A Note on the Estimation of Relative Risks of Rare Genetic Susceptibility Markers[1]

Colin B. Begg[2] and Marianne Berwick

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10021

Abstract

The comparison of an incident case series with an incident series of second primary cancers, using either a case-control or follow-up study design, is proposed as an efficient method for evaluating the relative risk of a rare genetic susceptibility marker and its prevalence in the population, and for evaluating gene-environment interactions. The relative efficiency of this design *versus* a conventional case-control study is highly dependent on the population prevalence of the marker and its relative risk. However, for relatively rare but highly penetrant genes, the relative efficiency can be very high. In an example presented regarding a planned study of the *p16* gene and its role in melanoma, a conventional case-control study may require up to 70 times as many subjects to achieve equivalent precision to the study of second primaries. The use of second primary cancers in this way requires assumptions about the validity of the classification of a new tumor as a second primary, the extent to which risk of a second cancer is influenced by treatment of the first cancer, and the nature and extent of surveillance bias. However, the problems of ascertaining a valid series of population controls are avoided. The study of second cancers represents an important and underused tool in molecular and genetic epidemiology.

Introduction

It is becoming increasingly evident that genetic predisposition has an important role in cancer incidence. Some of the major genes identified to date include the *p53* gene, responsible for a variety of cancer types, BRCA 1 and BRCA 2, which cause breast and ovarian cancers, and *p16*, suspected in the etiology of melanoma. These genes have been identified through family studies, in which clusters of the cancers have been observed, and then via linkage analysis. Although the study of cancer-prone families represents the classical approach for establishing linkage and locating the gene responsible for the cancers, such studies are not ideal for estimating the parameters relevant to the epidemiological impact of the gene, namely its population prevalence and relative risk (1). Unfortunately, the predominant epidemiological tool for estimating such population-based statistics, the case-control study, is also inadequate when dealing with very rare risk factors, because in any study with a feasible sample size there will be few, if any, controls identified with the genetic mutation, and thus little, if any, information available for estimating either the population prevalence or the relative risk.

The thesis of this article is that these parameters can be estimated much more efficiently by using an alternative case-control study design, one in which the "cases" are incident second primaries of the disease and the "controls" are incident first primaries of the disease. This kind of study design has been used increasingly in recent years to examine risk factors for second primaries, primarily for head and neck cancers (2–4) and breast cancer (5). Frequently, these studies have been cohort studies, in which a cohort of cases have been followed for incidence of second primaries. However, their utility for genetic epidemiology has not been fully realized.

The general rationale is as follows. Consider a rare cancer gene with high penetrance. The rarity means that its population prevalence (denoted hereafter as p) is very low, and thus a conventional control group is uninformative. However, the high penetrance ensures that the gene prevalence among cases (denoted q) is substantially higher, and the prevalence among patients with second primaries (denoted r) is higher still. The much higher prevalences among cases and second primaries permits a much more efficient study design for estimating the relative risk, and in turn for imputing the population prevalence.

In the remaining sections, the assumptions required for this approach are explained, and formulas are presented for estimating the prevalence and the relative risk, and their variances. An expression is obtained for the relative efficiency of the second cancers design relative to a conventional case-control study with equivalent sample sizes, and it is shown that in certain circumstances, gains in efficiency greater than an order of magnitude are possible. It is also shown that the second cancers study has a corresponding improved efficiency for evaluating gene-environment interactions. An example of a proposed study of risk factors for melanoma is presented.

Materials and Methods

Consider a genetic germ line mutation that has population prevalence p and confers a relative risk ψ for a specific type of cancer, adjusted for confounding factors. Furthermore, assume that any observed second primary of this cancer represents a truly independent occurrence of the cancer; *i.e.*, we assume that the second primary is biologically independent of the first primary. [We note that the differential diagnosis of a second primary from metastatic spread of the first primary is usually subjective at the pathological level (see "Discussion").] By

Received 5/28/96; revised 5/28/96; accepted 11/6/96.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] This research was supported by the National Cancer Institute, Department of Health and Human Services Award CA69396.

[2] To whom requests for reprints should be addressed, at Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021.

