American Journal of Epidemiology

Vol. 154, No. 3
Printed in U.S.A.

New Estimator of the Genotype Risk Ratio for Use in Case-Parental Control Studies

W. Dana Flanders,[1] Fengzhu Sun,[2] and Quanhe Yang[3]

Estimation of the genotype risk ratio can be an important part of studying the role of genetics in disease causation. For example, one might estimate risk among persons with genotype *DD* compared with risk among those with genotype *Dd*, where the candidate locus has alleles *D* and *d*, with *D* representing the disease susceptibility allele. In this paper, the authors propose a modified method of analysis for case-parental control studies that can improve efficiency. They show how investigators can use information from families in which both parents are observed to improve the estimator created by Sun et al., which applies when only one parent and an affected offspring have been observed. Since this information is not used by the conditional approach of Schaid and Sommer, the authors' approach allows for more complete use of available information, leading to a smaller mean squared error of the genotype risk ratio estimators. The authors also suggest a way to combine estimates from families in which one parent and one offspring are observed and estimates from families in which both parents and one offspring are observed. *Am J Epidemiol* 2001;154:259–63.

case-control studies; epidemiologic methods; genetics; genotype; odds ratio

In studying the role of genetics in disease causation, an important goal is estimation of the genotype risk ratio comparing risk among persons with a particular genotype to risk among persons without that genotype. For example, if the candidate locus has alleles *D* and *d*, with *D* representing the susceptibility allele, this goal typically involves estimating risk among persons with genotype *DD* compared with risk among those with genotype *dd* (R_2), and perhaps also estimating risk among persons with genotype *Dd* compared with risk among those with genotype *dd* (R_1). Researchers can estimate these genotype risk ratios using a case-control study design, but it can be challenging to select appropriate control subjects so as to avoid bias and confounding, such as that due to population stratification.

Several new methods allow genetic epidemiologists to use observations of nuclear families to test for and estimate the association between alleles at a particular locus and disease status (1–5). The transmission disequilibrium test, which uses parents or siblings as control subjects, provides a way to test for an association between a candidate gene and a disease that does not tend to produce biased results due to population stratification (6). Schaid and Sommer (7) provided a method of analyzing observations of parents and an affected offspring to estimate the risk ratio using maximum likelihood techniques. Their method can remain valid even without the assumption of Hardy-Weinberg equilibrium by conditioning the likelihood on parental genotype (the "conditional on parental genotype" method). Weinberg (8) presented a likelihood-based method for estimating the association between a candidate gene and disease risk for families with one or two parents missing, but the method requires symmetrical mating probabilities. Sun et al. (9, 10) also provided a method of analysis for use when observations of only one parent and an affected offspring are available; their method requires neither the assumption of Hardy-Weinberg equilibrium nor the assumption of random mating.

We propose a modified method of analysis for case-parental control studies that can improve efficiency. First, we show that the genotype risk ratio estimator of Sun et al. (9), which applies when only one parent and an affected offspring have been observed, can be obtained as the parameter value which maximizes a particular (weighted) pseudolikelihood. Second, we show how one can use information from families in which both parents are observed to improve the estimator of Sun et al. (9). Since this information is not used in the same way by the conditional approach of Schaid and Sommer (7), our approach allows more complete use of available information, leading to a smaller mean squared error of the genotype risk ratio estimators. Finally, we suggest a way to combine estimates from families in which one parent and one offspring are observed and estimates from families in which both parents and one offspring are observed.

Received for publication January 20, 2000, and accepted for publication October 5, 2000.

Abbreviation: GAW, Genetic Analysis Workshop.

[1] Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA.
[2] Department of Mathematics, University of Southern California, Los Angeles, CA.
[3] National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, GA.
Reprint requests to Dr. W. Dana Flanders, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322 (e-mail: wflande@sph.emory.edu).

MATERIALS AND METHODS

We assume a case-parental control study in which we study a random sample of N new case subjects and their parents. For M of these cases, only the mother is available; for F of these cases, only the father is available; and for the remaining cases (N-M-F), both parents are available. We assume that the availability of parents of these case subjects does not depend on their genotype. We obtain the genotypes of case subjects and their parents at the locus of interest, with relevant data summarized as in table 1.

