Computing Differential Sample Size for Case-Control Studies of Gene-Environment Interaction

Jimmy Thomas Efird, PhD, MSc; Mi-Kyung Hong, MPH

The rates for diseases such as cancer, cardiovascular disease, and diabetes are known to differ by ethnic/racial groups. However, neither genetic nor environmental factors fully explain the observed differences. Failure to account for genetic expression in the absence or presence of an environmental factor, and vice-versa, may lead to erroneous conclusions regarding the importance of these factors in disease etiology. We present a novel method for computing sample size for case-control studies involving the interaction of genetic and environmental factors. The method is based on an indirect estimate of the odds ratio for gene-environment interaction given only the odds ratio for environmental exposure and population genotype frequency. A table is presented providing sample sizes required for detecting a minimum odds ratio for gene-environment interaction given varying genotype frequencies and environmental exposure odds ratio values. Sample size increases proportionately with genotype frequency for a given environment exposure odds ratio. (*Ethn Dis.* 2008;18[Suppl 2]:S2-25–S2-29)

Key Words: Gene-Environment Effect Modification, Gene-Environment Interaction, Genotype Frequency, Odds Ratio

From the Biostatistics and Data Management Facility, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, Hawaii (JE); Center for Tobacco Control Research and Education, Institute for Health Policy Studies, University of California, San Francisco, San Francisco, California (MKH).

Address correspondence and reprint requests to: Jimmy Thomas Efird, Director, Biostatistics and Data Management Facility; 650 Ilalo St, Biosciences Bldg 320-B; Honolulu, HI 96822; 650-248-8282; 808-692-1979; jimmy.efird@stanfordalumni.org

Introduction

Epidemiologic study designs have traditionally been employed to examine etiologic factors in disease causation that involve environmental exposures associated with disease risk. With the advent of technological advances in understanding genetic variation in human populations and its impact on phenotypic differences, the task of assessing the interaction between environmental exposures and genetic traits in disease etiology is of concern.[1] Public health and epidemiology as a study of populations and frequency of disease must address variations in disease predisposition attributable to causative genes specific to racial identity. Racial variations in putative genes are of significance as they may be implicated in the pathogenesis of common diseases such as cancer, coronary heart disease, and birth defects.[1] Complex interactions between racial genotypes and environmental factors govern many relationships between disease risk and exposures of interest, and gene-environment interaction is present in these causal mechanisms of disease.[1,2] The case-control study design is often employed in the field of genetic epidemiology and is a powerful epidemiologic method. Many methodologic issues arise in designing case-control studies, and application of the case-control method involves care. Interaction as a significant phenomenon in case-control studies must be addressed to accurately conclude associations of disease occurrence. We present a numerical algorithm for computing differential sample size in a case-control study of gene-environment interaction on the odds ratio (OR) scale and indicate its function in studies of genetic variation in differences of phenotypic expression among ethnic groups.

We examine the simple case of no main genetic and environmental effects; however, direct extension of the algorithm provides the framework for estimating differential sample size in more complex cases involving partial genetic or environmental effects. The method is based on an indirect estimate of the OR for gene-environment interaction given only the OR for environmental exposure and the population genotype. The algorithm involves expressing a Z-statistic in terms of the cell frequencies of an environmental exposure 2×2 contingency table, wherein power is denoted as the Z-statistic minus the α critical region of the standard normal distribution. Sample size corresponding to a desired power is found iteratively by providing a starting "a-cell" frequency count, computing the remaining cell frequencies for the exposure OR and lower confidence limit, and then determining the differential sample size assuming that the exposure OR is fixed. An example is presented illustrating the increase in sample size needed to be powered at $\geq$80% to detect a specified OR for gene-environment interaction at the α=.05 level of statistical significance given the sample size and OR computed in an environmental exposure case-control study. A corollary method for indirectly estimating gene-environment interaction and power given only genotype frequency and OR of environmental exposure has been discussed elsewhere.[3]

