Cytochrome *P*450 1A1 Polymorphism and Childhood Leukemia: An Analysis of Matched Pairs Case-Control Genotype Data

Steve Selvin

School of Public Health, University of California at Berkeley, Berkeley, California

Abstract

The association between the genotypic frequencies of the cytochrome *P*450 1A1 polymorphism and the risk of childhood leukemia is explored with the data from a matched case-control study. The data are displayed in a 3 × 3 case-control array, and the discordant pair counts are assessed for quasi-independence, homogeneity, and symmetry. This statistical approach is contrasted to the more typical analysis of matched data based on a conditional logistic model and estimated odds ratios. The statistical analysis of 175 matched pairs (part of a large study of potential environmental/genetic influences on the risk of childhood leukemia) showed no evidence of an association between cytochrome *P*450 1A1 genotype frequencies and case-control status. (Cancer Epidemiol Biomarkers Prev 2004;13(8):1371–4)

Introduction

The possible association between cytochrome *P*450 1A1 (CYP1A1) genetic polymorphism and disease has been explored by several investigators. Some examples are the investigations of Ladonna et al. (1) on Spanish toxic oil syndrome, Ishibe et al. (2) on breast cancer, and Kim et al. (3) on cervical cancer and several reports on the association with childhood leukemia (e.g., refs. 4, 5). These analyses, as well as most analyses of matched case-control genetic data, are summarized in terms of ratios of discordant pairs of observations (odds ratios) estimated from conditional logistic regression models. The following statistical approach to the analysis of the CYP1A1/leukemia matched data is based on a 3 × 3 case-control array that allows an assessment of three fundamental statistical issues (i.e., independence, homogeneity, and symmetry of the discordant pairs). In addition, this approach is contrasted to the more typical use of odds ratios estimated from a conditional logistic model.

Data

To describe the CYP1A1 polymorphism and the risk of acute lymphoblastic leukemia, data collected as part of the Northern California Childhood Leukemia Study are used. These matched case-control observations were abstracted from a large number of genetic/environmental variables that potentially influence the risk of childhood leukemia. The cases are children ages 0 to 14 years old with newly diagnosed leukemia (1995 to 1999) obtained from major hospitals in the San Francisco Bay Area. Comparison with California State Cancer Registry data shows that >90% of the eligible children were ascertained. The control children were randomly selected from birth certificate records and matched to cases with respect to sex, age, race, and county of birth. A more extensive description of this far-ranging study is found elsewhere (6). The CYP1A1/leukemia data consisting of 175 matched pairs of acute lymphoblastic leukemia cases and their controls (117 concordant and 58 discordant pairs) are given in Table 1.

Quasi-Independence

Genotype data classified into a square array are quasi-independent when the categorical variables (row and columns) are independent with respect to only the discordant pairs. No restrictions are placed on the concordant pairs (the main diagonal of the case-control array). In symbols, the six expected cell counts of the discordant pairs (denoted F_{ij}) are

$$F_{ij} = Np_iq_j \qquad \text{for } i \neq j = 1, 2, \text{ and } 3.$$

The values p_i represents the case genotypic frequencies, and q_j represents the control genotypic frequencies. The quantity N represents the ''total number of pairs'' that would have occurred if the genotypic frequencies were independent and the data were randomly sampled. Specifically, the value $N = n / \Sigma p_iq_j$ for $i \neq j = 1, 2$, and 3 where n represents the total observed number of discordant pairs.

The maximum likelihood estimates of the p_i and q_j frequencies are found by iterative techniques (7) or an application of a specialized log-linear model (8). Both procedures are designed to estimate the genotypic frequencies excluding the concordant pairs from consideration

Received 9/24/03; revised 12/22/03; accepted 2/9/04.

Grant support: U.S. Environmental Health Sciences research grants R01 ES09137 and PS42 ES04705.

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.