Motivated by the weighted approach of Sun et al. (9), we consider the pseudolikelihood:

$$L_1 = \frac{x_{01}^{W_{01}}\, x_{10}^{W_{10}}\, x_{11}^{W_{11}}\, x_{12}^{W_{12}}\, x_{21}^{W_{21}}}{[x_{01} + x_{10} + x_{11} + x_{12} + x_{21}]^T}, \tag{1}$$

where W_{ij} is the weighted average $W_{ij} = N \times (F \times M_{ij} + M \times F_{ij})$; $x_{01} = (p_{10} + p_{01} + p_{11})/2$; $x_{10} = R_1[(p_{10} + p_{01})/2 + (p_{20} + p_{02})]$; $x_{11} = R_1[(p_{10} + p_{01}) + 2p_{11} + (p_{21} + p_{12})]/2$; $x_{12} = R_1[(p_{20} + p_{02}) + (p_{21} + p_{12})/2]$; $x_{21} = R_2[p_{11} + (p_{21} + p_{12})]/2$; $T = W...$; p_{ij} is the proportion of matings in which the father and mother had (i,j) copies of allele D, respectively, for $i,j = 0, 1, 2$; and a dot in the subscript denotes summation over the corresponding index. Equation 1 would be a multinomial likelihood if the W_{ij}'s were actual counts rather than weighted counts. We give two justifications below after simplifying the expression. We can eliminate two parameters, because $x_{10} + x_{11} - x_{12} = 2R_1x_{01}$ and $x_{11} + x_{12} - x_{10} = 2(R_1/R_2)x_{21}$. Dividing numerator and denominator by x_{10}, we can eliminate one additional parameter, to obtain

$$L_1 = \frac{R_1^{W_{10}+W_{11}+W_{12}}(1 + a - b)^{W_{01}}\; a^{W_{11}}\, b^{W_{12}}\; R_2^{W_{21}}(a + b - 1)^{W_{21}}}{[(1 + a - b)/2 + R_1(1 + a + b) + R_2(a + b - 1)/2]^T}, \tag{2}$$

where $a = x_{11}/x_{10}$ and $b = x_{12}/x_{10}$. We offer two justifications for use of the partial derivatives of equation 2 as estimating equations. First, it is easily shown that the partial derivatives have expectation 0, making them unbiased estimating equations. Second, taking first partial derivatives and solving gives, as the solution for R_1 and R_2,

$$R_1 = (W_{11} + W_{10} - W_{12})/(2W_{01}).$$

$$R_2 = R_1(2W_{21})/(W_{11} + W_{12} - W_{10}). \tag{3}$$

These solutions are the same as the estimators previously derived by Sun et al. (9) through their consideration of cross-products of expected cell counts. They also showed that these estimators were consistent for R_1 and R_2.

Showing that the previously proposed estimators arise as the solutions of estimating equations has the following advantage: If we can estimate the other parameters in the estimating equation, perhaps using other data, then the estimation of R_1 and R_2 may become more stable. This approach is possible if we have information from families in which we have observed both parents. Thus, using observations of both parents as summarized in the third section of table 1, we consider the likelihood of B_{10}, B_{11}, B_{12}, and B_{13}, conditional on B_1:

$$L_2 = \frac{c^{B_{10}}\, d^{B_{11}}\, e^{B_{12}}\, f^{B_{13}}}{(c/2 + d/2 + e/2 + f)^{B_1}}, \tag{4}$$

where $c = (p_{10} + p_{01})/2$; $d = p_{11}$; $e = (p_{21} + p_{12})/2$; and $f = (p_{20} + p_{02})$. We can reparameterize in terms of a, b, and $s = (p_{02} + p_{20})/(p_{10} = p_{01})$ to obtain

$$L_2 = \frac{((a - b)(1 + 2s) + 2s - 1)^{B_{11}}\; (b(1 + 2s) - 2s)^{B_{12}}\; s^{B_{13}}}{(1/4 + (a + b)(1/4 + s/2) + s/2)^{B_1}}. \tag{5}$$

To combine information from families in which we observe only one parent (equation 3) with information from families in which we observe both parents to estimate a and b, we find the values of R_1, R_2, a, b, and s. which maximize L, where L is given by

$$L = L_1 \times L_2. \tag{6}$$

We estimate the covariance matrix of the parameter estimators, V, using the "sandwich" estimator:

$$V = dL^{-1} \times V_L \times dL^{-1}, \tag{7}$$

TABLE 1. Distribution of disease (*D*) alleles among case subjects and their parents in an estimation of genotype risk ratio

