American Journal of Epidemiology
Vol. 151, No. 9
Printed in U.S.A.
Copyright © 2000 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Bias Associated with Study Protocols in Epidemiologic Studies of Disease Familial Aggregation

Yan Bai,[1] Stephanie Sherman,[2] Muin J. Khoury,[3] and W. Dana Flanders[1]

The effect of selection bias has not been well evaluated in epidemiologic studies which focus on familial aggregation. The authors illustrate this type of bias for a reconstructed cohort study. With the reconstructed cohort design, cases and controls are first selected from the population and their relatives form the exposed and unexposed cohorts, respectively. The recurrence risk ratio (RRR) is calculated to assess and measure familial aggregation. The ways of utilizing information from relatives affects the estimate of RRR, and the authors show that a traditional method used in epidemiologic studies can yield a severely biased estimate of the RRR. However, this traditional approach can give approximately unbiased estimates under special conditions. A novel selection approach is proposed which yields an unbiased estimate of RRR. In conclusion, when relatives are identified through cases or controls, they should be included and counted in the study cohorts each time a case or control is selected, even if they or other family members have already been included. *Am J Epidemiol* 2000;151:927–37.

cohort studies; genetics; research design; selection bias

The study of disease familial aggregation is central to genetic epidemiology (1, 2). Familial aggregation of disease refers to higher disease occurrence among relatives of a case compared with that among relatives of a healthy person or among the general population. Traditional epidemiologic methods such as case-control and cohort approaches have been used to estimate recurrence risk and recurrence risk ratio in assessment of familial aggregation of diseases such as birth defects and cancers (3–12).

Selection bias is an important problem in epidemiologic studies. It traditionally refers to the distortion of a measure of association, e.g., the risk ratio, that results from the way that subjects are selected for and participate in the study (13). Inappropriate selection can distort results so that the estimated association differs from the true association. Effects of selection bias have been well demonstrated in traditional epidemiologic studies, for which subjects are *individually* selected from a well-defined population. In epidemiologic studies of disease familial aggregation, subjects with disease (cases) or without disease (controls) are often selected and the information about their relatives is gathered (6, 7, 9). This design is particularly favored for the study of rare genetic disorders.

With this design, however, the study population is not a random sample of families or of individuals from the underlying population, because families with multiple cases could be overrepresented. In particular, when more than one proband is identified from a family, it is not clear from previous studies whether relatives should be included more than once. For example, in a study of familial risk of melanoma, Aitken et al. (6) removed duplicate records of relatives to include each relative only once, but did not give a justification. Similarly, in a study of cardiac malformations by Pierpont et al. (7) and in a study of central nervous system cancers by Farwell and Flannery (9), only one affected child in a family was considered as the proband even if more than one child was identified. On the other hand, in a study of major congenital heart defects, Sanchez-Cascos (8) counted the relatives in those families twice if there were two probands from those families. These examples show that in epidemiologic studies of disease familial aggregation in which families are selected through cases and controls, different methods are being used to enumerate families.

Received for publication October 26, 1998, and accepted for publication May 6, 1999.

Abbreviations: CNS, central nervous system; RRR, recurrence risk ratio.

[1] Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA.
[2] Department of Genetics, Emory University School of Medicine, Atlanta, GA.
[3] Office of Genetics and Disease Prevention, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA.
Reprint requests to Dr. Yan Bai, Genetic Epidemiology Branch, National Cancer Institute, Executive Plaza South, Room 7106, 6120 Executive Blvd., MSC 7236, Bethesda, MD 20852-7236.

The potential bias related to these selection methods has not been well studied.

This problem, selection bias in studies of familial aggregation, has a counterpart in genetic epidemiology: ascertainment bias. It refers to the bias in the estimate of genetic parameters, such as the segregation ratio, when families are ascertained through affected individuals, or probands. Ascertainment bias has been recognized and studied, primarily in the context of segregation analysis (14, 15). A key difference between selection bias and ascertainment bias that we will address here is that the former involves selection of families through two types of subjects—cases and controls—but the latter involves selection only through cases.

In this paper, we will first illustrate the bias associated with estimation of the recurrence risks and recurrence risk ratio using a selection protocol similar to the protocol which has been used in some published studies of familial aggregation (6, 7, 16, 17, 19). We then propose an alternative method which we show leads to valid estimates.

METHODS

We motivate our work by considering a simple hypothetical study of birth defects in which the population has four sibships, each with two siblings as summarized in table 1. We identify all cases of birth defects (n = 4 families) and study risk of disease in relatives of the index case. We have two index cases in family 1, so a simple and seemingly natural approach might be to include this family only once in estimating the recurrence risk for birth defects. With this approach, the sibling of the proband in family 1 is diseased, the sibling of the proband in family 2 is not diseased, and the sibling of the proband in family 3 is also not diseased, so that the estimate of recurrence risk would then be 1/3. However, this estimate is clearly biased, as we can see by considering $\Pr(\text{sib}_2 = 1|\text{sib}_1 = 1)$: of the two sibships in which $\text{sib}_1 = 1$, only one sibship has a diseased sib_2, so that the recurrence risk is 2 (the same estimate holds for $\Pr(\text{sib}_1 = 1|\text{sib}_2 = 1)$). Similar bias arises with this simple approach, in estimating recurrence risk for a negative family history ($\Pr(\text{sib}_2 = 1|\text{sib}_1 = 0)$). Thus, the simple approach leads to biased estimates of the recurrence risk. This potential bias in this simple example illustrates the need to investigate sampling and analysis strategies for reconstructed cohort studies.

