Am. J. Hum. Genet. 63:252–258, 1998

Inflation of Sibling Recurrence-Risk Ratio, Due to Ascertainment Bias and/or Overreporting

Sun-Wei Guo

Institute of Human Genetics and Division of Epidemiology, University of Minnesota, Minneapolis

Summary

One widely used measure of familial aggregation is the sibling recurrence-risk ratio, which is defined as the ratio of risk of disease manifestation, given that one's sibling is affected, as compared with the disease prevalence in the general population. Known as λ_S, it has been used extensively in the mapping of complex diseases. In this paper, I show that, for a fictitious disease that is strictly nongenetic and nonenvironmental, λ_S can be dramatically inflated because of misunderstanding of the original definition of λ_S, ascertainment bias, and overreporting. Therefore, for a disease of entirely environmental origin, the λ_S inflation due to ascertainment bias and/or overreporting is expected to be more prominent if the risk factor also is familially aggregated. This suggests that, like segregation analysis, the estimation of λ_S also is prone to ascertainment bias and should be performed with great care. This is particularly important if one uses λ_S for exclusion mapping, for discrimination between different genetic models, and for association studies, since these practices hinge tightly on an accurate estimation of λ_S.

Introduction

Familial aggregation of diseases is generally taken as evidence for the existence of a genetic etiologic mechanism, environmental factors common to family members, or a combination of both (e.g., see MacMahon 1978). An important goal for many genetic epidemiological studies is to demonstrate familial aggregation of a disease (Khoury et al. 1993). Once the evidence for familiality is well established, family history can be used to identify high-risk individuals. In addition, further analysis, such as segregation analyses, can be performed to delineate the genetic component and the mode of inheritance, which often are followed by linkage analysis to map the gene(s) responsible for the disease.

Since gene mapping, especially that based on genomewide screening, often is costly and time-consuming, it is vital, as a first step toward the understanding of the genetic mechanism underlying the disease of interest, to measure the familial aggregation as accurately as possible. One widely used measure for familial aggregation is the sibling recurrence-risk ratio, which is defined as the ratio of risk of disease manifestation, given that one's sibling is affected, compared with the disease prevalence in the general population. A significant deviation from unity in this measure suggests familial aggregation.

The idea of using the sibling recurrence-risk ratio to gauge familial aggregation or hereditary background of a disease can be traced back to Penrose (1953), and its use in the mapping of complex traits has been greatly extended by Risch's three seminal papers (Risch 1990*a*, 1990*b*, 1990*c*). Known as "λ_S," it now has been used extensively in discriminating between different genetic models underlying complex diseases such as non–insulin-dependent diabetes mellitus (Rich 1990), multiple sclerosis (Sadovnick 1994), and cleft lip with or without cleft palate (Farrall and Holder 1992) and in the mapping, inclusion, and exclusion of complex traits (e.g., see Kruglyak and Lander 1995; Risch and Merikangas 1996).

It is self-evident that "familial" is not necessarily "genetic" or "hereditary"; for example, scurvy, kuru, hepatitis B, and sudden infant death syndrome were once mistakenly thought to be hereditary because of their tendency to "run in families," but later this was found to be untrue. However, Khoury et al. (1988) demonstrate, on theoretical grounds, that, for a disease with a strong familial aggregation, environmental risk factors alone are *unlikely* to account for such strong aggregation, unless the presumed environmental risk factors are associated with enormous risk (which should be easy to detect in the first place).

The conclusion reached by Khoury et al. (1988), however, is based on three critical assumptions. First, human populations consist of exclusively nuclear families, each

Received February 10, 1998; accepted for publication May 14, 1998; electronically published June 19, 1998.

Address for correspondence and reprints: Dr. Sun-Wei Guo, Division of Epidemiology, School of Public Health, University of Minnesota, Minneapolis, MN 55454-1015. E-mail: swguo@med.umn.edu

0002-9297/98/6301-0035$02.00

with two offspring only; or, alternatively, for a nuclear family with more than two offspring, once the family has been ascertained, the affection status of a randomly selected sibling is determined and taken into the estimation of the recurrence-risk ratio. Second, sampling from the population is strictly random and is free of biases of any kind. Third, the affection status in relatives of an index subject can be determined without error.