No. of *D* alleles in offspring	No. of *D* alleles in parent(s)			
	No. of *D* alleles in mother			
	0	1	2	
Only mother and offspring observed				
0	M_{00}	M_{01}		
1	M_{10}	M_{11}	M_{12}	
2		M_{21}	M_{22}	
	No. of *D* alleles in father			
	0	1	2	
Only father and offspring observed				
0	F_{00}	F_{01}		
1	F_{10}	F_{11}	F_{12}	
2		F_{21}	F_{22}	
	No. of *D* alleles in both parents (mother, father)			
	(0,1) or (1,0)	(1,1)	(1,2) or (2,1)	(0,2) or (2,0)
Both parents and offspring observed				
0	B_{00}	B_{01}		
1	B_{10}	B_{11}	B_{12}	B_{13}
2		B_{21}	B_{22}	

where dL^{-1} is the inverse of the matrix of second partial derivatives of L with respect to the parameters, evaluated at the estimated parameter values, and V_L is the empirical estimator of the covariance matrix. Entry i,j of V_L is given by $\Sigma\Sigma\partial L/\partial p_i \times \partial L/\partial p_j$ evaluated at the estimated parameter values, where the summation is over the N families and p_i denotes the ith parameter.

We use information from both parents in a different way to estimate R_1 and R_2 when maximizing L (equation 6) than when obtaining the maximum likelihood estimate of Schaid and Sommer (one difference, for example, is our use of homozygous parents (B_{13}), whereas these mating types are not used in the conditional approach of Schaid and Sommer) (7). Thus, a summary estimator which combines the new estimate with that of Schaid and Sommer (7) should be more stable than either estimator alone. This conjecture is supported by the results of Monte Carlo simulations below. Thus, our overall estimate is a weighted average, on the logarithmic scale, of the maximum likelihood estimate of Schaid and Sommer (7) and the new estimate of R_1 and R_2 obtained above that maximizes equation 7, with weights inversely proportional to the estimated variance.

RESULTS

To illustrate application of the new estimator, we first present results of Monte Carlo simulations in which we compare the mean squared error of the new estimator with the mean squared errors of other estimators. We then present results of analyses of data from Genetic Analysis Workshop 9 (GAW9) (11) and compare results and estimated standard deviations obtained from several different methods of analysis.

Monte Carlo simulations

In a series of Monte Carlo experiments, we compared the mean squared error of our new estimator with that of the Schaid and Sommer (7) estimator and the Sun et al. (9) estimator. We also evaluated the mean squared error of a summary estimator that was the weighted average of our new estimator and the Schaid and Sommer (7) estimator.

In each Monte Carlo experiment, we analyzed genotype information randomly generated for 1,000 families, each consisting of one affected offspring and one parent or one affected offspring and two parents. We generated observations using the SAS pseudorandom number generator (SAS Institute, Inc., Cary, North Carolina), using one of the four sets of parameters summarized in table 2, where p_{ij} is the proportion of mating in which the father and mother had (i,j) copies of allele D, respectively, in subpopulation 1; pM is the proportion of mothers in subpopulation 1 expected to be available for analysis; pF is the proportion of fathers in subpopulation 1 expected to be available for analysis; and pN is the expected proportion of all cases from subpopulation 1. The primed parameters indicate the corresponding values for subpopulation 2. These values represent several scenarios with different disease risks and with subpopulations that differ with respect to allele frequency and relative availability of mothers and fathers and their representation in the study. However, the availability of parents should be independent of their genotype.

Results of the simulations, shown in table 3, suggest the following pattern. For estimation of R_1, the mean squared error of the new estimator is approximately the same as that of the Sun et al. (9) estimator; the mean squared error of the

TABLE 2. Parameter sets used to generate Monte Carlo experiments in an estimation of genotype risk ratio