TABLE 1. Simple example with four families to illustrate different ways of calculating recurrence risks

Family no.	Sibship*	
	sib_1	sib_2
1	1	1
2	1	0
3	0	1
4	0	0

* 1 for yes, 0 for no.

Parameter definition

We will study the effect of different selection protocols on the recurrence risk and recurrence risk ratio (RRR), measures of familial aggregation. The RRR is the cumulative incidence of disease through a specific age among relatives of a person with the disease (recurrence risk) divided by that among relatives of a person without the disorder. We consider the reconstructed cohort design described by Susser and Susser (2). In a reconstructed cohort study, investigators gather information from different types of relatives of cases and controls and treat them as cohorts—relatives of cases form the exposed cohort and relatives of controls form the unexposed cohort. These cohorts can be analyzed in a life table with birth as a point of entry (20, 21). For simplicity, we will limit our study population to the siblings (first-degree relatives) of the cases and controls. We assume that the disease under study occurs from birth through a certain age or is present at birth in a fixed cohort, that all family members have passed through the risk period, and that competing risks can be ignored. We also assume that cases and controls are selected randomly from the diseased and nondiseased populations, respectively, at the end of the risk period.

For simplicity, we assume that each sibship selected for study consists of three siblings. Assume the underlying population has N_T sibships, and let N_{3j} be the number of sibships in which j of the three siblings are affected by a certain disorder (j = 0 to 3). We use $\text{sib}_j = 1$ to denote that sibling j is affected and $\text{sib}_j = 0$ to denote that sibling j is not affected. We refer to the siblings as sib_1 to sib_3 and assume that birth order and numbering are independent of risk. Obviously, the underlying population has $3N_T$ people. Table 2 summarizes the information about the sibships in the underlying population.

Our purpose is to study the effect on the RRR of using different selection protocols to study the association between family history and disease occurrence.

Population frequencies

The expected proportions for each type of sibship can be expressed in terms of risks. We use the following general notation for risks: P denotes the background risk in the population, and P_{11} denotes the risk given that a person has two affected siblings, which is

TABLE 2. Expected frequency of sibships with varied number of affected individuals in a fixed cohort in the underlying population

No. of sibship N_{3j} (j = 3 to 0)	Sibship*			Expected sibship proportion
	sib_1	sib_2	sib_3	
N_{33}	1	1	1	PP_1P_{11}
N_{32}	1	1	0	$PP_1(1 - P_{11}) + P(1 - P_1)P_{10} + (1 - P)P_0P_{10}$
N_{31}	1	0	0	$P(1 - P_1)(1 - P_{10}) + (1 - P)(1 - P_0)P_{00} + (1 - P)P_0(1 - P_{10})$
N_{30}	0	0	0	$(1 - P)(1 - P_0)(1 - P_{00})$
N_T				1

* 1 for yes, 0 for no.

expressed as $P_{11} = \Pr(\text{sib}_a = 1|\text{sib}_b = \text{sib}_c = 1)$. P_{10} is the risk given that a person has one affected sibling and one unaffected sibling, which is expressed as $P_{10} = \Pr(\text{sib}_a = 1|\text{sib}_b = 1 \text{ and } \text{sib}_c = 0, \text{ or } \text{sib}_b = 0 \text{ and } \text{sib}_c = 1)$. P_{00} is the risk given that a person has two unaffected siblings, which is expressed as $P_{00} = \Pr(\text{sib}_a = 1|\text{sib}_b = \text{sib}_c = 0)$. We also define $P_1 = \Pr(\text{sib}_a = 1|\text{sib}_b = 1) = \Pr(\text{sib}_a = 1|\text{sib}_c = 1)$, and $P_0 = \Pr(\text{sib}_a = 1|\text{sib}_b = 0) = \Pr(\text{sib}_a = 1|\text{sib}_c = 0)$. The frequencies of sibships can be expressed using conditional probabilities.

With this notation, a rational definition for RRR is

$$\text{RRR} = P_1/P_0, \tag{1}$$

which compares the risk among those with a positive family history to that among those with a negative family history. This measure has been commonly used in previous epidemiologic studies on birth defects and cancers.

We first express P, P_1, and P_0 in terms of P_{10}, P_{11}, and P_{00} in the three-sibling sibships.

Using the law of total probability,

$$\begin{aligned} P_1 &= P(\text{sib}_3 = 1|\text{sib}_1 = 1) \\ &= P(\text{sib}_3 = 1|\text{sib}_1 = 1, \text{sib}_2 = 1) \times P(\text{sib}_2 = 1|\text{sib}_1 = 1) + P(\text{sib}_3 = 1|\text{sib}_1 = 1, \text{sib}_2 = 0) \\ &\quad \times P(\text{sib}_2 = 0|\text{sib}_1 = 1) \\ &= P_{11}P_1 + P_{10}(1 - P_1), \end{aligned} \tag{2}$$

and

$$\begin{aligned} P_0 &= P(\text{sib}_3 = 1|\text{sib}_1 = 0) \\ &= P(\text{sib}_3 = 1|\text{sib}_1 = 0, \text{sib}_2 = 1) \times P(\text{sib}_2 = 1|\text{sib}_1 = 0) + P(\text{sib}_3 = 1|\text{sib}_1 = 0, \text{sib}_2 = 0) \\ &\quad \times P(\text{sib}_2 = 0|\text{sib}_1 = 0) \\ &= P_{10}P_0 + P_{00}(1 - P_0). \end{aligned} \tag{3}$$

