Last year (July, 1962) Diamond and Lilienfeld in a paper published in the Journal called attention to the effect of misclassification on the apparent association between two variables as a common problem in epidemiology. Following discussion of this problem, Keys and Kihlberg endeavor to clarify the issues in this contribution.

EFFECT OF MISCLASSIFICATION ON ESTIMATED RELATIVE PREVALENCE OF A CHARACTERISTIC: PART I. TWO POPULATIONS INFALLIBLY DISTINGUISHED. PART II. ERRORS IN TWO VARIABLES

Ancel Keys, Ph.D., F.A.P.H.A., and Jaakko K. Kihlberg, Ph.D., M.P.H.

I. Two Populations Infallibly Distinguished

RECENT discussions about the effect of misclassification on the apparent association between two variables call attention to a common problem in epidemiology. Diamond and Lilienfeld[1] were surprised to find that their analysis of self-reports of the presence of an attribute among diseased and control women indicated an association between the attribute and the disease in question though true data on the attribute (from direct examination) failed to show any such association. They reported this finding as a matter of theoretical interest because it was "not consistent with several statistical papers, all of which have indicated that misclassification tends to decrease any true difference." In disagreement, Newell set forth an algebraic model to prove that "we can now reinstate the proposition that misclassification always tends to reduce the apparent difference between two proportions."[2] Finally, Diamond and Lilienfeld produced additional numerical examples, with hypothetical data, to demonstrate that "misclassification can produce spurious association."[3]

This issue needs clarification because the controversy, as it now stands in print, must confuse many readers. As will be seen, in their first paper Diamond and Lilienfeld should not have been surprised at their results while Newell's criticism apparently did not clarify the problem enough to prevent further confusion in the third paper.[3]

Analysis of the Problem

If there truly is an association between two variables, introduction of errors in measurement or identification of the variables must dilute or attenuate the degree of association between the variables provided that the errors are independent of any relationship between the variables. Diamond and Lilienfeld failed to note this important

limitation and Newell's algebraic analysis is concentrated on a single fourfold table and the special case of equal error probabilities. But the problem of Diamond and Lilienfeld is not such a special case.

The particular problem about which controversy arises is the estimation of the relative prevalence of a specified attribute in one population group, e.g., "diseased" persons, as compared with that in another, e.g., "healthy" persons, when recognition or report of the presence or absence of the attribute is subject to error and the error probabilities are not necessarily the same in the two groups. No question is raised about error in classifying persons into the two categories with regard to health and disease so analysis is limited to the simplest form of the misclassification problem, i.e., when only two all-or-none variables are involved and only one of these is subject to misclassification.

This problem is susceptible to strict algebraic analysis with the basic fourfold table set forth here as Table 1. If d is the proportion of the diseased population who truly possess the attribute in question, then $1-d$ represents the proportion who truly do not possess the attribute. And if P_1 is the probability of reporting correctly the presence of the attribute and P_2 is the probability of correctly reporting its absence in the diseased group, the reported prevalence of the attribute, d^*, as a proportion of all the diseased group, is $d^*=dP_1+(1-d)(1-P_2)$. The ratio of reported to true prevalence is, of course, d^*/d in this population group.

Exactly the same relationships apply in the "healthy" population. If h represents the true prevalence of the attribute, P_3 and P_4 the probabilities of correctly reporting presence and absence, respectively, of the attribute in the healthy group, then we have the reported prevalence $h^*=hP_3+(1-h)(1-P_4)$. The ratio h^*/h corresponds, for the healthy group, to d^*/d for the diseased group. It is the comparison of these two ratios that concerns us because on this comparison will depend whether the true relative prevalence of the attribute in the diseased population group, as compared with that in the healthy group, will be under- or overestimated.

