Statistical Analysis of Molecular Epidemiology Studies Employing Case-Series[1]

Colin B. Begg[2] and Zuo-feng Zhang

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

Abstract

The case-series design is being used increasingly to explore associations between environmental risk factors and genetic markers. It is demonstrated that the odds ratio derived from a case-series study is the ratio of the relative risk for developing marker-positive disease to the relative risk for developing marker-negative disease. This parameter is an empirical manifestation of etiological heterogeneity with respect to the risk factor under study, and it can be used to construct a statistical significance test. Presence of etiological heterogeneity, as reflected in departures of this parameter from unity, could be a result of either the presence of distinct causal mechanisms for the two categories of cases, or a different strength of effect via the same mechanism. The case-series approach represents an efficient and valid approach for evaluating gene-environment associations, especially in referral centers where it is difficult to identify a valid control group.

Introduction

Recent advances in technology for identifying genetic mutations and their products have led to an interest in correlating these biological markers with clinical and epidemiological factors with a view to better understanding the natural history and the etiology of diseases. There is a concomitant need for research into the most appropriate study designs for investigating these issues, and to determine the relevant statistical methods for data analysis and interpretation. Our purpose in this article is to examine the relevant methodology for studying the relationship between cancer biomarkers and environmental risk factors for cancer.

Our motivating example concerns the possible role of smoking in causing bladder cancers that are characterized by *p53* mutations. Several recent studies have identified correlations between smoking and *p53* mutations in tumor samples from a variety of tumor types, including lung cancer (1, 2), head and neck cancer (3), esophageal cancer (4), and bladder cancer (5). The results of these studies imply a distinct etiology for *p53*+ cancers in which smoking plays a stronger role than in *p53*– cancers.

Each of these studies has involved a case-series design in which tumor samples from a series of cancer patients are evaluated. That is, the studies have not involved control groups of patients without cancer, as in conventional studies of etiological factors. Most commonly, the relationship between the risk factor (smoking) and the marker (*p53*) has been characterized by the odds ratio. In contrast, Taylor *et al.* (6) conducted a study of the relationship between occupational exposures and the activation of the *ras* oncogene in the etiology of acute myeloid leukemia, employing a conventional control group identified by random digit dialing. In this study, in addition to calculating unadjusted and adjusted odds ratios characterizing the correlation between the occupational factors and *ras* mutations, the investigators also assessed the relative risks of the occupational risk factors for incidence of *ras*+ and *ras*– tumors separately.

In this article we will clarify the yield from the case-series design in relation to the information that can be obtained from the conventional case-control approach. Our particular focus will be on the interpretation of the parameters from the statistical models commonly employed, that is, Mantel-Haenszel techniques and logistic regression.

Methods

We are interested primarily in the hypothesis that the two categories of cases, distinguished by the presence or absence of the tumor marker, are characterized by etiological heterogeneity. That is, we are testing the hypothesis that the strength of effect of one or more risk factors differs for the two case groups. Such an effect could be because the causal pathway differs, or it could merely reflect a different magnitude of effect via the same mechanism. Empirical evidence of such etiological heterogeneity with respect to one or more risk factors would provide strong justification for more detailed investigations of the specific mechanisms of action.

Case-Series Design. This study design consists of a series of incident cases. Ideally, this would be a consecutive series of population-based incident cases. If the ascertainment is not complete, or if the study is, say, hospital based, we must assume that case selection for the two disease categories is not influenced differentially by the risk factors.

Suppose that Y is the risk factor of primary interest, assumed for simplicity to be binary, and that W denotes the set of remaining risk factors, where $Y+$ indicates presence of the risk factor and $Y-$ indicates its absence. Let $X+(X-)$ denote the presence (absence) of the tumor marker. Furthermore, let $\psi(W)$ be the odds ratio relating Y and X, conditional on W. In the context of our bladder cancer example, Y represents smoking status, X represents the presence or absence of *p53* mutations in the tumor samples, and W represents the remaining risk factors.

Received 8/10/93; revised 11/22/93; accepted 11/23/93.

[1] This research was supported by National Institute of Environmental Health Services Grant ES-06718, and by National Cancer Institute Grant CA-47538 from the NIH, Department of Health and Human Services.

