Misclassification in epidemiological surveys has recently received a good deal of attention. This paper offers an alternative solution to this important problem.

MISCLASSIFICATION IN EPIDEMIOLOGICAL SURVEYS

Dean Harper, Ph.D.

A number of recent papers have pointed out and discussed the problem of making inferences about the relationship between disease and presumed etiological factors (e.g., cancer of the cervix and lack of circumcision in the husband) from data derived from statements made by persons being interviewed in an epidemiological survey. 1-5 In such surveys the interviewee or patient may give a false answer to a question which then leads to a misclassification of that individual in respect to that particular question. For example, in a study of cancer of the cervix, the patient may say that her husband has not been circumcised when in fact he has been; then that husband will be falsely classified as not circumcised. If a number of individuals give such false information then clearly the relationship which is found between the disease and the etiological factor will be suspect. The true relationship will be disguised by the fact that some individuals have been misclassified. Diamond and Lilienfeld have indicated by examples that the relationship obtained from survey data may be greater or less than the true relationship. 3,5 Thus, a crucial problem in epidemiological surveys is to obtain results and relationships which are free of misclassification error.

Because this is a problem, Diamond and Lilienfeld suggest that any epidemiological survey should incorporate some way of dealing with the problem of misclassification. 3 They indicate that studies where the amount of misclassification error has been determined can be used to correct the results obtained in other surveys where such error is not directly investigated. However, since the amount of misclassification error may vary from study to study this solution is not always adequate. They further suggest that in an epidemiological survey a subsample of individuals be given medical examinations; from this the amount of misclassification for that subsample could be determined. The results for the larger sample could be corrected then by the results obtained from the physical examination of the subsample of individuals.

In this paper an alternative solution to the problem of misclassification will be suggested. This will not involve direct physical examination of a subsample of patients; rather it will involve asking the patient or interviewee, at least twice, whether he has the condition thought to be etiological for the disease.

Let the situation be described in the following way. A population of individuals is being surveyed in respect to the presence or absence of some characteristic or experience held to be relevant in the etiology of some disease. (By assuming a population, we avoid problems of statistical inference.) The population can be classified without error into cases, those who have the disease, and controls, those who do not have the disease. Therefore, let us consider just one category of individuals, say cases. Everything that is said can be extended to controls. The population cannot be classified without error, how-
ever, into those who have the etiological factor and those who do not; error arises in determining the proportion of the population who have the etiological factor. This error arises because individuals in the population are asked if they have the etiological factor and some of them give false information. Let us assume that those who have the etiological factor have some probability, denote it by $r$, of saying that they do not have the etiological factor and that those who do not have the etiological factor have some probability, denote it by $q$, of saying that they do have the etiological factor. These are probabilities of giving false information. (Thus, $r$ is the rate of false-negatives and $q$ is the rate of false-positives; see Newell.⁴) Denote the true proportion of individuals who have the etiological factor as $X$ and the reported proportion as $P$. Then it is clear that

\[(1) \quad P = X(1 - r) + (1 - X)q.\]

That is, the proportion of people who say that they have the etiological factor is equal to the proportion who really do, $X$, times their probability of saying that they do, $1 - r$, plus the proportion of individuals who do not have the factor, $1 - X$, times their probability of saying they do, $q$. This is equivalent to the first equation stated on page 2106 of Diamond and Lilienfeld.⁵

Let us further assume that the question concerning the presence of the etiological factor is asked two or more times of each individual during the course of the interview. Then, let $r_1$ and $q_1$ represent the probabilities of false answers on the first asking of the question, $r_2$ and $q_2$ the probabilities of false answers on the second asking of the question, and so on. Now equation (1) can be written:

\[(2) \quad P = X(1 - r_i) + (1 - X)q_i,\]

where $i = 1, 2, \ldots .

That is, for the $i^{th}$ asking of the question we have the proportion, $P_i$, of people who say they have the etiological factor and this can be expressed as being composed of the quantities indicated in equation (2). However, we also have some new quantities, $P_{12}$, $P_{13}$, $P_{123}$, and so on; in general, let these be referred to by $P_{ij}$, $P_{ijk}$, etc., where $i$, $j$, and $k$ denote different askings of the same question. That is, $P_{ij}$ represents the proportion of people who say they have the etiological factor on both the $i^{th}$ and the $j^{th}$ askings of the question. It is a joint probability.

Let one further assumption be made. For those who have the etiological factor, let us assume that those who give false information on the $i^{th}$ asking of the question are no more likely to give false information on the $j^{th}$ asking of the question than are those who gave correct information on the $i^{th}$ asking of the question. That is, the fact that an individual gives false information one time makes him no more likely to give false information a second time than one who gives correct information the first time. There is statistical independence between responses to different askings of the question. The same assumption will be made for those who do not, in fact, have the etiological factor. This assumption will be referred to as the assumption of intra-class independence. Within each of the two classes, i.e., those who have the etiological factor and those who do not, there is statistical independence in the responses to different askings of the question.