In many genetic epidemiological studies, these three assumptions often are not satisfied, either individually or jointly. Sampling is, by choice and necessity, often anything but random. A sibship with multiple affected individuals may be more likely to enter into the sample. The affection statuses in relatives of the index subject typically are determined on the basis of his or her report, because of constraints in time and resources. Index subjects, if affected, may be more likely to be aware of the diagnosis in their relatives and may be more diligent in their search for additional cases. They also may be more likely to misinterpret, say, benign tumor as cancer, or angina as myocardial infarction, in their relatives. Further compounding the potential problems in the three assumptions discussed above is the misunderstanding and misuse of λ_S, as we shall see below.

In this paper, I will show that, for a fictitious disease that is strictly nongenetic and nonenvironmental, λ_S can be dramatically inflated because of misunderstanding of the original definition of λ_S, ascertainment bias, and overreporting. Therefore, for a disease of entirely environmental origin, the λ_S inflation due to ascertainment bias and/or overreporting is expected to be more prominent if the risk factor also is familially aggregated. This suggests that, like segregation analysis, the estimation of λ_S also is prone to ascertainment bias and should be performed with great care. This is particularly important if we use λ_S for exclusion mapping, for discriminating between different genetic models, and for association studies.

Methods

Consider a disease with population frequency p. Suppose that each individual in the population is equally liable to succumb to the disease and that the disease can be detected at any time after birth (i.e., no variable age at onset). Suppose also that there is no birth-order effect, no cohort effect, and no sex difference. In addition, suppose that the disease or trait has nothing to do with genes or environmental risk factors. Thus, the affection status of any individual is entirely independent of those of others, including his siblings. Conceptually, one can view the affection status of any individual as being determined, effectively, by flipping a biased coin with a head probability of p.

If D denotes that an individual is affected with the disease, the sibling recurrence-risk ratio is defined as:

$$\lambda_S = \frac{P(\text{sib2 } D|\text{sib1 } D)}{p}$$

(Khoury et al. 1988; Risch 1990*a*). If X_i denotes the affection status of sib i ($i = 1,2$), with $X_i = 1$ if sib i is affected or $X_i = 0$ if sib i not, then the definitions given above can written more compactly, as

$$\lambda_S = \frac{P(X_2 = 1|X_1 = 1)}{p} . \qquad (1)$$

For the fictitious disease that we are considering, λ_S should be unity if the three critical assumptions of Khoury et al. (1988) are invoked.

In practice, however, any population, at the nuclear-family level, is composed of families having zero, one, two, or more offspring, and the distribution of sibship size can be empirically estimated from, say, census data. Clearly, for any genetic epidemiological study, any ascertained family with zero or one child would have to be discarded for the purpose of estimating λ_S. For families with exactly two offspring, the estimation of λ_S can be made accurately as long as these families are both randomly ascertained and free of overreporting and the designation of "sib1" or "sib2" is independent of their affection statuses.

For families with k ($k > 2$) offspring, however, the definition given above is somewhat ambiguous. Which sib should be designated as "sib1" or as "sib2"? Which sib should be included if there are multiple affected sibs? Without careful consideration of these issues, the resultant estimates would be meaningless.

A more serious problem is that sibships of different sizes are usually ascertained with different probabilities. In particular, sibships with a greater number of affected individuals usually have a higher probability of being ascertained, simply because they are more conspicuous or easier to ascertain or, in the case of a cross-sectional survey, because of their eagerness to participate or to seek medical attention. Worse yet, the event "sib1 D" is sometimes taken as the index case and the event "sib2 D" is taken as "at least one affected among the rest of the sibship," or the other way around.

In many genetic epidemiological studies, one typical procedure is to take a group of individuals known to have the disease or trait in question and to determine the frequency of its occurrence among relatives within a specified degree of kinship—in particular, siblings—as Haenszel observed as early as 1959 (Haenszel 1959). This is the case, for example, in two genetic epidemiological studies of homosexuality (Pillard and Weinrich 1986; Bailey and Benishay 1993), in which homosexual

subjects were recruited through advertisements and were used as index subjects and in which the sexual orientations of their same-sex siblings subsequently were determined. Thus, in effect, definition (1) becomes

$$\lambda_S^* = \frac{P\left(\sum_{i=2}^{k} X_i \geqslant 1 \middle| X_1 = 1, \mathcal{A}\right)}{p}, \qquad (2)$$

where sib1 is designated as an index individual, $\mathcal{A}$ denotes the event that this sibship has been ascertained, and the event $\sum_{i=2}^{k} X_i \geqslant 1$ denotes that there is at least one affected sibling of sib1.