Using rules of conditional probabilities, the expected sibship proportions can be expressed in terms of risks, P, P_1, P_0, P_{00}, P_{10}, and P_{11}. These expressions are given in table 2 (column "Expected sibship proportion"). For example, the expected proportion of sibships in which one of three siblings is affected is:

$$P(\text{sib}_1 = 1) \times P(\text{sib}_2 = 0|\text{sib}_1 = 1) \times P(\text{sib}_3 = 0|\text{sib}_1 = 1, \text{sib}_2 = 0) +$$

$$P(\text{sib}_1 = 0) \times P(\text{sib}_2 = 1|\text{sib}_1 = 0) \times P(\text{sib}_3 = 0|\text{sib}_1 = 0, \text{sib}_2 = 1) +$$

$$P(\text{sib}_1 = 0) \times P(\text{sib}_2 = 0|\text{sib}_1 = 0) \times P(\text{sib}_3 = 1|\text{sib}_1 = 0, \text{sib}_2 = 0)$$
$$= P(1 - P_1)(1 - P_{10}) + (1 - P)(1 - P_0)P_{00} + (1 - P)P_0(1 - P_{10}).$$

Conceptually, the background risk in the population, P, is expressed by

$$P = (3N_{33} + 2N_{32} + N_{31})/3N_T.$$

Using the expressions in table 3 for N_{3j}, we can now express P in terms of the risks.

$$P = (P_0 + P_{00} + P_0P_{10} - P_0P_{00})/(2 + P_0 + P_{00} + P_0P_{10} + P_1P_{10} - P_1 - P_{10} - P_1P_{11} - P_0P_{00}). \quad (4)$$

Combining results from equations 2 to 4, we can now express the background risk P, P_1, and P_0 in terms of P_{11}, P_{10}, and P_{00}.

$$P = P_{00}(1 + P_{10} - P_{11})/(1 - P_{11} - P_{10} + 2P_{00} - 2P_{11}P_{00} + P_{10}P_{11} + P_{10}P_{00}), \quad (5)$$

$$P_1 = P_{10}/(1 - P_{11} + P_{10}), \quad (6)$$

and

$$P_0 = P_{00}/(1 - P_{10} + P_{00}). \quad (7)$$

We can thus express the proportions of sibships in terms of P_{11}, P_{10}, and P_{00} only. These expressions are given in the Appendix.

Cases are selected by taking a simple random sample from those with the disease, and controls are selected by taking a simple random sample from those without the disease. In particular, selection is independent of the prior selection of a subject's family members. We define the overall selection probabilities, π for cases and ε for controls. Families are selected through affected or unaffected subjects or both. The chance that a sibship is selected depends on the number of affected and unaffected subjects it has. For example, if j affected and $3 - j$ unaffected siblings are present in a sibship, the probability that this sibship is selected through an affected or unaffected subject would approximately equal $1 - (1 - \pi)^j$ and $1 - (1 - \varepsilon)^{3-j}$, respectively, if the population is not too small. For example, if a sibship has two affected siblings and one unaffected sibling, the probability that this sibship is selected through one affected subject is $1 - (1 - \pi)^2$ and the probability that this sibship is selected through an unaffected subject is ε. Each sibship could be selected through one, two, or three members of this sibship. Table 3 summarizes the selection probabilities for each type of sibship.

To evaluate the potential bias, we also need the probabilities (summarized in table 3, columns 4–11) that a sibship is identified through different numbers of affected or unaffected subjects.

Different protocols for selection and counting of subjects

First, we will evaluate a method of counting subjects and analyzing data which might seem acceptable to many epidemiologists initially, but which we will see is biased. We will then propose an alternative method not previously described which yields an unbiased estimator of the RRR.

TABLE 3. Selection probabilities of sibships through varied numbers of affected and unaffected members

Sibship*			No. of sibship	Affected†				Unaffected†			
sib_1	sib_2	sib_3		1	2	3	Overall	1	2	3	Overall
1	1	1	N_{33}	$3\pi(1-\pi)^2$	$3\pi^2(1-\pi)$	π^3	$1-(1-\pi)^3$	–	–	–	–
1	1	0	N_{32}	$2\pi(1-\pi)$	π^2	–	$1-(1-\pi)^2$	ε	–	–	$1-(1-\varepsilon)$
1	0	0	N_{31}	π	–	–	$1-(1-\pi)$	$2\varepsilon(1-\varepsilon)$	ε^2	–	$1-(1-\varepsilon)^2$
0	0	0	N_{30}	–	–	–	–	$3\varepsilon(1-\varepsilon)^2$	$3\varepsilon^2(1-\varepsilon)$	ε^3	$1-(1-\varepsilon)^3$

* 1 for yes, 0 for no.
† Probabilities that a sibship is selected through one, two, or three affected or unaffected siblings.

Protocol I

In Protocol I, we randomly select cases from those with the disease in the defined population. Each time we select a case, we add his/her siblings to the exposed cohort if and only if that sibship has *not* previously been selected through a case. This approach has been used in previous studies (6, 7, 9). Similarly, we select controls from those without disease. For each control selected, we add his/her siblings to the unexposed cohort if and only if that sibship has *not* previously been selected through a control. Under this protocol, any sibship will contribute either 0 or 2 siblings to each cohort.

With this protocol, we use the information from a sibship in each cohort at most once, even if we select multiple affected or unaffected siblings from that sibship. We can calculate the probability of selection for each sibship type, as summarized in table 3 (columns "Overall"). Combining selection probabilities in table 3 and the sibship proportions summarized in table 2, we calculate the expected numbers of diseased and nondiseased subjects in the cohorts under this protocol (table 4).