Table 1 Frequency data for second primaries study[a]

	First primaries	Second primaries
Gene carriers	a $[n_1q]$	b $[n_2r]$
Noncarriers	c $[n_1(1-q)]$	d $[n_2(1-r)]$
	n_1	n_2

[a] Expected frequencies based on the true prevalences are shown in parentheses. q and r are true gene prevalences among first and second primary cases, respectively, and n_1 and n_2 are the corresponding sample sizes. a–d represent observed frequencies.

definition, it follows that after adjusting for confounding factors, the relative risk of a new (second) primary among gene carriers is also ψ.

It is easily shown (6) that the gene prevalence among incident cases with first primaries is given by

$$q = \frac{p\psi}{1 - p + p\psi}. \quad \text{(A)}$$

By a similar argument, the prevalence of the gene among patients with incident second primaries is given by:

$$r = \frac{p\psi^2}{1 - p + p\psi^2} \quad \text{(B)}$$

i.e., the relative risk is, equivalently,

$$\psi = \frac{q(1-p)}{p(1-q)} = \frac{r(1-q)}{q(1-r)}. \quad \text{(C)}$$

Consider now a case-control study of second primaries comprising n_1 cases (the control group) and n_2 second primaries (the case group). Let the observed frequencies be as displayed in Table 1. It is easy to show that the maximum likelihood estimates of p and ψ, denoted $\hat{p}$ and $\hat{\psi}$, are given by:

$$\hat{p} = \frac{a^2d}{a^2d + c^2b}$$

and

$$\hat{\psi} = \frac{bc}{ad}.$$

Confidence intervals can be obtained by using the following approximations:

$$\hat{\text{var}}\left[\log\left(\frac{\hat{p}}{1-\hat{p}}\right)\right] = \frac{2}{a} + \frac{2}{c} + \frac{1}{b} + \frac{1}{d} \quad \text{(D)}$$

and

$$\hat{\text{var}}[\log(\hat{\psi})] = \frac{1}{a} + \frac{1}{c} + \frac{1}{b} + \frac{1}{d}.$$

The first of these is derived using the fact that:

$$\log[\hat{p}/(1-\hat{p})] = 2\log[\hat{q}/(1-\hat{q})] - \log[\hat{r}/(1-\hat{r})],$$

which follows directly from Eq. C, where $\hat{q}$ and $\hat{r}$ are the sample estimates of q and r, respectively.

It is important to clarify that this method permits estimation of population-based parameters; *i.e.*, if the n_1 cases are consecutive incident cases from a defined population base and the n_2 second primaries are consecutive incidences from the same population base, then the estimates of both the gene prevalence and the relative risk are indirect estimates of the same parameters that would be obtained by a valid conventional case-control study of the same population base.

We can evaluate the relative efficiency of this design to the conventional case-control study by equating the sample sizes and obtaining the ratio of the variances. Let the variances of the estimate of ψ be denoted v_2 and v_1 for these two designs, respectively. Then

$$v_2 \simeq (n_1q)^{-1} + [n_1(1-q)]^{-1} + (n_2r)^{-1} + [n_2(1-r)]^{-1},$$

and

$$v_1 \simeq (n_0p)^{-1} + [n_0(1-p)]^{-1} + (n_1q)^{-1} + [n_1(1-q)]^{-1},$$

where n_0 is the number of conventional controls.
[Note: These are due to Woolf (7). Although they are not the most accurate approximations available for data analysis (8), their closed form permits a simple expression for the relative efficiency, and thus for illustrating the relative merits of the designs.] Setting $n_0 = n_1 = n_2$, it is easily shown that the relative efficiency of the two designs for estimating ψ is given by:

$$\rho_\psi = \frac{v_1}{v_2} = \frac{\psi^2 + \psi(1 - p + p\psi)^2}{(1 - p + p\psi^2)^2 + \psi(1 - p + p\psi)^2}.$$

These relative efficiencies are displayed in Table 2 for a variety of gene prevalences and relative risks. As an example, on the second row it can be seen that a risk factor with a prevalence of 5% and a relative risk of 5 will have a prevalence of 21% among cases and 57% among second cancers, and so the second cancers design is more efficient by a factor of 2.7 to 1. The results show that the second cancers design increases in relative efficiency as the prevalence is reduced and can be vastly more efficient in certain circumstances. For example, a gene with relative risk of 50 and a population prevalence of 1 in 1000 would require a conventional case-control study 40 times as large as a study of second primaries. For other configurations, the efficiency gains are less dramatic, but still substantial. Table 2 also shows configurations for which the second cancers design is less efficient than the conventional case-control study. In general, it is less efficient if $p(1 + \psi) > 1$.