On the other hand, some studies, especially cross-sectional surveys, determine the affection status of the index subject through medical examinations and estimate the disease prevalence in the index subjects, given that one or more siblings are affected (e.g., see Monroe et al. 1995; Narod et al. 1995). Consequently, definition (1) effectively becomes

$$\lambda_S^{**} = \frac{P\left(X_1 = 1 \middle| \sum_{i=2}^{k} X_i \geqslant 1, \mathcal{A}\right)}{p}. \qquad (3)$$

It should be noted that definitions (2) and (3) do not agree with definition (1), even if $k = 2$ (for two-offspring families), because of potential ascertainment bias. For $k > 2$, definitions (2) and (3) for sibling recurrence-risk ratio are not, even without ascertainment bias, intended in the original definition of λ_S (Risch 1990*a*), but they nonetheless have been used in some genetic epidemiological studies.

Impact of Ascertainment Bias on λ_S

Although there are countless ways to ascertain a sibship, I consider the scenario in which case there is a constant probability π that any affected individual becomes an index case and that ascertainments of different affected individuals are assumed to be independent (Bailey 1951). Under this scheme, which is termed "multiple ascertainment" (Ewens 1991), the probability that a sibship with r affected individuals contains at least one index case and thus enters the sample is $1 - (1 - \pi)^r$. In other words, $P(\mathcal{A} | \sum_{i=1}^{k} X_i = r) = 1 - (1 - \pi)^r$.

Now, if $\delta = 1 - \pi$ and $q = 1 - p$,

$$\begin{aligned}
\lambda_S^* &= \frac{1}{p} P\left(\sum_{i=2}^{k} X_i \geqslant 1 \middle| X_1 = 1, \mathcal{A}\right) \\
&= \frac{1}{p} \frac{P\left(X_1 = 1, \sum_{i=2}^{k} X_i \geqslant 1, \mathcal{A}\right)}{P(X_1 = 1, \mathcal{A})} \\
&= \frac{1}{p} \frac{\sum_{j=1}^{k-1} P\left(X_1 = 1, \sum_{i=2}^{k} X_i = j, \mathcal{A}\right)}{\sum_{j=0}^{k-1} P\left(X_1 = 1, \sum_{i=2}^{k} X_i = j, \mathcal{A}\right)} \\
&= \frac{1}{p} \frac{\sum_{j=1}^{k-1} P\left(\mathcal{A} | X_1 = 1, \sum_{i=2}^{k} X_i = j\right) P\left(X_1 = 1, \sum_{i=2}^{k} X_i = j\right)}{\sum_{j=0}^{k-1} P\left(\mathcal{A} | X_1 = 1, \sum_{i=2}^{k} X_i = j\right) P\left(X_1 = 1, \sum_{i=2}^{k} X_i = j\right)} \\
&= \frac{1}{p} \frac{\sum_{j=1}^{k-1} (1 - \delta^{j+1}) p \binom{k-1}{j} p^j q^{k-1-j}}{\sum_{j=0}^{k-1} (1 - \delta^{j+1}) p \binom{k-1}{j} p^j q^{k-1-j}} \\
&= \frac{1 - \pi q^{k-1} - (1 - \pi)(1 - p\pi)^{k-1}}{p[1 - (1 - \pi)(1 - p\pi)^{k-1}]}.
\end{aligned}$$

Similarly, if definition (3) is used, it can be shown that

$$\begin{aligned}
\lambda_S^{**} &= \frac{1}{p} P\left(X_1 = 1 \middle| \sum_{i=2}^{k} X_i \geqslant 1, \mathcal{A}\right) \\
&= \frac{1 - \pi q^{k-1} - (1 - \pi)(1 - p\pi)^{k-1}}{1 - (1 - p\pi)^k - \pi p q^{k-1}}.
\end{aligned}$$

Since the sample size of many genetic epidemiological studies is small relative to the size of the population from which the sample is drawn, and since the frequency of most diseases is not high (say, <10%), π is usually small, unless all cases are completely ascertained. Consequently, the ascertainment probabilities for a sibship with 0,1, . . . ,k affected individuals are proportional to $0{:}\pi{:}[1 - (1 - \pi)^2]{:}\cdots{:}[1 - (1 - \pi)^k]$, which approach $0{:}1{:}2{:}\cdots{:}k$ as $\pi \to 0$. In human genetics, this special case is referred to as "single ascertainment" (Ewens 1991), because there is only one index case in each ascertained sibship.