Using the expected values, we have recurrence risks and $\widehat{\mathrm{RRR}}_I$:

$$\hat{P}_1 = \{2[1-(1-\pi)^3]N_{33} + [1-(1-\pi)^2N_{32}\}/\{2[1-(1-\pi)^3N_{33} + 2[1-(1-\pi)^2]N_{32} + 2\pi N_{31}\}\}. \quad (8)$$

$$\hat{P}_0 = \{[1-(1-\varepsilon)^2]N_{31} + 2\varepsilon N_{32}\}/\{2[1-(1-\varepsilon)^3]N_{30} + 2[1-(1-\varepsilon)^2]N_{31} + 2\varepsilon N_{32}\}. \quad (9)$$

$$\widehat{\mathrm{RRR}}_I = \{2[1-(1-\pi)^3]N_{33} + [1-(1-\pi)^2]N_{32}\}\{[1-(1-\varepsilon)^3]N_{30} + [1-(1-\varepsilon)^2]N_{31} + \varepsilon N_{32}\}/\{\{[1-(1-\varepsilon)^2]N_{31} + 2\varepsilon N_{32}\}\{[1-(1-\pi)^3]N_{33} + [1-(1-\pi)^2]N_{32} + \pi N_{31}\}\}. \quad (10)$$

The risk ratio is somewhat complicated and depends on the risks and selection probabilities. If both π and ε are small (below 0.1), the above risk ratio approximates P_1/P_0, an unbiased estimate of the RRR. Otherwise the *RRR* will tend to be biased, as illustrated in the "Results" section.

Protocol II

Although the above type of selection (counting and analyzing) protocol yielded a biased estimate of recurrence risk and RRR, an approach like this has been used in some published studies (6, 7). We now propose and evaluate another method which we will show yields unbiased estimates.

In Protocol II, we also start by selecting sibships or families through affected (cases) or unaffected (controls) subjects from each sibship as in Protocol I. Each time we select a case, we include his/her siblings in the exposed cohort, *even if some members of this sibship have already been selected for the study*. We also randomly select controls from those who are healthy in the population. Each time we find a control, we include his/her siblings in the unexposed cohort, *even if some members of this sibship have already been selected for the study*. Like in Protocol I, selection of cases or controls is independent of previous selections with respect to family relationship. This protocol has two important characteristics. First, sibships which include multiple affected or multiple unaffected people may be selected and included more than once. For example, suppose a sibship contains two affected siblings and one unaffected sibling. When we select the exposed cohort, members from this sibship may be included and counted 0, 1, or 2 times depending on the number of affected(s) that we have selected from the sibship. When we select the unexposed cohort, members from this sibship may be included 0 or 1 times depending on the number of

TABLE 4. Expected numbers of diseased and nondiseased subjects who are exposed to a positive family history under Protocol I

Sibship*	Exposed	Unexposed
Diseased	$2[1-(1-\pi)^3]N_{33} + [1-(1-\pi)^2]N_{32}$	$[1-(1-\varepsilon)^2]N_{31} + 2\varepsilon N_{32}$
Undiseased	$[1-(1-\pi)^2]N_{32} + 2\pi N_{31}$	$2[1-(1-\varepsilon)^3]N_{30} + [1-(1-\varepsilon)^2]N_{31}$
Total	$2[1-(1-\pi)^3]N_{33} + 2[1-(1-\pi)^2]N_{32} + 2\pi N_{31}$	$2[1-(1-\varepsilon)^3]N_{30} + 2[1-(1-\varepsilon)^2]N_{31} + 2\varepsilon N_{32}$

unaffected we have selected from the sibship. Second, some subjects from a sibship may be included in the analysis as exposed and as unexposed (when a sibling is identified and selected as a case and another sibling is identified and selected as a control). Using the same example, if one of the cases in the sibship is selected, other members will be included in the exposed cohort. If the control is selected, other members will be included in the unexposed cohort. Obviously, one member in this sibship would be included in both exposed and unexposed cohorts. A similarity of this protocol to Protocol I is that the same member in a sibship can be included in both the exposed cohort and the unexposed cohort if he/she has both an affected and an unaffected sibling whom we select. Even though use of subjects from the same sibship more than once may seem unconventional, we will show that this protocol yields an unbiased estimate of the RRR.

In table 5, we summarize the expected numbers of diseased and nondiseased in the exposed and unexposed cohorts in terms of the risks and selection probabilities, based on selection Protocol II. For example, the expected number of diseased in the exposed cohort is:

$$[2 \times 3\pi(1 - \pi)^2 + 4 \times 3\pi^2(1 - \pi) + 6\pi^3]N_{33} + [2\pi(1 - \pi) + 2\pi^2]N_{32}.$$

The recurrence risks and RRR based on these expected values are:

$$\hat{P}_1 = (3N_{33} + N_{32})/(3N_{33} + 2N_{32} + N_{31}). \tag{11}$$

$$\hat{P}_0 = (N_{32} + N_{31})/(3N_{30} + 2N_{31} + N_{32}). \tag{12}$$

$$\widehat{\mathrm{RRR}}_{II} = (3N_{33} + N_{32})(3N_{30} + N_{32} + 2N_{31})/(3N_{33} + 2N_{32} + N_{31})(N_{32} + N_{31}) \tag{13}$$

$$= P_{10}(1 - P_{10} + P_{00})/[P_{00}(1 - P_{11} + P_{10})] = P_1/P_0.$$

Thus, the recurrence risk and RRR based on Protocol II provide consistent estimates.

