

# The Estimation of Sibling Genetic Risk Parameters Revisited

Guohua Zou and Hongyu Zhao\*

*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut*

This report points out that some sibling genetic risk parameters can be regarded as the ratios of the characteristic values in the ascertainment subpopulation. Based on this observation, we reconsider Olson and Cordell's ([2000] *Genet. Epidemiol.* 18:217–235) and Cordell and Olson's ([2000] *Genet. Epidemiol.* 18:307–321) estimators, and re-derive these estimators. Furthermore, we provide the closed-form variance estimators. Simulation results suggest that our proposed estimators perform very well, and single ascertainment may be better than complete ascertainment for estimating these genetic parameters. © 2004 Wiley-Liss, Inc.

**Key words:** locus-specific relative risk; sibling recurrence risk; unbiased estimator; variance estimation

Grant sponsor: National Institutes of Health; Grant number: GM59507.

\*Correspondence to: Hongyu Zhao, Ph.D., Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

Received 12 August 2003; Accepted 11 December 2003

Published online 25 February 2004 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/gepi.10322

## INTRODUCTION

The sibling genetic risk parameters such as locus-specific sibling relative risk and sibling recurrence risk are basic quantities in genetic epidemiologic studies and are useful in genetic counseling. Usually, these parameters are estimated by the maximum likelihood (ML) method. Because there are no closed forms for ML estimators in general, the Expectation-Maximization (EM) algorithm [Dempster et al., 1977] is often used to obtain the ML estimates. Moreover, the corresponding variance estimators are complicated [Cordell and Olson, 1997, 2000]. Recently, Olson and Cordell [2000] proposed counting estimators for these sibling genetic risk parameters. Their estimators are of closed forms and hence are intuitive and can be calculated easily. The authors showed that their estimators are consistent. However, the corresponding variance estimators are either complicated or not provided. We note that the parameters of interest, locus-specific sibling relative risks and sibling recurrence risk, in fact depend only on the ascertainment subpopulation, which consists of all sibships with at least one affected siblings. Therefore, it is possible to approach their estimation problem by making use of standard sampling theory for finite populations [Cochran, 1977]. Based on this ob-

servations, we can give a simple derivation of Olson and Cordell's [2000] and Cordell and Olson's [2000] estimators and show some other merits of these estimators such as approximate unbiasedness except for consistency. Furthermore, we can obtain approximately unbiased variance estimators.

There are two commonly used ascertainment schemes: complete ascertainment and single ascertainment. A general question is which one should be used for estimating the specific genetic risk parameters. Based on the expressions of the variances for the estimators derived here, it is possible to compare different ascertainment schemes. In fact, our simulation results suggest that for estimating the sibling genetic risk parameters, single ascertainment may be better than complete ascertainment.

## METHODS

### THE ESTIMATION OF THE PROBABILITY THAT AN AFFECTED SIB PAIR SHARES $k$ ALLELES IBD

First we consider the estimation of the probability that an affected sib pair shares  $k$  alleles identical-by-descent (IBD) at the disease locus because it is an important quantity and closely

related to the locus-specific sibling relative risk. These probabilities are denoted by  $z_k$ , where  $k=0, 1, \text{ and } 2$ , in our following discussion.

Let  $N_a$  be the total number of sibships with  $a$  affected siblings,  $N_{kaj}$  be the total number of affected sib pairs sharing  $k$  alleles IBD in the  $j$ th sibship with  $a$  affected siblings in the total population. For the ascertained sample, let  $n_a$  denote the number of sibships with  $a$  affected siblings, and  $n_{kaj}$  denote the number of affected sib pairs sharing  $k$  alleles IBD in the  $j$ th sibship with  $a$  affected siblings in the sample. As in Olson and Cordell [2000], we consider a marker locus completely linked to the disease locus and assume that the marker is fully informative in the sense that the exact number of shared alleles can always be determined.

The key step for our treatment is to express  $z_k$  in the form of

$$z_k = \frac{\sum_{a=2}^{\infty} \sum_{j=1}^{N_a} N_{kaj}}{\sum_{a=2}^{\infty} \binom{a}{2} N_a}, \quad k = 0, 1, \text{ and } 2. \quad (1)$$

Note that the denominator in formula (1) means the total number of affected sib pairs in the population, and the numerator in (1) is the number of affected sib pairs sharing  $k$  alleles IBD. It is clear that  $z_k$  depends only on the ascertainment subpopulation  $A$ .

Now we consider the subpopulation  $A_{-1}$  of the ascertainment subpopulation  $A$ , which is formed by the sibships with at least two affected siblings. Let  $Y_i$  denote the number of affected sib pairs sharing  $k$  alleles IBD in the  $i$ th sibship of the subpopulation  $A_{-1}$ , and  $X_i = \binom{a_i}{2}$ , where  $a_i$  is the number of affected siblings in the  $i$ th sibship of  $A_{-1}$ . Then formula (1) can be written as

$$z_k = \frac{\sum_{i=1}^{N_{-1}} Y_i}{\sum_{i=1}^{N_{-1}} X_i}, \quad (2)$$

where  $N_{-1}$  is the size of subpopulation  $A_{-1}$  and equals  $\sum_{a=2}^{\infty} N_a$ .