Methods

Assuming that 1) genotype (G) is independent of environmental exposure (E), 2) neither genetic nor environmental effects act alone to influence disease, and 3) disease (D) is rare in both exposed and unexposed populations, then an indirect estimate for the OR of a gene-environment effect is given as

$$\mathrm{OR(GE|D)} = [\mathrm{OR(E|D)} - 1 + \mathrm{g}]/\mathrm{g}, \quad [1]$$

where (g) corresponds to the genotype frequency. The study power for [1] may be derived from the formula

$$\mathrm{Zpower} = \left| \{\mathrm{OR(E|D)} - 1 + \mathrm{g}\} \bullet \log\{[\mathrm{OR(E|D)} - 1 + \mathrm{g}]/\mathrm{g}\} \middle/ \mathrm{OR(E|D)}\sqrt{\left(\frac{1}{\mathrm{a}} + \frac{1}{\mathrm{b}} + \frac{1}{\mathrm{c}} + \frac{1}{\mathrm{d}}\right)} \right| - (Z_{\alpha/2}), \quad [2]$$

where $Z_{\alpha/2}$ denotes the critical region of the standard normal distribution, and (a, b, c, d) are the cell frequencies for a 2×2 table of environmental exposure by case-control status. When only the OR(E|D) and (1-α)% confidence interval (CI) are known, then the 2×2 cell frequencies may be iteratively derived from the equation for the CI in terms of the point estimate and marginal case-control frequencies for the OR, ie,

$$\mathrm{OR(E|D)} \bullet \exp\left\{ \pm Z_{\alpha/2} \bullet \sqrt{\frac{1}{a} + \frac{1}{n_2 - d} + \frac{1}{n_1 - a} + \frac{1}{\left(n_1/[a/(c \times OR(E \mid D))]\right)}} \right\}, \quad [3]$$

where n_1 and n_2 denote the number of cases and controls and $c = n_1 - a$.

In planning a future case-control study to directly estimate OR(GE|D), in which information is simultaneously collected on environmental and genetic factors, one may use equation [2] to ascertain the sample size required to detect a minimum OR(GE|D) for a specified power or greater at the α level of statistical significance. The technique involves incrementally multiplying each 2×2 cell frequency by a constant such that OR(E|D) remains fixed until the desired power is achieved.

RESULTS

In the table, for genotype frequencies ranging from .2 to .8, we provide the sample size required to detect a minimum OR(GE|D) when planning a gene-environment case-control study for power ≥80% and 90%. The computations are based on values for OR(E|D) and corresponding 95% CI obtained from an environmental exposure case-control study of the same population as the planned gene-environment case-control study. The 2×2 cell frequencies were iteratively derived by using equation [3].

For example, 324 cases and controls each would be required in a future gene-environment case-control study to detect a minimum OR(GE|D) of 1.8 with power ≥80% at the α=.05 level of statistical significance, given (g)=.6, OR(E|D)=1.5 and 95% CI for OR(E|D) of 1.0526–2.1377. This reflects a difference of an additional 74 cases and controls compared with the environmental exposure case-control study involving only 250 cases and controls. However, in other examples shown in the table, the required sample size for the gene-environment case-control study is less than the reference environmental exposure case-control study.

As seen in the table, we observe that the required sample size for a gene-environment case-control study increases proportionately with the genotype frequency for a given OR(E|D). Furthermore, overall sample sizes in general tend to be lower for higher values of OR(E|D).

DISCUSSION

In nature, the process of how genes respond to the environment is readily illustrated by the example of coloration pattern in the African butterfly *Bicyclus*. During the rainy season, this butterfly takes on a colorful pattern. In contrast, a dull brown coloration is expressed in the dry season.[4] The consequences of gene-environment interaction may be equally important in humans, underlying the pathologic origins of many diseases of public health importance. The interplay between genes and the environment also plays an important role in how individuals respond differently to different drugs. Consider for example the antihypertensive agent debrisoquine. Debrisoquine's metabolism is altered by a functional polymorphism governing the activity of the cytochrome P45 CYP2D6 enzyme. This polymorphism may lead to compromised drug efficacy in certain groups, as the incidence of the cytochrome CYP2D6 enzyme is known to vary widely by different ethnic populations.[5] In fact, differential response by racial or ethnic groups has been reported for several classes of drugs including angiotensin-converting enzyme inhibitors, vasodilator antihypertensives, β-adrenoceptor blockers, calcium-channel blockers, anticoagulants, and glucocorticoids.[6]

The need to accurately estimate the required sample size to detect an interaction effect of genetic variation and environmental exposure is critical in the design of a case-control study examining this effect. We have presented a statistical method that can be used to compute differential sample size in a case-control study of gene-environment interaction given only a population genotype frequency and OR of environmental exposure based on prior studies. The implications for this novel technique are notable in that many low-penetrance susceptibility genes alone cannot confer disease risk. For example, in studying the etiology of cancer, polymorphisms in enzymes central to carcinogen metabolism do not appear to impart risk; however, in conjunction with environmental stress, cancer risk is elevated.[7] In the event of an interaction between environmental and genetic components, the failure to account for a modification of joint effects leads to bias in the risk estimation.[7] Of interest in our discussion, ethnic variation in such diseases as cancer may be attribut-