Under single ascertainment, since

$$\begin{aligned}
&P\left(\sum_{i=2}^{k} X_i \geqslant 1 \middle| X_1 = 1, \mathcal{A}\right) \\
&= \frac{P\left(X_1 = 1, \sum_{i=2}^{k} X_i \geqslant 1, \mathcal{A}\right)}{P(X_1 = 1, \mathcal{A})} \\
&= \frac{p^2(k + q + q^2 + \cdots + q^{k-2})}{p^2(k + q + q^2 + \cdots + q^{k-2}) + pq^{k-1}};
\end{aligned}$$

then,

$$\lambda_S^* = \frac{k + q + \cdots + q^{k-2}}{k - (k-1)q} \ .$$

If definition (4) is used, it can be shown that

$$\lambda_S^{**} = \frac{k + q + \cdots + q^{k-2}}{k - q^{k-1}} \ .$$

It is easy to see that both λ_S^* and λ_S^{**} are >1, regardless of the value of p.

Impact of Ascertainment Bias and/or Overreporting Error on λ_S

In many epidemiological studies, the affection statuses in relatives of an index subject are usually determined by his or her report. It is likely that the subject would overreport the affection status in his or her siblings. For example, angina could be mistakenly construed as myocardial infarction. This overreporting is more likely to happen if adult siblings live in different geographic areas and/or if the disease is not severe (J. T. Bensen, A. D. Liese, J. T. Rushing, M. Province, A. R. Folsom, and D. Arnett, personal communication).

Now, suppose that, for the siblings of the index subject, their affection status is denoted as Y_i, $(i = 2,...,k)$ and that the overreporting, or false-positive, rate is α; that is, $P(Y_i = 1|X_i = 0) = \alpha$. Suppose also that the affection status of the index subject is always determined without error. This happens when the index subjects can be examined thoroughly or are known to be affected. Moreover, the ascertainment scheme is such that the probability that a sibship is ascertained depends only on the number of apparently affected individuals. In particular, I continue to consider the scheme in which

$$P\left(\mathscr{A}\Big|Y_1 = 1, \sum_{i=2}^{k} Y_i = j\right) \propto (j+1) \ .$$

Hence,

$$\begin{aligned}
&P\left(\sum_{i=2}^{k} Y_i = j\right) \\
&= \sum_{l=0}^{j} P\left(\sum_{i=2}^{k} X_i = l\right) P\left(\sum_{i=2}^{k} Y_i = j \Big| \sum_{i=2}^{k} X_i = l\right) \\
&= \sum_{l=0}^{j} \binom{k-1}{l} p^l q^{k-1-l} \binom{k-1-l}{j-l} \alpha^{j-l} (1-\alpha)^{k-1-j} \\
&= \binom{k-1}{j} q^{k-1-j} \alpha^j (1-\alpha)^{k-1-j} \sum_{l=0}^{j} \binom{j}{l} \left(\frac{p}{\alpha}\right)^l q^{j-l} \\
&= \binom{k-1}{j} (p + q\alpha)^j [q(1-\alpha)]^{k-1-j} \ .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&P\left(\sum_{i=2}^{k} Y_i \geq 1 \Big| Y_1 = 1, \mathscr{A}\right) \\
&= \frac{\sum_{j=1}^{k-1} P\left(Y_1 = 1, \sum_{i=2}^{k} Y_i = j, \mathscr{A}\right)}{\sum_{j=0}^{k-1} P\left(Y_1 = 1, \sum_{i=2}^{k} Y_i = j, \mathscr{A}\right)} \\
&= \frac{\sum_{j=1}^{k-1} (j+1) \binom{k-1}{j} (p + q\alpha)^j [q(1-\alpha)]^{k-1-j}}{\sum_{j=0}^{k-1} (j+1) \binom{k-1}{j} (p + q\alpha)^j [q(1-\alpha)]^{k-1-j}} \ .
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\lambda_S^* &= \frac{1}{p} P\left(\sum_{i=2}^{k} Y_i \geqslant 1 \Big| Y_1 = 1, \mathscr{A}\right) \\
&= \frac{(k-1)(p + q\alpha) + 1 - q^{k-1}(1-\alpha)^{k-1}}{p[(k-1)(p + q\alpha) + 1]} \ .
\end{aligned}$$