RESULTS

To illustrate the magnitude of bias associated with use of Protocol I, we show how the bias in the RRR varies with different values of selection probabilities, π and ε, and disease occurrence. We also illustrate the bias associated with Protocol I by using two examples: the study of central nervous system tumors in children and the study of perinatal death.

Illustration of possible bias for various selection probabilities

We first compare the estimate of RRR under Protocol I for different selection probabilities of cases and controls (figures 1 and 2). We fixed the underlying risks so that the true RRR approximately equals 21. The bias increases as either π or ε increases. For example, when $\pi = 0.01$ and $\varepsilon = 0.01$, the estimated RRR is about 21. However, when $\pi = 0.5$ and $\varepsilon = 0.01$, the estimated RRR is about 16. For the situations considered, the effect of π was slightly greater than that of ε. For example, when $\pi = 0.25$ and $\varepsilon = 0.01$, the estimated RRR is 18.4, but when $\pi = 0.01$ and $\varepsilon = 0.25$, the estimated RRR is 18.6.

TABLE 5. Expected numbers of diseased and nondiseased subjects who are exposed to a positive family history under Protocol II

	Exposed	Unexposed
Diseased	$[6\pi(1 - \pi)^2 + 12\pi^2(1 - \pi) + 6\pi^3]N_{33} + [2\pi(1 - \pi) + 2\pi^2]N_{32}$	$2\varepsilon N_{32} + [2\varepsilon(1 - \varepsilon) + 2\varepsilon^2]N_{31}$
Nondiseased	$2\pi N_{31} + [2\pi(1 - \pi) + 2\pi^2]N_{32}$	$[6\varepsilon(1 - \varepsilon)^2 + 12\varepsilon^2(1 - \varepsilon) + 6\varepsilon^3]N_{30} + [2\varepsilon(1 - \varepsilon) + 2\varepsilon^2]N_{31}$
Total	$6\pi N_{33} + 4\pi N_{32} + 2\pi N_{31}$	$6\varepsilon N_{30} + 4\varepsilon N_{31} + 2\varepsilon N_{32}$

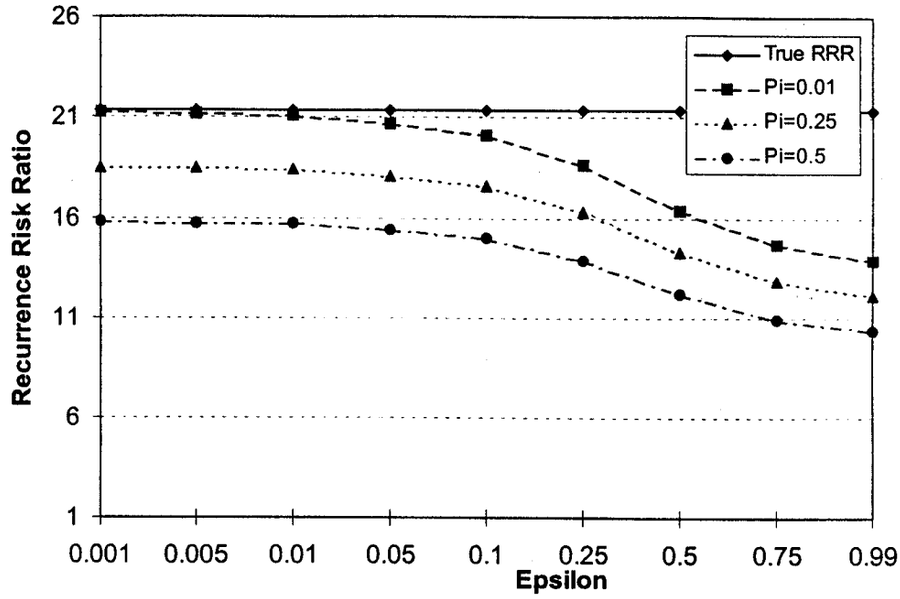

FIGURE 1. Comparison of biased and true recurrence risk ratios (RRRs) for varying control selection probabilities (ε) under low, medium, and high case selection probabilities (π) (0.01, 0.25, and 0.5; true RRR = 21.33).

Illustration of possible bias for various disease risks

To study the effect of disease risk on the magnitude of bias, we fixed the ratios of P_{11}/P_{00} and P_{10}/P_{00} at 5 and 2, respectively, and fixed the selection probability for controls (ε) at 0.1. We illustrate the bias associated with RRR under various disease risks and selection probabilities for the case (0.1, 0.3, and 0.8) in figure 3.

