Quantification of Agreement in Psychiatric Diagnosis Revisited

Patrick E. Shrout, PhD; Robert L. Spitzer, MD; Joseph L. Fleiss, PhD

Eighteen years ago in this journal, Spitzer and colleagues[1] published "Quantification of Agreement in Psychiatric Diagnosis," in which they argued that a new measure, Cohen's κ statistic,[2] was the appropriate index of diagnostic agreement in psychiatry. They pointed out that other measures of diagnostic reliability then in use, such as the total percent agreement and the contingency coefficient, were flawed as indexes of agreement since they either overestimated the discriminating power of the diagnosticians or were affected by associations among the diagnoses other than strict agreement. The new statistic seemed to overcome the weaknesses of the other measures. It took into account the fact that raters agree by chance alone some of the time, and it only gave a perfect value if there was total agreement among the raters. Furthermore, generalizations of the simple κ statistic were already available. This family of statistics could be used to assess classifications into two or more categories, and could be tailored to situations when different disagreements were unequally serious.

In the years following that first article on κ in the ARCHIVES, its message lost novelty as the κ statistic became widely known. Indeed, κ has become the standard method for assessing diagnostic agreement in psychiatry and other medical specialties[1,3,4] and the mathematical details involved in its calculation and application are now available in statistics and methods textbooks.[5,6] Despite its widespread endorsement, several authors have argued that κ has weaknesses that limit its usefulness. One set of criticisms[7,8] centers on the definition of chance agreement employed by κ, although, as has been discussed by Grove et al,[9] the alternative definitions of chance agreement are far less attractive. A different criticism of κ was raised in the same article by Grove et al[9]; using a model developed by Kraemer,[10] they argued that κ is not useful when base rates of disorders are low, such as in community studies. Concern with this so-called base rate problem has led Spitznagel and Helzer[11] to suggest that κ be replaced with an old statistic, Yule's Y,[12] even though it has many of the weaknesses of the measures that were being used at the time the article by Spitzer et al[1] was published.

In the present article we first review basic concepts of diagnostic reliability that are often misunderstood and discuss the advantages that the family of κ statistics has in its quantification. We then examine the asserted problem with κ when base rates are low and point out serious flaws in the proposal by Spitznagel and Helzer[11] to use Yule's Y as a measure of agreement. Finally, we discuss the important implications of reliability and its mismeasurement in terms of recent empirical studies of diagnostic reliability in general community populations. Except for a small number of well-marked paragraphs, we will avoid mathematical technicalities in our discussion.

Accepted for publication Jan 31, 1986.

From the New York State Psychiatric Institute, New York (Drs Shrout, Spitzer, and Fleiss); and the Division of Biostatistics, School of Public Health (Drs Shrout and Fleiss), and the Department of Psychiatry (Dr Spitzer), Columbia University, New York.

Reprint requests to Division of Biostatistics, Columbia University School of Public Health, 600 W 168th St, New York, NY 10032 (Dr Shrout).

RELIABILITY AND ITS IMPORTANCE

Reliability in the psychometric sense is the reproducibility of distinctions made between some aspects of persons. It is important to note that mere replication without discrimination is not enough; the replication must be in terms of ordering, categorizing, or otherwise discriminating among the persons or objects. While a car that always starts may be reliable in lay terms, a clinician who always gives the same diagnosis is not reliable in psychometric terms. Unless he or she distinguishes among patients, psychometric reliability cannot be demonstrated.

Subtly implied by the requirement that persons or objects must be distinguished for reliability to be defined is that the reliability of a measure is specific to a population. A measure that is reliable when used in a heterogeneous population may not be reliable in a more homogeneous population. For example, an IQ test that reliably distinguishes mentally retarded from normal adolescents may be very unreliable in ranking college-bound students according to their cognitive aptitudes.