When $\alpha = 0$, this equation reduces to

$$\lambda_S^* = \frac{(k-1)p + 1 - q^{k-1}}{p[(k-1)p + 1]} \ ,$$

which agrees with definition (4). By means of definition (3), it can be shown that

$$\begin{aligned}
\lambda_S^{**} &= \frac{1}{p} P\left(Y_1 = 1 \Big| \sum_{i=2}^{k} Y_i \geq 1, \mathscr{A}\right) \\
&= \frac{(k-1)(p + q\alpha) + 1 - q^{k-1}(1-\alpha)^{k-1}}{(k-1)(p + q\alpha) + p[1 - q^{k-1}(1-\alpha)^{k-1}]} \ .
\end{aligned}$$

Since $q(1-\alpha) < 1$, λ_S^{**} is always >1.

Results

Table 1 shows λ_S^* and λ_S^{**} under the multiple-ascertainment scheme when the ascertainment probability is $\pi = .05$. As expected, under this ascertainment scheme, $\lambda_S^* = \lambda_S^{**}$ if the sibship size is 2 but is $\lambda_S^* > \lambda_S^{**}$ if the sibship size is >2. In view of the independence in the affection status, the probability that at least one sibling of the index subject is affected increases with the sibship size, given that the index subject is affected and that the sibship is ascertained. Thus λ_S^* increases as the sibship size increases. On the other hand, the probability that the index subject is affected decreases with the sibship size, given that at least one sibling of the rest of the sibship is affected and that the sibship is ascertained.

In general, the inflation of either λ_S^* or λ_S^{**} becomes

Table 1

Sibling Recurrence-Risk Ratios When Ascertainment Bias Is Introduced

	Value for Sibship Size of				
p and Recurrence Risk	2	3	4	5	6
.30:					
λ_S^*	1.52	2.29	2.71	2.95	3.10
λ_S^{**}	1.52	1.45	1.40	1.35	1.32
.20:					
λ_S^*	1.64	2.68	3.36	3.83	4.15
λ_S^{**}	1.64	1.58	1.53	1.49	1.45
.10:					
λ_S^*	1.78	3.19	4.32	5.24	5.98
λ_S^{**}	1.78	1.74	1.71	1.68	1.65
.05:					
λ_S^*	1.86	3.51	4.99	6.30	7.48
λ_S^{**}	1.86	1.84	1.82	1.80	1.78
.03:					
λ_S^*	1.90	3.66	5.30	6.84	8.27
λ_S^{**}	1.90	1.88	1.87	1.86	1.85
.01:					
λ_S^*	1.93	3.82	5.66	7.45	9.21
λ_S^{**}	1.93	1.93	1.92	1.92	1.91

Note.–The ascertainment probability is assumed to be $\pi = .05$.

more prominent as the disease frequency (p) decreases. The magnitude of the inflation can be quite large, especially for λ_S^*, when sibship size is moderate and p is small.

Table 2 lists λ_S^* and λ_S^{**} under the single-ascertainment scheme, and the inflation is more prominent. The results are similar to those of the multiple-ascertainment scheme with $\pi = .05$. It is interesting to note that, even for $k = 2$, the inflation can be substantial.

Table 3 shows the effect of ascertainment bias and/or overreporting. Compared with λ_S^{**}, λ_S^* is very sensitive to overreporting bias, and the effect of overreporting can be quite dramatic, especially for smaller p. For example, for a false-positive rate of only 5%, λ_S^* increased by fivefold for a sibship size of 2. Interestingly, λ_S^{**} is not very sensitive to overreporting. In fact, its value decreases slightly if there is overreporting.

