RESEARCH COMMUNICATION

Minimal Sizes of Cases with a Susceptible Genotype and Minimal Odds Ratios among Susceptible Individuals in Case-control Studies

Nobuyuki Hamajima[1], Hironori Mutoh[2], Hidetaka Eguchi[3], Hiroyuki Honda[2]

Abstract

Objective: Disease risk elevation due to an environmental factor only for individuals with a susceptible genotype is a typical example of gene-environment interaction. In order to identify risk factors interacting with susceptible genotypes in case-control studies, presumptions on minimal size of cases with the susceptible genotype (S_{min}) and odds ratio (OR) among the susceptible individuals ($OR_{susceptible}$) are useful.

Model: Proportion of exposed cases (P_1) and OR for whole cases (OR_{whole}) statistically detectable in a case-control study can be calculated in a conventional method. P_1 was assumed to be a weighted sum of the exposed among cases with the genotype (P_x) and cases without the genotype (equal to proportion of the exposed among controls, P_0), i.e., $S P_x + (1 - S) P_0$, where S is the size (proportion) of cases with the genotype. For each calculated P_1, S became the minimum (S_{min}) in case of $P_x = 1$. $OR_{susceptible}$ was calculated by $\{P_x (1 - P_0)\} / \{(1 - P_x) P_0\}$.

Results: S_{min} and $OR_{susceptible}$ were listed for the combinations of the above components. For example, a detectable P_1 was 0.638 for P_0=0.5 in a case-control study with 200 cases (N_1) and 200 controls (N_0), when α error of a two-sided test was 0.05 with an 80% of power. In case of P_1=0.638, OR_{whole} was 1.77, producing S_{min}=0.277 for infinite $OR_{susceptible}$. It indicates that an environmental factor cannot be detected in case that a high-risk genotype frequency is less than 0.277.

Interpretation: If the size of cases with a susceptible genotype is expected to be less than S_{min}, case-control studies are unlikely to detect a significant OR of the environmental factor.

Key Words: gene-environment interaction – genetic polymorphism – sample size – case-control studies

Asian Pacific J Cancer Prev, **6**, 165-169

Introduction

Recent development of genotyping methods allows us to examine the hypothesis that environmental factors cause a disease for individuals with a susceptible genotype. Although not perfect, it was exemplified by the finding that smoking causes lung cancer more frequently in those with low enzyme activity genotypes of carcinogen detoxification enzyme genes (Kiyohara et al., 2002; Mohr et al., 2003). Epidemiologically, such phenomena are termed as a gene-environment interaction, which is defined with a relative risk ratio of environmental exposure for those with a genotype relative to those without it, or a relative risk ratio of genotype for the exposed relative to the unexposed (Khoury and Flanders, 1996; Hamajima et al., 1999; Brennan, 2002). Since the elucidation of the interactions is useful for individualized disease prevention, researches on the interactions have been becoming popular in the field of epidemiology (Mucci et al., 2001; Kang, 2003). The targeted genotypes are selected from commonly observable ones, which are called "polymorphism" genotypes.

When the genotype interacting with an environmental factor is known, a sample size to detect the odds ratio (OR) of the factor in a case-control study can be calculated based

[1] Department of Preventive Medicine / Biostatistics and Medical Decision Making, Nagoya University Graduate School of Medicine, Nagoya, Japan [2] Department of Biotechnology, School of Engineering, Nagoya University, Nagoya, Japan. [3]Department of Radiobiology/ Molecular Epidemiology, Radiation Effects Research Foundation, Hiroshima, Japan.
Corresponding to: Nobuyuki Hamajima, M.D., Ph.D., M.P.H., Department of Preventive Medicine / Biostatistics and Medical Decision Making, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8550 Japan, TEL:+81-52-744-2133, FAX:+81-52-744-2971, e-mail: nhamajim@med.nagoya-u.ac.jp

on the genotype frequency with a conventional method (Hwang et al., 1994; Garcia-Closas and Lubin, 1999). On the contrary, the sample size cannot be calculated in case that the genotype frequency is unknown. In order to detect environmental factors in case-control studies including both subjects with and without the susceptible genotype, we had better have presumptions on the size (proportion) of individuals with the genotype and the OR among them. This paper aims to demonstrate minimal size of cases with the susceptible genotype to detect a significant environmental factor in case-control studies, as well as minimal required OR for individuals with the susceptible genotype.