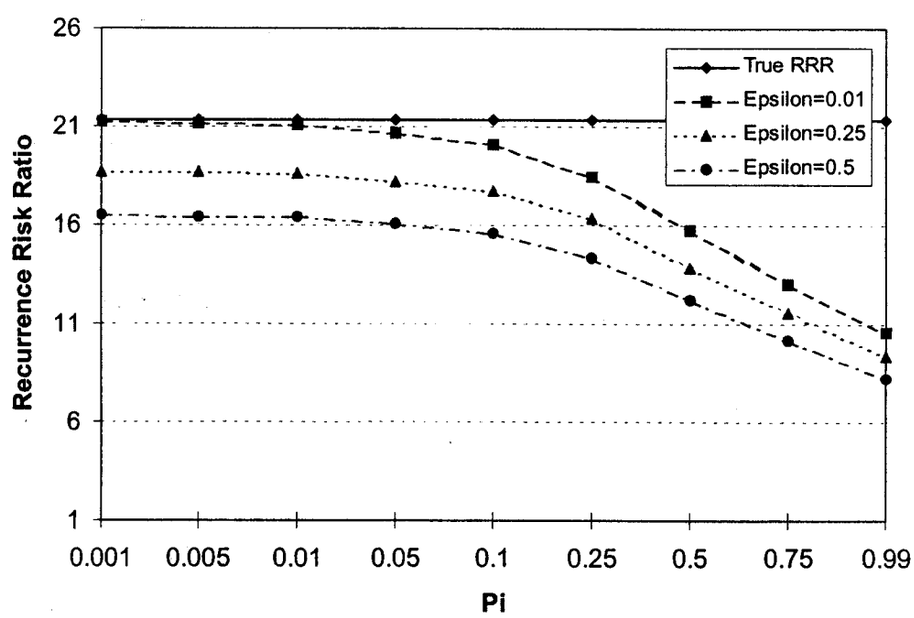

FIGURE 2. Comparison of biased and true recurrence risk ratios (RRRs) for varying case selection probabilities (π) under low, medium, and high control selection probabilities (ε) (0.01, 0.25 and 0.5; true RRR = 21.33).

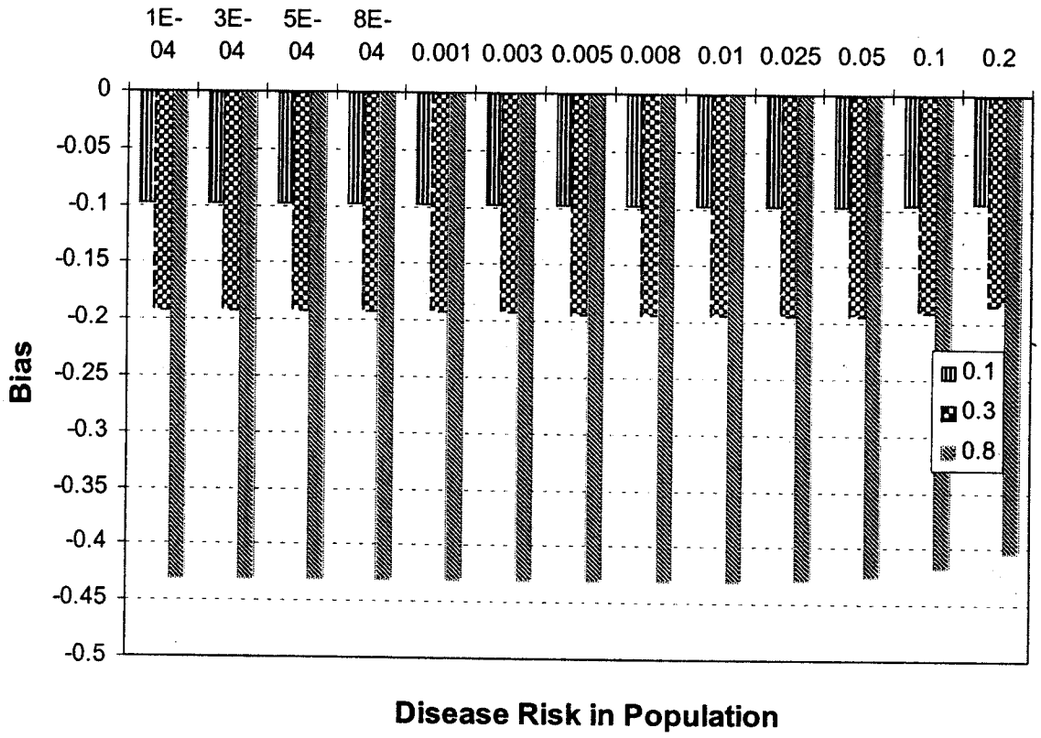

FIGURE 3. Bias associated with the recurrence risk ratio under different rates of disease occurrence and various selection probabilities for the case (0.1, 0.3, and 0.8), with control selection probability equal to 0.1 and P_{11}/P_{00} and P_{10}/P_{00} fixed at 5 and 2, respectively.

Bias is calculated as the difference between the observed and expected RRRs.

When the ratios of P_{11}/P_{00} and P_{10}/P_{00} are fixed, the true RRR changes with the baseline risk P. Figure 3 shows that the bias is affected little by disease risk. In other words, whether the disorder is rare or common seems to have little impact on the magnitude of the bias. This figure shows that the bias, however, is affected phenomenally by the selection probability for cases.

We only showed the situation in which the RRR from Protocol I is biased toward the null. However, we have not studied situations such as that of the RRR from Protocol I being in the reverse direction under certain selection probabilities and disease risks.

Example 1. In their study of cancers in relatives of children with central nervous system (CNS) tumors, Farwell and Flannery (9) reported an increased risk of CNS tumors among the relatives of a CNS tumor case (RRR = 8), compared with risks calculated from the general population. In two families where more than one case was seen (one family with two cases and the other family with three cases), they left only one sibling in the case series and considered others as relatives with cancer instead of treating them as probands (Protocol I). Had this analysis been performed by our proposed protocol, the RRR would have been 18. This example illustrates that bias can be substantial under the usual approach (Protocol I).

Example 2. To further illustrate the bias in the expected RRR under Protocol I, we use the risks taken from a previous study of perinatal death in Norway (18). Based on that study, P_{11}, P_{10}, and P_{00} are equal to 0.09, 0.03, and 0.01, respectively, and the RRR is 2.9. Assuming these values to represent the population parameters, we show how the expected RRR under Protocol I depends on the selection probabilities. In figure 4, when both of the selection probabilities are small, the risk ratio from Protocol I approximates the true risk ratio, 2.9. The bias increases as the selection probabilities increase, and the association is substantially underestimated (RRR = 1.6) when selection probabilities for cases and controls are 0.8 and 0.2, respectively.