Failure to reproduce a series of diagnoses or measurements usually implies that the assessments are affected by some source of variation other than that of the subject attribute under study. In the case of diagnostic measurement there are a variety of sources of variation that may result in unreliability.[13] One is error during the information-gathering phase of diagnosis (information variance): a respondent may provide incorrect information due to misunderstanding, lapse of concentration, or intentional resistance, and a diagnostician may err in the choice and phrasing of questions and recording of responses. Another may be the instability of the clinical phenomenon being measured (occasion variance): the respondent may respond truthfully to a well-posed question, but the answer may change each time the question is asked as the condition of the respondent changes. Yet another possible source of variation is an idiosyncratic set of diagnostic criteria employed by the diagnostician (criterion variance). If different clinicians have different concepts of a disorder, then the diagnostic measure will change depending on which clinician is chosen to make the diagnosis. Finally, variability may result from uncareful, inconsistent, or incompetent inference on the part of the diagnostician.

In most clinical and epidemiological diagnostic research, all sources of nonsubject (ie, error) variation in diagnoses are considered when assessing reliability. The reproducibility of the diagnosis that is of interest includes reassess-

Table 1.—Hypothetical Results From a Reliability Study*

Clinician 2	Clinician 1, No. of Cases (Cell): Diagnosis Positive	Clinician 1, No. of Cases (Cell): Diagnosis Negative	Total No. of Cases
Diagnosis positive	3 (a)	3 (b)	6
Diagnosis negative	3 (c)	91 (d)	94
Total No. of Cases	6	94	100

*Observed agreement: $P_o = 0.940$. Expected chance agreement: $P_c = 0.887$. $\kappa = (P_o - P_c) / (1 - P_c) = 2 \times (ad - bc) / [2 \times (ad - bc)] + [N \times (b + c)] = 0.468$. Odds ratio $\omega = ad / bc = 30.333$. Yule's $Y = \sqrt{\omega} - 1 / \sqrt{\omega} + 1 = 0.693$.

ment by another diagnostician on another day in another setting; the ideal is that diagnosticians are completely interchangeable, as would be instruments for physical measurement manufactured by the same company. The ideal of interchangeability implies that consistent differences between examiners are also considered nonsubject variation; if consistent differences exist, then one or more examiners are said to have a constant bias. For example, one might observe that when one clinician makes a diagnosis of borderline personality disorder, another clinician tends to make a diagnosis of affective disorder. The need generally to take account of constant biases in assessing reliability is discussed in an article by two of us (P.E.S. and J.L.F.).[14]

In theory, the assessment of reliability calls for independent applications of a measurement procedure so that agreement can be determined. In practice, completely independent measurements are rarely possible, since the respondent is usually affected by the diagnostic interview. For example, the respondent may misunderstand that the second assessment is supposed to collect new information, or may deliberately attempt to be consistent across interviews. Another practical difficulty is that if much time elapses between the first and second measurement, a clinical phenomenon of interest may actually change; in this case it is not possible to measure reliability properly. The strengths and weaknesses of different designs for assessing reliability have been discussed elsewhere[9,15,16] and are not an issue here. In our discussion below, we will assume a reliability study design that calls for two separate diagnostic assessments of relatively stable psychiatric conditions.

It is a psychometric truism that the validity of a measure is limited by its reliability.[17] This is obviously the case if the measure is totally unreliable, since its values are completely random by definition. If the measure has mediocre or poor reliability, its validity will suffer to some degree. Because this principle of measurement has been questioned in this journal,[15,18] we briefly review the statistical literature that documents the effects of employing fallible measures in research. For continuous or quantitative measures, correlations between fallible measures and validation criteria are systematically attenuated by unreliability,[17] and multivariate statistical procedures produce biased results.[19] Prevalence estimates based on fallible classifications are biased, and for diseases that are rare the bias is usually in the direction of overestimation.[20,21] Assessments of risk factors using unreliable diagnostic variables can produce either overestimation or underestimation of the strength of association, depending on the pattern of unreliability in the exposed and nonexposed groups.[22-24] Large sample sizes are no protection against these systematic biases that sometimes hide strong associations and other times create associations when there are none. In practical terms, unreliable diagnosis can result in the wrong treatment of patients, intensive study of the wrong groups of persons in prospective or retrospective studies, and granting of a clean bill of health to the wrong persons in mental health screening exercises.