Statistical Models

We recognized that there was a subgroup of cases with a genotype susceptible to an environmental factor. In order to calculate minimal detectable odds ratios of the environmental factor among those with the genotype ($OR_{susceptible}$), the following steps were made, as shown in Chart.

2.1. A proportion of exposed cases (P_1) producing a significant result in a case-control study with N_0 controls and N_1 cases was calculated based on a significance level (α), statistical power (1-β), and proportion of exposed controls (P_0), using the below conventional formula for a sample size calculation (Donner, 1984).

$$N_0 = \frac{\left[Z_\alpha \sqrt{(1+M)\,P\,(1-P)} + Z_\beta \sqrt{M\,P_0\,(1-P_0) + P_1\,(1-P_1)} \right]^2}{M\,(P_0 - P_1)^2}$$

where P is defined with $(P_0 + M\,P_1) / (1 + M)$, M with the ratio of N_1 / N_0, and Z_α and Z_β with the values derived from a normal distribution with mean=0 and variance=1 for a given significance level (α) and statistical power (1-β), respectively.

2.2. Odds ratio for whole subjects (OR_{whole}) was obtained by $P_1\,(1-P_0) / P_0\,(1-P_1)$.

2.3. P_1 was also defined with a weighted average calculated by $S\,P_x + (1 - S)\,P_0$, as shown in Fig 1. In this formula, P_x and P_0 were the proportions for the exposed in cases with and without the susceptible genotype, respectively. S was the size in proportion for cases with the genotype. It was assumed that the environmental exposure does not elevate the risk of disease for cases without the genotype. Accordingly, the proportion of the exposed among them was set to be the same as that among the controls, i.e., P_0.

2.4. S_{min} was defined as the S in case of P_x=1. It was the minimum of S, because P_x was the maximum at 1.

2.5. $OR_{susceptible}$ was calculated with $\{P_x\,(1-P_0)\}/\{P_0\,(1-P_x)\}$.

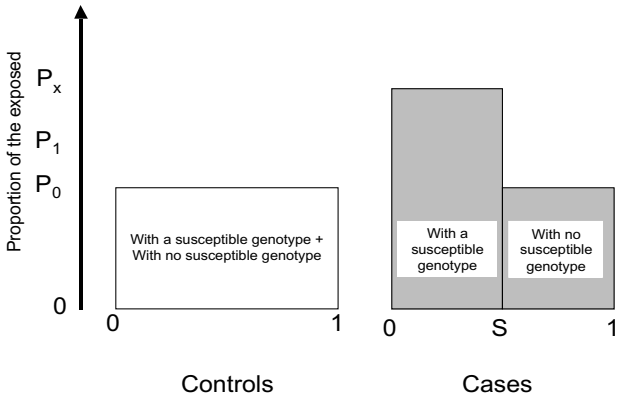


Figure 1. Proportions of the Exposed among Controls (P_0) and Cases (P_1). P_1 is the Average Proportion for Cases with a Susceptible Genotype (P_x) and Cases with no Susceptible genotype (P_0). The Area Surrounded by a Dotted Line is the Same as the Shadowed Areas. S is the Size in Proportion of Cases with a Susceptible Genotype.

Results

Since a large number of combinations exist, those with α=0.05 in a two-sided test (Z_α=1.96), 1-β=0.80 (Z_β=0.842), and $N_0=N_1$ (M=1) were calculated as examples. Table 1 shows the calculated P_1, OR_{whole}, and S_{min}, when N_0 is fixed to be 200, 500, 1,000, or 2,000, and P_0 to be 0.05, 0.1, 0.3, 0.5 or 0.8. For example, a detectable P_1 was 0.638 for P_0=0.5 in a case-control study with 200 cases (N_1) and 200 controls (N_0), when α error of a two-sided test was 0.05 with an 80% of power. In case of P_1=0.638, OR_{whole} was 1.77, producing S_{min}=0.277 for infinite $OR_{susceptible}$. It indicates that an environmental factor cannot be detected in case that a high-risk genotype frequency is less than 0.277. Figure 2 depicts the relationship between S_{min} and N_0 for given P_0. The minimal size of cases with the genotype (S_{min}) increased with the proportion of the exposed in controls (P_0) and decreased with the number of controls (N_0).