No general statement about under- or overestimation of the association between the attribute and the disease is possible; the result will depend on the values of d, h, P_1, P_2, P_3, and P_4 that apply. Effects of given error probabilities on estimates of prevalence for given true prevalences are indicated in Figure 1, which covers probabilities of correct reporting from 0.5 to 0.99; few situations will be of interest if the probability of correct reporting is less than 50 per cent. Figure 1 is concerned with the estimation of the values of

Table 1—The Fourfold Table Expressed in Proportion of the Population and Probabilities of Correctly Reporting Presence (P_1) and Absence (P_2) of the Attribute

True Status	Reported Status: Yes	Reported Status: No	Total
Yes	dP_1	$d(1-P_1)$	d
No	$(1-d)(1-P_2)$	$(1-d)P_2$	$1-d$
Total	$dP_1+(1-d)(1-P_2)$	$d(1-P_1)+(1-d)P_2$	1

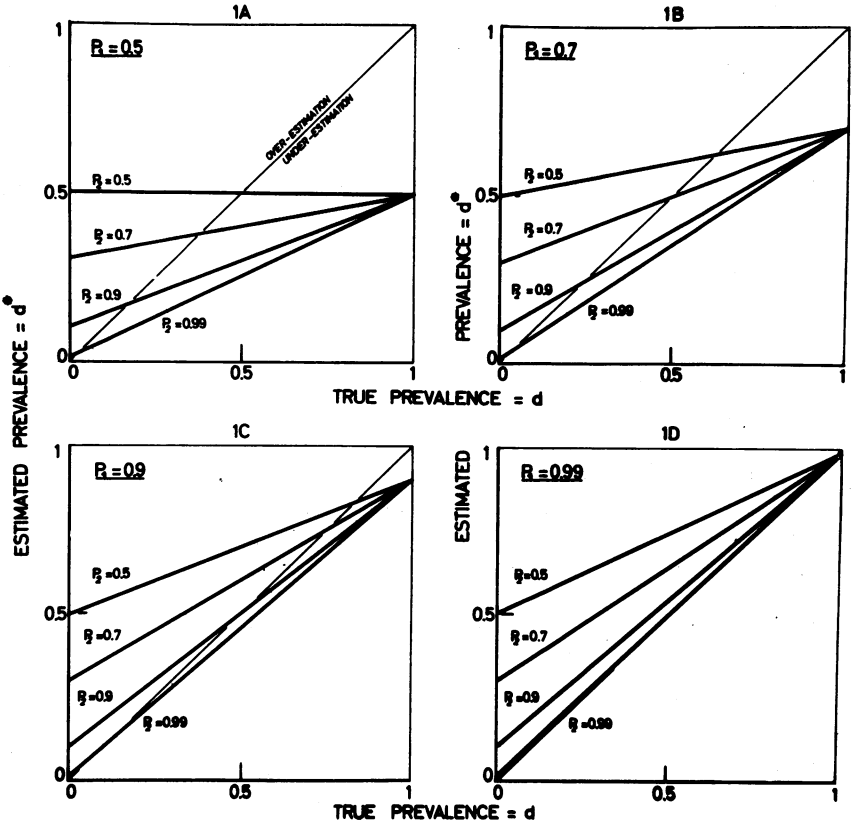

Figure 1

d* (or h*); comparison of the ratios d*/d and h*/h indicates the effect of misclassification on estimation of relative prevalence or of the degree of association of the attribute with the disease in question.

Where probability of correct classification is at least 0.5, overestimation of prevalence dominates when true prevalence is very low, while underestimation is more common when true prevalence is very high. The degree of association between the attribute and disease (or other characteristic used to separate the population into two categories) is judged from the ratio of the two prevalences; the effect of error in classification with respect to the attribute can produce anything from grossly overestimated (spurious) association to gross underestimation of the true association. This holds for the model considered here and for all of the large variety of practical situations that correspond to the model.