DISCUSSION

We have investigated unique aspects of selection bias in studies of familial aggregation. We assumed that the cases and controls selected were representative of the diseased and nondiseased subjects, respectively,

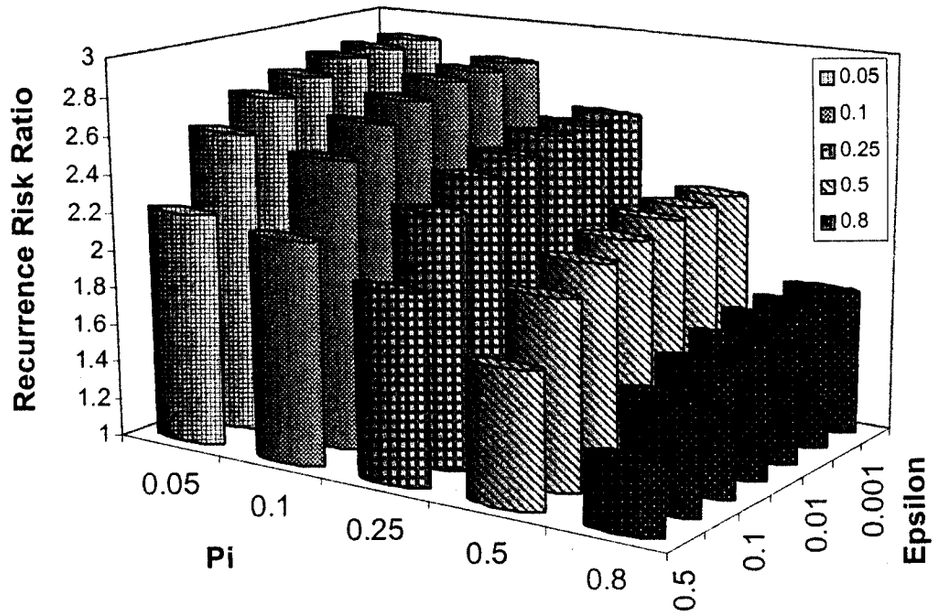

FIGURE 4. Recurrence risk ratio of perinatal death and family history for varying selection probabilities of cases and controls (true recurrence risk ratio = 2.95).

in the underlying population. This assumption allowed us to focus on selection bias which arises from the methods used to select and count family members of cases and controls and to define cohorts in a reconstructed cohort study design.

We have shown that one method (Protocol I) of subject selection, counting, and analysis yields biased estimates of the recurrence risks and risk ratio in reconstructed cohort studies of familial aggregation. This protocol has been used in published studies (6, 7, 9) and might seem to be consistent with ordinary epidemiologic practice, since each family contributes members to each cohort at most once. The RRR with this approach will approximate the true measure only when the selection probabilities are small for both affected and unaffected subjects. Even if this protocol provides an approximately unbiased estimate of the RRR under certain conditions, it will also tend to be an infeasible substitute for Protocol II if disease is rare. Specifically, one needs to have a low selection probability for cases to obtain an unbiased estimate. However, a small selection probability coupled with a rare disease would make it difficult to accrue adequate numbers of cases.

We suggest that if a reconstructed cohort design is used to study familial aggregation—that is, if relatives of affected people are classified as exposed and relatives of unaffected people are classified as unexposed—family members should be counted and included in the study as many times as a family member is identified as a case or control. This approach (Protocol II) yields a consistent estimator of the RRR.

In practice, one can select subjects for a reconstructed cohort study in several ways. For example, in the Minnesota breast cancer study (6), the exposed subjects consisted of relatives of breast cancer cases collected by cancer clinics, and unexposed subjects consisted of spouses of the male relatives of the cases. The occurrence of breast cancer was studied among these relatives. In the studies of Aitken et al. (6) and Farwell and Flannery (9), subjects with disease were selected from disease registries. By considering the relatives of the cases as exposed and the relatives of the controls as unexposed, these examples approximate the sampling method discussed here, and therefore are subject to potential bias if relatives are included only once.

As we described above, the definition of the exposure status of family members depends on which type of subjects in this family came to the attention of the investigator. In particular, it is possible that family members can be included in both exposed and unexposed cohorts. This definition of exposure differs from that usually used in epidemiologic studies. For example, in their study of the bias associated with using family history as a risk factor in case-control studies, Khoury and Flanders (22) classified people as exposed if they had at least one affected relative and unexposed if they had no affected relatives. Further

studies may identify improved definitions of exposure which lead to unbiased estimates and more efficient designs.

Although they share some features, selection bias in epidemiologic studies of familial aggregation and ascertainment bias in segregation analysis are fundamentally different. In epidemiologic studies of familial aggregation, such as those noted previously, a control group was selected in a manner analogous to that used to select the case group. In particular, in the reconstructed cohort study, families with no affected members are eligible for inclusion. On the other hand, neither inclusion of families without diseased members nor inclusion of relatives of controls generally occurs in segregation analysis. Thus, ascertainment bias in segregation analysis and selection bias have essential differences. On the other hand, our solution to avoiding bias is similar to that used to avoid ascertainment bias in some situations, i.e., to include families once each time a family member is selected as a proband (23).

The selection bias discussed here also differs from the usual selection bias in epidemiologic studies in that this type of selection bias arises in the *selection and counting* of families through individuals, whereas in traditional epidemiologic studies selection bias arises in the selection of individuals or just selection of groups instead. In our illustrations, selection bias in studies of familial aggregation pointed toward the null. However, the direction of bias depends on both selection probabilities and disease risks, and thus prediction of the direction may be difficult. Further work needs to be done to derive the variance estimator for the RRR under the correct selection protocol.