Requests for reprints: Steve Selvin, School of Public Health, University of California at Berkeley, Berkeley, CA 94720. Phone: 510-642-3241; Fax: 510-643-5163. E-mail: selvin@stat.berkeley.edu

Table 1. The observed numbers of matched pairs by case-control status and CYP1A1 genotypes

	Control: AA*	Control: AG	Control: GG	Total
Case: AA*	103	26	2	131
Case: AG	23	14	2	39
Case: GG	1	4	0	5
Total	127	44	4	175

*AA, CYP1A1 homozygotic wild-type.

(truncated). For the Northern California Childhood Leukemia Study data (Table 1), these estimated genotypic frequencies are

$$\hat{p}_1 = 0.339,\ \hat{p}_2 = 0.617,\ \hat{p}_3 = 0.044,$$

$$\hat{q}_1 = 0.305,\ \hat{q}_2 = 0.659,\ \hat{q}_3 = 0.035.$$

A natural estimate of the expected number of case-control discordant pairs based on the quasi-independence model is $\hat{F}_{ij} = \hat{N}\hat{p}_i\hat{q}_j$ where $\hat{N} = n \,/\, \Sigma\hat{p}_i\hat{q}_j$ for $i \neq j = 1, 2,$ and 3. These expected counts estimated from the data in Table 1 are displayed in Table 2.

The Pearson χ^2 goodness-of-fit test statistic (1 *df*) summarizing the deviations of the observed values from the expected values (Table 1 versus Table 2) is $X^2_Q = 0.712$ ($P = 0.399$). The *df* values are the number of observations (off-diagonal cell frequencies = 6) minus the number of independent parameters necessary to estimate the expected values or to specify the appropriate log-linear model. In the case of quasi-independence, five independent estimated parameters establish the expected values in Table 2 making the *df* equal to 1 (6 − 5 = 1).

In the context of the analysis of matched case-control data, another form of this same χ^2 test is called a ''test for the consistency of the odds ratios'' (9). The non-informative concordant pairs are excluded because the increased correlation within pairs due to the matching process tends to increase the number of concordant pairs relative to the number predicted by a model postulating independence.

Marginal Homogeneity

Quasi-independence does not imply marginal homogeneity (identical case and control genotypic frequency distributions). The issue of marginal homogeneity in general is the subject of several statistical articles (e.g., refs. 10-12). The expected row/column totals, under the conjecture of marginal homogeneity, are estimated by

$$\hat{N}_i = \frac{n_{i.} + n_{.i}}{2}$$

where $n_{i.} = \Sigma_j f_{ij}$, $n_{.j} = \Sigma_i f_{ij}$ and f_{ij} represents the number of pairs in the $(i,j)^{\text{th}}$ cell. For the CYP1A1/leukemia data, the estimated homogeneous marginal totals are

$$\hat{N}_1 = 129.0,\ \hat{N}_2 = 41.5,\ \text{and}\ \hat{N}_3 = 4.5.$$

Symmetry

When the case-control discordant pairs are independent and the marginal frequencies are homogeneous, the expected counts of case-control pairs create a symmetrical array; that is, when case-control status is unrelated to the genotypic frequencies, the counts within the three kinds of discordant pairs (f_{ij} versus f_{ji}) differ by chance alone. Under these conditions (independence + homogeneity = symmetry), the maximum likelihood estimates of the genotypic frequencies (denoted $\hat{P}_i$) are

$$\hat{P}_i = \frac{f_{\{i,j\}} f_{\{i,k\}}}{f_{\{1,2\}} f_{\{1,3\}} + f_{\{1,2\}} f_{\{2,3\}} + f_{\{1,3\}} f_{\{2,3\}}} \text{ for } j \neq k$$

where $f_{\{ij\}} = f_{ij} + f_{ji}$. Because the number of independent parameters equals the number of independent observations (2), the maximum likelihood estimates $\hat{p}_i$ can be derived by equating expected and observed frequencies (13). The specific estimated genotype frequencies from the acute lymphoblastic leukemia data are