Table. Sample size required to detect indicated OR or greater for gene-environment interaction [OR(GE|D)] with power ≥80% and 90% at α=.05 for selected genotype frequencies (g)*

Exposure (E) Case-control Study No. Cases (total)	OR(E|D) (95% CI)†	2×2 Cell Frequencies (a, b, c, d)‡	Genotype Frequency (g)	OR(GE|D) (95% CI)§	Planned GxE Study Power≥80% | 90% No. Cases (total)¶
250 (500)	1.25 (.8690–1.7988)	(99, 86, 151, 164)	.2	2.3 (.8199–6.1820)	791 (1582) | 1059 (2118)
			.4	1.6 (.8077–3.2719)	1058 (2116) | 1417 (2834)
			.6	1.4 (.8300–2.4194)	1203 (2406) | 1610 (3220)
			.8	1.3 (.8514–2.0243)	1293 (2586) | 1731 (3462)
	1.5 (1.0526–2.1377)	(125, 100, 125, 150)	.2	3.5 (1.6383–7.4772)	188 (376) | 251 (502)
			.4	2.3 (1.2467–4.0606)	271 (542) | 362 (724)
			.6	1.8 (1.1310–2.9719)	324 (648) | 434 (868)
			.8	1.6 (1.0798–2.4455)	362 (724) | 485 (970)
	3.0 (2.0853–4.3179)	(158, 91, 92, 159)	.2	11.0 (6.6993–18.0734)	22 (44) | 29 (58)
			.4	6.0 (3.8083–9.4587)	33 (66) | 44 (88)
			.6	4.3 (2.8483–6.5963)	42 (84) | 56 (112)
			.8	3.5 (2.3705–5.1703)	49 (98) | 66 (132)
	5.0 (3.2096–7.7931)	(108, 33, 142, 217)	.2	21.0 (12.3892–35.6173)	15 (30) | 21 (42)
			.4	11.0 (6.6471–18.2141)	23 (46) | 30 (60)
			.6	7.7 (4.7354–12.4194)	29 (58) | 38 (76)
			.8	6.0 (3.7811–9.5262)	34 (68) | 45 (90)
500 (1000)	1.25 (.9565–1.6337)	(169, 145, 331, 355)	.2	2.3 (1.0698–4.7327)	859 (1718) | 1150 (2300)
			.4	1.6 (.9712–2.7191)	1148 (2296) | 1537 (3074)
			.6	1.4 (.9557–2.1001)	1305 (2610) | 1746 (3492)
			.8	1.3 (.9544–1.8051)	1403 (2806) | 1878 (3756)
	1.5 (1.1676–1.9270)	(250, 200, 250, 300)	.2	3.5 (2.0462–5.9866)	188 (376) | 251 (502)
			.4	2.3 (1.4821–3.4158)	271 (542) | 362 (724)
			.6	1.8 (1.3029–2.5798)	324 (648) | 434 (868)
			.8	1.6 (1.2171–2.1696)	362 (724) | 485 (970)
	3.0 (2.1026–4.2804)	(125, 50, 375, 450)	.2	11.0 (6.7747–17.8605)	42 (84) | 56 (112)
			.4	6.0 (3.8476–9.3565)	63 (126) | 84 (168)
			.6	4.3 (2.8755–6.5304)	80 (160) | 107 (214)
			.8	3.5 (2.3915–5.1223)	94 (188) | 126 (252)
	5.0 (3.8242–6.5396)	(345, 154, 155, 346)	.2	21.0 (15.2621–28.9067)	11 (22) | 15 (30)
			.4	11.0 (8.1112–14.9233)	17 (34) | 22 (44)
			.6	7.7 (5.7286–10.2641)	21 (42) | 28 (56)
			.8	6.0 (4.5380–7.9357)	25 (50) | 33 (66)
750 (1500)	1.25 (0.9614–1.6252)	(150, 125, 600, 625)	.2	2.3 (1.0853–4.6647)	1239 (2478) | 1658 (3316)
			.4	1.6 (.9809–2.6919)	1656 (3312) | 2217 (4434)
			.6	1.4 (.9630–2.0840)	1882 (3764) | 2519 (5038)
			.8	1.3 (.9603–1.7939)	2023 (4046) | 2709 (5418)
	1.5 (1.2226–1.8404)	(375, 300, 375, 450)	.2	3.5 (2.2581–5.4250)	188 (376) | 251 (502)
			.4	2.3 (1.6001–3.1639)	271 (542) | 362 (724)
			.6	1.8 (1.3871–2.4231)	324 (648) | 434 (868)
			.8	1.6 (1.2834–2.0575)	362 (724) | 485 (970)