While reliability is necessary for validity, high reliability is not sufficient to guarantee validity. The scientific process of validation should begin with the reliability study and should continue well beyond the documentation of good reliability. Carey and Gottesman[18] have discussed some of the difficulties involved in reconciling reliability and validity results.

MEASURING DIAGNOSTIC AGREEMENT WITH κ

Suppose 100 community respondents are assessed by two clinicians, each of whom makes a diagnosis of any *DSM-III* mental disorder.[25] Table 1 shows a hypothetical set of results from such an exercise. In this example, each clinician makes a *DSM-III* diagnosis in only 6% of the cases; the vast majority are called noncases by both. It is tempting to note that in 94 of the 100 ratings the clinicians agreed, but as pointed out by Cohen[2] and by Spitzer et al,[1] chance agreement can produce very high values of total percent agreement. For example, if neither clinician interviewed any of the subjects but both simply randomly assigned 6% of them to the case group—perhaps because they expected that the prevalence of a current *DSM-III* disorder in a general population would be low—they would be expected to agree on the noncases 88.4% (which is 0.94^2) of the time, and they would be expected to agree on the cases (any *DSM-III* diagnoses) about 0.4% of the time. Thus, with a base rate of 6%, chance agreement would produce an overall rate of agreement of about 88.8%.

In the example in Table 1, the clinicians actually do better than what would be expected by chance. The difference $94\% - 88.8\% = 5.2\%$ represents their improvement over chance. The best improvement possible is 11.2% ($100\% - 88.8\%$). The κ statistic is defined as the proportion of the best possible improvement actually obtained by the clinicians; in this case, $\kappa = 5.2/11.2 = 0.46$. Thus, chance-corrected agreement in Table 1 is about half of what is possible. (This value of κ differs slightly due to rounding error from the result shown in the bottom of Table 1.)

The example in Table 1 illustrates the simplest application of κ. Using the basic definition, $\kappa = (P_o - P_c)/(1 - P_c)$, where P_o is the proportion of observed agreement and P_c is the proportion of agreement expected by chance, a whole family of κ statistics has been defined. These include indexes for assessing agreement when several rather than two diagnostic categories are used,[2] when several rather than two clinicians are used,[26,27] and when disagreements vary in diagnostic importance.[28] An article by one of us (J.L.F.)[6] contains details about these and other κ statistics.

BASE RATES AND RELIABILITY

We made the point in the first section that reliability is more difficult to attain in homogeneous than in heterogeneous populations. This fact can be shown formally by expressing reliability as the proportion of the total variance that is not attributable to random error. For a fixed amount of error, this proportion will decrease as the total variance decreases, ie, as more homogeneous populations are sampled. As an illustration, consider a thermometer whose temperature readings have an error of up to 1°C. While it may be quite reliable for discriminating cold weather from hot weather, it obviously would not be reliable for detecting

Table 2.—Model for Observed Agreement Between Two Clinicians*

Clinician 2	Clinician 1: Diagnosis Positive	Clinician 1: Diagnosis Negative	Column Total
Diagnosis positive	$B(S_n)^2 + (1-B)(1-S_p)^2$	$B(S_n)(1-S_n) + (1-B)(1-S_p)(S_p)$	$B(S_n) + (1-B)(1-S_p)$
Diagnosis negative	$B(S_n)(1-S_n) + (1-B)(1-S_p)(S_p)$	$B(1-S_n)^2 + (1-B)(S_p)^2$	$B(1-S_n) + (1-B)(S_p)$
Row total	$B(S_n) + (1-B)(1-S_p)$	$B(1-S_n) + (1-B)(S_p)$	1.0

*Each clinician is assumed to have the same sensitivity (S_n) and specificity (S_p) when the base rate is B.