Discussion

It should be pointed out that there is nothing wrong in the definition of the sibling recurrence-risk ratio. In fact, it is a splendid idea—a significant deviation from unity can be "plausibly interpreted as indicating that the trait in question has some hereditary background" (Penrose 1953, p. 257), especially in the absence of evidence that there is an environmental factor. It also has the appeal of both quantifying the genetic effect without knowing exactly the mode of inheritance and having the apparent capability to discriminate between different genetic models if certain assumptions are satisfied (Risch 1990*a*); and it appears to be measurable, if done with care. The flip side is that it can easily be misused, as has been shown above, and is quite sensitive to ascertainment bias and/or overreporting.

Accurate measurement of the sibling recurrence-risk ratio, a trivially easy task in experimental species, can be difficult for human geneticists, since random sampling can be either too expensive or impractical, and since people do have biases in recalling the events that have happened in their lives. Although work on ascertainment bias in human genetics has a long history (for an excellent review of this subject, see Ewens 1991), dating from Galton (1904), Weinberg (1912), and Fisher (1934), we still find genetic epidemiological reports being published in which the problem is altogether ignored. This is unfortunate, since the usefulness of many statistical methods for mapping—especially for exclusion mapping—hinges tightly on an accurate estimation of the measure.

In this paper, I have considered a fictitious disease that has neither a genetic component nor an environmental component. In fact, the manifestation of the disease is entirely stochastic: the results are similar to those produced by the tossing of a biased coin. Obviously, such a disease, if one exists, would be very rare. Most diseases, especially the chronic ones, have known or suspected risk factors of environmental origin. If this is the case,

Table 2

Sibling Recurrence-Risk Ratios under the Single-Ascertainment Scheme

	Value for Sibship Size of				
p and Recurrence Risk	2	3	4	5	6
.30:					
λ_S^*	1.54	2.31	2.73	2.97	3.11
λ_S^{**}	1.54	1.47	1.42	1.37	1.33
.20:					
λ_S^*	1.67	2.71	3.40	3.86	4.18
λ_S^{**}	1.67	1.61	1.56	1.51	1.47
.10:					
λ_S^*	1.82	3.25	4.39	5.31	6.06
λ_S^{**}	1.82	1.78	1.75	1.71	1.68
.05:					
λ_S^*	1.90	3.59	5.09	6.42	7.62
λ_S^{**}	1.90	1.88	1.86	1.84	1.82
.03:					
λ_S^*	1.94	3.75	5.42	6.99	8.44
λ_S^{**}	1.94	1.93	1.91	1.90	1.89
.01:					
λ_S^*	1.98	3.91	5.80	7.63	9.43
λ_S^{**}	1.98	1.98	1.97	1.97	1.96

Table 3

Effect of Ascertainment Bias and/or Overreporting, on Sibling Recurrence-Risk Ratio, under the Single-Ascertainment Scheme

	Value for Sibship Size of				
p and Recurrence Risk	2	3	4	5	6
.30:					
λ_S^*	1.67	2.45	2.84	3.05	3.17
λ_S^{**}	1.54	1.47	1.41	1.36	1.31
.20:					
λ_S^*	1.94	3.05	3.72	4.15	4.42
λ_S^{**}	1.67	1.60	1.54	1.49	1.44
.10:					
λ_S^*	2.53	4.33	5.64	6.62	7.35
λ_S^{**}	1.82	1.76	1.71	1.67	1.63
.05:					
λ_S^*	3.55	6.37	8.63	10.45	11.95
λ_S^{**}	1.90	1.86	1.82	1.79	1.75
.03:					
λ_S^*	4.85	8.87	12.22	15.04	17.43
λ_S^{**}	1.94	1.91	1.87	1.84	1.81
.01:					
λ_S^*	11.23	20.95	29.41	36.80	43.29
λ_S^{**}	1.98	1.95	1.92	1.90	1.87

Note.—The false-positive error is assumed to be $\alpha = .05$.

then λ_S^* and λ_S^{**} would be measurably higher, as can be seen from the following numerical example.