$$\hat{P}_1 = 0.320,\ \hat{P}_2 = 0.641,\ \text{and}\ \hat{P}_3 = 0.039.$$

These estimated genotypic frequencies produce an estimate of the expected counts of matched case-control discordant pairs (denoted F_{ij}') where

$$\hat{F}'_{ij} = \hat{N}'\,\hat{P}_i\hat{P}_j = \frac{f_{\{ij\}}}{2}$$

The estimated ''sample size'' is $\hat{N}' = n \,/\, \Sigma\hat{P}_i\hat{P}_j$ for $i \neq j = 1, 2,$ and 3. The Northern California Childhood Leukemia Study matched pairs data (Table 1) produce the estimates displayed in Table 3. These expected discordant pairs are quasi-independent, and the marginal frequencies are homogeneous.

Table 2. The expected numbers of discordant matched pairs when case-control status is exactly quasi-independent of the CYP1A1 genotypic frequencies

	Control: AA*	Control: AG	Control: GG	Total
Case: AA	103	26.582	1.418	131
Case: AG	22.418	14	2.582	39
Case: GG	1.582	3.418	0	5
Total	127	44	4	175

*AA, CYP1A1 homozygotic wild-type.

Table 3. The expected number of discordant matched pairs when case-control status is exactly unrelated to the CYP1A1 genotype frequencies

	Control: AA*	Control: AG	Control: GG	Total
Case: AA	103	24.5	1.5	129.0
Case: AG	24.5	14	3.0	41.5
Case: GG	1.5	3.0	0	4.5
Total	129.0	41.5	4.5	175

*AA, CYP1A1 homozygotic wild-type.

The Pearson goodness-of-fit test statistic (Table 1 versus Table 3) is $X_S^2 = 1.184$ ($P = 0.757$) and has an approximate χ^2 distribution with 3 ***df*** when the observed counts randomly differ from the expected counts. The comparison of these observed and expected numbers of discordant pairs is identical to the sum of three McNemar-like test statistic (14). In symbols,

$$X_S^2 = \Sigma \frac{(f_{ij} - f_{ji})^2}{f_{ij} + f_{ji}} \text{ for } i \neq j = 1, 2, \text{ and } 3.$$

The χ^2 test for symmetry requires three independent estimated parameters giving a test statistic with 3 *df* (6 − 3 = 3). In addition, this test statistic partitions into three independent components each with 1 *df*.

In addition, the three estimated genotypic frequencies $\hat{P}_i$ lead to a compact form of a χ^2 test statistic to evaluate marginal homogeneity in a 3 × 3 case-control array. The test statistic is

$$X_H^2 = \frac{1}{2\hat{N}'} \Sigma \frac{(n_{i.} - n_{.i})^2}{\hat{P}_i} \text{ for } i = 1, 2, \text{ and } 3.$$

The test statistic X_H^2 summarizes deviations from marginal homogeneity and has an approximate χ^2 distribution with 2 *df* when the expected marginal frequencies are homogeneous. The χ^2 expression for testing homogeneity requires four independent estimated parameters yielding 2 *df* (6 − 4 = 2). For the Northern California Childhood Leukemia Study data, the χ^2 value is $X_H^2 = 0.479$ ($P = 0.787$).

Conditional Logistic Model

The additive conditional logistic model applied to genotype frequency data collected in a matched design yields estimates of the logarithms of the three ratios of discordant pairs (denoted b_i). When the genotypic frequencies are the same for both cases and controls, the model estimated ratio within all three kinds of discordant pairs is 1 ($b_1 = b_2 = b_3 = 0$). Furthermore, the additive model requires that $b_1 + b_3 = b_2$. For the CYP1A1/leukemia data, these estimated log-ratios are

$\hat{b}_1 = -0.170$ for AA/AG discordant pairs,
$\hat{b}_2 = 0.110$ for AA/GG discordant pairs, and
$\hat{b}_3 = 0.280$ for AG/GG discordant pairs.