For those who have the etiological factor, let $r_{ij}$ denote the probability that they will give false information, i.e., say they do not have the factor, on both the $i^{th}$ and the $j^{th}$ askings of the question. By virtue of the assumption of intra-class independence, the following relationships can be stated:

\[(3) \quad r_{ij} = r_ir_j,\]

etc.
Now the following equations can be written.

\[ P_{ij} = X(1 - r_i)(1 - r_j) + (1 - X)q_iq_j, \]

(4) \[ P_{ijk} = X(1 - r_i)(1 - r_j)(1 - r_k) + (1 - X)q_iq_jq_k, \]

etc.

\[ P_i, P_{ij}, \text{ and } P_{ijk} \] are determined from the responses of individuals who were questioned in the epidemiological survey; \( X \) represents the true proportion of the population which has the etiological factor. It is \( X \) which we would like to determine. The question is: can \( X \) be determined from the \( P_i, P_{ij}, P_{ijk}, \) and so on? It can be if we make some further simplifying assumptions. Two different situations will be discussed.

**Situation 1: Constant Probabilities**—This first situation represents the most stringent case. It is assumed that all individuals, regardless of which of the two classes they are in and regardless of which time the question is asked, will have the same probability of giving a false answer. That is,

(5) \[ r_i = q_i = r, \] for all \( i. \)

Then equations (2) and (4) become

(6) \[ P_i = X(1 - r) + (1 - X)r \]

(7) \[ P_{ij} = X(1 - r)^2 + (1 - X)r^2, \]

\[ = X(1 - 2r) + r^2. \]

It can be shown that

(8) \[ r = \frac{1 \pm \sqrt{1 - 4(P_i - P_{ij})}}{2}. \]

(See Appendix A for this.)

There are two solutions for \( r. \) If we assume that individuals are more likely to give correct information than false information we can select the smaller value for \( r \) as the correct probability. Denote this solution as \( r^* \). Then

(9) \[ X = \frac{P_i - r^*}{1 - 2r^*}. \]

Here, then, is a solution for the true proportion of individuals who have the etiological factor.

**Situation 2: Constant Intraclass Probabilities**—The assumptions made above in Situation 1 may be too stringent. The research of Lilienfeld and Graham suggests that those with a given etiological factor may not have the same probability of giving false information as those without the etiological factor. They found that those who were circumcised were more likely to give false information concerning circumcision than those who were not. Let us then assume that those who have the etiological factor do not necessarily have the same probability of giving false information as those who do not have the etiological factor. That is, giving false information is a function of having the etiological factor. However, we will still assume that those in each class will have the same probabilities of giving false information regardless of when the question is asked. That is,

(10) \[ r_1 = r_2 = r_3 = r, \]

\[ q_1 = q_2 = q_3 = q. \]

Then equations (2) and (4) become

(11) \[ P_i = X(1 - r) + (1 - X)q, \]

\[ P_{ij} = X(1 - r)^2 + (1 - X)q^2, \]

\[ P_{ijk} = X(1 - r)^3 + (1 - X)q^3. \]

Here are three equations in three unknowns \( X, r, \) and \( q. \) Solutions for these unknowns are obtained in the following way.

First, solutions for \( r \) and \( q \) are obtained from

\[ Y = \frac{P_iP_{ij} - P_{ijk}}{2(P_iP_{ij} - P_{ijk})}, \]

\[ Z = \sqrt{Y^2 - 2YP_i + P_{ij}}, \]

where

\[ 1 - r = Y + Z, \]

\[ q = Y - Z. \]

The values obtained for \( r \) and \( q \) are inserted into the first equation of (11) to give a solution for \( X, \) the true proportion of those who have the etiological factor. This solution is

(12) \[ X = \frac{P_i - q}{1 - r - q}. \]
As indicated above, these procedures outlined in Situation 1 and Situation 2 are applied separately to the two groups—cases and controls. The proportion of cases who give false information can first be determined; then the proportion of controls who give false information can be determined.

It should be noted that Situation 2 with less stringent assumptions than Situation 1 requires that the question be asked three times rather than two times. This requirement of asking the question several times is not too crucial. In a hospital setting patients are accustomed to being asked the same question several times by different interviewers. Where individuals are interviewed in their homes then the questionnaire can be constructed so that the crucial question is asked two or three times, worded somewhat differently each time and separated by other questions. The repetition of items in this fashion is done not infrequently in a survey.

It should also be noted that this procedure does not allow for a determination of which individuals have given the false information. Rather, it merely indicates the proportion of people who have given false information.