In this paper, we have pointed out that one selection protocol tends to yield a biased estimate of the association between family history and disease. The bias is substantial under certain conditions because the observed RRR can be lowered by 50 percent or more. We conclude that when relatives are identified through cases or controls in a reconstructed cohort study, they should be included and counted in the study cohorts each time a case or control is selected, even if they or other family members have already been included.

REFERENCES

1. Khoury MJ, Beaty TH, Cohen BM. Fundamentals of genetic epidemiology. New York, NY: Oxford University Press, 1993.
2. Susser E, Susser M. Familial aggregation studies: a note on their epidemiologic properties. Am J Epidemiol 1989;129:23–30.
3. Gold EB, Leviton A, Lopez R, et al. The role of family history in risk of childhood brain tumors. Cancer 1994;73:1302–11.
4. Colditz GA, Willett WC, Hunter DJ, et al. Family history, age, and risk of breast cancer: prospective data from the Nurses' Health Study. JAMA 1993;270:338–43.
5. Allan LD, Crawford DC, Chita SK, et al. Familial recurrence risk of congenital heart disease in a prospective series of mothers referred for fetal echocardiography. Am J Cardiol 1986;58:334–7.
6. Aitken JF, Duffy DL, Green A, et al. Heterogeneity of melanoma risk in families of melanoma patients. Am J Epidemiol 1994;140:961–73.
7. Pierpont ME, Gobel JW, Moller JH, et al. Cardiac malformations in relatives of children with truncus arteriosus or interruption of the aortic arch. Am J Cardiol 1988;61:423–7.
8. Sanchez-Cascos A. The recurrence risk in congenital heart disease. Eur J Cardiol 1978;7:197–210.
9. Farwell J, Flannery JT. Cancer in relatives of children with central-nervous-system neoplasms. N Engl J Med 1984;311:749–53.
10. Steinberg GD, Carter BS, Beaty TH, et al. Family history and the risk of prostate cancer. Prostate 1990;17:337–47.
11. Sellers TA, Anderson VE, Potter JD, et al. Epidemiologic and genetic follow-up study of 544 Minnesota breast cancer families: design and methods. Genet Epidemiol 1995;12:417–29.
12. Zhao LP, Hsu L, Davidov O, et al. Population-based family study designs: an interdisciplinary research framework for genetic epidemiology. Genet Epidemiol 1997;14:365–88.
13. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven, 1998.
14. Elston RC. Segregation analysis. Adv Hum Genet 1981;11:63–120.
15. Morton NE. Genetic tests under incomplete ascertainment. Am J Hum Genet 1959;11:1–16.
16. Sattin RW, Rubin GL, Webster LA, et al. Family history and the risk of breast cancer. JAMA 1985;253:1908–13.
17. Claus EB. Genetic epidemiology of breast cancer in younger women. J Natl Cancer Inst Monogr 1994;16:49–53.
18. Bracken MB. Perinatal epidemiology. New York, NY: Oxford University Press, 1984:129–30.
19. Schottenfeld D, Fraumeni JF Jr. Cancer epidemiology and prevention. 2nd ed. New York, NY: Oxford University Press, 1996.
20. Schupf N, Kapell D, Lee JH, et al. Increased risk of Alzheimer's disease in mothers of adults with Down's syndrome. Lancet 1994;344:353–6.
21. Mayeux R, Ottman R, Tang MX, et al. Genetic susceptibility and head injury as risk factors for Alzheimer's disease among community-dwelling elderly persons and their first-degree relatives. Ann Neurol 1993;33:494–501.
22. Khoury MJ, Flanders WD. Bias in using family history as a risk factor in case-control studies of disease. Epidemiology 1995;6:511–19.
23. Spence MA, Hodge SE. Segregation analysis. In: Rimoin DL, Connor JM, Pyeritz RE, eds. Principles and practice of medical genetics. New York, NY: Churchill Livingstone,1996:103–9.

APPENDIX

In this paper, the proportions of sibships in which 3, 2, 1, and 0 affected siblings are present are expressed in terms of diseases risks.

$$N_{33}/N_T = PP_1P_{11} = P_{11}P_{10}P_{00}/(1 - P_{11} - P_{10} + 2P_{00} - 2P_{11}P_{00} + P_{10}P_{11} + P_{10}P_{00}).$$

$$\begin{aligned} N_{32}/N_T &= PP_1(1 - P_{11}) + P(1 - P_1)P_{10} + (1 - P)P_0P_{10} \\ &= 3P_{10}P_{00}(1 - P_{11})/(1 - P_{11} - P_{10} + 2P_{00} - 2P_{11}P_{00} + P_{10}P_{11} + P_{10}P_{00}). \end{aligned}$$

$$\begin{aligned} N_{31}/N_T &= P(1 - P_1)(1 - P_{10}) + (1 - P)(1 - P_0)P_{00} + (1 - P)P_0(1 - P_{10}) \\ &= 3P_{00}(1 - P_{10})(1 - P_{11})/(1 - P_{11} - P_{10} + 2P_{00} - 2P_{11}P_{00} + P_{10}P_{11} + P_{10}P_{00}). \end{aligned}$$

$$\begin{aligned} N_{30}/N_T &= (1 - P)(1 - P_0)(1 - P_{00}) \\ &= (1 - P_{00})(1 - P_{10})(1 - P_{11})/(1 - P_{11} - P_{10} + 2P_{00} - 2P_{11}P_{00} + P_{10}P_{11} + P_{10}P_{00}). \end{aligned}$$