Suppose that the population consists only of two-sibling nuclear families. Suppose also that $p = .001$ and that the frequency of exposure to a risk factor, with a relative risk of 2, is .35. Suppose further that the familial correlation in risk exposure is .75 and that the probabilities for ascertaining the families—$\{X_1 = 0, X_2 = 0\}$, $\{X_1 = 0, X_2 = 1\}$, $\{X_1 = 1, X_2 = 0\}$, and $\{X_1 = 1, X_2 = 1\}$—are .125, .125, .25, and .50, respectively. Then, $\lambda_S^* = 4.26$ and $\lambda_S^{**} = 2.18$, when there is no overreporting, and $\lambda_S^* = 97.12$ and $\lambda_S^{**} = 3.99$, when the rate of overreporting is 5%, both of which are substantially greater than the 1.09 reported by Khoury et al. (1988, fig. 2 *A*, p. 680). Thus, for a disease with known or yet to be identified risk factor(s) of environmental origin, the effect of ascertainment bias and/or overreporting would be more dramatic.

This paper has considered only one or two ascertainment schemes. There are, of course, many other plausible schemes. Once the scheme is known, adjustment should be easy to make. In reality, however, the exact ascertainment scheme may not even be clear or known to the investigator. For example, in a prostate cancer study, 26,781 men age $\geqslant 45$ years initially were selected from an electoral list and were sent a written invitation to participate in the study. Of this group, only 27.2% indicated their willingness to participate (Narod et al. 1995). It is difficult to determine exactly what motivated these men to participate; it is even more difficult to determine exactly what motivated others not to do so, in order to devise an adjustment for ascertainment bias.

Recall bias in epidemiological studies is common (Raphael 1987; Floderus et al. 1990). Overreporting in genetic epidemiological studies also is quite common (Enterline and Capt 1959; Hastrup et al. 1985; Herrmann 1985; Hunt et al. 1986; Kee et al. 1993). One well-documented example of such overreporting or recall bias is a case-control study of rheumatoid arthritis, in which affected individuals are more likely than their unaffected siblings to report that one or both of their parents also are affected (Schull and Cobb 1969). In another epidemiological study, the false-positive rate was reported to be 11% (Whittemore et al. 1995). In general, the overreporting rate varies with many factors, such as the perceived severity of the disease of interest and the definition of the disease (J. T. Bensen, A. D. Liese, J. T. Rushing, M. Province, A. R. Folsom, and D. Arnett, personal communication).

Although this article has focused only on the sibling recurrence-risk ratio, the results can be generalized to the recurrence-risk ratio for other relative pairs. As is true for the difference in sibship size, different individuals may have different numbers of aunts, uncles, and offspring.

It should be noted that some genetic epidemiological studies, such as those conducted by Bailey and Benishay (1993) and Pillard and Weinrich (1986), do not specifically use the sibling recurrence-risk ratio to measure familial aggregation. However, because the sampling is always prone to both ascertainment bias (especially for traits that carry a social stigma) and overreporting, the potential for misinterpretation of data always exists.

The correct way to estimate λ_S is, of course, to try to correct for ascertainment bias and/or overreporting error. This may not be an easy task. What one could do is to avoid the use of definitions (2) or (3), by specifying a priori which sib will be taken to be sib2. For example, one can specify, in advance, that either the immediately older sibling is to be ascertained or, if the latter is not available, the immediately younger sibling.

It is somewhat surprising that the ascertainment bias, coupled with overreporting, can substantially inflate the sibling recurrence-risk ratio. This result demonstrates that, just like segregation analysis, the estimation of the ratio is not immune to ascertainment bias. It further underscores the difficulty in establishing compelling evidence that the disease or disorder has a genetic component. In an era when gene mapping relies increasingly on brute-force procedures such as genomewide scanning (whether in an association study or not), this should cause sobering alarm, since what have been hunted might well be phantom genes.

This also raises the question of the extent to which we should trust the estimates of recurrence-risk ratios

reported in some genetic epidemiological studies of complex diseases, studies that apparently have not made any attempt to correct for ascertainment bias and/or overreporting. Perhaps extreme caution is needed when, on the basis of these estimates, one either performs exclusion mapping or discriminates between different genetic models. It is perhaps timely to heed the advice given by the late R. A. Fisher in his well-known paper on ascertainment bias, published 65 years ago: "It is a statistical commonplace that the interpretation of a body of data requires a knowledge of how it was obtained" (Fisher 1934, p. 13).