abnormalities in body temperature. In the physical sciences it is well accepted that the expected variation of measurements (expressed in terms of the range) must be taken into account when evaluating an instrument's error level. An instrument that is quite acceptable in one context may be totally unacceptable in another.[29]

The same principle holds for diagnostic judgments. When a dichotomous characteristic is measured, the homogeneity of the population is determined by the proportion of respondents who possess the characteristic. In epidemiologic studies this proportion is termed the *prevalence* of the disorder, and in clinical studies it has been called the *base rate*.[18] A population with maximum heterogeneity is one with a base rate of 50%. As the rate approaches either 0% or 100% the population becomes more homogeneous, and the same number of diagnostic disagreements will have a greater impact on the unreliability of the diagnosis. A major strength of κ is precisely that it does weigh disagreements more when the base rate approaches 0% or 100%.

The difficulty in obtaining reliable diagnoses in homogeneous populations (those with low prevalences) was termed the *base rate problem* by Carey and Gottesman.[18] Consistent with our discussion above, they warned that because a diagnostic procedure has been shown to be reliable in a heterogeneous population (such as a clinical sample), it cannot be assumed to be reliable in a more homogeneous population (such as a community sample). Three years later, Grove et al,[9] using a characterization of κ based on work by Kraemer,[10] presented a different interpretation of the base rate problem: they concluded that the problem was not the demand placed on a procedure when it is used to study rare disorders but that the problem was the reliability statistic, κ. As we see it, they, in effect, invoked the centuries-old rule of killing the messenger who brings bad news. They also argued that since the reliability of a given diagnostic procedure tends to go down when that procedure is applied to a population with a low base rate, reliability should not be studied or reported when the base rate is less than 5%. This would rule out the study of reliability of all but a handful of diagnoses in the community.

In a recent article, Spitznagel and Helzer[11] built on the misinterpretation by Grove et al.[9] Beginning with their title, which includes the phrase "the base rate problem in the kappa statistic," they perpetuated the incorrect idea that it is κ that is afflicted with a base rate problem. They proposed that κ be reserved for conditions with a relatively high prevalence rate, and that it be replaced with Yule's Y,[12] a long-established measure of *association*, whenever the prevalence rate is low. Because their article will undoubtedly attract special attention as a result of its citation in reports of the National Institute of Mental Health Epidemiologic Catchment Area program,[30] we will provide a detailed critique of their analysis, and in a subsequent section will examine the usefulness of Yule's Y for measuring reliability.

Spitznagel and Helzer[11] reported an analysis of κ under two psychometric models. The first, which they termed the *validity model*, assumes that the true clinical status is known and that only one fallible diagnosis is under study. Under this model they showed that for fixed sensitivity (the proportion of true cases correctly classified as cases by the fallible procedure) and fixed specificity (the proportion of true noncases correctly classified as noncases by the fallible procedure), the value of κ for the agreement between the true and the fallible diagnoses will vary as the base rate for the true classification varies. They also showed that Yule's Y does not vary and that it gives exactly the same value regardless of the base rate of the true classification, when specificity and sensitivity are fixed.

Spitznagel and Helzer's mathematics is correct, but their analysis is irrelevant for reliability. A validity model is simply not appropriate for evaluating a reliability statistic such as κ. When a diagnostic criterion is available (eg, when a diagnostic screening procedure is tested using an extensive clinical evaluation as the criterion), the proper statistics are sensitivity, specificity, positive predictive value (the proportion of putative cases who are truly cases), and negative predictive value (the proportion of putative noncases who are truly noncases).[31] Neither κ nor Yule's Y, nor any other measure of agreement or association, provides information about false-positives and false-negatives, yet this information is precisely what is sought in a validity study. Since neither κ nor Yule's Y should be used when a criterion is available, the variation of these statistics with the criterion base rate is moot.