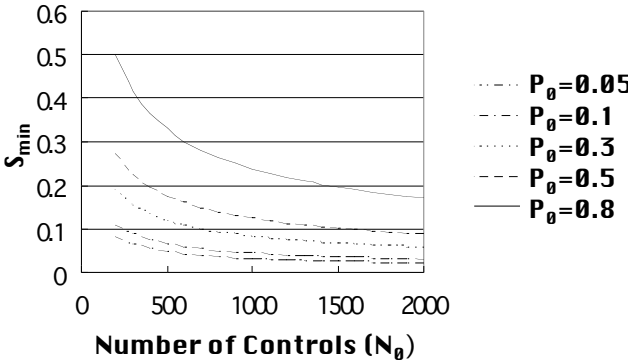


Figure 2. Minimal Size of Susceptible Cases Enabling to Detect a Significant Odds Ratio (S_{min}) According to Sample Sizes (N_0, in Case of $N_0=N_1$) and Proportion of the Exposed among Controls (P_0)

Table 1. Detectable Proportion of the Exposed among Cases (P_1), Odds Ratio for Whole Subjects (OR_{whole}), Minimal Size of Cases with a Susceptible Genotype (S_{min}) according to Number of Controls (N_0) and Proportion of Exposed Controls (P_0), under a Significance Level (α) = 0.05 for a Two-sided Test with Statistical Power (1-β) =0.8

N_0	P_0=0.05	P_0=0.1	P_0=0.3	P_0=0.5	P_0=0.8
			P_1		
200	0.130	0.200	0.435	0.638	0.900
500	0.096	0.160	0.384	0.588	0.866
1,000	0.081	0.141	0.359	0.563	0.848
2,000	0.071	0.128	0.341	0.544	0.834
			OR_{whole}		
200	2.84	2.25	1.79	1.77	2.25
500	2.02	1.71	1.45	1.43	1.62
1,000	1.67	1.47	1.31	1.29	1.39
2,000	1.46	1.32	1.21	1.19	1.26
			S_{min}		
200	0.084	0.111	0.192	0.277	0.499
500	0.048	0.066	0.120	0.176	0.330
1,000	0.033	0.045	0.084	0.125	0.239
2,000	0.022	0.031	0.059	0.088	0.171

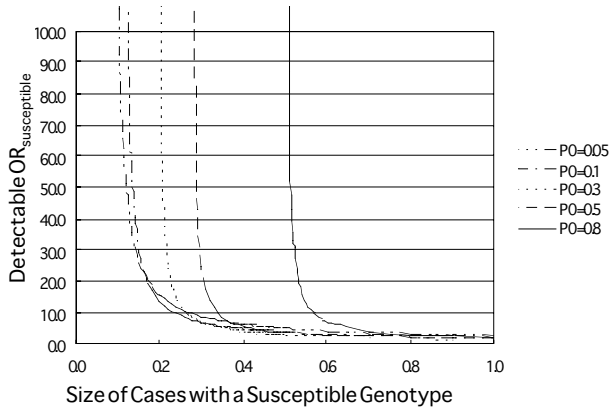
Figure 3 shows $OR_{susceptible}$ in a case-control study with 200 cases and 200 controls according to size of cases with the genotype (S) and proportion of the exposed controls (P_0). Since all the cases with the genotype were to be the exposed at S_{min}, the $OR_{susceptible}$ was infinite at S_{min}. In case of $S > S_{min}$, the $OR_{susceptible}$ decreased with S, and was equal to OR_{whole} at S=1. Figure 4 shows $OR_{susceptible}$ in case of P_0= 0.5 according to N_0 (=N_1). As N_0 was larger, $OR_{susceptible}$ was smaller in a given S. Table 2 lists the detectable $OR_{susceptible}$ according to S for different P_0 and N_0.

The above results can be used for the following examples. 1) When a case-control study has only 200 cases (N_1) and 200 controls (N_0), smoking can not be evaluable as a risk factor of male colon cancer in the following condition. Those with the susceptible genotype (S) are assumed to be 20% among the cases, and smokers are 50% among the controls (P_0). Table 1 provides S_{min} = 0.277 for N_0=N_1=200 and P_0=0.5, which is larger than the assumed S (0.2). 2) When a 30% of male colon cancer cases (S) have a genotype susceptible to smoking, $OR_{susceptible}$ more than 3.85 would be detected in a case-control study with 500 male cases (N_1) and 500 male controls (N_0), in an area where smokers are 50% among the male population (P_0) as indicated in Table 2.