The corresponding estimated odds ratios ($e^{\hat{b}_1}$) are 0.843, 1.116, and 1.323, respectively.

The estimated log-ratios are directly related to the expected cell counts generated by the quasi-independence model. In symbols, the estimates from the quasi-independence model give

$$\frac{\hat{F}_{21}}{\hat{F}_{12}} = e^{\hat{b}_1}, \quad \frac{\hat{F}_{31}}{\hat{F}_{13}} = e^{\hat{b}_2}, \quad \text{and} \quad \frac{\hat{F}_{32}}{\hat{F}_{23}} = e^{\hat{b}_3}$$

In other words, both models generate identical expected counts contained in a 3 × 3 case-control array (Table 2).

From another prospective, both models necessarily conform to the relationship that

$$D = \frac{F_{21}F_{13}F_{32}}{F_{12}F_{31}F_{23}} = e^{b_1 - b_2 + b_3} = 1.0$$

or

$$\begin{aligned}\log(D) &= [\log(F_{21}) + \log(F_{13}) + \log(F_{32})] \\ &\quad -[\log(F_{12}) + \log(F_{31}) + \log(F_{23})] \\ &= b_1 - b_2 + b_3 = 0.0.\end{aligned}$$

The requirement that $D = 1$ is essentially the definition of quasi-independence or model additivity (no interaction) within a 3 × 3 array.

The distribution of the estimated value

$$\begin{aligned}\log(\hat{D}) &= [\log(f_{21}) + \log(f_{13}) + \log(f_{32})] \\ &\quad -[\log(f_{12}) + \log(f_{31}) + \log(f_{23})]\end{aligned}$$

has estimated variance given by

$$\hat{v} = \Sigma \frac{1}{f_{ij}} \quad \text{for } i \neq j = 1, 2, \text{and } 3.$$

These two estimates provide a computationally easy and additional assessment of the correspondence between the observed and the expected counts generated by the quasi-independence or the additive conditional logistic models.

The estimate of $\log(\hat{D})$ and its estimated variance from the CYP1A1/leukemia data are $\log(\hat{D}) = -1.264$ and $\hat{v} = 2.332$ yielding the test statistic

$$X_{Q'}^2 = \frac{[\log(\hat{D})]^2}{\hat{v}} = \frac{(-1.264)^2}{2.332} = 0.685 (P = 0.408)$$

The value $X_{Q'}^2$ has an approximate χ^2 distribution with 1 *df* when the data (Table 1) randomly differ from the expected counts generated by the quasi-independence or the conditional logistic models (Table 2). In general, this χ^2 statistic ($X_{Q'}^2$) and the χ^2 statistic (X_Q^2) will be similar, particularly for case-control arrays with many discordant pairs.

It should also be noted that the likelihood ratio test based on the additive conditional logistic model addresses only the hypothesis that $b_1 = b_2 = b_3 = 0$ or the ratios of the corresponding discordant pairs are 1.0. The score likelihood ratio test statistic is identical to the previously described test for marginal homogeneity (X_H^2).

Discussion

A symmetrical case-control array of matched pairs data indicates that no association likely exists between case-control status and genotypic frequencies. When statistical evidence emerges of nonsymmetry, two issues become important (i.e., independence and marginal homogeneity).

That is, two reasons exist for a significant lack of symmetry: the failure of the genotypic frequencies to be independent (failure of the additive conditional logistic model to reflect the data) and the failure of the matched data to have the same case and control genotypic frequencies or both. These two sources of deviation for a symmetrical array are easily identified and indicate different dimensions of the association between genotypic frequencies and disease risk.

In general, the likelihood ratio statistic estimated from an additive conditional logistic model addresses only the issue of marginal homogeneity of the case-control array because an additive model explicitly requires the discordant matched pairs to be independent. That is, substantial differences in the ratios of discordant pairs can exist in a 3×3 case-control array with perfectly homogeneous marginal frequencies $X_H^2 = 0$ when the genotypic frequencies that determine the numbers of discordant pairs are not independent. Without an assessment of the quasi-independence model X_Q^2 or equivalently the consistency of the odds ratios ($b_1 + b_3 = b_2$), inferences from matched case-control data are potentially biased and even potentially misleading. As with statistical models in general, goodness-of-fit is a critical issue.