Even with identical error probabilities operating in the two population groups, gross error may result in the estimation of the degree of association. For example, if $P_1=P_2=P_3=P_4=0.7$ and $d=0.01$ while $h=0.05$, we find $d^*/d=30.40$ and $h^*/h=6.40$; i.e., the prevalence of the attribute is overestimated by 30-fold for the diseased group and by sixfold for the healthy group, and the association of the at-

tribute with the disease will be grossly overestimated.

Berkson[4] used numerical examples to point out limitations in the application of fourfold table analysis of medical data where selective forces operate. Though error and misclassification were not involved in Berkson's analysis, the distorting influence of the selection situations considered in Berkson's paper are similar in effect to the error situation discussed in the present paper. And Berkson showed how a spurious correlation could result in some cases while in others a true correlation could be underestimated.

Cause of the Controversy

Why, then, were Diamond and Lilienfeld surprised with their results, why did Newell disagree with them, and why did Diamond and Lilienfeld fail thereafter to remove the basis for further controversy?

One cause of the controversy seems to be in Diamond and Lilienfeld's failure to realize that in the model set up in Table 1 there are only three essential parameters, namely p, P_1, and P_2. Any attempt to rely on further parameters, notably the conditional probabilities

$$R_1=(1-d)(1-P_2)/[dP_1+(1-d)(1-P_2)]$$
$$R_2=d(1-P_1)/[d(1-P_1)+(1-d)P_2]$$

will fail to bring anything new into the picture, since these probabilities already are functions of p, P_1, and P_2. This is pointed out correctly by Newell who, however, is not very successful in demonstrating the issue.

Both sides of the controversy refer to a paper by Bross,[5] but it should be noted that Bross investigated only the special case where the error probabilities are identical in the two populations being compared. In practice, such equality of probabilities must be uncommon and it does not exist in the examples used in the discussion. In another communication[6] it is demonstrated what happens when this restriction is lifted.

To some extent confusion has been promoted by differences in terminology and form of computation used by the protagonists. Diamond and Lilienfeld chose a numerical example rather than an algebraic model and, for reasons inexplicable to us, considered what they called "relative risk." In the terminology we use, this relative risk for "stated percentages" is $R_S=\frac{h^*}{1-h^*}\cdot\frac{1-d^*}{d^*}$ and for "true percentages" is $R_T=\frac{h}{1-h}\cdot\frac{1-d}{d}$. The question as to whether the true "relative risk" is under- or overestimated by R_S then is simply the question as to whether the statistic:

$$T=\frac{R_S}{R_T}=\left(\frac{h^*}{1-h^*}\cdot\frac{1-h}{h}\right)\left(\frac{d}{1-d}\cdot\frac{1-d^*}{d^*}\right)$$

is less than, equal to, or greater than 1. Besides theoretical objections that can be raised, this formulation, even when clarified as above, is clumsy and unnecessarily complicated.

Newell's formulation, though commendably in algebraic form, is also complicated with nine terms, plus two more symbols for ratios, for a single fourfold table, where, as shown in Table 1, three symbols will do. Further, in the discussion of his example of two numerical fourfold tables, Newell chose to emphasize the difference in the proportions of false positives in the two sexes rather than noting the ratios of the reported prevalences, $d^*/h^*=1.47$ and true prevalences, $d/h=2.10$, which to us are more meaningful and are easily obtainable from his Table 3. (Here we use d to refer to the males and h to the females, and prevalence refers to chronic bronchitis.)

These figures from Newell's data show directly that true prevalence of bronchitis is underestimated in both sexes but more so in the males and that the true association of bronchitis

with maleness is underestimated with these particular error-containing reports. Newell suggests that error bias may have arisen because the men were more reluctant than the women to admit bronchitic symptoms. Suppose that query about bronchitic symptoms were made by officers calling up a draft for heavy labor. In that case we can imagine that the men would be more apt to exaggerate than to deny their symptoms and the result then would be that the true association between maleness and bronchitis could well be overestimated from the reports.