Discussion

We know intuitively that risk factors affecting a small proportion of individuals may not be detected in a study, because of the effect dilution. Accordingly, even with a high penetrance, rare genotypes are not examined in association studies. As Shpilberg et al stated, "A twofold risk for 1000 exposed versus nonexposed people could be an average twofold risk for all 1000 exposed or a 20-fold risk for 100 exposed individuals" (Shpilberg et al., 1997). In case-control studies, however, there were no reference tables on the proportion of susceptible individuals. To date, several papers have been reporting required sample sizes for unmatched case-control studies to detect a gene-environment or gene-gene interaction (Hwang et al., 1994; Garcia-Closas and Lubin, 1999; Gauderman, 2002a; Gauderman, 2002b, Selinger-Leneman et al., 2003). But, their view is different from the present report. Tables and Figures presented in this paper provide useful information to avoid studies impossible to detect the significant results. The newly introduced concept, S_{min}, is an important measure when case-control studies are planned taking account of a susceptible subgroup in the study subjects.

In the present paper, the size of susceptible cases was



Figire 3. Detectable Minimal $OR_{subgroup}$ in a Case-control Study with 200 Cases and 200 Controls According to Size of Cases with a Susceptible Genotype (S) and Proportion of the Exposed among Controls (P_0)

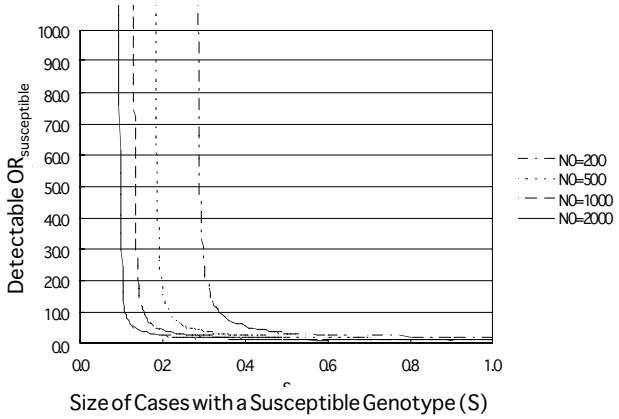


Figure 4. Detectable Minimal $OR_{subgroup}$ in a Case-control Study with Half of the Controls Exposed (P_0=0.5), According to Size of Cases with a Susceptible Genotype (S) and Number of Controls (N_0)

Table 2. Detectable OR for Individuals with a Genotype Susceptible to Environmental Factor ($OR_{susceptible}$) according to Size of Cases with the Susceptible Genotype (S), Proportion of Exposed Controls (P_0), and Number of Controls (N_0), under a Significance Level (α) = 0.05 for a Two-sided Test with Statistical Power (1-β) =0.8

N_0	S=0.1	S=0.2	S=0.3	S=0.5	S=0.7	S=1
P_0=0.05						
200	107	15.5	8.80	5.05	3.73	2.84
500	19.8	7.40	4.86	3.15	2.49	2.02
1,000	10.7	4.90	3.44	2.40	1.98	1.67
2,000	6.72	3.50	2.60	1.93	1.66	1.46
P_0=0.1						
200	N.E.	13.4	6.86	3.85	2.88	2.25
500	20.5	5.94	3.83	2.52	2.04	1.71
1,000	9.28	3.93	2.78	2.00	1.69	1.47
2,000	5.55	2.86	2.16	1.67	1.47	1.32
P_0=0.3						
200	N.E.	85.4	6.96	3.09	2.26	1.79
500	N.E.	6.00	3.22	2.05	1.69	1.45
1,000	18.6	3.42	2.30	1.67	1.46	1.31
2,000	5.81	2.40	1.82	1.45	1.31	1.21
P_0=0.5						
200	N.E.	N.E.	24.7	3.48	2.31	1.77
500	N.E.	15.9	3.85	2.09	1.67	1.43
1,000	N.E.	4.33	2.43	1.67	1.43	1.29
2,000	16.4	2.59	1.84	1.43	1.29	1.19
P_0=0.8						
200	N.E.	N.E.	N.E.	584	4.10	2.25
500	N.E.	N.E.	N.E.	3.43	2.12	1.62
1,000	N.E.	N.E.	5.85	2.14	1.65	1.39
2,000	N.E.	8.43	2.66	1.65	1.41	1.26

N.E.: ORsusceptible does not exist.

used, not of susceptible controls which represent the population without disease under study. Generally, the size of susceptible cases is larger than the size of susceptible controls ($S_{control}$). Although Tables and Figures could similarly be made using $S_{control}$, the size of susceptible cases (S) was adopted here. The S seems easier to be understood and estimated by clinicians, who are faced with patients.