The second psychometric model presented by Spitznagel and Helzer[11] represents a formulation of a genuine reliability study in which two fallible diagnostic procedures are compared. Suppose that dichotomous diagnostic evaluations are made by two clinicians who both have the same sensitivity and specificity relative to an unmeasured criterion. If it can be assumed that within the case group and within the noncase group the raters make independent random errors, then the data in the cross-classification table can be represented as in Table 2. This representation is the same as given by Carey and Gottesman[18] and is consistent with mathematical models of latent classes.[32] Using this model and assuming certain fixed values for sensitivity and specificity, Spitznagel and Helzer showed (in their Fig 3) that both κ and Yule's Y vary as the base rate of the disorder varies. Our Fig 1 (after Fig 3 of Spitznagel and Helzer) shows how the measures vary for a sensitivity of 0.80 and a specificity of 0.98. Our Fig 2 shows how they vary for a sensitivity of 0.40 and a specificity of 0.98; these latter values are similar to those reported by two sites of the Epidemiological Catchment Area Survey for major depression.[33,34]

Our figures indicate that both κ and Yule's Y will tend to go down if reliability studies are performed in populations with base rates near 1% or 0%. From our discussion of reliability and base rates, the results for κ—a true measure of reliability[35]—are as expected. Yule's Y seems also to be

table, and as such has all of the strengths of the odds ratio as a measure of association.[5(pp61-67)] In the notation of Table 1, the odds ratio, ω, is (ad)/(bc), and the Y statistic is

$$Y = (\sqrt{\omega} - 1)/(\sqrt{\omega} + 1).$$

The definition of Yule's Y reveals its most obvious problem: its lack of interpretability. The Y statistic is a function of the odds ratio, but the function involves taking the square root of ω. This nonlinear transformation of ω has no apparent intuitive appeal. Spitznagel and Helzer[11] claim that Y can be interpreted as a κ value when (1) the clinicians each have the same sensitivity and specificity relative to a diagnostic criterion, (2) the sensitivities are equal to the specificities, and (3) the base rate is 50%. While this may be true, these conditions are so restrictive that they almost never will apply. When they do not apply, it is not possible to interpret Y as a true reliability coefficient.

When two clinicians are in good agreement, they tend to give the same diagnoses to the same persons, and their diagnostic rates tend to be similar. In such instances, there are few disagreements (ie, relatively few entries in cells b and c of Table 1), the counts in b and c are similar in magnitude, the table displays symmetry, and both κ and Yule's Y may be expected to be large. A problem with Y is that it (but not κ) may be large if one of the cells, b or c, has many entries while the other is empty or nearly empty.

To appreciate this undesirable property of Yule's Y, consider a variation of Table 1 in which the six disagreements are in cell c and none are in cell b. The value of Yule's Y for the modified table is 1.0, which reflects the fact that, for this table, clinician 1 gives a positive diagnosis whenever clinician 2 does. No one would cite this as an example of perfect agreement, though, since clinician 2 agrees on fewer than half of the subjects given a positive diagnosis by clinician 1. Spitznagel and Helzer[11] acknowledge this problem with the Y statistic and recommend an adjustment method[6(p432)] for reducing the value of Y when the count in cell b or c is 0 (their adjustment method yields the value Y = 0.675 for the modified table). There is nothing wrong with statistical adjustments when they are mathematically necessary,[6(p64)] but we suggest that Spitznagel and Helzer's adjustment in the measurement of reliability is an unnecessary complication brought about by their recommendation to use an unnecessarily complicated statistic, Yule's Y. The κ statistic, in contrast, requires no adjustment and yields a value for the modified table, 0.476, that is virtually identical to the value for the original data in Table 1.