More Complex Problems

In their second paper,[3] Diamond and Lilienfeld indicate that they are looking into the much more complex situation where both variables, presence of the attribute and diagnosis of the disease, are subject to error in reporting. The analysis of this situation is given in a separate communication[6]; here it is enough to say that 15 parameters are involved and numerical analysis, if not theoretically hopeless, will rarely be practicable.

These situations, in which misclassification occurs in respect to two mutually exclusive categories, are special cases of the general theory of errors. With continuously distributed variables it is necessary to consider also the distribution of errors. Further, inquiry may be made not merely about the fact of association but also about its character—linear, curvilinear, and so forth. We shall present a more general analysis of error systems elsewhere.[7]

Summary

Diamond and Lilienfeld reported that misclassification can produce a spurious association between an attribute and a disease though previous statistical papers agreed that errors in classification must produce underestimation of an association. But Newell insisted that the analysis was incorrect.

It is shown that it is necessary to consider the different effects of different kinds of errors. Gross over- or underestimations of an association may result from misclassification and there is no conflict between this fact and the generalization that random errors attenuate a correlation. The basic algebraic model is set forth and the numerical possibilities are indicated in graphs.

REFERENCES

1. Diamond, Earl L., and Lilienfeld, Abraham, M. Effects of Errors in Classification and Diagnosis in Various Types of Epidemiological Studies. A.J.P.H. 52:1137–1144 (July), 1962.
2. Newell, David J. Errors in the Interpretation of Errors in Epidemiology. Ibid. 52:1925–1928 (Nov.), 1962.
3. Diamond, Earl L., and Lilienfeld, Abraham, M. Misclassification Errors in 2x2 Tables with One Margin Fixed: Some Further Comments. Ibid. 52:2106–2110 (Dec.), 1962.
4. Berkson, J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. Biometrics Bull. 2:47–53 (June), 1946.
5. Bross, Irwin. Misclassification in 2x2 Tables. Ibid. 10:478–486 (Dec.), 1954.
6. Kihlberg, Jaakko K., and Keys, Ancel. Effect of Misclassification on Estimated Relative Prevalence of a Characteristic. II. Errors in Two Variables. (See Part II, this paper.)
7. Kihlberg, Jaakko K., and Keys, Ancel. Misclassification and Misclassification Probabilities. (In preparation.)

II. Errors in Two Variables

A treatment of the effects of errors, in the case where one variable only is subject to misclassification, revealed many of the complexities involved.[1] Very often, however, the situation is not as simple as that, and, for example, we must consider conditions where both a disease and some other characteristic are simultaneously subject to misclassification.

In order to build up a mathematical model for this situation, let us state that nature would classify the subjects into four categories, or boxes, according to their true status in respect to the disease and the characteristic:

1. (DC): disease and characteristic present;
2. (Dc): disease present, characteristic absent;
3. (dC): disease absent, characteristic present;
4. (dc): both disease and characteristic absent.

Suppose that the true prevalences in these four boxes are a, b, c, e, respectively, so that $a+b+c+e=1$; this part of the model, therefore, is determined in terms of three essential parameters (any three of the prevalences). The purpose of the diagnosis maker is to estimate these unknown prevalences, and having performed his classification operation, he comes out with estimated prevalences a^*, b^*, c^*, and e^*, $a^*+b^*+c^*+e^*=1$.

Since the investigator may make errors in recognizing the presence or absence of the disease and the characteristic, these estimators may also then be in error, and in the following we shall inquire into the nature of these errors. For this purpose, consider those subjects who truly possess the disease and the characteristic and therefore truly belong to the category (DC) (prevalence a), which we may think of as box 1. In the identification procedure, many of these subjects will be correctly classified as (DC), say a fraction P_{11} of a. However, some will be erroneously classified as (Dc), (dC), or (dc), say fractions P_{12}, P_{13}, and P_{14} of a. In other words, those truly belonging in box 1 (DC) will be classified into four boxes in the following manner:

$a_{11}=aP_{11}$ into (DC), correctly,
$a_{12}=aP_{12}$ into (Dc), "characteristic error,"
$a_{13}=aP_{13}$ into (dC), "disease error,"
$a_{14}=aP_{14}$ into (dc), both errors.