Note that (2) is of the form of ratio for  $Y$ -value and  $X$ -value of the subpopulation  $A_{-1}$ . So for complete ascertainment, a natural estimator is

$$\hat{z}_k = \frac{\sum_{i=1}^{n_{-1}} y_i}{\sum_{i=1}^{n_{-1}} x_i} = \frac{\sum_{a=2}^{\infty} \sum_{j=1}^{n_a} n_{kaj}}{\sum_{a=2}^{\infty} \binom{a}{2} n_a}, \quad (3)$$

where  $n_{-1} = \sum_{a=2}^{\infty} n_a$  is the number of sibships with at least two affected siblings in the ascertained sample with size  $n$ .

The corresponding variance and its estimator are given by [Cochran, 1977],

$$V(\hat{z}_k) \approx \frac{1}{n_{-1}} \cdot \frac{1}{\bar{X}_{-1}^2} \frac{1}{N_{-1} - 1} \sum_{a=2}^{\infty} \sum_{j=1}^{N_a} \left[ N_{kaj} - z_k \binom{a}{2} \right]^2, \quad (4)$$

and

$$v(\hat{z}_k) = \frac{1}{n_{-1} \bar{x}_{-1}^2} \cdot \frac{1}{n_{-1} - 1} \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \left[ n_{kaj} - \hat{z}_k \binom{a}{2} \right]^2, \quad (5)$$

where  $\bar{X}_{-1} = \sum_{a=2}^{\infty} \binom{a}{2} N_a / N_{-1}$  is the mean of the subpopulation  $A_{-1}$  and  $\bar{x}_{-1} = \sum_{a=2}^{\infty} \binom{a}{2} n_a / n_{-1}$  is the corresponding sample mean. It should be noted that there is a slight difference between the case we consider here and the standard method given in Cochran [1977] since  $n_{-1}$  is a random variable, not a constant. Of course, if we draw sibships directly from the subpopulation  $A_{-1}$ , i.e., the ascertainment subpopulation is  $A_{-1}$  not  $A$ , then  $n_{-1}$  will be a constant. Using standard techniques in sampling theory for finite populations, we can show that both the target estimator  $\hat{z}_k$  and its variance estimator  $v(\hat{z}_k)$  are approximately unbiased.

Similar conclusions can be drawn for general ascertainment schemes. For example, for single ascertainment, the estimator of  $z_k$  is

$$\hat{z}_k = \frac{\sum_{i=1}^{n_{-1}} y_i / \pi_i}{\sum_{i=1}^{n_{-1}} x_i / \pi_i} = \frac{\sum_{a=2}^{\infty} \sum_{j=1}^{n_a} n_{kaj} / \pi_a}{\sum_{a=2}^{\infty} \binom{a}{2} n_a / \pi_a}, \quad (6)$$

where  $\pi_i$  is the probability that the  $i$ th sibship in the subpopulation  $A_{-1}$  is ascertained, which equals  $\pi_a$  ( $\propto a$ ) if the sibship has  $a$  affected siblings. Note that here the sibship is ascertained through affected individuals, which is a commonly used sampling scheme in practice and is called single ascertainment of individuals by Olson and Cordell [2000]. (If a sibship is ascertained through pairs of affected siblings, a scheme Olson and Cordell [2000] call single ascertainment of pairs, then  $n_{-1}=n$  and  $\pi_a \propto \binom{a}{2}$ ). The corresponding variance and its estimator are

given by

$$V(\hat{z}_k) \approx \frac{1}{n_{-1}} \cdot \frac{1}{\left[\sum_{a=2}^{\infty} \binom{a}{2} N_a\right]^2} \cdot \sum_{a=2}^{\infty} \sum_{j=1}^{N_a} \frac{[N_{kaj} - z_k \binom{a}{2}]^2}{\pi_a} \tag{7}$$

and

$$v(\hat{z}_k) = \frac{n_{-1}}{(n_{-1} - 1) \left(\sum_{a=2}^{\infty} \binom{a}{2} n_a / \pi_a\right)^2} \cdot \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \left[n_{kaj} - \hat{z}_k \binom{a}{2}\right]^2 / \pi_a^2 \tag{8}$$

Likewise, the estimators  $\hat{z}_k$  and  $v(\hat{z}_k)$  given in (6) and (8) are both approximately unbiased. Note that (3) and (6) are just the estimators proposed by Olson and Cordell [2000].