Table. Continued

Exposure (E) Case-control Study No. Cases (total)	OR(E|D) (95% CI)†	2×2 Cell Frequencies (a, b, c, d)‡	Genotype Frequency (g)	OR(GE|D) (95% CI)§	Planned GxE Study Power≥80% | 90% No. Cases (total)¶
1000 (2000)	3.0 (2.4301–3.7036)	(450, 250, 300, 500)	.2	11.0 (8.2531–14.6611)	22 (44) | 29 (58)
			.4	6.0 (4.6108–7.8078)	33 (66) | 44 (88)
			.6	4.3 (3.3982–5.5259)	42 (84) | 56 (112)
			.8	3.5 (2.7927–4.3864)	50 (100) | 67 (134)
	5.0 (3.9351–6.3531)	(375, 125, 375, 625)	.2	21.0 (15.7902–27.9287)	13 (26) | 18 (36)
			.4	11.0 (8.3790–14.4409)	20 (40) | 26 (52)
			.6	7.7 (5.9094–9.9465)	25 (50) | 34 (68)
			.8	6.0 (4.6752–7.7002)	30 (60) | 40 (80)
	1.25 (1.0365–1.5075)	(351, 302, 649, 698)	.2	2.3 (1.3373–3.7858)	841 (1682) | 1126 (2252)
			.4	1.6 (1.1335–2.3297)	1125 (2250) | 1506 (3012)
			.6	1.4 (1.0756–1.8660)	1278 (2556) | 1711 (3422)
			.8	1.3 (1.0502–1.6404)	1374 (2748) | 1839 (3678)
	1.5 (1.2565–1.7907)	(500, 400, 500, 600)	.2	3.5 (2.3946–5.1157)	188 (376) | 251 (502)
			.4	2.3 (1.6749–3.0226)	271 (542) | 362 (724)
			.6	1.8 (1.4400–2.3342)	324 (648) | 434 (368)
			.8	1.6 (1.3246–1.9935)	362 (724) | 485 (970)
	3.0 (2.3333–3.8572)	(250, 100, 750, 900)	.2	11.0 (7.8081–15.4967)	42 (84) | 56 (112)
			.4	6.0 (4.3824–8.2148)	63 (126) | 84 (168)
			.6	4.3 (3.2425–5.7912)	80 (160) | 107 (214)
			.8	3.5 (2.6737–4.5816)	94 (188) | 126 (252)
	5.0 (4.1264–6.0586)	(625, 250, 375, 750)	.2	21.0 (16.7083–26.3940)	12 (24) | 15 (30)
			.4	11.0 (8.8434–13.6825)	17 (34) | 23 (46)
			.6	7.7 (6.2223–9.4462)	21 (42) | 29 (58)
			.8	6.0 (4.9122–7.3287)	25 (50) | 34 (68)

OR = odds ratio, CI = confidence interval.

* The SAS code for this table is available at http://www.pbrc.hawaii.edu/bdmf/GXE_sascode.html.

† Univariate odds ratio for environmental exposure given disease.

‡ Cell frequencies were iteratively derived using equation [3].

§ Indirect estimate of odds ratios for gene-environment interaction given disease.

¶ Sample size required for planned gene-environment case-control study to directly estimate OR(GE|D) for power≥80% and 90% at $\alpha=0.05$.

able to differences in genetic susceptibility polymorphisms as well as differences in environmental and dietary exposures,[7] and recognition of these gene-environment interactions could aid in risk prediction for subgroups.

Variations in genotypic frequencies of key regulatory genes have been documented as occurring in different ethnic populations, and these variations have paralleled differences in gene expression phenotypes.[8] This finding is of considerable consequence as some genetic polymorphisms may govern gene expression, and allelic frequency differences of these polymorphisms could indicate variable population differences in disease prevalence as regulated by gene expression phenotypes.[8] Such advances in the understanding of complex diseases and genetic variants have import for studies examining effects of etiologically significant environmental exposures such as smoking, ionizing radiation, vitamin use, alcohol use, intake of dietary antioxidants, and exogenous hormone use.[9] An appreciation of the joint effects of genetic factors and environmental exposures has greatened with the recent technological advances used in genetic epidemiology and new statistical methods are concomitant with these innovations. Moreover, interpretations of genetic contributions by ethnicity in relation to environment and disease outcome are captured by these emerging epidemiologic methods, and the model we have presented is of particular relevance.