Here, $P_{11}+P_{12}+P_{13}+P_{14}=1$ so that of these transition probabilities three are essential parameters. The situation is illustrated in Figure 1.

Now each of the four true boxes is subject to an error system of its own, with transition probabilities which may or may not be identical from box to box. Since we have three essential parameters for each box, there are $4 \times 3=12$ essential parameters to describe the transition system, and adding the three needed to describe the true prevalences, our model contains 15 essential parameters.

A comprehensive analysis of such a situation is a formidable task and is not attempted here, but we have brought forward the model in order to show how easily different authors may come to differing conclusions if they have omitted the necessary steps

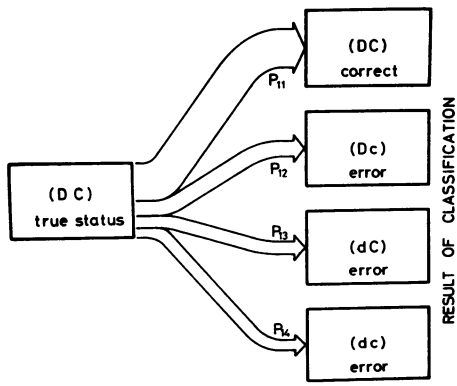

Figure 1—Schematic Example of Misclassification, and Transition Probabilities P_{11}, P_{12}, P_{13}, P_{14}

in model building.[2-5] However, we are able to write down the algebraic formulas for the four estimated prevalences:

$$\begin{aligned} a^* &= a_{11}+b_{21}+c_{31}+e_{41} \\ b^* &= a_{12}+b_{22}+c_{32}+e_{42} \\ c^* &= a_{13}+b_{23}+c_{33}+e_{43} \\ d^* &= a_{14}+b_{24}+c_{34}+e_{44}. \end{aligned} \tag{1}$$

It is now clear that each estimator is an algebraically simple but practically complex function of the several parameters defined in the foregoing. It should also be obvious that it is impossible to say anything about the erroneous behavior of these estimators without specifying the several parameters involved.

If it is insisted that the model be put into a more workable form, some simplifying assumptions must be made which will reduce the number of parameters. Pertinent to recent discussions, let us assume that the presence or absence of disease is known infallibly. Then, the parameter situation is changed as follows:

$a+b$ is known, and
$c+e$ is known;
$P_{13}=P_{14}=P_{23}=P_{24}=P_{31}=P_{32}=P_{41}=P_{42}=0$;
$P_{12}=1-P_{11}$;
$P_{21}=1-P_{22}$;
$P_{34}=1-P_{33}$;
$P_{43}=1-P_{44}$.

We have only four transition parameters left, and the estimators now are:

$$\begin{aligned} a^* &= aP_{11}+b(1-P_{22}), \\ b^* &= bP_{22}+a(1-P_{11}), \\ c^* &= cP_{33}+e(1-P_{44}), \\ e^* &= eP_{44}+c(1-P_{33}). \end{aligned} \tag{2}$$

Now this is nothing but the situation of two fourfold tables discussed previously.[1] To link with the previous notation we observe

$$\begin{aligned} d^* &= a^*/(a^*+b^*), \\ h^* &= c^*/(c^*+e^*), \end{aligned} \tag{3}$$

as estimated prevalences of the characteristic in the diseased and healthy subjects, respectively.