A third problem with Yule's Y is that it is limited to analyses of fourfold tables and cannot be generalized to other reliability designs. Even if Y were interpretable as a reliability statistic, it would be imprudent for researchers to abandon the family of κ statistics, which includes forms that are applicable to reliability designs that involve multiple diagnostic categories, multiple raters, and even varying numbers of rates.[5(pp215-232)]

The most important reason to avoid Yule's Y as an index of reliability is that it inevitably will mislead researchers into thinking that measurement error is not a problem when in fact it is. As shown in Spitznagel and Helzer's own figures and in our Figs 1 and 2, Y gives consistently higher values than κ for the same level of error. Since κ is interpretable as a reliability coefficient, the difference between Y and κ must be regarded as bias—that is, Y consistently overstates the true reliability. This bias is especially pronounced when the base rate is low, and thus is likely to be most problematic when Y is applied to reliability results from epidemiologic surveys. Regrettably, it is this most misleading application that is strongly endorsed by Spitznagel and Helzer.

One of the important practical features of κ is its interpretability in qualitative as well as quantitative terms. Values greater than approximately 0.75 are generally taken to indicate excellent agreement beyond chance, values below approximately 0.40 are generally taken to represent poor agreement beyond chance, and values in between are generally taken to represent fair to good agreement beyond chance.[5(p218),37] Comparable standards do not exist for Yule's Y, so there is no way of judging, for example, whether the value Y = 0.693 for the data in Table 1 represents poor, fair, or good chance-corrected agreement. The reader who takes such a value as indicating good agreement beyond chance will have been misled. If the reliability data in Table 1 are from a subset of subjects in a substantive study in which the positives will serve to determine the numerator of an estimated prevalence rate, or will constitute the cases in a case-control study, the results of that study might be suspect: only half the subjects identified as positive by one of the raters will be so identified by the other. The Y value of 0.693 may incorrectly suggest to the investigator that reliability is adequate, whereas the κ value of 0.468 correctly warns the investigator that reliability is mediocre.

COMMENT

The recommendations made by Spitzer et al[1] for the quantification of diagnostic agreement have, with few exceptions, been well received by the psychiatric research community over the past 18 years. The family of κ statistics[2] has proved to be extremely useful and versatile in the testing and development of diagnostic procedures and diagnostic criteria.

As psychiatric researchers turn their attention to evaluating mental disorders in nonclinical populations, obtaining diagnostic reliability will prove to be even more challenging than before. The source of this challenge is the relatively low rate of disorder in nonclinical populations. Since few true-positive cases are expected, even a small number of false-positives may undermine the overall reliability of the procedure. The low rate of disorder also inspires modifications in the design of reliability studies. To provide a sufficient number of positive cases to obtain stable measurement of agreement, some investigators[33,34] have oversampled for the reliability study respondents who were diagnosed as cases in the first assessment. Since this sampling plan artificially increases the base rate in the reliability subsample, it is necessary to reconstruct through weighting of the likely pattern of agreement in the original population before a reliability statistic is computed. The κ statistic can be adapted to stratified reliability designs and, when properly computed, accurately reflects the challenge to reliability inherent in the study of rare disorders.

Contrary to the arguments by Grove et al[9] and Spitznagel and Helzer,[11] there is no base rate problem with the κ statistic. Across all base rates, the maximum κ value is 1.0, indicating perfect agreement. Actual examples of acceptable κ values obtained in samples with very low base rates of certain disorders are available in Appendix F of the *DSM-III*.[25] Table 1 of that appendix, which lists κ statistics and base rates for adult disorders from two phases of reliability studies involving pairs of clinicians, includes the following results: mental retardation (phase 1 κ = .80, phase 1 base rate = 1.8%; phase 2 κ = .83, phase 2 base rate = 2.1%), dementias arising in the senium and presenium (κ_1 = .85, base $rate_1$ = 2.4%; κ_2 = .91, base $rate_2$ = 1.8%), and psychosexual disorder (κ_1 = .92, base $rate_1$ = 2.1%; κ_2 = .75, base $rate_2$ = 1.5%). These results demonstrate that there is no mathematical necessity for small κ values with low sample

base rates, as implied by the discussions of the alleged base rate problem with κ.