**THE ESTIMATION OF LOCUS-SPECIFIC RELATIVE RISKS**

It is well known that the probability that an affected sib pair shares  $k$  alleles IBD calculated above,  $z_k$ , can be used to obtain the locus-specific relative risk parameters  $\lambda_s$  (for sibling),  $\lambda_o$  (for offspring), and  $\lambda_m$  (for monozygotic twin). In fact, assuming a one-locus disease model, Risch [1987, 1990] showed the following relationship between  $z_k$  and the relative risk parameters:

$$\lambda_s = \frac{1}{4z_0}, \lambda_o = \frac{z_1}{2z_0}, \text{ and } \lambda_m = \frac{z_2}{z_0}.$$

Thus,  $\lambda_s$ ,  $\lambda_o$ , and  $\lambda_m$  can be naturally estimated by

$$\hat{\lambda}_s = \frac{1}{4\hat{z}_0}, \hat{\lambda}_o = \frac{\hat{z}_1}{2\hat{z}_0}, \text{ and } \hat{\lambda}_m = \frac{\hat{z}_2}{\hat{z}_0} \tag{9}$$

As shown by Cordell and Olson [2000] through simulations, even when the estimates of  $z_k$  are unbiased, the estimates  $\hat{\lambda}_s$ ,  $\hat{\lambda}_o$ , and  $\hat{\lambda}_m$  of  $\lambda_s$ ,  $\lambda_o$ , and  $\lambda_m$  may have large biases when the sample size is not large enough. The adjustment for these three estimators in the small-sample case can be obtained as in Cordell and Olson [2000],

$$\hat{\lambda}'_s = \frac{1}{4} \left[ \frac{1}{\hat{z}_0} - \frac{v(\hat{z}_0)}{\hat{z}_0^3} \right],$$

$$\hat{\lambda}'_o = \frac{1}{2} \left[ \frac{\hat{z}_1}{\hat{z}_0} + \frac{\text{cov}(\hat{z}_0, \hat{z}_1)}{\hat{z}_0^2} - \frac{v(\hat{z}_0)\hat{z}_1}{\hat{z}_0^3} \right],$$

and

$$\hat{\lambda}'_m = \frac{\hat{z}_2}{\hat{z}_0} - \frac{v(\hat{z}_0) + \text{cov}(\hat{z}_0, \hat{z}_1)}{\hat{z}_0^2} - \frac{v(\hat{z}_0)\hat{z}_2}{\hat{z}_0^3},$$

where  $v(\hat{z}_0)$  is given by formula (5) for complete ascertainment or formula (8) for single ascertainment, and  $\text{cov}(\hat{z}_0, \hat{z}_1)$  is the estimator of the covariance between  $\hat{z}_0$  and  $\hat{z}_1$ . Interestingly, using the approach here, the estimator of the covariance between  $\hat{z}_0$  and  $\hat{z}_1$  has a simple form:

$$\text{cov}(\hat{z}_0, \hat{z}_1) = \frac{1}{n_{-1}\bar{x}_{-1}^2} \cdot \frac{1}{n_{-1} - 1} \cdot \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \left[ n_{0aj} - \hat{z}_0 \binom{a}{2} \right] \cdot \left[ n_{1aj} - \hat{z}_1 \binom{a}{2} \right] \tag{10}$$

for complete ascertainment, and

$$\text{cov}(\hat{z}_0, \hat{z}_1) = \frac{n_{-1}}{(n_{-1} - 1) \left(\sum_{a=2}^{\infty} \binom{a}{2} n_a / \pi_a\right)^2} \cdot \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \left[ n_{0aj} - \hat{z}_0 \binom{a}{2} \right] \cdot \left[ n_{1aj} - \hat{z}_1 \binom{a}{2} \right] / \pi_a^2 \tag{11}$$

for single ascertainment. Formulas (10) and (11) can be obtained by using the fact that the covariance between  $\hat{z}_0$  and  $\hat{z}_1$  is equal to

$$\frac{1}{2} \cdot [V(\hat{z}_0 + \hat{z}_1) - V(\hat{z}_0) - V(\hat{z}_1)],$$

and  $\hat{z}_0$  and  $\hat{z}_1$  have the same denominator.

Note that we have given the closed form expressions for the variance and covariance estimators, our adjustment will be much easier to calculate than those given by Cordell and Olson [1997, 2000]. On the other hand, by aid of formula (1), we can express the relative risk parameters as

$$\lambda_s = \frac{1}{4} \cdot \frac{\sum_{a=2}^{\infty} \binom{a}{2} N_a}{\sum_{a=2}^{\infty} \sum_{j=1}^{N_a} N_{0aj}},$$

$$\lambda_o = \frac{1}{2} \cdot \frac{\sum_{a=2}^{\infty} \sum_{j=1}^{N_a} N_{1aj}}{\sum_{a=2}^{\infty} \sum_{j=1}^{N_a} N_{0aj}},$$

and

$$\lambda_m = \frac{\sum_{a=2}^{\infty} \sum_{j=1}^{N_a} N_{2aj}}{\sum_{a=2}^{\infty} \sum_{j=1}^{N_a} N_{0aj}}.$$