Let us now write

$P_{11}=P_1$,
$P_{22}=P_2$,
$P_{33}=P_3$,
$P_{44}=P_4$,

and

$a+b=1$,
$c+e=1$,

so that

$a=d$,
$c=h$,

and we are, in this notation, in full accord with the previous presentation.[1]

Now the "relative risk" used by Diamond and Lilienfeld[2] is

$$R_S=h^*(1-d^*)/d^*(1-h^*) \tag{4}$$

for the stated prevalences, and

$$R_T=h(1-d)/d(1-h) \tag{5}$$

for the true prevalences. It is obvious that $R_T=1$ if and only if $d=h$, that is, the true prevalences in the diseased and healthy groups are the same, which simply is the null hypothesis we usually want to test. Under the null hypothesis, R_S will be $=1$ if and only if $h^*=d^*$. Clearly, this will be the case if $P_1=P_3$ and $P_2=P_4$, and in some extraordinary combinations of the parameters, but otherwise R_S will be larger or smaller than 1, entirely depending on the parameter configuration. A more comprehensive analysis of the behavior of R_S will be presented elsewhere[5,6]; here it is sufficient to note that if $h^*>d^*$, then $R_S>1$, and if $h^*<d^*$, then $R_S<1$. In other words, Diamond and Lilienfeld's "relative risk" may be an overestimation as well as underestimation of the true "relative risk."

Instead of following Diamond and Lilienfeld we prefer to consider as relative risk the simple and logical quantity

$$\rho_S=d^*/h^* \tag{6}$$

for the stated relative prevalences, and

$$\rho_T=d/h \tag{7}$$

for the true relative prevalences. Again, if $d=h$ (the null hypothesis),

$\rho_T=1$, but ρ_S can be <1, $=1$, or >1, depending on the parameters.

A closer algebraic analysis can be performed by noticing that $d^*=h^*$ implies

$$(8) \qquad dP_1+(1-d)(1-P_2) = dP_3+(1-d)(1-P_4)$$

where we have set $d=h$. Working out this implication we obtain

$$(9) \qquad d/(1-d)=(P_2-P_4)/(P_1-P_3)$$

as a necessary condition for having $d^*=h^*$. In the case $P_2=P_4$, $P_1=P_3$, the above remains indeterminate, but we have already seen that in such a case $d^*=h^*$.

These considerations should suffice to demonstrate that the presence of errors may lead to either overestimation or