The relative efficiency of the designs for estimating the population prevalence (p), denoted ρ_p, can be evaluated in an analogous manner, leading to:

$$\rho_p = \frac{\psi^2}{2\psi(1 - p + p\psi)^2 + (1 - p + p\psi^2)^2}.$$

[Note: ρ_p is here defined to be the ratio of the variance estimates of $\log [p/(1-p)]$ from the two designs.] These values are also provided in Table 2. Again, substantial efficiency gains are possible, although somewhat less than for the estimation of ψ. The reason for the lesser gains in efficiency are in part due to the fact that the random errors in the estimation of both q and r are embodied in the derived estimate of p from the second cancers design, using Eq. D, whereas in a conventional case-control study p is estimated directly from the observed relative frequency of the gene in the controls.

Turning to the evaluation of gene-environment interactions, the relative efficiency of the designs is influenced by a number of factors, including the prevalences of the two interacting factors, their individual relative risks, and their statistical dependence in the control group. A multiplicative interaction, hereafter denoted ϕ, is defined to be the ratio of the odds ratio linking the gene x and the environmental factor y in the case group to the corresponding odds ratio in the control group. A

Table 2 Relative efficiencies for estimating the relative risk

Relative risk (ψ)	Gene prevalences			Relative efficiency[a]	
	Controls (p)	Cases (q)	Second primaries (r)	ρ_ψ	ρ_p
5	0.25	0.63	0.89	0.65	0.28
	0.05	0.21	0.57	2.7	1.3
	0.01	0.05	0.20	4.4	2.0
	0.001	0.005	0.02	4.9	2.3
10	0.25	0.77	0.97	0.27	0.11
	0.05	0.34	0.84	2.1	1.3
	0.01	0.09	0.50	7.1	3.6
	0.001	0.01	0.09	9.7	4.6
20	0.25	0.87	0.99	0.10	0.03
	0.05	0.51	0.95	0.92	0.68
	0.01	0.17	0.80	8.1	4.9
	0.001	0.02	0.29	18.5	9.2
50	0.25	0.94	1.00	0.03	0.01
	0.05	0.72	0.99	0.19	0.15
	0.01	0.34	0.96	3.3	2.8
	0.001	0.05	0.71	39.7	21.5
100	0.05	0.84	1.00	0.05	0.04
	0.01	0.50	0.99	1.0	0.9
	0.001	0.09	0.91	41.9	27.6

[a] Efficiency of case-control study of second cancers relative to conventional case-control study. ρ_ψ and ρ_p are the relative efficiencies for estimating ψ and p, respectively.

convenient premise for our comparisons is to assume that x and y are statistically independent in the population controls (9); *i.e.*, the odds ratio linking x and y in the control group is 1. In this case, under the null hypothesis of no interaction, x and y will also be statistically independent in the case series, and the variance of the estimate of the interaction parameter (ϕ) in the conventional case-control study when the true value of ϕ is in the region of the null value of 1 is thus approximated by:

$$w_1 = \frac{1}{n_0 p_x p_y (1 - p_x)(1 - p_y)} + \frac{1}{n_1 q_x q_y (1 - q_x)(1 - q_y)},$$

where p_x and p_y are the control prevalences of the gene and the environmental factor, respectively, and q_x and q_y are the corresponding prevalences among the cases. [Note: This approximation deteriorates as ϕ departs from the null value of 1.] Under the null hypothesis of no interaction, the two factors will remain independent among the series of second primaries, and so the second cancers design leads to a corresponding variance estimate:

$$w_2 = \frac{1}{n_1 q_x q_y (1 - q_x)(1 - q_y)} + \frac{1}{n_2 r_x r_y (1 - r_x)(1 - r_y)},$$

