Quantification of Agreement in Psychiatric Diagnosis Revisited

Patrick R. Shroot, Ph.D.; Robert L. Spitzer, M.D.; Joseph L. Fleiss, Ph.D.

Eighteen years ago in this journal, Spitzer and colleagues published "Quantification of Agreement in Psychiatric Diagnosis," in which they argued that a new measure, Cohen's \( \kappa \) statistic, 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 in indexes of agreement since they either overestimated the discriminating power of the diagnostician 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 \( \kappa \) 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 \( \kappa \) in the ARCHIVES, its message lost novelty as the \( \kappa \) statistic became widely known. Indeed, \( \kappa \) has become the standard method for assessing diagnostic agreement in psychiatry and other medical specialties and the mathematical details involved in its calculation and application are now available in statistics and methods textbooks. Despite its widespread endorsement, several authors have argued that \( \kappa \) has weaknesses that limit its usefulness. One set of criticisms centers on the definition of chance agreement employed by \( \kappa \), although, as has been discussed by Grove et al., the alternative definitions of chance agreement are far less attractive. A different criticism of \( \kappa \) was raised in the area of diagnostic categories, namely a concern that the measures of agreement that were being used at the time the article by Spitzer et al. 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 \( \kappa \) statistics has in its quantification. We then examine the asserted problem with \( \kappa \) when observed rates are low and point out flaws in the proposal by Epstein and Helzer to use \( \kappa \) as a measure of agreement. Finally, we discuss the implications of reliability and its measurement in terms of recent empirical studies of diagnostic reliability in general community populations. Except for a smaller number of well-marked paragraphs, we will avoid mathematical technicalities in our discussion.

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 his 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 sources 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 be unreliable, uninformative, or irrelevant: 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 over time; the question may be asked at the condition of the respondent changes. Yet another possible source of variation is idiosyncratic set of diagnostic criteria employed by the diagnostician (criterion variance). If all of these 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 unhelpful, inconsistent, or incompetent inference on the part of the diagnostician.

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

Accepted for publication June 30, 1986.

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

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

172 Arch Gen Psychiatry—Vol 44, Feb 1987
goal, by another diagnostician on another day in another setting, the ideal is that clinicians 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 variables. If consistent differences exist, then one or more examiners are said to have a consistent bias. For example, it might be observed that when one clinician makes a diagnosis of borderline personality disorder, another clinician fails to make a diagnosis of affective disorder. The need literally to take account of consistent biases in assessing reliability is discussed in an article by two of us (P.E.S. and T.F.)[a].

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 subject 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 strength and weaknesses of different designs for assessing reliability have been discussed elsewhere[a] and are not in the current focus here. In our discussion below, we will assume a reliaility 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. 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,[a][b] we briefly review the statistical literature that documents the effects of involving faulty measures in research. For continuous or quantitative measures, correlations between feasible measures and validation criteria are systematically attenuated by unreliability, and multi variate statistical procedures produce biased results.[c] Prevalence estimates based on reliable classifications are biased, and for diseases that are rare the bias is usually in the direction of overestimation. 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.[d] Large sample sizes are the protection against these systematic biases that sometimes hide strong associations and other times create associations when there are none. In practical terms, unreliable diagnoses 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 validating studies begin with uncovering a base reliability study and should continue well beyond the documentation of good reliability. Casey and Gotfredsen[5] have discussed some of the difficulties involved in reconfirming reliability and validity results.

MEASURING DIAGNOSTIC AGREEMENT WITH K

Suppose 100 community respondents are assessed by two clinicians, each of whom makes a diagnosis of any DSM-III mental disorder.[2] Table 1 shows a hypothetical set of results from such an exercise. In this example, each clinician makes a DSM-III diagnosis in only 62% 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[6] and by Fleiss et al.,[2] 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 0.96) of the time, and they would be expected to agree on the cases (any DSM-III diagnoses) about 0.8% of the time. Thus, with a base rate of 6%, chance agreement would produce an overall rate of agreement of about 88.8%.[7]

In the example in Table 1, the clinicians actually do better than what would be expected by chance. The difference 4%—88.8%—0.8% represents their improvement over chance. The best improvement possible is 11.2% (2006—88.8%). The statistic $k$ is defined as the proportion of the best possible increase in agreement explained by the clinicians; in this case, $k=5.212/0.96$—0.66. Thus, chance-corrected agreement in Table 1 is about half of what is attributable. (This value of $k$ differs slightly due to rounding error from the result shown in the bottom row of Table 1.)