In conclusion, this paper provided the useful figures when case-control studies on environmental factors interacting with genotypes are designed. These figures are applicable for OR of a genotype interacting with environmental factors, and also for gene-gene interactions to be derived from case-control studies based on high-throughput SNP analysis (Marnellos, 2003; McLeod and Yu, 2003).

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Special Priority Areas of Cancer from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Chart for the Calculation Steps

1. Calculation of P_1 to obtain a significant result from given P_0, N_0, N_1, significance level, and statistical power.
2. Calculation of OR_{whole} from P_0 and P_1.
3. Calculation of P_x from P_0, P_1, and given S.
4. Calculation of S_{min} in case of P_x = 1.
5. Calculation of $OR_{susceptible}$ from P_0, P_x, and S.

N_0: Number of controls
N_1: Number of cases
P_0: Proportion of the exposed among controls
P_x: Proportion of the exposed among cases with a susceptible genotype
P_1: Proportion of the exposed among cases, which is defined with $S P_x + (1 – S) P_0$
S : Size (proportion) of cases with the susceptible genotype
S_{min}: The minimal S, i.e., S in case of P_1=1
OR_{whole}: Odds ratio for whole cases
$OR_{susceptible}$: Odds ratio for individuals with the susceptible genotype.

References

Brennan P (2002). Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis*, **23**, 381-7.

Donner A (1984). Approaches to sample size estimation in the design of clinical trials – a review. *Stat Med* , **3**, 199-214.

Garcia-Closas M, Lubin JH (1999). Power and sample size calculations in case-control studies of gene-environment interactions: comments of different approaches. *Am J Epidemiol*, **149**, 689-92.

Gauderman WJ (2002a). Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med*, **21**, 35-50.

Gauderman WJ (2002b). Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol*, **155**, 478-84.

Hamajima N, Yuasa H, Matsuo K, Kurobe Y (1999). Detection of gene-environment interaction by case-only studies. *Jpn J Clin Oncol*, **29**, 490-3.

Hwang SJ, Beaty TH, Liang KY, Coresh J, Khoury MJ (1994). Minimum sample size estimation to detect gene-environment interaction in case-control studies. *Am J Epidemiol*, **140**, 1029-37.

Kang D (2003). Genetic polymorphisms and cancer susceptibility of breast cancer in Korean women. *J Biochem Mol Biol*, **36**, 28-34.

Khoury MJ, Flanders WD (1996). Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol*, **144**, 207-13.

Kiyohara C, Otsu A, Shirakawa T, Fukuda S, Hopkin JM (2002). Genetic polymorphisms and lung cancer susceptibility: a review. *Lung Cancer*, **7**, 241-56.

Marnellos G (2003). High-throughput SNP analysis for genetic association studies. *Curr Opin Drug Discov Devel*, **6**, 317-21.

McLeod HL, Yu J (2003). Cancer pharmacogenomics: SNPs, chips, and the individual patient. *Cancer Invest*, **21**, 630-40.

Mohr LC, Rodgers JK, Silvestri GA (2003). Glutathione S-transferase M1 polymorphism and the risk of lung cancer. *Anticancer Res*, **23**, 2111-24.

Mucci LA, Wedren S, Tamimi RM, Trichopoulos D, Adami HO (2001). 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*, **249**, 477-93.

Selinger-Leneman H, Genin E, Norris JM, Khlat M (2003). Does accounting for gene-environment (G_E) interaction increase the power to detect of a gene in a multifactorial disease? *Genet Epidemiol*, **24**, 200-7.

Shpilberg O, Dorman JS, Ferrell RE, et al (1997). The next stage: molecular epidemiology. *J Clin Epidemiol*, **50**, 633-8.