Acknowledgments

This research was supported by National Institutes of Health grants R29-GM52205 and R01-GM56515. The author thanks Drs. Pak Sham, Aaron Folsom, and Donna Arnett for their helpful comments on an earlier version of this paper.

References

Bailey JM, Benishay DS (1993) Familial aggregation of female sexual orientation. Am J Psychiatry 150:272–277

Bailey NTJ (1951) The estimation of the frequencies of recessive with incomplete multiple selection. Ann Eugenics 16: 215–222

Enterline PE, Capt KG (1959) A validation of information provided by household respondents in health surveys. Am J Public Health 49:205–212

Ewens WJ (1991) Ascertainment biases and their resolution in biological surveys. In: Rao CR, Chakraborty R (eds) Handbook of statistics. Vol 8. Elsevier Science, New York, pp 29–61

Farrall M, Holder S (1992) Familial recurrence-pattern analysis of cleft lip with or without cleft palate. Am J Hum Genet 50:270–277

Fisher RA (1934) The effect of methods of ascertainment upon the estimation of frequencies. Ann Eugenics 6:13–25

Floderus B, Barlow L, Mack TM (1990) Recall bias in subjective reports of familial cancer. Epidemiology 1:318–321

Galton F (1904) Average number of kinfolk in each degree. Nature 70:529, 626

Haenszel W (1959) Some problems in the estimation of familial risks of disease. J Natl Cancer Inst 23:487–505

Hastrup JL, Hotchkiss AP, Johnson CA (1985) Accuracy of knowledge of family history of cardiovascular disorders. Health Psychol 4:291–306

Herrmann N (1985) Retrospective information from questionnaires. I. Comparability of primary respondents and their next-of-kin. Am J Epidemiol 121:937–947

Hunt SC, Williams RR, Barlow GK (1986) A comparison of positive family history definitions for defining risk of future disease. J Chronic Dis 39:809–821

Kee F, Tiret L, Robo JY, Nicaud V, McCrum E, Evans A, Cambien F (1993) Reliability of reported family history of myocardial infarction. Br Med J 307:1528–1530

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Oxford University Press, New York

Khoury MJ, Beaty TH, Liang KY (1988) Can familial aggregation of disease be explained by familial aggregation of environmental risk factors? Am J Epidemiol 127:674–683

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

MacMahon B (1978) Epidemiologic approaches to family resemblance. In: Morton NE, Chung CS (eds) Genetic epidemiology. Academic Press, New York, pp 3–11

Monroe KR, Yu MC, Kolonel LN, Coetzee GA, Wilkens LR, Ross RK, Henderson BE (1995) Evidence of an X-linked or recessive genetic prostate cancer risk. Nat Med 1:827–829

Narod SA, Dupont A, Cusan L, Diamond P, Gomez J-L, Suburu R, Labrie F (1995) The impact of family history on early detection of prostate cancer. Nat Med 1:99–101

Penrose LS (1953) The genetic background of common diseases. Acta Genet 4:257–265

Pillard RC, Weinrich JD (1986) Evidence of familial nature of male homosexuality. Arch Gen Psychiatry 43:808–812

Raphael K (1987) Recall bias: a proposal for assessment and control. Int J Epidemiol 16:167–170

Rich SS (1990) Mapping genes in diabetes: genetic epidemiological perspective. Diabetes 39:1315–1319

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

Risch N (1990*b*) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

Risch N (1990*c*) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Sadovnick AD (1994) Genetic epidemiology of multiple sclerosis: a survey. Ann Neurol 36 Suppl 2:S194–S203

Schull WJ, Cobb S (1969) The intrafamilial transmission of rheumatoid arthritis. III. The lack of support for a genetic hypothesis. J Chronic Dis 22:217–222

Weinberg W (1912) Further contributions to the theory of heredity. V. On the inheritance of the predisposition to blood disease with methodological supplements to my sibship method. Arch Rassen-Gesellschaftsbiol 9:694–709

Whittemore AS, Wu AH, Kolonel LN, John EM, Gallagher RP, Howe GR, West DW, et al (1995) Family history and prostate cancer risk in black, white, and Asian men in the United States and Canada. Am J Epidemiol 141:732–740