The interpretation of quasi-independence of two variables is not different in principle from the interpretation in most contingency tables. Quasi-independence becomes an issue when specific cell frequencies are truncated from consideration. In a matched pairs design, the frequencies on the diagonal cells of the case-control array (the concordant pairs) are not included in the analysis. Nevertheless, two variables are not independent (or not quasi-independent) when the occurrence of one changes the probability of the occurrence of the other. For example, case-control status and genotypic frequencies are not quasi-independent when $P(\text{AA} \mid \text{AG is a control})$ is not equal to $P(\text{AA case})$. A phenotype frequency among the matched pair cases will not be quasi-independent when, for example, cases and controls have differing racial compositions and the phenotypic frequencies under investigation differ among races. In fact, nonindependence potentially arise whenever the controls fail to be a random sample of the population from which the cases were selected.

The χ^2 test of symmetry is a simultaneous evaluation of both independence and homogeneity. Applied to the CYP1A1/leukemia data, this test produces no evidence of an association between genotypic frequencies and case-control status ($X_S^2 = 1.184$ with $P = 0.757$). It then becomes a foregone conclusion that χ^2 tests of quasi-independence and marginal homogeneity ($X_Q^2 = 0.712$ and $X_H^2 = 0.479$) consists of two nonsignificant pieces.

References

1. Ladonna MG, Izuierdo-Martinez M, Posada de la Paz M, et al. Pharmacogeneic profile of xenobiotic enzyme metabolism in survivors of the Spanish toxic oil syndrome. Environ Health Perspect 2001;109:369-75.
2. Ishibe N, Hankinson SE, Coditz GA, et al. Cigarette smoking, cytochrome ***P***450 1A1 polymorphism and breast cancer in the Nurses' Health Study. Cancer Res 1998;58:667-71.
3. Kim Jk, Lee CG, Park YG, et al. Combined analysis of germline polymorphisms of p53, GSTMI, GSTTI, CYP1A1, and CYP2E1: relation to the incidence rate of cervical carcinoma. Cancer 2000; 88:2082-91.
4. Krajiinovic M, Labuda D, Richer C, Karimi S, Simrett D. Susceptibility of childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTMI and GSTTI genetic polymorphism. Blood 1999;93:1496-501.
5. Infante-Rivard C, Krajinovic M, Labuda D, Sinnet D, Parental smoking, CYP1A1 genetic polymorphism and childhood leukemia. Cancer Causes & Control 2000;11:547-53.
6. Ma X, Buffler PA, Selvin S, Matthay KK, Wiencke JL, Reynolds P. Daycare attendance and risk of childhood acute lymphoblastic leukemia. Br J Cancer 2002;86:1419-24.
7. Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis: theory and practice. Cambridge (MA): MIT Press; 1975.
8. Freeman DH. Applied categorical data analysis. New York (NY): Marcel Dekker, Inc.; 1987.
9. Breslow NE, Day NE. Statistical methods in cancer research. Vol 1a. Lyon (France): IARC Scientific Publication; 1908. No. 32.
10. Mandansky A. Test of homogeneity for correlated samples. J Am Stat Assoc 1963;58:97-119.
11. Bhapkar VP. On tests of marginal symmetry and quasi-symmetry in two and three-dimensional contingency tables. Biometrics 1979;35: 417-26.
12. Stuart A. A test for homogeneity of the marginal distribution in a two-way classification. Biometrika 1955;42:412-6.
13. Bailey NTJ. Testing the solubility of maximum likelihood equations in the routine application of scoring methods Biometrics 1951;7: 268-74.
14. Bowker AH. A test for symmetry in contingency tables. J Am Stat Assoc 1948;42:572-4.