where r_x and r_y are defined in an analogous manner. The relative efficiency of the two designs for estimating ϕ, denoted ρ_ϕ, depends not only on the prevalences of the gene, but also on the changing prevalences of the environmental factor y. However, if we make the approximate simplifying assumption that the factor is not a particularly strong risk factor and that changes in its prevalences are such that $p_y(1 - p_y) \simeq q_y(1 - q_y) \simeq r_y(1 - r_y)$, then it is easily shown that:

$$\rho_\phi \simeq \rho_\psi.$$

i.e., the relative efficiency for estimating a gene-environment interaction is similar to the relative efficiency for estimating the relative risk of the gene, and so the results in Table 2 for ρ_ψ also apply to the study of gene-environment interactions.

When planning a study, use of the preceding approximations leads to the following simple formula for calculating the power for detecting a specified interaction, where w_i represents either w_1 or w_2 depending on which study design is used:

$$P_i = \Phi[-z_{\alpha/2} + w_i^{-1/2} \log \phi],$$

where α is the (two-sided) significance level, and $\Phi(\cdot)$ represents the standard normal distribution.

Results

As an example, consider a recently planned case-control study of risk factors for melanoma. This study involves a conventional case-control study of 1650 incident cases and 1650 population controls. It is anticipated that during the accrual period 165 second primary melanomas will be ascertained; *i.e.*, $n_0 = 1650$, $n_1 = 1650$, and $n_2 = 165$. The conventional case-control study will be used to evaluate environmental risk factors and suspected prevalent susceptibility markers, such as mutagen sensitivity (10). We focus here on the use of the case-control study of second primaries to evaluate the relative risk (ψ) of *p16* mutations, the prevalence of *p16* mutations in the population (p_x), and the interaction of *p16* and sun exposure, the major suspected environmental risk factor. It is assumed that moderate-to-heavy sun exposure has a prevalence in the controls of 0.4 and a relative risk of 1.5 (11). The following projections are based on the assumption that the prevalence of *p16* among cases (1st primaries) is 1–2%, where 5–10% of subjects are assumed to have familial melanoma, 20% of which are due to *p16* (12). It is speculated that the relative risk may range from 20 to 100. On the basis of these sample sizes and values of q and ψ, the relative yield of the two designs can be compared (Table 3).

The second cancers study is seen to be greater than an order of magnitude more efficient for all six configurations, ranging in relative efficiency approximately from 10 to 72. The problem in using the conventional study for estimating the population prevalence of *p16*, denoted p, is highlighted by the

Table 3 Planned melanoma case-control studies

Gene prevalence (q)[a]	Relative risk (ψ)	No. of gene carriers[b]	Relative efficiency (ρ_ψ)	Power[c]	
				P_1	P_2
0.01	20	0.8	12.2	0.10	0.70
0.01	50	0.3	35.0	0.07	0.78
0.01	100	0.2	71.6	0.05	0.79
0.02	20	1.7	10.6	0.17	0.91
0.02	50	0.7	28.0	0.09	0.93
0.02	100	0.3	52.5	0.07	0.92

[a] Gene prevalence among incident first primary cases.
[b] Expected number of gene carriers among 1650 controls ($1650p_s$).
[c] Power to detect an interaction odds ratio of 5, at the 5% significance level (two-sided): P_1 corresponds to conventional case-control study, and P_2 corresponds to second cancers study.

fact that the expected number of gene carriers in the controls ranges from 0.2 to 1.7; *i.e.*, in our series of 1650 population controls, we only expect to find approximately one gene carrier. In the second cancers study, p is estimated indirectly, but is based on the "richer" prevalences in cases and second cancers. The power to detect an interaction odds ratio of 5 is presented for each configuration. It is clear from the numbers (last two columns of Table 3) that the conventional study has very low power to detect such an interaction, whereas the study of second cancers has good power.