They are all of the form of ratio like (1). So these relative risk parameters can be estimated directly in the same fashion as that for the probability  $z_k$  instead of using the estimates of  $z_k$  like (9). For convenience to applied researchers, we provide

the corresponding estimators as follows: For complete ascertainment, the estimators are given by

$$\hat{\lambda}_s^* = \frac{1}{4} \cdot \frac{\bar{x}_{-1}}{\bar{y}_{-1}^{(0)}}, \hat{\lambda}_o^* = \frac{1}{2} \cdot \frac{\bar{y}_{-1}^{(1)}}{\bar{y}_{-1}^{(0)}}, \text{ and } \hat{\lambda}_m^* = \frac{\bar{y}_{-1}^{(2)}}{\bar{y}_{-1}^{(0)}},$$

where  $\bar{y}_{-1}^{(k)} = \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} n_{kaj} / n_{-1}$  is the sample mean,  $k=0, 1$ , and  $2$ . Although  $\hat{\lambda}_s^*$  and  $\hat{\lambda}_o^*$ ,  $\hat{\lambda}_o^*$  and  $\hat{\lambda}_m^*$  are actually equal (here we use the notation “\*” to show the distinction in expressions of the same estimators), respectively, the variance estimators for the former are easier to obtain. Furthermore, the adjustments for the estimators  $\hat{\lambda}_s^*$ ,  $\hat{\lambda}_o^*$ , and  $\hat{\lambda}_m^*$  in the small-sample case are simpler:

$$\hat{\lambda}_s^{*'} = \frac{1}{4} \left[ \frac{\bar{x}_{-1}}{\bar{y}_{-1}^{(0)}} + \frac{cov(\bar{y}_{-1}^{(0)}, \bar{x}_{-1})}{(\bar{y}_{-1}^{(0)})^2} - \frac{v(\bar{y}_{-1}^{(0)})\bar{x}_{-1}}{(\bar{y}_{-1}^{(0)})^3} \right],$$

$$\hat{\lambda}_o^{*'} = \frac{1}{2} \left[ \frac{\bar{y}_{-1}^{(1)}}{\bar{y}_{-1}^{(0)}} + \frac{cov(\bar{y}_{-1}^{(0)}, \bar{y}_{-1}^{(1)})}{(\bar{y}_{-1}^{(0)})^2} - \frac{v(\bar{y}_{-1}^{(0)})\bar{y}_{-1}^{(1)}}{(\bar{y}_{-1}^{(0)})^3} \right],$$

and

$$\begin{aligned} \hat{\lambda}_m^{*'} &= \frac{\bar{y}_{-1}^{(2)}}{\bar{y}_{-1}^{(0)}} \\ &+ \frac{cov(\bar{y}_{-1}^{(0)}, \bar{x}_{-1}) - v(\bar{y}_{-1}^{(0)}) - cov(\bar{y}_{-1}^{(0)}, \bar{y}_{-1}^{(1)})}{(\bar{y}_{-1}^{(0)})^2} \\ &- \frac{v(\bar{y}_{-1}^{(0)})\bar{y}_{-1}^{(2)}}{(\bar{y}_{-1}^{(0)})^3}, \end{aligned}$$

where

$$v(\bar{y}_{-1}^{(0)}) = \frac{1}{n_{-1}} \cdot \frac{1}{n_{-1} - 1} \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} (n_{0aj} - \bar{y}_{-1}^{(0)})^2,$$

$$\begin{aligned} cov(\bar{y}_{-1}^{(0)}, \bar{x}_{-1}) &= \frac{1}{n_{-1}} \cdot \frac{1}{n_{-1} - 1} \\ &\times \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} (n_{0aj} - \bar{y}_{-1}^{(0)}) \\ &\times \left( \binom{a}{2} - \bar{x}_{-1} \right), \end{aligned}$$

and

$$\begin{aligned} cov(\bar{y}_{-1}^{(0)}, \bar{y}_{-1}^{(1)}) &= \frac{1}{n_{-1}} \cdot \frac{1}{n_{-1} - 1} \\ &\times \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} (n_{0aj} - \bar{y}_{-1}^{(0)}) (n_{1aj} - \bar{y}_{-1}^{(1)}). \end{aligned}$$

For single ascertainment, the estimators are

$$\hat{\lambda}_s^{**} = \frac{1}{4} \cdot \frac{x_{-1}^*}{y_{-1}^{(0)*}}, \hat{\lambda}_o^{**} = \frac{1}{2} \cdot \frac{y_{-1}^{(1)*}}{y_{-1}^{(0)*}}, \text{ and } \hat{\lambda}_m^{**} = \frac{y_{-1}^{(2)*}}{y_{-1}^{(0)*}},$$

where

$$x_{-1}^* = \frac{1}{n_{-1}} \sum_{a=2}^{\infty} \frac{\binom{a}{2} n_a}{\pi_a},$$

and

$$y_{-1}^{(k)*} = \frac{1}{n_{-1}} \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \frac{n_{kaj}}{\pi_a}, \quad k = 0, 1, 2.$$

