Effects of Misclassification in Epidemiologic Studies

WARREN H. GULLEN, M.D., M.P.H., JACOB E. BEARMAN, Ph.D., and EUGENE A. JOHNSON, Ph.D.

A FREQUENTLY used procedure in epidemiology and other disciplines involves the comparison of disease rates in two populations, communities, or groups. While epidemiologists desire, of course, to know the true population rates, they must, of necessity, be content with observed rates, often derived from samples of the populations. These observed rates may differ from the true rates of the samples as a result of error introduced by misclassification according to disease status. In other words, some persons with the disease will be erroneously classified by the study procedure as being without the disease and some persons without the disease will be classified, in error, as having the disease. As well described by Rubin and coworkers (*1*), the apparent difference in sample rates is related to the true difference in sample rates by the formula

$$(P'_1 - P'_2) = (P_1 - P_2)(1 - \phi - \theta) \qquad [1]$$

where:

- P_1 is the true rate for the sample of population 1,
- P_2 is the true rate for the sample of population 2,
- P'_1 is the apparent rate for the sample of population 1,
- P'_2 is the apparent rate for the sample of population 2,
- ϕ is the proportion of persons in the sample with the disease erroneously classified as without disease,
- θ is the proportion of persons in the sample without the disease erroneously classified as with disease.

Unmodified use of formula 1 is based on two assumptions: (*a*) the "ϕ error" is of the same magnitude for both samples and, likewise, the "θ error" is the same for both samples and (*b*) there is no error in assigning a person to the correct population. The relationship just described may be of aid in considering, for example, a comparison of prevalence rates in New York and Chicago.

If the samples are "properly" drawn from the relevant populations, certain statistical procedures can be applied, and it is likely that probability statements would be made. For instance, a test of significance might be performed on the contrast $(P'_1 - P'_2)$. Unless the investigator takes into account the misclassification errors, he will probably assume that $(P'_1 - P'_2)$ is identical with $(P_1 - P_2)$ in performing the tests. It is obvious that a "significant" result derived employing $(P'_1 - P'_2)$ implies that a "significant" result would have been derived had $(P_1 - P_2)$ been used since $|P'_1 - P'_2|$ cannot be larger than $|P_1 - P_2|$ as $|1 - \theta - \phi| \leq 1$.

Dr. Gullen is with the department of community medicine, Medical College of Georgia, at Augusta. Dr. Bearman is with the division of biometry, University of Minnesota School of Public Health, Minneapolis. Dr. Johnson is with the University's biomedical data processing center.

On the other hand, it is clear that unless ϕ and θ are both 0, the power of a test using $(P_1'-P_2')$ would be less than that of a test using (P_1-P_2) if the sample sizes, and so forth, were held fixed. Tables have been constructed (*1*) to assist in selecting a multiplier of sample size to be used to compensate for the decreased power. In other words, a certain sample size, n, used with the contrast (P_1-P_2) would yield, if one assumes a certain type I error probability, and so forth, a given power at certain alternatives. The table provides a factor which is a function of θ and ϕ and which is multiplied by n to derive another sample size which would yield the same power if $(P_1'-P_2')$ were used, as it often must be.

In this paper, we will consider error not only in assigning a person according to the presence of disease but also error in assigning a person to one population or another. This possibility has not been elaborated previously. The two populations may be considered as being persons with and without some factor (for example, smokers and nonsmokers) or persons with and without some other disease (for example, diabetics and nondiabetics). Thus, we might want to compare coronary disease rates among smokers and nonsmokers or among diabetics and nondiabetics. There can be error not only in classifying subjects according to presence of coronary disease but also in classifying subjects according to presence of the factor; that is, a smoker may be classified in error as a nonsmoker and vice versa.

Before considering the formulation of this more complex situation, certain definitions need to be specified.

N = total sample.

P = true overall prevalence rate of the disease in the sample.

P_1 = true prevalence rate of disease among those in the sample with the factor.

P_2 = true prevalence rate of disease among those in the sample without the factor.

Π = true proportion of the sample with the factor.

α = probability that a person with disease will be classified as without disease.

β = probability that a person without disease will be classified as having the disease.

γ = probability that a person with the factor will be classified as without the factor.

δ = probability that a person without the factor will be classified as having the factor.

P_1' = apparent sample prevalence rate of the disease among those classified with the factor (that is, observed prevalence rate as influenced by the misclassification errors).

P_2' = apparent sample prevalence rate of the disease among those classified without the factor.

These definitions are subject to the following assumptions.