By contrast, the evaluation of a prevalent marker, such as mutagen sensitivity, is accomplished more efficiently via the conventional case-control study. Although this is a continuous biological marker measuring the mean number of chromosome breaks in cells exposed to bleomycin, "sensitivity" has been arbitrarily defined as greater than 0.8 breaks per cell, for which the population prevalence has been estimated at approximately 45%, and the corresponding relative risk for aerodigestive cancers is estimated to be in the range of 2–4 (13). Using these as the basis for power calculations for the proposed melanoma study, the relative efficiency of the second cancers design, based on equivalent sample sizes, ranges from 0.84 to 0.46. Furthermore, because the available numbers of first primary cancers and controls are much greater, the actual efficiency of the conventional case-control study greatly exceeds that of the second cancers design.

Discussion

The principal message of this article is that the estimation of population-based epidemiological parameters pertaining to rare but highly penetrant cancer susceptibility genes, although difficult and expensive using conventional epidemiological study designs, may be much more feasible in a design in which incident second primary cancers are compared with patients with first primaries. Although the relative efficiencies depend on the population prevalence of the gene and its relative risk, and the magnitude of these can only be speculated on at the design stage of the study, the ranges of values that are plausible for known cancer genes include combinations in which the relative efficiencies range from one to two orders of magnitude for estimating the relative risk. Furthermore, second cancers designs with feasible sample sizes permit the exploration of gene-environment interactions, a task that is not feasible for rare genes using the conventional case-control design. The use of designs that are based solely on incident cases of cancer eliminates the potential biases that accompany the use of a population-based control group (14). A population-based study thus can be accomplished using only information that is easily ascertainable on a population basis, *i.e.*, consecutive series of incident cancers and second primary cancers. By contrast, conventional case-control studies are critically dependent on the ascertainment of random population controls, a task that is difficult to accomplish and impossible to validate. The trade-off in the proposed design is the assumption that the second primary cancers are truly biologically independent of the first primaries, a task that is currently accomplished by pathologists subjectively on the basis of physiological and histological factors. In the context of melanoma, the subject of our example, the classification is based primarily on a judgment about whether the lesion originated in the dermis (metastasis) or the epidermis (second primary), which in turn is based on the anatomic configuration of the tumor (15). Any misclassifications of this nature are more likely to attenuate the estimate of the relative risk than to inflate it (16).

We must also assume that the survival from diagnosis of cases is unaffected by the risk factors under investigation and that the risk of diagnosing a second primary is not affected artifactually by patient factors, notably the status of the gene or genes of interest, and that treatment for the first primary has no confounding influence. In fact, it has been speculated that radiotherapy treatment would be selectively carcinogenic in individuals with important germ line mutations. This is unlikely to be a problem in the melanoma study discussed earlier, but it could be important in other disease sites.

Melanoma represents an ideal model for the study of second primaries in that very little of the target tissue is removed by surgery, and so an individual is still fully at risk for a second cancer. Clearly, for many organ-defined cancers, many if not all of the cells at risk may be removed by surgery for the first primary, fundamentally reducing the absolute risk of a second primary. However, with the conditions noted above, there is no particular reason to suppose that this would alter the *relative* risk among gene carriers, in which case the second cancers design would remain valid for estimating the relative risk, although the incidence of second cancers may be reduced to the extent that the design is impractical. In fact, studies of risk factors for second primaries are often conducted in a prospective cohort of patients with first primaries (2–5). In this setting, projection of the expected numbers of second cancers as a function of follow-up time is necessary, and the published tables of incidence rates of second cancers are useful for this purpose (17).

A further cautionary note is that the use of unadjusted relative risks involves an implicit aggregation of risks across age and other risk factors. In fact, the key equations linking the three prevalences, Eqs. A and B, are valid only if the relative risk of the gene is constant across age categories (6). In practice, this is unlikely to be true, and an analysis stratified by age

at diagnosis is advisable; *i.e.*, age at diagnosis of first primaries should be frequency-matched with age at diagnosis of second primaries. However, the efficiency gains will still be apparent, since Eqs. A and B will be approximately valid within suitably narrow categories of age stratification.

In an earlier article, Rothman *et al.* (18) catalogued the study designs used in research on the epidemiology of biomarkers. The study of second primaries, either using the case-control method or a follow-up of a cohort of first primaries, deserves a prominent place in this epidemiological toolbox.