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

We can evaluate $\psi(W)$ using standard statistical methods such as the Mantel-Haenszel procedure or logistic regression. A test of the hypothesis that $\psi(W) = 1$ is a test of the hypothesis that the strength of Y as a risk factor is different for the two case groups (*e.g.*, *p53+* and *p53–*).

Case-Control Design. A more conventional approach is to use a case-control design, in which a control series is assembled in addition to the preceding case-series. We will assume that the control group is sampled randomly from the source population of the cases, as opposed to using a matched design. In this setting the conventional analytic strategy is to use polychotomous logistic regression (7). In this model the relationships between marker-positive cases and controls, and between marker-negative cases and controls, are both modeled concurrently using two separate (logistic) regression functions. Let β_1 be the coefficient of the primary risk factor in the logistic regression relating marker-positive cases and controls, and let β_2 be the corresponding parameter relating marker-negative cases and controls. If there are no interactions between Y and W, then β_1 is the conditional log odds ratio of the risk factor on marker-positive disease, and β_2 is the conditional log odds ratio of the risk factor on marker-negative disease. To test the hypothesis that the two diseases possess etiological heterogeneity with respect to the risk factor, one can test the hypothesis that $\beta_1 = \beta_2$, that is, that the two odds ratios are equal. Such a comparison can be accomplished by using, for example, a likelihood ratio test. Quantitative evidence of the degree of departure from the hypothesis can be characterized by the difference in these coefficients, $\beta_1 - \beta_2$. This is the logarithm of the ratio of the two adjusted relative risks of the risk factor, that is, the relative risk with respect to marker-positive and marker-negative cases, respectively. Specifically, if $\theta_1(W)$ and $\theta_2(W)$ are the respective relative risks, then

$$\log(\theta_1(W)/\theta_2(W)) = \beta_1 - \beta_2.$$

Relationship of Case-Series and Case-Control Approaches. The odds ratio derived from the case-series study, $\psi(W)$, is the same parameter as the ratio of relative risks obtained from the polychotomous model, that is,

$$\psi(W) = \theta_1(W)/\theta_2(W).$$

However, the fact that different statistical models are used in the two approaches means that different (consistent) estimators will be obtained, depending on whether $\psi(W)$ is estimated directly from the case series, or indirectly from estimates of $\theta_1(W)$ and $\theta_2(W)$. When no adjustments are made for the remaining risk factors, the two methods do in fact produce identical estimators, and this will be illustrated by example in the next section.

The theoretical equivalence of the two approaches is demonstrated easily by considering the component probabilities of the various odds ratios. Let $Z+$(Z–) denote case (control) status. In the case series study all subjects are cases, and so all terms in the odds ratio $\psi(W)$ are conditional upon $Z+$ and W. Specifically,

$$\psi(W) = \frac{P(Y+ \mid X+, Z+, W)/P(Y- \mid X+, Z+, W)}{P(Y+ \mid X-, Z+, W)/P(Y- \mid X-, Z+, W)}.$$

In the case-control setting the combination $(X+, Z+)$ represents a marker-positive case, while $Z-$ represents a control. (Note: the tumor marker is not measured in the controls because they have no tumor.) Therefore, the odds ratio (relative risk) for the marker-positive cases is

$$\theta_1(W) = \frac{P(Y+ \mid X+, Z+, W)/P(Y- \mid X+, Z+, W)}{P(Y+ \mid Z-, W)/P(Y- \mid Z-, W)}.$$

Correspondingly,

$$\theta_2(W) = \frac{P(Y+ \mid X-, Z+, W)/P(Y- \mid X-, Z+, W)}{P(Y+ \mid Z-, W)/P(Y- \mid Z-, W)}.$$

It is clear by inspection that $\psi(W) = \theta_1(W)/\theta_2(W)$. The denominators of $\theta_1(W)$ and $\theta_2(W)$ are the same because the same control population is relevant for each case series. This is true in principle, although in practice, if one were to use separate control groups, as would be the case for a pair-matched design, the estimates of $\psi(W)$ from the two methods would not be equivalent even when we do not condition on the remaining risk factor, W.

Table 1 Frequencies by *p53* mutations and smoking status

Smoking Status	Cases		Controls
	p53+	*p53–*	
Smoker	34	43	81
Nonsmoker	10	21	64

Example

We illustrate the method using data from our own case-series study of the relationship between smoking and *p53* mutations in patients with bladder cancer, treated at Memorial Sloan-Kettering Cancer Center (8). The raw frequencies are contained in Table 1. For illustrative purposes we have employed a control group consisting of patients with other cancers believed to be unrelated to smoking, although this would not be an ideal control group for a case-control study in general.