1. The probability of misclassification according to disease is independent of the probability of misclassification according to the factor (that is, the probability of being misclassified according to presence of the disease is the same for both the group with the factor and the group without the factor, and so forth).

2. Exclude $(\gamma=0, \delta=1)$ and $(\gamma=1, \delta=0)$. These combinations of γ and δ would be vacuous with reference to the problem since everybody would be classified as with factor in the first instance and everybody would be classified as without factor in the second instance. In either instance there would be no comparison to be made.

3. Exclude $\Pi=0$ or 1. At those values there would be no true comparison since either nobody would have the factor or everybody would.

Following is a descriptive table of the sample if one assumes there were no misclassification errors.

Status	*Factor present*	*Factor absent*	*Total*
Disease present	$P_1\Pi N$	$P_2(1-\Pi)N$	PN
Disease absent	$(1-P_1)\Pi N$	$(1-P_2)(1-\Pi)N$	$(1-P)N$
Total	ΠN	$(1-\Pi)N$	N

The corresponding table allows for the various misclassification errors.

Status	*Factor present*	*Factor absent*	*Total*
Disease present	a	b	c
Disease absent	d	e	f
Total	g	h	j

where:

$$a=P_1\Pi N-[\alpha(1-\gamma)P_1\Pi N+\gamma(1-\alpha)P_1\Pi N+\alpha\gamma P_1\Pi N]+[\beta(1-\gamma)(1-P_1)\Pi N+$$

$$+\delta(1-\alpha)P_2(1-\Pi)N$$
$$+\beta\delta(1-P_2)(1-\Pi)N],$$
$$g=\Pi N-\gamma\Pi N+\delta(1-\Pi)N,$$
$$b=P_2(1-\Pi)N-[\delta(1-\alpha)P_2(1-\Pi)N$$
$$+\alpha(1-\delta)P_2(1-\Pi)N+\delta\alpha P_2(1-\Pi)N]$$
$$+[\gamma(1-\alpha)P_1\Pi N+\beta(1-\delta)(1-P_2)(1-\Pi)N$$
$$+\beta\gamma(1-P_1)\Pi N],$$
$$h=(1-\Pi)N-\delta(1-\Pi)N+\gamma\Pi N.$$

Similar constructions could be formulated for c, d, e, f, and j but these quantities are not necessary in the sequel.

Now, $P_1'=\frac{a}{g}$ and $P_2'=\frac{b}{h}$. Therefore, $(P_1'-P_2')=\left[\frac{a}{g}-\frac{b}{h}\right]=\left[\frac{ha-bg}{gh}\right]$. A not minor exercise in elementary algebra reveals the following equation:

$$(P_1'-P_2')=\frac{(P_1-P_2)\Pi(1-\Pi)(1-\alpha-\beta)(1-\gamma-\delta)}{[\delta+\Pi(1-\gamma-\delta)][1-\{\delta+\Pi(1-\gamma-\delta)\}]}. \quad [2]$$

Equation 2 can be written in the form

$$(P_1'-P_2')=(P_1-P_2)R. \quad [3]$$

It remains to investigate the behavior of the quantity R. In the simple equation 1 described at the beginning of the paper, that is, no misclassification on the basis of the factor, $R=(1-\theta-\phi)$, and it is clear that $|R|\leq 1$. It would be useful to know that R in the more complex equation 2 behaves similarly. Such can be shown to be true.

In accordance with equations 2 and 3 denote

$$R=\Pi(1-\Pi)(1-\alpha-\beta)A, \quad [4]$$

where:

$$A=\frac{(1-\gamma-\delta)}{[\delta+\Pi(1-\gamma-\delta)][1-\{\delta+\Pi(1-\gamma-\delta)\}]}. \quad [5]$$

It can be shown that

$$\frac{\partial A}{\partial\gamma}=\frac{-\Pi^2(1-\gamma-\delta)^2-\delta(1-\delta)}{[\delta+\Pi(1-\gamma-\delta)]^2[1-\{\delta+\Pi(1-\gamma-\delta)\}]^2}. \quad [6]$$

The denominator is always positive because of assumption 2 and the fact that both major factors are squared. The numerator is always <0 since $\Pi^2>0$, $(1-\gamma-\delta)^2\geq 0$, and $[\delta(1-\delta)]\geq 0$, and the two terms cannot vanish simultaneously because of assumption 2. Thus, $\frac{\partial A}{\partial\gamma}$ is always negative.