References

1. Khoury, M. J., Beaty, T. H., and Cohen, B. M. Fundamentals of Genetic Epidemiology, pp. 164–199. Oxford, United Kingdom: Oxford University Press, 1993.

2. Franco, E. L., Kowalski, L. P., and Kanda, J. L. Risk factors for second cancers of the upper respiratory and digestive systems: a case-control study. J. Clin. Epidemiol., *44:* 615–625, 1991.

3. Day, G. L., Blot, W. J., Shore, R. E., McLaughlin, J. K., Austin, D. F., Greenberg, R. S., Liff, J. M., Preston-Martin, S., Sarkar, S., Schoenberg, J. B., and Fraumeni, J. F. Second cancers following oral and pharyngeal cancers: role of tobacco and alcohol. J. Natl. Cancer Inst., *86:* 131–137, 1994.

4. Barbone, F., Franceschi, S., Talamini, R., Barzan, L., Franchin, G., Favero, A., and Carbone, A. A follow-up study of determinants of second tumor and metastasis among subjects with cancer of the oral cavity, pharynx, and larynx. J. Clin. Epidemiol., *49:* 367–372, 1996.

5. Bernstein, J. L., Thompson, W. D., Risch, N., and Holford, T. R. Risk factors predicting the incidence of second primary breast cancer among women diagnosed with a first primary breast cancer. Am. J. Epidemiol., *136:* 925–936, 1992.

6. Begg, C. B., Zhang, Z-F., Sun, M., Herr, H., and Schantz, S. P. Methodology for evaluating the incidence of second primary cancers with application to smoking-related cancers from the Surveillance, Epidemiology and End Results (S. E. E. R.) Program. Am. J. Epidemiol., *142:* 653–665, 1995.

7. Woolf, B. On estimating the relationship between blood group and disease. Ann. Hum. Genet., *19:* 251–253, 1955.

8. Breslow, N. E., and Day, N. E. Statistical Methods in Cancer Research. The Analysis of Case-Control Studies. IARC Scientific Publ. No. 32, pp. 134–135. Lyon, France: IARC, 1980.

9. Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat. Med., *13:* 153–162, 1994.

10. Spitz, M. R., Hoque, A., Trizna, Z., Schantz, S. P., Amos, C. I., King, T. M., Bondy, M. L., and Hsu, T. C. Mutagen sensitivity as a risk factor of second malignant tumors following malignancies of the upper aerodigestive tract. J. Natl. Cancer Inst., *86:* 1681–1684, 1994.

11. Berwick, M., and Chen, Y. T. Reliability of reporting sunburn history in a case-control study of cutaneous malignant melanoma. Am. J. Epidemiol., *141:* 1033–1037, 1995.

12. Hussussian, C. J., Struewing, J. P., Goldstein, A. M., Higgins, P. A. T., Ally, D. S., Sheahan, M. D., Clark, W. H., Tucker, M. A., and Dracopoli, N. C. Germline p16 mutations in familial melanoma. Nat. Genet., *8:* 15–21, 1994.

13. Spitz, M. R., Fueger, J. J., Halabi, S., Schantz, S. P., Sample, D., and Hsu, T. C. Mutagen sensitivity in upper aerodigestive tract cancer: a case-control analysis. Cancer Epidemiol., Biomarkers & Prev., *2:* 329–333, 1993.

14. Khoury, M. J., and Flanders, W. D. Non-traditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am. J. Epidemiol., *144:* 207–213, 1996.

15. Barnhill, R. L. Pathology of melanocytic nevi and malignant melanoma. Boston, MA: Butterworth and Heinemann, 1995.

16. Newell, D. J. Misclassification in 2 × 2 tables. Biometrics, *19:* 187–188, 1963.

17. Boice, J. D., Storm, H. M., and Curtis, R. E. Multiple primary cancers in Connecticut and Denmark. Washington, D. C.: United States Department of Health, Education and Welfare, Public Health Service, National Cancer Institute Monograph 68, 1985.

18. Rothman, N., Stewart, W. F., and Schulte, P. A. Incorporating biomarkers into cancer epidemiology: a matrix of biomarker and study design categories. Cancer Epidemiol., Biomarkers & Prev., *4:* 301–311, 1995.