The adjustments for the estimators  $\hat{\lambda}_s^{**}$ ,  $\hat{\lambda}_o^{**}$ , and  $\hat{\lambda}_m^{**}$  in the small-sample case are given by

$$\hat{\lambda}_s^{**'} = \frac{1}{4} \left[ \frac{x_{-1}^*}{y_{-1}^{(0)*}} + \frac{cov(y_{-1}^{(0)*}, x_{-1}^*)}{(y_{-1}^{(0)*})^2} - \frac{v(y_{-1}^{(0)*})x_{-1}^*}{(y_{-1}^{(0)*})^3} \right],$$

$$\begin{aligned} \hat{\lambda}_o^{**'} &= \frac{1}{2} \left[ \frac{y_{-1}^{(1)*}}{y_{-1}^{(0)*}} + \frac{cov(y_{-1}^{(0)*}, y_{-1}^{(1)*})}{(y_{-1}^{(0)*})^2} \right. \\ &\quad \left. - \frac{v(y_{-1}^{(0)*})y_{-1}^{(1)*}}{(y_{-1}^{(0)*})^3} \right], \end{aligned}$$

and

$$\begin{aligned} \hat{\lambda}_m^{**'} &= \frac{y_{-1}^{(2)*}}{y_{-1}^{(0)*}} \\ &+ \frac{cov(y_{-1}^{(0)*}, x_{-1}^*) - v(y_{-1}^{(0)*}) - cov(y_{-1}^{(0)*}, y_{-1}^{(1)*})}{(y_{-1}^{(0)*})^2} \\ &- \frac{v(y_{-1}^{(0)*})y_{-1}^{(2)*}}{(y_{-1}^{(0)*})^3}, \end{aligned}$$

where

$$v(y_{-1}^{(0)*}) = \frac{1}{n_{-1}} \cdot \frac{1}{n_{-1} - 1} \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \left( \frac{n_{0aj}}{\pi_a} - y_{-1}^{(0)*} \right)^2,$$

$$\begin{aligned} cov(y_{-1}^{(0)*}, x_{-1}^*) &= \frac{1}{n_{-1}} \cdot \frac{1}{n_{-1} - 1} \\ &\times \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \left( \frac{n_{0aj}}{\pi_a} - y_{-1}^{(0)*} \right) \\ &\times \left( \frac{\binom{a}{2}}{\pi_a} - x_{-1}^* \right), \end{aligned}$$

and

$$\begin{aligned} cov(y_{-1}^{(0)*}, y_{-1}^{(1)*}) &= \frac{1}{n_{-1}} \cdot \frac{1}{n_{-1} - 1} \\ &\times \sum_{a=2}^{\infty} \sum_{j=1}^{n_a} \left( \frac{n_{0aj}}{\pi_a} - y_{-1}^{(0)*} \right) \left( \frac{n_{1aj}}{\pi_a} - y_{-1}^{(1)*} \right). \end{aligned}$$

**THE ESTIMATION OF SIBLING RECURRENCE RISK**

The sibling recurrence risk,  $K_s$ , can be defined in two equivalent ways [see, Olson and Cordell, 2000]:

1. The proportion of affecteds among all siblings of affecteds in a population;
2. The probability that a sibling of an affected individual is also affected.

First note that this parameter depends only on the ascertainment subpopulation  $A$ . To give our derivation for its estimator, we express the sibling recurrence risk in the form

$$K_s = \frac{\sum_{s=2}^{\infty} \sum_{a=1}^s a(a-1)N_{s(a)}}{\sum_{s=2}^{\infty} \sum_{a=1}^s a(s-1)N_{s(a)}}, \tag{12}$$

where  $N_{s(a)}$  is the number of sibships of size  $s$  with  $a$  affected siblings in the population. Expression (12) can be obtained by noting that for the sibship of size  $s$  with  $a$  affected siblings, the number of siblings of the affected individual is  $\binom{a}{1} \cdot (s-1)$ , and the number of affected siblings of the affected individual is  $\binom{a}{1} \cdot (a-1)$ .