The odds ratios and confidence intervals are presented in Table 2. The unadjusted odds ratios are calculated directly from the cross-products, as usual. That is, $\hat{\psi} = (34 \times 21)/(10 \times 43)$, $\hat{\theta}_1 = (34 \times 64)/(10 \times 81)$, $\hat{\theta}_2 = (43 \times 64)/(21 \times 81)$. The equivalence of $\hat{\psi}$ and $\hat{\theta}_1/\hat{\theta}_2$ is evident by inspection. Calculation of adjusted odds ratios involves the use of simple logistic regression for the case-series study, and polychotomous logistic regression for the case-control study. The estimates of $\psi(W)$ can differ for the two approaches because additional parameters are necessary in the polychotomous model, and so the models are slightly different. However, in this example the estimates and confidence intervals are virtually identical. To compute the confidence interval for $\psi(W)$ derived from the polychotomous model, it is necessary to recognize the covariance between the estimates of $\log \hat{\theta}_1(W)$ and $\log \hat{\theta}_2(W)$. If this covariance is c, and the individual variances of the estimates are v_1 and v_2 respectively, then

$$\text{var} \log[\hat{\theta}_1(W)/\hat{\theta}_2(W)] = v_1 + v_2 - 2c.$$

These terms are available in most computer software packages.

The statistical modelling for evaluating $\psi(W)$ and its dependence on W is generally simpler for the case-series model. If any of the risk factors modify the influence of the primary risk factor, Y, then this can be identified by studying first order interactions in the logistic model. By contrast,

Table 2 Odds ratio estimates

Design	Parameter	Estimate (95% CI)	
		Unadjusted	Adjusted[a]
Case-control	$\theta_1(W)$	2.69 (1.23–5.87)	3.61 (1.41–9.29)
	$\theta_2(W)$	1.62 (0.87–3.00)	2.11 (0.98–4.54)
	$\theta_1(W)/\theta_2(W) \equiv \psi(W)$	1.66 (0.69–4.01)	1.71 (0.66–4.43)
Case-series	$\psi(W)$	1.66 (0.69–4.01)	1.71 (0.63–4.66)

[a] Adjusted for age using logistic regression for the case-series approach and polychotomous logistic regression for the case-control approach.

these modifications are represented by second order interactions in the polychotomous model, and so it would be easier to overlook important interactions when using the polychotomous approach.

Discussion

We have shown that the odds ratio relating an environmental risk factor to the presence of a biological marker is an appropriate measure for characterizing the degree of etiological heterogeneity between the disease groupings defined by the marker. This parameter has been shown to be the ratio of the relative risk of the factor in causing marker-positive disease to the relative risk in causing marker-negative disease. Moreover, it can be estimated directly from an appropriately designed case-series study without the need for a control group. The use of a control group is necessary if we wish to estimate the individual relative risks. These observations legitimize the common recent practice of exploring gene-environment associations in case-series studies.

In carrying out a case-series study, it is important nonetheless to observe the same epidemiological principles of case selection as is appropriate for case-control studies. A consecutive series of incident cases is ideal, preferably from a defined population base. For example, selection of cases may be dependent on the availability of a sufficiently large tumor sample to permit the marker studies, and this could lead to bias if tumor size at presentation is related to the risk factors, the prevalence of the marker, or both. Similarly, the use of prevalent cases can lead to biased estimates if the risk factors under study are associated with patient survival. Accession of tumor specimens from prevalent cases with advanced disease is common in hospital settings where these laboratory studies often are dependent on "found" specimens. These observations are important, because our impression is that the issue of case selection is rarely addressed in these types of study.

The use of logistic regression on a case-series is a convenient analytic approach for evaluating the presence of gene-environment associations. The proper role of these studies is exploratory, to examine the nature of the effects of known risk factors and to identify hitherto unrecognized risk factors that may act on rare genetic traits, and which could be overlooked in the typical case-control or cohort study because of lack of power.