The example in Table 1 illustrates the simplest application of $k$. Using the basic definition, $k=(P_k-P)/P(1-P)$, where $P_k$ is the proportion of observed agreement and $P$ is the proportion of agreement expected by chance, a whole family of $k$ statistics has been defined. These include indices for assessing agreement when either of two dichotomous categories are used,[7][a] when several rather than two clinicians are used, and when disagreements vary in diagnostic importance. An article by one of us (J.E.F.) contains details about these and other $k$ 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 accounted for by error. For a fixed proportion of error, this proportion will decrease as the total variance decreases, i.e., 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 would not be reliable for detecting

Arch Gen Psychiatry—Vol 44, Feb 1987

Aphakism in Diagnosing—S-Pfrou et al

173
abnormalities in body temperature. In the physical sciences it is well accepted that the expected variation of measurement (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.\

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. A population with maximum heterogeneity is one with a base rate of 50%. As the base rate approaches 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 a study 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 prevalence) was termed the base rate problem by Carey and Gottesman. 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. using a characterization of x based on work by Kraemer, 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, x. As we see it, they, in effect, inverted 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 built on the misinterpretation by Grove et al. Beginning with their title, which includes the phrase "the base rate problem in the kappa statistic," they perpetuated the incorrect idea that it is that is affected with a base rate problem. They proposed that x be reserved for conditions with a relatively high prevalence rate, and that it be replaced with Y, 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, we will provide a detailed critique of their analysis, and in a subsequent section, we will discuss the reliability of Yule's Y for measuring reliability.

Spitznagel and Helzer reported an analysis of x under two psychometric models. The first, which they termed the "validity model," assumes that the true clinical status is known and that only one diagnostic procedure is under study. Under this model they showed that for fixed sensitivity (the proportion of true cases correctly classified as cases by the flexible procedure) and fixed specificity (the proportion of true noncases correctly classified as noncases by the flexible procedure), the value of x for the agreement between the true and the false 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 results regardless of the base rate of the true classification, who specificity and sensitivity are fixed.

Spitznagel and Helzer's mathematics is correct, but their analysis is irrelevant for reliability. A validity model simply not appropriate for evaluating a reliability statistic such as x. When a diagnostic criterion is available (e.g., 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 true cases), and negative predictive value (the proportion of putative noncases who are true noncases). Neither x nor Yule's Y, nor any other measure of agreement or association, provides information about false-positives and false-negatives; this information is precisely what is sought in a validity study. Since neither x 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 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 a properly constructed table can be represented as in Table 2. This representation is the same as that given by Carey and Gottesman and is consistent with mathematical models of latent classes. Using this model and assuming certain fixed values for sensitivity and specificity, Spitznagel and Helzer showed (in their Fig 5) that both x 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 0.80. Our Fig 2 shows how they vary for a sensitivity of 0.40 and a specificity of 0.98; these last values are similar to those reported in two studies of the Epidemiological Catchment Area Survey for major depression.

Our figures indicate that both x and Yule's Y will tend to go down if reliability studies are performed in populations with base rates of 10% or 5%. From our discussion, however, the true rate of reliability and base rates, the results for—true measure of reliability—are as expected. Yule's Y seems also to be
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 $q$. This nonlinearity transforms $q$ in a way that has no apparent intuitive appeal. Spitzner and Heitzeber char that the odds ratio can be interpreted as a $q$ 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 will almost never 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 (i.e., relatively few entries in cells $b$ or $c$ of Table 1), the counts in $b$ and $e$ are similar in magnitude, the table displays symmetry, and both $e$ and Yule's $Y$ may be expected to be large. A problem with $Y$ is that it is: (but not always) 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 $e$ 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, clinicians $1$ 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 $1$. Spitzner and Heitzeber acknowledge this problem with the $Y$ statistic and recommend an adjustment method for reducing the value of $Y$ when the count in cell $b$ or $c$ is $0$ (their adjustment method yields the value $e - 0.075$ for the modified table). There is nothing wrong with statistical adjustments when they are mathematically necessary to meet the conditions of the statistic. To make an 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 $k$ statistic, in contrast, requires no adjustment and yields a value of $0.50$, which is the typical value for the statistic when the base rate is $0.50$ in the modified table.