Parameter*	Set 1	Set 2	Set 3	Set 4
R_1	1.0	1.5	1.5	1.0
R_2	1.0	2.0	2.0	1.0
p_{00}, p'_{00}	0.40, 0.39	0.40, 0.39	0.44, 0.44	0.44, 0.44
p_{01}, p'_{01}	0.15, 0.19	0.15, 0.19	0.15, 0.14	0.15, 0.14
p_{10}, p'_{10}	0.21, 0.20	0.21, 0.20	0.21, 0.20	0.21, 0.20
p_{11}, p'_{11}	0.10, 0.09	0.10, 0.09	0.10, 0.09	0.10, 0.09
p_{02}, p'_{02}	0.020, 0.020	0.020, 0.020	0.020, 0.020	0.020, 0.020
p_{20}, p'_{20}	0.030, 0.026	0.030, 0.026	0.030, 0.026	0.030, 0.026
p_{12}, p'_{12}	0.025, 0.020	0.025, 0.020	0.025, 0.020	0.025, 0.020
p_{21}, p'_{21}	0.020, 0.030	0.020, 0.030	0.020, 0.030	0.020, 0.030
pM, pM'	0.40, 0.40	0.40, 0.40	0.20, 0.20	0.20, 0.20
pF, pF'	0.20, 0.20	0.20, 0.20	0.40, 0.40	0.40, 0.40
pN	0.33	0.33	0.33	0.33

* p_{ij} are the proportions of the mating in which the father and mother had (i,j) copies of allele D in the subpopulation; pM is the proportion of mothers in subpopulation 1 expected to be available for analysis; pF is the proportion of fathers in subpopulation 1 expected to be available for analysis; the primed parameters indicate the corresponding values for subpopulation 2; and pN is the expected proportion of all cases from subpopulation 1.

TABLE 3. Mean squared errors from Monte Carlo simulations in an estimation of genotype risk ratio

Parameter set	Estimator							
	Sun et al. (9)*		New†		Summary‡		Schaid and Sommer (7)§	
	R_1	R_2	R_1	R_2	R_1	R_2	R_1	R_2
1	0.027	0.242	0.025	0.119	0.012	0.060	0.023	0.102
2	0.030	0.186	0.028	0.086	0.014	0.049	0.024	0.079
3	0.030	0.221	0.033	0.126	0.014	0.049	0.026	0.077
4	0.031	0.353	0.030	0.193	0.013	0.064	0.023	0.103

* Given by equation 3 (maximizes equation 2).
† Maximizes equation 6.
‡ Weighted average of the new estimator and the Schaid and Sommer estimator.
§ Maximum likelihood estimator.

summary estimator which combines the new estimator with the Schaid and Sommer (7) estimator has substantially reduced mean squared error in comparison with that of all other estimators evaluated. Convergence was attained in 95 percent or more of the simulated data sets (when convergence was not attained, we used the Sun et al. (9) estimator in place of the new estimator).

The pattern is slightly different for estimation of R_2. The mean squared error of the new estimator is substantially lower than that of the Sun et al. (9) estimator, and the mean squared error of the summary estimator has a substantially lower mean squared error than the other estimators.

Example

We illustrate the approach using publicly available information from GAW9 (data on chromosome 1, marker 31, allele 8 (D1G31M8)) (11). The data we analyzed, summarized in table 4, were generated from those in GAW 9 by: 1) randomly selecting a subset of 200 families and treating them as though the father's genotype were unavailable (first section of table 4); 2) selecting another 200 families and treating them as though the mother's genotype were unavailable (second section of table 4); and 3) selecting 105 families and using information on the genotype of both parents (third section of table 4).

Analyses show that the estimated standard deviation of the new estimator and the standard deviation of the summary estimators are substantially smaller than the other estimators (table 5).

DISCUSSION

We have proposed a new estimator for analyses of case-parental control studies. In developing this estimator, we have shown that an estimator previously proposed by Sun et al. (9) for analysis of case-parental control studies in which only one parent is observed actually arises as the solution of unbiased estimating equations. We have also shown how to use additional information from observations of both parents to help estimate nuisance parameters that appear in these estimating equations. The new estimator, when combined as a weighted average with a maximum likelihood estimator used for analysis of studies in which both parents are observed (7), leads to a new summary estimator. Finally, we have evaluated the mean squared error of the new estimator and the new summary estimator and have found a pattern which suggests that their use can be associated with substantially reduced mean squared error.