The proposed use of a case-series design to estimate the ratio of relative risks is mathematically equivalent to the use of the mortality odds ratio, which is based only on persons who have died to estimate the proportional increase in mortality as a result of an environmental risk factor, as demonstrated by Miettinen and Wang (9). The relationship of this model to the logistic regression model was explained subsequently by Breslow and Day (10). Prentice and Pyke (11) have supplied further theoretical justification for the use of logistic regression models to study the simultaneous incidence rates of multiple disease types.

The choice of statistical model is often made on the grounds of expediency, in recognition of the fact that the model may be only an approximation of the true relationships between the variables under study. However, if the polychotomous logistic model is an exact reflection of the relationship between the two (or more) distinct disease entities and the control group, then we do lose some information about $\psi(W)$ when we discard the controls. However, theoretical studies of this model have demonstrated that the loss of information is usually small, especially for the estimation of individual regression parameters (12, 13). Consequently, given the costs and potential biases inherent in selecting a control group, the case-series design represents an efficient and valid methodology for studying gene-environment associations.

References

1. Suzuki, H., Takahashi, T., Kuroishi, T., Suyame, M., Ariyoshi, Y., and Ueda, R. *p53* mutations in non-small cell lung cancer in Japan: association between mutations and smoking. Cancer Res., *52:* 734–736, 1992.

2. Kondo, K., Umemoto, A., Akimoto, S., Uyama, T., Hayashi, K., Ohnishi, Y., and Monden, Y. Mutations in the *p53* tumor suppressor gene in primary lung cancer in Japan. Biomed. Biophys. Res. Commun., *183:* 1139–1146, 1992.

3. Field, J. K., Spandidos, D. A., Malliri, A., Gosney, J. R., Yiagnisis, M., and Stell, P. M. Elevated p53 expression correlates with a history of heavy smoking in squamous cell carcinoma of the head and neck. Br. J. Cancer., *64:*473–577, 1991.

4. Hollstein, M. C., Peri, L., Mandard, A. M., et al. Genetic analysis of human esophageal tumors from two high incidence geographic areas: frequent *p53* base situations and absence of *ras* mutations. Cancer Res., *51:* 4102–4106, 1991.

5. Spruck, C. H., Rideout, W. M., Olumi, A. F., Ohneseit, P. F., Yang, A. S., Tsai, Y. C., Nichols, P. W., Horn, T., Hermann, G. G., Steven, K., Ross, P. K., Yu, M. C., Jones, P. A. Distinct pattern of *p53* mutations in bladder cancer: relationship to tobacco usage. Cancer Res., *53:* 1162–1166, 1993.

6. Taylor, J. A., Sandler, D. P., Bloomfield, C. D., Shore, D. L., Ball, E. D., Neubauer, A., McIntyre, O. R., Liu, E. *ras* Oncogene activation and occupational exposures in acute myeloid leukemia. J. Natl. Cancer Inst., *84:*1626–1632, 1992.

7. Dubin, N., Pasternak, B. S. Risk assessment for case-control subgroups by polychotomous logistic regression. Am. J. Epidemiol., *123:*1101–1117, 1986.

8. Zhang, Z. F., Sarkis, A. S., Cordon-Cardo, C., Dalbagni, G., Melamed, J., Aprikian, A., Pollack, D., Herr, H. W., Fair, W. R., Reuter, V. E., and Begg, C. B. Tobacco smoking, occupation, and p53 nuclear overexpression in early stage bladder cancer. Cancer Epidemiol., Biomarkers & Prev., *3:*19–24, 1994.

9. Miettinen, O. S., and Wang, J. D. An alternative to the proportionate mortality ratio. Am. J. Epidemiol., *114:* 144–148, 1981.

10. Breslow, N. E., and Day, N. E. Statistical Methods in Cancer Research: Volume 2—The Design and Analysis of Cohort Studies (IARC Scientific Pub. No. 82). Lyon, France: International Agency for Research on Cancer, 1987.

11. Prentice, R. L., and Pyke, R. Logistic disease incidence models and case-control studies. Biometrika, *66:* 403–412, 1979.

12. Begg, C. B., and Gray, R. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika, *71:* 11–18, 1984.

13. Bull, S. B., and Donner, A. A characterization of the efficiency of individualized logistic regressions. Can. J. Statist., *21:* 71–78, 1993.

Statistical analysis of molecular epidemiology studies employing case-series.

C B Begg and Z F Zhang

Cancer Epidemiol Biomarkers Prev 1994;3:173-175.