Now, to perform a similar procedure with respect to δ, first define $\Pi'=1-\Pi$. Substituting $(1-\Pi')$ for each Π in A, we find that

$$A=\frac{(1-\gamma-\delta)}{[\gamma+\Pi'(1-\gamma-\delta)][1-\{\gamma+\Pi'(1-\gamma-\delta)\}]}, \quad [7]$$

which is of exactly the same form as the original A except that γ and δ are interchanged and Π is replaced by Π' which is a constant in differentiations with respect to γ and δ.
It is now clear that

$$\frac{\partial A}{\partial\delta}=\frac{-(\Pi')^2(1-\gamma-\delta)^2-\gamma(1-\gamma)}{[\gamma+\Pi'(1-\gamma-\delta)]^2[1-\{\gamma+\Pi'(1-\gamma-\delta)\}]^2}, \quad [8]$$

which is always negative by an argument similar to that employed regarding equation 6.

Since $\frac{\partial A}{\partial\gamma}$ and $\frac{\partial A}{\partial\delta}$ are negative, A is a monotonically decreasing function of γ for fixed δ and Π and of δ for fixed γ and Π (that is, A decreases as γ or δ or both increase). Thus, for any fixed Π, A has its maximum at $\gamma=0$, $\delta=0$ and its minimum at $\gamma=1$, $\delta=1$.

Let us consider a fixed Π, $\Pi=\Pi_0$. We know from the preceding discussion that the maximum A will occur at $\gamma=\delta=0$. At this point $A=1/[\Pi_0\ (1-\Pi_0)]$, by direct substitution. Define

$$B=\Pi\ (1-\Pi)A. \quad [9]$$

Then at $\Pi=\Pi_0$, $\gamma=\delta=0$, $B=1$, and $R=(1-\alpha-\beta)$. Since A is minimum where $\gamma=\delta=1$, substitution of those values in the formulas yields $A=-1/[\Pi_0\ (1-\Pi_0)]$, $B=-1$, and $R=-(1-\alpha-\beta)$. Thus, the maximum value of B occurs at $\gamma=\delta=0$ and the minimum value of B occurs at $\gamma=\delta=1$; further, the maximum and minimum values of B are independent of Π since the results were developed for any $\Pi=\Pi_0$.

Since $|B|\leq 1$ and since it is obvious that $|(1-\alpha-\beta)|\leq 1$, $|R|\leq 1$ or R ranges from -1 to $+1$.

A couple of characteristics of R are worthy of special note, aside from the fact that it is always between -1 and $+1$. First, it is apparent that if α, β, γ, and δ are all less than 0.5 (a not unreasonable assumption since classification would indeed be poorly done if errors were more frequent than correct assignments), then R is positive. In this situation $(P'_1-P'_2)$ may be smaller than (P_1-P_2), but the difference will be in the same direction, that is, $(P'_1-P'_2)$ and (P_1-P_2) will be of the same sign. Thus, if $P_2 > P_1$, then $P'_2 > P'_1$ and vice versa. This may well offer some comfort as may the observation that the assumption on $(\alpha, \beta, \gamma, \delta)$ is sufficient but not necessary.

In the simple case of misclassification described in previous papers, it was assumed that there were no errors in classifying according to factor or population. In terms of the new formula, this assumption is equivalent to $\gamma=\delta=0$. If $\gamma=\delta=0$ in R, $R=(1-\alpha-\beta)$ no matter what Π is. This is exactly what has already been shown to be the correct formula in the simple case. Thus, the general case of equation 2 reduces to equation 1 in the simple case.

The amount of error introduced in the rate contrast by the various classification errors can be illustrated by the following not unreasonable numerical example. Suppose that $\Pi=0.5$, $\alpha=\beta=\gamma=\delta=0.1$. Then $R=0.64$ and the apparent difference in sample rates is less than two-thirds of the real difference in sample rates. This is not an extreme example. Thus, modest degrees of classification errors can lead to rather severe decreases in contrasts between rates.

An interesting sidelight of this example follows. In the simple equation 1, if $\theta=\phi=0.1$, $R=0.8$. In the present example, $R=0.64=(0.8)^2$. It can be demonstrated that, in the complex case, whenever $\Pi=0.5$ and $\alpha=\beta=\gamma=\delta$, R is the square of the value found in the simple case with $\theta=\phi=$the common value specified for the complex case.