Now we denote  $Y_i = a_i(a_i - 1)$ ,  $X_i = a_i(s_i - 1)$ ,  $i = 1, \dots, N_{A_1}$ , where  $s_i$  is the size of the  $i$ th sibship in the subpopulation  $A_1$  consisting of the sibships with size greater than one and at least one affected siblings in the population and  $a_i$  is the number of the affected siblings in it, and  $N_{A_1}$  is the size of  $A_1$  and equals  $\sum_{s=2}^{\infty} \sum_{a=1}^s N_{s(a)}$ . Then  $K_s$  given in formula (12) can be written as

$$K_s = \frac{\sum_{s=2}^{\infty} \sum_{a=1}^s \sum_{j=1}^{N_{s(a)}} a(a-1)}{\sum_{s=2}^{\infty} \sum_{a=1}^s \sum_{j=1}^{N_{s(a)}} a(s-1)} = \frac{\sum_{i=1}^{N_{A_1}} Y_i}{\sum_{i=1}^{N_{A_1}} X_i}.$$

Similarly, if the ascertainment probability of a sibship with  $a$  affected siblings in the subpopulation  $A_1$  is  $\pi_a$ , then an approximately unbiased estimator of  $K_s$  is given by

$$\begin{aligned} \hat{K}_s &= \frac{\sum_{s=2}^{\infty} \sum_{a=1}^s \sum_{j=1}^{n_{s(a)}} a(a-1)/\pi_a}{\sum_{s=2}^{\infty} \sum_{a=1}^s \sum_{j=1}^{n_{s(a)}} a(s-1)/\pi_a} \\ &= \frac{\sum_{s=2}^{\infty} \sum_{a=1}^s a(a-1)n_{s(a)}/\pi_a}{\sum_{s=2}^{\infty} \sum_{a=1}^s a(s-1)n_{s(a)}/\pi_a}, \end{aligned}$$

where  $n_{s(a)}$  is the number of sibships of size  $s$  with  $a$  affected siblings in the ascertained sample with size  $n$ .

The corresponding variance and its approximately unbiased estimator are given by,

$$\begin{aligned} v(\hat{K}_s) &\approx \frac{1}{n_1} \cdot \frac{1}{\left[ \sum_{s=2}^{\infty} \sum_{a=1}^s a(s-1)N_{s(a)} \right]^2} \\ &\cdot \sum_{s=2}^{\infty} \sum_{a=1}^s \frac{a^2 [(a-1) - K_s(s-1)]^2}{\pi_a} N_{s(a)}, \end{aligned}$$

and

$$\begin{aligned} v(\hat{K}_s) &= \frac{n_1}{\left[ \sum_{s=2}^{\infty} \sum_{a=1}^s a(s-1)n_{s(a)}/\pi_a \right]^2} \cdot \frac{1}{n_1 - 1} \\ &\times \sum_{s=2}^{\infty} \sum_{a=1}^s \frac{a^2 [(a-1) - \hat{K}_s(s-1)]^2}{\pi_a^2} n_{s(a)}, \end{aligned}$$

where  $n_1$  is the number of sibships with size greater than one and at least one affected siblings in the ascertained sample.

**RESULTS**

In this section, we conduct some simulations to calculate the variances and their estimates for estimating the probability that an affected sib pair shares  $k$  alleles IBD, locus-specific relative risk, and sibling recurrence risk. We generate a population with 100,000 sibships whose sizes are determined by virtue of the Poisson distribution with mean 3 and truncated at 1 and 6. A total of 200 sibships are drawn from the simulated population under complete and single ascertainment for dominant and recessive genetic models, respectively. We perform 200 simulations to calculate the averages of estimated parameters of interest and their standard deviations. We first estimate  $z_{kr}$ , the probability that an affected sib pair shares  $k$  alleles IBD using formula (3) under complete ascertainment and formula (6) under single ascertainment, and then calculate the corresponding variances and variance estimates

**TABLE I. The estimate of the probability that an affected sib pair shares  $k$  alleles IBD and its variance estimate under different ascertainment<sup>a</sup>**

| Genetic model        | Ascertainment scheme | $\hat{z}_0$ | True var. ( $\times 10^{-4}$ ) | Estimated var. ( $\times 10^{-4}$ ) | s.d. ( $\times 10^{-4}$ ) |
|----------------------|----------------------|-------------|--------------------------------|-------------------------------------|---------------------------|
| Dominant $z_0=0.081$ | Complete             | 0.080       | 1.42                           | 1.40                                | 0.02                      |
|                      | Single               | 0.080       | 1.21                           | 1.20                                | 0.01                      |
| Recessive $z_0=0.04$ | Complete             | 0.041       | 0.94                           | 0.96                                | 0.02                      |
|                      | Single               | 0.040       | 0.81                           | 0.81                                | 0.01                      |
| Genetic model        | Ascertainment scheme | $\hat{z}_1$ | True var. ( $\times 10^{-4}$ ) | Estimated var. ( $\times 10^{-4}$ ) | s.d. ( $\times 10^{-4}$ ) |
| Dominant $z_1=0.490$ | Complete             | 0.490       | 4.76                           | 4.69                                | 0.05                      |
|                      | Single               | 0.491       | 4.05                           | 4.05                                | 0.03                      |
| Recessive $z_1=0.32$ | Complete             | 0.322       | 5.33                           | 5.35                                | 0.06                      |
|                      | Single               | 0.319       | 4.60                           | 4.56                                | 0.03                      |
| Genetic model        | Ascertainment scheme | $\hat{z}_2$ | True var. ( $\times 10^{-4}$ ) | Estimated var. ( $\times 10^{-4}$ ) | s.d. ( $\times 10^{-4}$ ) |
| Dominant $z_2=0.427$ | Complete             | 0.430       | 4.66                           | 4.66                                | 0.05                      |
|                      | Single               | 0.428       | 3.96                           | 3.96                                | 0.03                      |
| Recessive $z_2=0.64$ | Complete             | 0.637       | 5.64                           | 5.68                                | 0.06                      |
|                      | Single               | 0.642       | 4.87                           | 4.84                                | 0.03                      |