This procedure is suggested as an alternative to that proposed by Diamond and Lilienfeld. It is superior in that it should be less costly than giving a subsample of individuals in the survey medical examinations. On the other hand, in the procedure proposed here there are obviously some very crucial assumptions—that of intraclass independence being the most crucial. This assumption would seem most valid when false information comes about because the patient or interviewee is lying or misunderstands the question. That is, it seems likely that an individual may lie to one interviewer but not necessarily to another, or may lie at one time but not necessarily another time. The same could be said about understanding a question. On the other hand, if the patient or interviewee understands the question each time it is asked, is not disposed to lie and believes that he knows the correct answer, then this assumption will not be valid; in these circumstances one who gives wrong information at one time or to one interviewer will do so at other times and to other interviewers. Thus, this method would seem valid to the extent that misinformation arises because interviewees lie or misunderstand the question—phenomena which have some degree of randomness. Nevertheless, it should be noted that this assumption of intraclass independence underlies various measuring instruments used by social scientists in social surveys.6,7 It is not, therefore, without precedent.

APPENDIX A

Consider the following two equations:

(A-1)  \[ P_1 = X(1-2r) + r, \]
(A-2)  \[ P_{11} = X(1-2r) + r^2. \]

Here are two equations in two unknowns, \( X \) and \( r \). The first can be subtracted from the second to give

(A-3)  \[ P_{11} - P_1 = r^2 - r \]

which can be written

(A-4)  \[ r^2 - r + (P_1 - P_{11}) = 0. \]

This is a quadratic equation; its solution is

(A-5)  \[ r = \frac{1 \pm \sqrt{1 - 4(P_1 - P_{11})}}{2} \]

APPENDIX B

Consider the following three equations:

(B-1)  \[ P_1 = X(1-r) + (1-X)q, \]
(B-2)  \[ P_{11} = X(1-r)^2 + (1-X)q^2, \]
(B-3)  \[ P_{11k} = X(1-r)^3 + (1-X)q^3. \]

Here are three equations in three unknowns: \( X, r, \) and \( q \). Let these equations be rewritten as follows:

(B-4)  \[ P_1 = X(1-r-q) + q, \]
(B-5)  \[ P_{11} = X(1-r^2-q^2) + q^2, \]
(B-6)  \[ P_{11k} = X(1-r^3-q^3) + q^3. \]
Now (B-4) can be solved for X.

\[ (B-7) \quad X = \frac{P_i - q}{1 - r - q} \]

This value for X can be inserted into (B-5) to give

\[ P_{ij} = \frac{P_i - q}{(1 - r - q)} \left[ (1 - r) - q \right] + q^2, \]
\[ = \frac{P_i - q}{(1 - r - q)} \left[ (1 - r) - q \right] \]
\[ = [P_i - q] \left[ (1 - r) + q + q^2 \right], \]
\[ P_{ij} = P_i(1 - r) + q - q(1 - r). \]

The value for X, given in (B-7), can also be inserted into equation (B-6) to give

\[ P_{ijk} = \frac{P_i - q}{(1 - r - q)} \left[ (1 - r)^3 - q^3 \right] + q^3, \]
\[ = \frac{P_i - q}{(1 - r - q)} \left[ (1 - r) - q \right] \]
\[ = [P_i - q] \left[ (1 - r)^2 + (1 - r)q + q^2 \right] + q^3, \]
\[ P_{ijk} = P_i \left[ (1 - r)^2 + (1 - r)q + q^2 \right] - q(1 - r)^2 - (1 - r)q^2. \]

Now, we have two equations, (B-8) and (B-9), in the two unknowns, q and r; they both contain quadratic terms. It can be noted, however, that these two equations are symmetric in respect to \((1 - r)\) and q and therefore can be solved. Let

\[ 1 - r = Y + Z, \]
\[ q = Y - Z. \]

Then equation (B-8) becomes

\[ P_{ij} = P_i \left[ (Y + Z) + (Y - Z) \right] - (Y + Z)(Y - Z), \]
\[ = P_i(2Y) - (Y^2 - Z^2), \]
which can be rewritten

\[ (B-11) \quad 2Y^2 = 2YP_i + P_i. \]

After substituting (B-10) into (B-9) we obtain

\[ P_{ijk} = P_i \left[ (Y + Z)^2 + (Y + Z)(Y - Z) + (Y - Z)^2 \right] - (Y + Z)(Y - Z)^2, \]
\[ = P_i(3Y^2 + Z^2) - 2Y(Y^2 - Z^2), \]

or

\[ (B-12) \quad 2Y(Y^2 - Z^2) = (3Y^2 + Z^2)P_i + P_{ijk}. \]

This value for \(Z^2\) in (B-11) can be inserted into (B-12) to give

\[ 2Y(Y^2 - 2YP_i + P_{ij}) = (3Y^2 + Z^2)P_i + P_{ijk} = 0. \]

This can be simplified to give

\[ 2Y(P_i^2 - P_{ij}) = P_i^2 + P_{ijk} = 0. \]

Then

\[ (B-13) \quad Y = \frac{P_iP_{ij} - P_{ijk}}{2P_i^2 - P_{ij}}. \]

This value for Y can be inserted into (B-11) to give a value for Z. From Y and Z, values of r and q can be obtained using equation (B-10) and finally a value for X, the true proportion who have the etiological factor can be obtained using equation (B-4).

REFERENCES