Other examples of this relationship are as follows (if one assumes $\Pi=0.5$). If $\alpha=\beta=\gamma=\delta=0.01$, $R=(1-0.01-0.01)^2=0.96$; if $\alpha=\beta=\gamma=\delta=0.05$, $R=(1-0.05-0.05)^2=0.81$; if $\alpha=\beta=\gamma=\delta=0.2$, $R=(1-0.2-0.2)^2=0.36$; and if $\alpha=\beta=\gamma=\delta=0.25$, $R=(1-0.25-0.25)^2=0.25$. These examples not only illustrate this special relationship between the simple and complex cases, but they demonstrate again the marked effect on the rate contrast of moderate percentages of classification error.

Since R ranges from -1 to $+1$, it is again clear that any significance test based on use of $(P'_1-P'_2)$ would have the same features as in the simple case. That is, a "significant" result does, indeed, imply a "significant" result; nevertheless, the power of the test is less than that of the corresponding test using (P_1-P_2).

Rather than the development of complex tables providing "build-up" factors for sample size to bring the power up to what would have been realized had there been no misclassification error, another procedure suggests itself. Use of the tables developed for the simple case demands specification of θ and ϕ, and it is very likely that use of any tables that could be constructed for the complex case would demand specification of the parameters in R. But if one can specify the values of those parameters, it would seem to be possible to calculate R and, hence, calculate P_1 and P_2 or (P_1-P_2) from P'_1 and P'_2. Thus, the usual statistical techniques could be used and the putative power would be realized without having to take larger samples. This advantage in not having to take the larger samples derives from a more efficient use of the information embodied in specified error values.

It is true that in most instances the values of the parameters in R would not be known (nor would the values of θ and ϕ be known in most instances of the simple case). However, reasonably decent estimates of the parameters or bounds on them might be available, yielding reasonable estimates of or bounds on R. For example, the classification of subjects in a field survey using a questionnaire might be compared with the classification of the same subjects using very intensive and sophisticated methods *for a subsample of the larger sample.* This procedure would not provide "true" values of the parameters but might provide useful estimates or provide the basis, with suitable multiplication, for bounds that would be accepted by many "experts" (that is, a statement that R is almost certainly not less than a given specified value).

It seems appropriate to point out what has

not been discussed in this paper and what should be considered further. Not considered is classification of a person according to a schema which has more than two classes (for example, disease absent, mild, moderate, or severe instead of merely present or absent). Likewise not considered is the use of nested classification (that is, classifying according to more than two variables). It seems likely on intuitive grounds that the formula would extend itself in a "reasonable" way and the "comforting" features would remain. Finally, not considered is the case in which the probability of misclassification according to factor is not independent of the probability of misclassification according to disease. But this eventuality involves a situation that is usually considered as being more than misclassification, possibly selection bias. It may be that no "nice" relationship exists in such situations.

Summary

Comparisons of disease rates are frequently made. The rates observed may be affected by classification errors. Some persons will be misclassified according to disease status or according to presence of an attribute, or both.

Under broad assumptions, the difference in observed sample prevalence rates is never larger than the difference in true sample prevalence rates, even if there is classification error in assigning persons to groups as well as to disease categories. Thus, even if the investigator cannot quantitate the classification error and adjust for it, or if he is ignorant of it, there may be some comfort in that the comparison is a conservative one and classification error never results in the apparent difference being larger than the real difference. There is a problem since unless one quantitates and adjusts for the classification errors, the apparent difference in rates may be substantially less than the true difference, and the investigator may well not detect a "significant" difference that really exists.

A second perhaps comforting feature is that if the percentage of each type of classification error is less than 50 percent, the apparent difference in sample rates and the true difference in sample rates are in the same direction and, hence, the correct group will have the larger apparent rate.

This discussion is concerned only with classification errors and their effects on differences in sample rates. Any inference from sample rates to population rates also involves the effect of sampling variability. Therefore, the difference in observed sample rates may, because of sampling variation, be larger than the difference in true population rates.

REFERENCE

(*1*) Rubin, T., Rosenbaum, J., and Cobb, S.: The use of interview data for the detection of associations in field studies. J Chronic Dis 4: 253–266, September 1956.

New Bulletin on Pesticides

The Pesticides Program of the Public Health Service has announced publication of a new monthly bulletin entitled *Health Aspects of Pesticides Abstract Bulletin.* The first bulletin was published in September 1968.

The purpose of this periodical is to foster current awareness of the major worldwide literature pertaining to the effects of pesticides on human beings. Each issue will carry approximately 200 English language abstracts scanned from at least 500 domestic and foreign journals.

Requests for subscriptions should be sent directly to the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 The subscription price is $2.75 a year, domestic; $3.50, foreign; and 30 cents for single copies.