A major advantage of the new estimators is that they can use the information more efficiently than previously proposed estimators when information is sometimes available

TABLE 4. Distribution of disease (*D*) alleles among case subjects and their parents in an estimation of genotype risk ratio using data from Genetic Analysis Workshop 9 (D1G31M8*)

No. of *D* alleles in offspring	No. of *D* alleles in parent(s)			
	No. of *D* alleles in mother			
	0	1	2	
Only mother and offspring observed				
0	$M_{00} = 135$	$M_{01} = 8$		
1	$M_{10} = 28$	$M_{11} = 21$	$M_{12} = 4$	
2		$M_{21} = 4$	$M_{22} = 0$	
	No. of *D* alleles in father			
	0	1	2	
Only father and offspring observed				
0	$F_{00} = 138$	$F_{01} = 5$		
1	$F_{10} = 23$	$F_{11} = 30$	$F_{12} = 0$	
2		$F_{21} = 4$	$F_{22} = 0$	
	No. of *D* alleles in both parents (mother, father)			
	(0,1) or (1,0)	(1,1)	(1,2) or (2,1)	(0,2) or (2,0)
Both parents and offspring observed				
0	$B_{00} = 8$	$B_{01} = 0$		
1	$B_{10} = 30$	$B_{11} = 1$	$B_{12} = 1$	$B_{13} = 2$
2		$B_{21} = 3$	$B_{22} = 0$	

* Chromosome 1, marker 31, allele 8.

TABLE 5. Results of analyses of data from Genetic Analysis Workshop 9 (D1G31M8*)

Parameter	Estimator							
	Sun et al. (9)		New		Combined		Schaid and Sommer (7)	
	RR†	SD†	RR	SD	RR	SD	RR	SD
R_1	3.8	0.30	3.8	0.29	3.8	0.25	3.9	0.39
R_2	14.3	2.5	10.4	0.78	10.8	0.54	11.4	1.0

* Chromosome 1, marker 31, allele 8.
† RR, relative risk; SD, estimated standard deviation.

for only one parent. As such, the new estimators should be useful for analysis of case-parental control studies. If a risk ratio estimate derived from other types of studies, such as case-control studies, is available, we can obtain a summary estimate as the weighted average of our new estimate and the other estimate. We have focused on estimation of genotype risk ratios, but our method leads naturally to a test for association. One simply divides the logarithm of the estimated risk ratio by the estimated standard error. Under the null hypothesis, this statistic should have an approximately standard normal distribution, and therefore it provides a method of testing for no association between the candidate gene and disease risk.

An important limitation of our approach and of many other approaches like ours is that we have not addressed the importance of age or other covariates. For conditions present at birth or for those which develop early in life, the limitation related to age might have little impact. However, for diseases with a late onset, this limitation could be important. For example, a particular allele that was associated with extended longevity might appear to be associated with increased risk simply because persons with that allele tended to live longer. These kinds of limitations should be addressed in further work, perhaps incorporating survival techniques and mixture models.

The more traditional case-control study design with population control subjects or with sibling control subjects is an attractive alternative to the parental control design. An important concern with population controls, however, is the possibility of confounding. For example, if allele frequencies differ by ethnic or population subgroup, and if disease risk also differs by ethnic or population subgroup, then use of population controls could lead to confounded estimates. Stratification in the analysis, of course, would reduce or eliminate this confounding, provided that one knew and had identified the appropriate subgroups on which to stratify. Alternatively, the degree of confounding might be minor if the differences in allele frequency or disease risk across subgroups were not too large (12). Nevertheless, the case-parental control design should be useful when the investigator suspects confounding by ethnic group or other subgroups of the population, and the newly proposed estimators can provide an improved method of analysis in this situation. A working version of a program that can be used to calculate our new estimator, modified from the version used for simulations, is available from the authors.

ACKNOWLEDGMENTS

Dr. Fengzhu Sun was partially supported by grant R29DK53392 from the National Institutes of Health.

REFERENCES

1. Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 1996;13:423–49.
2. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993; 52:506–16.
3. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 1996;59:983–9.
4. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–17.
5. Knapp M, Seuchter SA, Baur MP. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 1993;52:1085–93.
6. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 1995;57:455–64.
7. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 1993;53:1114–26.
8. Weinberg CR. Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 1999;64:1186–93.
9. Sun F, Flanders WD, Yang Q, et al. A new method for estimating the risk ratio in studies using case-parental control design. Am J Epidemiol 1998;148:902–9.
10. Sun F, Flanders WD, Yang Q, et al. Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. Am J Epidemiol 1999;150:97–104.
11. Speer MC, Terwilliger JD, Ott J. Data simulation for GAW9 problems 1 and 2. Genet Epidemiol 1995;12:561–4.
12. Flanders WD, Khoury MJ. Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. Epidemiology 1990;1:239–46.