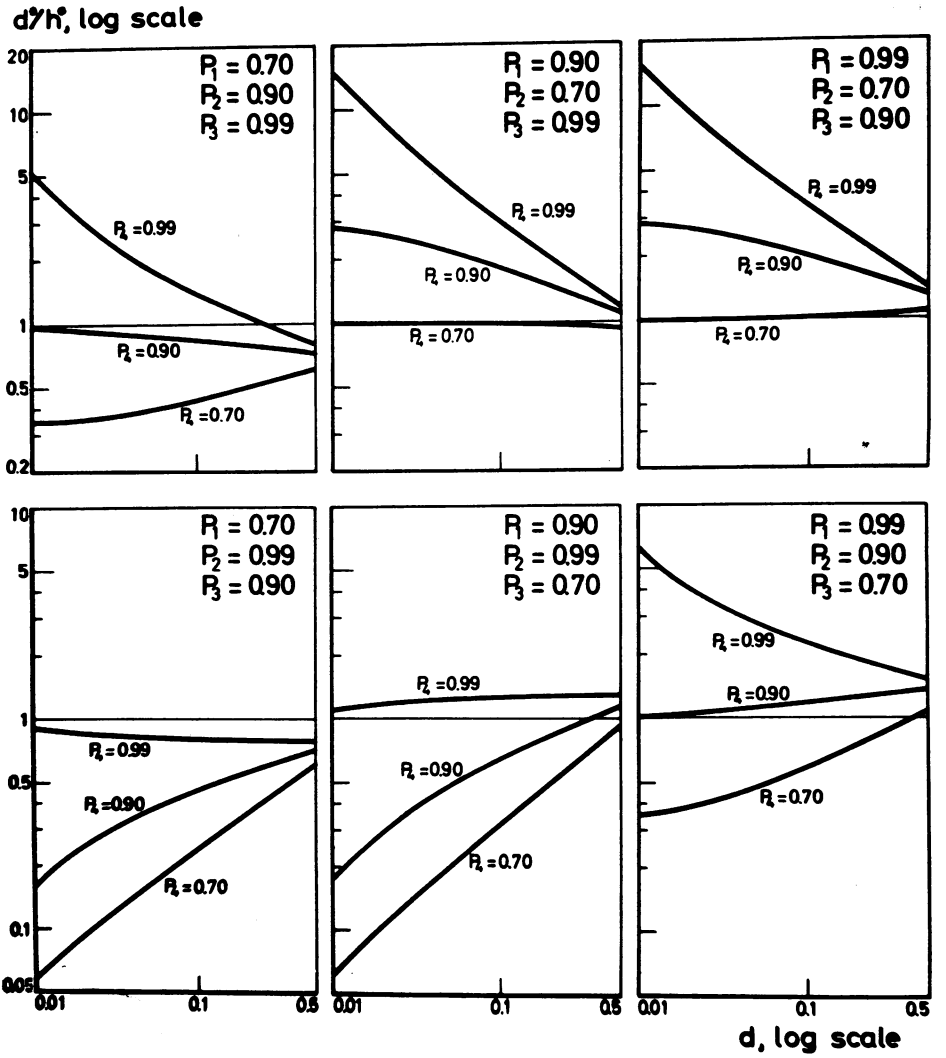

Figure 2—Examples of the Values of the Estimated Relative Risk, d^*/h^*, for Various Combinations of P_1, P_2, P_3, and P_4, and for True Prevalence $d=h$ (Null Hypothesis) from 0.01 Through 0.5

underestimation of a difference in prevalence rates. Unbiased estimation takes place only under specific (and often rare) conditions. These were explored to some extent by Bross.[5]

Figure 2 shows a few examples of the behavior of ρ_R as a function of d, P_1, P_2, P_3, and P_4, demonstrating the wide variations possible under realistic conditions.

In many cases the probability of recognizing presence or absence of a characteristic is independent of the presence or absence of the disease, and vice versa. (This will never be true, of course, if the recognition of presence of the characteristic is considered in diagnosing the disease.) In these situations the number of essential parameters is reduced, for we need only four probabilities to govern the transition system, these four applying to all nature's four categories:

P_D = probability of recognizing the disease;
P_d = probability of recognizing the healthy status;
P_C = probability of recognizing presence of the characteristic;
P_c = probability of recognizing absence of the characteristic.

Then, all 16 transition probabilities can be written in terms of these four parameters, for example:

$$\begin{aligned} P_{11} &= P_D \cdot P_C \\ P_{12} &= P_D \cdot (1-P_C) \\ P_{13} &= (1-P_D)P_C \\ P_{14} &= (1-P_D)(1-P_C). \end{aligned} \tag{10}$$

By appropriate rearrangement of the subscripts the other 12 probabilities can be developed in an analogous manner. The formulas for prevalence estimators still follow the general pattern already given, but with fewer parameters. For example:

$$a^* = aP_DP_C + bP_D(1-P_c) + c(1-P_d)P_C + e(1-P_d)(1-P_c). \tag{11}$$

Although the model has now been greatly simplified, we still have seven essential parameters left, and a detailed analysis is hardly feasible. However, if we again set the null hypothesis d=h, some algebra shows that

$$\begin{aligned} d^* &= a^*/(a^*+b^*) = dP_C + (1-d)(1-P_c) \\ &= c^*/(c^*+d^*) = h^* \end{aligned} \tag{12}$$

so that while the estimators themselves may be biased, the estimated relative risk will always be =1, if h=d, and if identification probability of disease is independent of the presence/absence of the characteristic, and vice versa. In other words, under these premises spurious association cannot occur. This follows algebraically from the restriction set, namely from the independence of probability of recognizing a quality irrespective of whether another quality is present or not. The statement in formula (12) simply is that the P-parameters shown appear in the same manner in both the estimator of d and the estimator of h, and that if d=h, an identity follows. In this form of estimation, the parameters P_D and P_d disappear from the picture.