<sup>a</sup>Dominant model: one-locus, fully penetrant, allele frequency=0.1. Recessive model: one-locus, fully penetrant, allele frequency=0.25. s.d.: Standard deviation of estimated variance.

**TABLE II. The estimate of locus-specific relative risk under different ascertainment<sup>a</sup>**

| Genetic model               | Ascertainment scheme | $\hat{\lambda}_s'$ | s.d.' | $\hat{\lambda}_s^{*'}'$ | s.d. <sup>*'</sup> |
|-----------------------------|----------------------|--------------------|-------|-------------------------|--------------------|
| Dominant $\lambda_s=3.08$   | Complete             | 3.111              | 0.033 | 3.111                   | 0.033              |
|                             | Single               | 3.104              | 0.029 | 3.104                   | 0.029              |
| Recessive $\lambda_s=6.25$  | Complete             | 6.108              | 0.103 | 6.107                   | 0.103              |
|                             | Single               | 6.290              | 0.105 | 6.290                   | 0.105              |
| Genetic model               | Ascertainment scheme | $\hat{\lambda}_o'$ | s.d.' | $\hat{\lambda}_o^{*'}'$ | s.d. <sup>*'</sup> |
| Dominant $\lambda_o=3.02$   | Complete             | 3.046              | 0.036 | 3.046                   | 0.036              |
|                             | Single               | 3.051              | 0.033 | 3.050                   | 0.033              |
| Recessive $\lambda_o=4.00$  | Complete             | 3.942              | 0.074 | 3.941                   | 0.074              |
|                             | Single               | 4.003              | 0.071 | 4.003                   | 0.071              |
| Genetic model               | Ascertainment scheme | $\hat{\lambda}_m'$ | s.d.' | $\hat{\lambda}_m^{*'}'$ | s.d. <sup>*'</sup> |
| Dominant $\lambda_m=5.26$   | Complete             | 5.351              | 0.067 | 5.352                   | 0.067              |
|                             | Single               | 5.315              | 0.056 | 5.315                   | 0.056              |
| Recessive $\lambda_m=16.00$ | Complete             | 15.547             | 0.275 | 15.545                  | 0.275              |
|                             | Single               | 16.155             | 0.287 | 16.154                  | 0.287              |

<sup>a</sup>Dominant model: one-locus, fully penetrant, allele frequency=0.1. Recessive model: one-locus, fully penetrant, allele frequency=0.25. s.d.': Standard deviations of  $\hat{\lambda}_s'$ ,  $\hat{\lambda}_o'$ , and  $\hat{\lambda}_m'$ ; s.d.<sup>\*'</sup>: standard deviations of  $\hat{\lambda}_s^{*'}'$ ,  $\hat{\lambda}_o^{*'}'$ , and  $\hat{\lambda}_m^{*'}'$ .

using formulas (4) and (5) for complete ascertainment and (7) and (8) for single ascertainment, respectively. The results are presented in Table I. It can be seen that the estimates of  $z_k$  where  $k=0, 1$ , and  $2$ , and their variance estimates are very accurate. The calculation results on the estimates of locus-specific relative risks and their standard deviations are given in Table II. Furthermore, we calculate the estimates of the sibling recurrence risk under the two ascertainment schemes. The values of these estimates and their variance

estimates are summarized in Table III. It is clear that our proposed estimators perform well.

Comparing the results under complete and single ascertainment, we see that the biases of the estimates of the parameters of interest and their variance estimates are all small under the two ascertainment schemes, especially single ascertainment. Also, the variances under complete ascertainment are usually greater than those under single ascertainment, especially for the estimation of sibling recurrence risk. This shows

**TABLE III. The estimate of sibling recurrence risk and its variance estimate under different ascertainment<sup>a</sup>**

| Genetic model         | Ascertainment scheme | $\hat{K}_s$ | True var. ( $\times 10^{-4}$ ) | Estimated var. ( $\times 10^{-4}$ ) | s.d. ( $\times 10^{-4}$ ) |
|-----------------------|----------------------|-------------|--------------------------------|-------------------------------------|---------------------------|
| Dominant $K_s=0.584$  | Complete             | 0.581       | 7.56                           | 7.56                                | 0.08                      |
|                       | Single               | 0.585       | 5.44                           | 5.46                                | 0.04                      |
| Recessive $K_s=0.391$ | Complete             | 0.386       | 10.41                          | 10.36                               | 0.18                      |
|                       | Single               | 0.392       | 5.86                           | 5.83                                | 0.04                      |

<sup>a</sup>Dominant model: one-locus, fully penetrant, allele frequency=0.1. Recessive model: one-locus, fully penetrant, allele frequency=0.25. s.d.: Standard deviation of estimated variance.

that for estimating the probability that an affected sib pair shares  $k$  alleles IBD, locus-specific relative risk, and sibling recurrence risk, using single ascertainment is better. Note that the premise of this conclusion is that the ascertainment scheme is known correctly.