The main advantages of the new method presented over other approaches are ease of use and greater precision when integrating conditional information from an environmental exposure case-control study.

Limitations of our statistical modeling of gene-environment interaction arise if the model does not properly reflect the true scale of biological action demonstrating the interaction on a biologic model.[2] Additionally, if the interactions are very complex and there are numerous genetic factors and large environmental contributions, the ability to characterize subgroup and ethnic differences in terms of genes and environment is challenging.[10] Although the model presented assumes an "unmatched" case-control design, the underlying algorithm may be easily adapted for other more complex study designs. A set of approximate solutions may be obtained in the case of limited decimal accuracy for OR and CI values or when the model contains multiple covariates. It also may be intriguing to examine the effect of model misspecification on the power and sample size; however, this is beyond the scope of the current paper. Further limitations of this method have been discussed in detail elsewhere.[3]

With these stated caveats, the motivation to quantify gene-environment interaction and disease risk for ethnic subgroups is nevertheless compelling. The field of pharmacogenomics, whereby individualization of drug therapy is dictated by genetic information, has implications for racial identity insomuch as it serves as a proxy for the genetic wildcard in the response to drug therapy.[11] A recent meta-analysis of ethnic differences in risks of adverse reactions to drugs used in cardiovascular medicine[12] found that individuals from different ethnic groups had differing risks for adverse drug-related events resulting from cardiovascular drug intake. One explanation for such differences in susceptibility to adverse drug reactions has attributed it to varying distributions of genetic polymorphisms in drug receptors or drug-metabolizing enzymes among different racial and ethnic subgroups.[11] The real-world implications of a genetic-environment interaction as indexed by race and ethnicity is demonstrated by these racial differences in response to drugs. Such outcomes will direct future efforts to characterize gene-environment interactions, and the methods to quantify them and will be a timely endeavor for communities of color in service to enhance public health practice.

ACKNOWLEDGMENTS

This manuscript was made possible by grant numbers P20 MD000173 from NCMHD and G12RR003061 from NCRR. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCMHD or NCRR.

REFERENCES

1. Khoury MJ, Beaty TH. Applications of the case-control method in genetic epidemiology. *Epidemiol Rev.* 1994;16(1):134–150.
2. Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev.* 1997;19(1):33–43.
3. Efird JT. Method for indirectly estimating gene-environment effect modification and power given only genotype frequency and odds ratio of environmental exposure. *Eur J Epidemiol.* 2005;20(5):389–393.
4. Marcus G. The role of genetics in the brain's nature. *The Nation.*2004:5A.
5. Bernard S, Neville KA, Nguyen AT, Flockhart DA. Interethnic differences in genetic polymorphisms of CYP2D6 in the US population: clinical implications. *Oncologist.* 2006;11(2): 126–135.
6. Tate SK, Goldstein DB. Will tomorrow's medicines work for everyone? *Nat Genet.* 2004;36(11 Suppl):S34–42.
7. Mucci LA, Wedren S, Tamimi RM, Trichopoulos D, Adami HO. The role of gene-environment interaction in the aetiology of human cancer: examples from cancers of the large bowel, lung and breast. *J Intern Med.* 2001;249(6):477–493.
8. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet.* 2007;39(2):226–231.
9. Millikan RC, Player J, de Cotret AR, et al. Manganese superoxide dismutase Ala-9Val polymorphism and risk of breast cancer in a population-based case-control study of African Americans and Whites. *Breast Cancer Res.* 2004;6(4):R264–274.
10. Mountain JL, Risch N. Assessing genetic contributions to phenotypic differences among "racial" and "ethnic" groups. *Nat Genet.* 2004;36(11 Suppl):S48–53.
11. Wood AJ. Racial differences in the response to drugs—pointers to genetic differences. *N Engl J Med.* 2001;344(18):1394–1396.
12. McDowell SE, Coleman JJ, Ferner RE. Systematic review and meta-analysis of ethnic differences in risks of adverse reactions to drugs used in cardiovascular medicine. *BMJ.* 2006;332(7551):1177–1181.