estimated, the balanced design appears to be as complex to implement as the counter-matching design. Cain and Breslow (17) investigated the efficiency of a balanced design versus a random sampling case-control design in estimating exposure-

covariate interaction. Similar to the counter-matching design, the balanced design was always more efficient than a random sampling design for estimating interaction. Limited direct comparisons between the “balanced” and “counter-matching” designs showed similar efficiencies in interaction estimation (9, 18). Additional research on the efficiency of the balanced design using different gene-environment interaction schemes, such as those presented in this paper, would facilitate better comparison of these two complex study designs.

The development of designs for specific gene-environment interaction studies should consider factors specific to each particular study. For instance, the frequencies of the genetic and environmental variables, their main effects, and the a priori supposed interaction effects would determine the efficiency and feasibility of various study designs. In addition, the costs of contacting and enrolling cases and controls, measuring surrogate variables, gene typing, measuring environmental variables, and establishing a reliable administrative structure with which to accurately implement a complex study design (e.g., multistage selection and data collection) should be assessed. In addition, if the specific genes of interest are highly prevalent metabolic genes for which relatively inexpensive phenotypic assays are available, one could consider a design that would screen large numbers of potential study subjects with a phenotypic assay and then genotype only those subjects selected for counter-matching. Alternatively, the rarity of the disease may suggest sampling strategies that differ from designs which would be used if the disease were common. If the disease is common, such that many cases are available, a one-control-per-case counter-matching design could be envisioned. Again, the specifics of a particular study will determine what type(s) of designs to consider. Whatever the design, the principle of using surrogate measures to inform the sampling may be considered in order to increase power for a gene-environment interaction study.

The increased interest in evaluating gene-environment interaction for many chronic diseases and the requirement of larger sample sizes for such studies have led to the evaluation of epidemiologic study designs that differ from traditional case-control or cohort designs. Counter-matching is one such alternative design. Although the efficiency of counter-matching relative to a 1:3 case-control study is greatest when the risk factors of interest are very rare, the study of such very rare factors is not realistic unless one is interested only in detecting very strong interaction effects (e.g.,  $RR_{int} > 10$ ). Nevertheless, a 1-1-1 counter-matching design appears to be more appropriate than most traditional epidemiologic methods for the study of gene-environment interaction involving rare genes or uncommon environmental exposures. However, both efficiency and feasibility must be evaluated before one considers using a counter-matching design, since it is more complex than that of a standard nested case-control study.

## ACKNOWLEDGMENTS

This project was supported by the Foundation Philippe, Inc. (Paris, France).

## REFERENCES

1. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;13:356-65.
2. 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.
3. Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev* 1997;19:33-43.
4. Andrieu N, Goldstein A. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. *Epidemiol Rev* 1998;20:137-47.
5. Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect* 1994;102(suppl 8):47-51.
6. Steenland K, Deddens JA. Increased precision using counter-matching in nested case-control studies. *Epidemiology* 1997;8:238-42.
7. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Stat Sci* 1996;11:35-53.
8. Cologne JB. Counterintuitive matching. (Editorial). *Epidemiology* 1997;8:227-9.
9. Langholz B, Borgan O. Counter-matching: a stratified nested case-control sampling method. *Biometrika* 1995;82:69-79.
10. Rotnitzky A. Efficiency and efficient estimators. In: Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Vol 2. New York, NY: John Wiley and Sons, Inc, 1998:1286-92.
11. Langholz B. Asymptotic variance and sample size expressions for simple and stratified nested case-control sampling in some specific cases. (Technical report no. 43). Los Angeles, CA: Department of Preventive Medicine, Biostatistics Division, University of Southern California, 1996.
12. Breslow NE, Day NE, eds. *Statistical methods in cancer research*. Vol 1. The analysis of case-control studies. Lyon, France: International Agency for Research on Cancer, 1980. (IARC scientific publication no. 32).
13. Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* 1999;149:693-705.
14. Caporaso N, Rothman N, Wacholder S. Case-control studies of common alleles and environmental factors. *J Natl Cancer Inst Monogr* 1999;26:25-30.
15. Siegmund KD, Langholz B, Thomas D. Stratified case-control sampling using related controls. (Abstract). *Genet Epidemiol* 1998;15:541.
16. Breslow NE. Case-control study, two phase. In: Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Vol 1. New York, NY: John Wiley and Sons, Inc, 1998:532-40.
17. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *Am J Epidemiol* 1988;128:1198-206.
18. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika* 1988;75:11-20.