## DISCUSSION

In this article, we have noted that some parameters of interest in genetic epidemiology such as locus-specific sibling relative risk, sibling recurrence risk, and the probability that an affected sib pair shares  $k$  alleles IBD, in fact, depend only on the ascertainment subpopulation. Correspondingly, we have given a re-derivation of Olson and Cordell's [2000] and Cordell and Olson's [2000] estimators by using standard sampling theory. Furthermore, the corresponding variances and their estimators are provided. Simulation studies show that these estimators are approximately unbiased and also suggest that single ascertainment is a better choice for estimating these genetic risk parameters. Naturally, as in Olson and Cordell [2000] and Cordell and Olson [2000], our methods require the knowledge of the sibship ascertainment scheme. It should be pointed out that for the estimations of the probability that an affected sib pair shares  $k$  alleles IBD and locus-specific sibling relative risk, we have assumed that the IBD sharing information can be exactly determined for each sib pair. This is, of course, often unrealistic in practice. However, our results can be readily extended to the case where we only have IBD sharing estimates. In fact, for this case, all the formulas except the expressions of the true variances (4) and (7) in the report hold if we replace the true numbers of IBD sharing by their estimates. Further, in the case where only IBD sharing estimates are available, formulas (4) and (7) should be changed to

$$V(\hat{z}_k) \approx \frac{1}{n_{-1}} \cdot \frac{1}{\bar{X}_{-1}^2} \frac{1}{N_{-1} - 1} \\ \times \sum_{a=2}^{\infty} \sum_{j=1}^{N_a} \left[ N_{kaj} - z_k \binom{a}{2} \right]^2 + \frac{1}{n_{-1}} \cdot \frac{\sigma_k^2}{\bar{X}_{-1}^2},$$

and

$$V(\hat{z}_k) \approx \frac{1}{n_{-1}} \cdot \frac{1}{\left[ \sum_{a=2}^{\infty} \binom{a}{2} N_a \right]^2} \cdot \\ \sum_{a=2}^{\infty} \sum_{j=1}^{N_a} \frac{[N_{kaj} - z_k \binom{a}{2}]^2}{\pi_a} \\ + \frac{1}{n_{-1}} \cdot \frac{\sigma_k^2}{\left[ \sum_{a=2}^{\infty} \binom{a}{2} N_a \right]^2} \sum_{a=2}^{\infty} \frac{N_a}{\pi_a},$$

respectively, where  $\sigma_k^2$  is the variance of the estimate of sharing  $k$  alleles IBD.

There may be some other parameters in genetic epidemiology that can be treated along the line of this report. An example is the estimation of the disease prevalence in the ascertainment subpopulation for which the estimation problem has been considered by Burton et al. [2000, 2002], Epstein et al. [2002], and Zou and Zhao (unpublished data, 2003) in the presence of hidden parameter heterogeneity.

## ACKNOWLEDGMENTS

The authors thank Professors Heather Cordell and Jane Olson and two reviewers for their constructive comments.

## REFERENCES

- Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC. 2000. Ascertainment adjustment: where does it take us? *Am J Hum Genet* 67:1505–1514.
- Burton PR, Palmer LJ, Keen KJ, Olson JM, Elston RC. 2002. Letter to the editor: Response to Epstein et al. *Am J Hum Genet* 71:441–442.
- Cochran WG. 1977. *Sampling techniques*, 3rd ed., New York: John Wiley & Sons.

- Cordell HJ, Olson JM. 1997. Confidence intervals for relative risk estimates from affected sib pair data. *Genet Epidemiol* 14: 593–598.
- Cordell HJ, Olson JM. 2000. Correcting for ascertainment bias of relative risk estimates obtained using affected-sib-pair linkage data. *Genet Epidemiol* 18:307–321.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via EM algorithm. *J R Stat Soc Ser B* 39:1–38.
- Epstein MP, Lin X, Boehnke M. 2002. Ascertainment-adjusted parameter estimates revisited. *Am J Hum Genet* 70:886–895.
- Olson JM, Cordell HJ. 2000. Ascertainment bias in the estimation of sibling genetic risk parameters. *Genet Epidemiol* 18:217–235.
- Risch N. 1987. Assessing the role of HLA-linked and unlinked determinants of disease. *Am J Hum Genet* 40:1–14.
- Risch N. 1990. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 46:222–228.