Abandoning the use of a standard reliability statistic can result in confusion in the psychiatric research literature. For example, in the abstract of a recent report on the ECA reliability trials in Saint Louis, Helzer et al[33] reported that when lifetime diagnoses made by the ECA's Diagnostic Interview Schedule[38] (DIS) were compared with *DSM-III* diagnoses made by psychiatrists, "chance corrected concordance was 0.60 or better for eight of the 11 diagnoses." The reader is left with the impression that the DIS *was generally in good agreement* with the clinical diagnoses. In another article on the ECA reliability trials in Baltimore, Anthony et al[34] reported that "the chance-corrected degree of agreement between the DIS and psychiatrists' one month diagnosis was moderate for . . . [one diagnosis] and lower for the other mental disorder categories." In this case the reader is certainly left with the impression that the DIS *was generally not in agreement* with the clinical diagnoses. While one might begin to speculate that the time frames used with the DIS or the form of the structured clinical interview might account for the different results, such a substantive analysis would be premature. The abstract of the article by Helzer et al[33] referred to results obtained with Yule's Y, while the abstract of the article by Anthony et al[34] referred to results obtained with κ. The text of the article by Helzer et al did contain the κ results comparable with those obtained by Anthony et al. From these statistics one finds that only one of the 11 diagnoses has a reliability at the 0.60 level or better. We suggest that the impression left by Anthony et al[34] is the correct one: for most of the diagnoses studied, the agreement in community samples between the DIS and clinical diagnoses is poor.

Some readers might quarrel with this application of κ since the diagnoses being compared are not from the same diagnostic method. While it can be argued that κ is appropriate because both diagnostic methods are fallible and hence should be treated equivalently in the analysis, a more satisfying analysis of the reliability of the DIS itself would come from a design in which the same respondent was interviewed twice by lay interviewers using the DIS. To our knowledge, no such reliability study has yet been reported on a community sample. One approximation to this pure reliability study is the test-retest study of the DIS reported by Helzer et al[33] in which the respondent was first interviewed by a lay interviewer and then by a psychiatrist, both using the DIS. The results of this study, while somewhat better than the agreement between the lay DIS and clinical diagnoses, still indicated that the majority of the diagnoses are not reliable in the community. More than one half of the κ values are less than 0.30 and only two are better than 0.60. The DIS represents an advance in structured diagnostic methods that can be applied to community samples, but more work is needed to improve the reliability of diagnoses in these samples. The κ statistic will provide investigators working on this problem with a valid quantification of chance-corrected diagnostic agreement in the general population.

This study was supported by National Institute of Mental Health grants MH 37393 and MH 30906.