We have stated that if d=h, then $d^*=h^*$ within the stated restrictions. However, in instances where $d^*=h^*$, we also may have $d \neq h$, and the reader is advised to take careful notice of this. To illustrate, and emphasizing that the probability independence restriction is there:

From theory to observations:
If d=h, then $d^*=h^*$,
If $d \neq h$, then $d^*<h^*$,
or $d^*=h^*$,
or $d^*>h^*$.

From observations to theory:
If $d^*=h^*$, then d<h,
or d=h,
or d>h;
If $d^* \neq h^*$, then d<h,
or d=h,
or d>h.

Finally, it is to be noted that our discussion deals with classification errors only. The effect of random varia-

tion (in our case multinomial) has not been taken into account, but it is clear that such variation will further blur the picture. A comprehensive analysis of this is a complicated problem in theoretical statistics, although useful results in certain restricted situations can be readily obtained.[5]

REFERENCES

1. Keys, Ancel, and Kihlberg, Jaakko K. Effect of Misclassification on Estimated Relative Prevalence of a Characteristic. I. Two Populations Infallibly Distinguished. (See Part I, this paper.)
2. Diamond, Earl L., and Lilienfeld, Abraham, M. Effects of Errors in Classification and Diagnosis in Various Types of Epidemiological Studies. A.J.P.H. 52:1137–1144 (July), 1962.
3. Newell, David J. Errors in the Interpretation of Errors in Epidemiology. Ibid. 52:1925–1928 (Nov.), 1962.
4. Diamond, Earl L., and Lilienfeld, Abraham, M. Misclassification Errors in 2x2 Tables with One Margin Fixed: Some Further Comments. Ibid. 52:2106–2110 (Dec.), 1962.
5. Bross, Irwin. Misclassification in 2x2 Tables. Biometrics Bull. 10:478–486 (Dec.), 1954.
6. Kihlberg, Jaakko K., and Keys, Ancel. Misclassification and Misclassification Probabilities. (In preparation.)

The authors are associated with the Laboratory of Physiological Hygiene, University of Minnesota School of Public Health, Minneapolis, Minn.

This study was aided by a grant from the U. S. Public Health Service, No. HE-04997-03.

Credo of a Public Health Practitioner

At the alumni dinner honoring Hugh R. Leavell, M.D., retiring professor of public health practice at the Harvard School of Public Health, he summed up his years of public health as follows:

"I believe in:

The dignity and importance of putting public health into practice;

The need for the kind of teamwork that is built on respect for the essential contributions of other professions, and on understanding of their professional strivings;

The real value of group discussion and decision, recognizing that times come when the leader must decide;

The concept of comprehensive health care as a unifying idea to coordinate the work of many and sometimes divergent forces;

The importance of the social sciences in helping the practitioner do a more effective job; in systematizing knowledge of the community; and in providing research tools and concepts which can add new knowledge;

Health education as the channel which brings to the people the fruits of the laboratory;

Mental Health as a key to improved interpersonal relationships and as a support to the weak—and to the strong in their weak moments;

Community diagnosis to discover what our community patient needs, what he wants, and how he may be reached;

International health as not only a fascinating career, but also as a bridge for two-way traffic bringing nations together;

The inescapable responsibility of the health officer to view his community as patient, and to be impatient with any who persistently refuse to play their proper roles in the total community health enterprise."

(From August, 1963, "Harvard School of Public Health Alumni Bulletin.")