References

1. Spitzer RL, Cohen J, Fleiss JL, Endicott J: Quantification of agreement in psychiatric diagnosis. *Arch Gen Psychiatry* 1967;17:83-87.
2. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46.
3. Koran LM: The reliability of clinical methods, data and judgments: I. *N Engl J Med* 1975;293:642-646.
4. Koran LM: The reliability of clinical methods, data and judgments: II. *N Engl J Med* 1975;293:695-701.
5. Fleiss JL: *Statistical Methods for Rates and Proportions*, ed 2. New York, John Wiley & Sons Inc, 1981.
6. Bishop YM, Fienberg SE, Holland PW: *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass, MIT Press, 1975.
7. Maxwell AE: Coefficients of agreement between observers and their interpretation. *Fr J Psychiatry* 1977;130:79-83.
8. Janes CL: Agreement measurement and the judgment process. *J Nerv Ment Dis* 1979;167:343-347.
9. Grove WM, Andreasen NC, McDonald-Scott P, Keller MB, Shapiro RW: Reliability studies of psychiatric diagnosis: Theory and practice. *Arch Gen Psychiatry* 1981;38:408-413.
10. Kraemer HC: Ramifications of a population model for kappa as a coefficient of reliability. *Psychometrika* 1979;44:461-472.
11. Spitznagel EL, Helzer JE: A proposed solution to the base rate problem in the kappa statistic. *Arch Gen Psychiatry* 1985;42:725-728.
12. Yule GU: On the methods of measuring association between two attributes. *J R Statist Soc* 1912;75:581-642.
13. Spitzer RL, Endicott J, Robins E: Clinical criteria and *DSM-III*. *Am J Psychiatry* 1975;132:1187-1192.
14. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979;86:420-428.
15. Robins LN: Reflections on testing the validity of psychiatric interviews. *Arch Gen Psychiatry* 1985;42:918-924.
16. Helzer JE, Robins LN, Taibleson M, Woodruff RA, Reich T, Wish ED: Reliability of psychiatric diagnoses: I. A methodological review. *Arch Gen Psychiatry* 1977;34:129-133.
17. Lord FM, Novick MR: *Statistical Theories of Mental Test Scores*. Reading, Mass, Addison-Wesley Publishing Co, 1968.
18. Carey G, Gottesman II: Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. *Arch Gen Psychiatry* 1978;35:1454-1459.
19. Fleiss JR, Shrout PE: The effect of measurement error on some multivariate procedures. *Am J Public Health* 1977;67:1184-1189.
20. Bross I: Misclassification in 2 × 2 tables. *Biometrics* 1954;10:478-486.
21. Keys A, Kihlberg JK: The effect of misclassification on estimated relative prevalence of a characteristic. *Am J Public Health* 1963;53:1656-1665.
22. Goldberg JD: The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J Am Stat Assoc* 1975;70:561-567.
23. Diamond EL, Lillienfeld AM: Effects of errors in classification and diagnosis in various types of epidemiological studies. *Am J Public Health* 1962;52:1137-1144.
24. Diamond EL, Lillienfeld AM: Misclassification errors in 2 × 2 tables with one margin fixed: Some further comments. *Am J Public Health* 1962;52:2106-2110.
25. American Psychiatric Association, Committee on Nomenclature and Statistics: *Diagnostic and Statistical Manual of Mental Disorders*, ed 3. Washington, DC, American Psychiatric Association, 1980.
26. Fleiss JL, Nee JC, Landis JR: The large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 1979;86:974-977.
27. Davies M, Fleiss JL: Measuring agreement for multinomial data. *Biometrics* 1982;38:1047-1051.
28. Fleiss JL, Cohen J, Everitt BS: Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969;72:323-327.
29. Harris FK: Measurement error, in Considine DM, Ross SD (eds): *Handbook of Applied Instrumentation*. New York, McGraw-Hill International Book Co, 1964, pp 12-27.
30. Regier DA, Myers JK, Kramer M, et al: The NIMH epidemiologic catchment area program. *Arch Gen Psychiatry* 1984;41:934-941.
31. Baldessarini RJ, Finklestein S, Arana GW: The predictive power of diagnostic tests and the effect of prevalence of illness. *Arch Gen Psychiatry* 1983;40:569-573.
32. Lazarsfeld PF, Henry NW: *Latent Structure Analysis*. Boston, Houghton Mifflin Co, 1968.
33. Helzer JE, Robins LN, McEvoy LT, Spitznagel EL, Stoltzman RK, Farmer A, Brockington IF: A comparison of clinical and DIS diagnoses: Physician reexamination of lay interviewed cases in the general population. *Arch Gen Psychiatry* 1985;42:657-666.
34. Anthony JC, Folstein M, Romanoski AJ, Von Korff MR, Nestadt GN, Chahal R, Merchant A, Brown CH, Shapiro S, Kramer M, Gruenberg EM: Comparison of lay DIS and standardized psychiatric diagnosis: Experience in eastern Baltimore. *Arch Gen Psychiatry* 1985;42:667-675.
35. Fleiss JL, Cohen J: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973;33:613-619.
36. Endicott J, Spitzer RL, Fleiss JL: The Global Assessment Scale. *Arch Gen Psychiatry* 1976;33:766-771.
37. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
38. Robins LN, Helzer JE, Croughan J, Ratcliff KS: The NIMH Diagnostic Interview Schedule: Its history, characteristics and validity. *Arch Gen Psychiatry* 1981;38:381-389.