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 $k$ statistics, which includes forms that are applicable to reliability designs that involve multiple diagnostic categories, multiple raters, and even varying numbers of raters.

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 not. Consider Table 2, which shows new figures and in our Figs 1 and 2, Y values consistently greater than $e$ for the same level of error. Since $e$ is an interpretable as a reliability coefficient, the difference between $Y$ and $k$ must be regarded as bias. This is, consistently overestimating the reliability of a diagnostic system 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 Spitzner and Heitzeber.

One of the important practical features of $k$ 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 between 0.40 and 0.75 are generally taken to represent fair to good agreement beyond chance. Comparable standards do not exist for Yule's $Y$, so there is no way of judging, for example, whether the $Y$ value of $0.695$ 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 a larger study 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 0.695 may incorrectly suggest to the investigator that reliability is adequate, whereas the $k$ value of 0.609 correctly warns the investigator that reliability is mediocre.

COMMENT

The recommendations made by Spitzner et al. 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 $k$ statistics 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 for 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 means that proportions of minorities studied. To estimate a sufficient number of positive cases to obtain stable measurement of agreement, some investigators have sampled for the reliability study respondents were diagnosed as cases in the first assessment. Since the sampling plan artificially increases the base rate in the reliability subsample, it is necessary to reconstruct the weightings of the likeliness pattern of agreement in the original population before a reliability statistic is computed. The $k$ statistic can be adjusted to stratified reliability designs and, when properly computed, accurately reflects the challenge to reliability inherent in the study of rare disorder.

Contrary to the arguments by Grove et al. and Spitzner and Heitzeber, there is no base rate problem with the $k$ statistic. Across all base rates, the maximum $k$ value is 1.0, indicating perfect agreement. Actual examples of acceptable $k$ values obtained in samples with very low base rates of certain disorders are available in Appendix F of the DSM-IV. The $k$ statistic can be adjusted to stratified reliability designs and, when properly computed, accurately reflects the challenge to reliability inherent in the study of rare disorder.
As stated by the implications of the alleged base problem with x.

Understanding the use of a standard reliability statistic can be a confusion in the psychiatric research literature. For example, in the abstract of a recent report on the ECA reliability trials in Saint Louis, Fischer et al. reported that "if disease diagnoses made by the ECA's Diagnostic Interview Schedule (DIMS) were compared with DSM-III diagnoses made by psychiatrists, "change corrected concordance was 0.60 or better for eight of the 11 diagnoses." The paper also noted the impression that the DIS was very good agreement with the clinical diagnoses. In another article on the ECA reliability trials in Baltimore, Albus et al. reported that "the change-corrected correlation between the DIS and psychiatrists' one month follow up diagnosis for... four... diagnoses was lower for more... neurological disorders categories." In this case, the paper is certainly left with the impression that the DIS was generally not in agreement with the clinical diagnoses. It might one also begin to speculate that the time frame used with the DIS as a form of the diagnostic interview schedule might be responsible for the different results, such as the clinical analysis would be premature. The abstract of another article by Seibert et al. referred to results obtained with the DIS and the article by the authors of the DIS cited a reference to results obtained with the DIS. A text of the article by Seibert et al. cited a reference to results obtained with the DIS. Although the concept of a bias is not well supported by the statistics and the results of the correlation of the DIS with the clinical diagnoses, it is clear that the agreement in community samples between the DIS and clinical diagnoses is poor.

Some readers might quarrel with this application of x, since the diagnoses being compared are not from the same diagnostic method. While it can be argued that x is applicable because both diagnostic methods are fallible and hence should be treated equivalently in the analysis, a more stinging criticism 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 study has yet been reported in a community sample. One approximation to this pure test-retest reliability study is the test-retest study of the DIS reported by Ecker et al. in which the respondent was first interpreted 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 DIS and clinical diagnoses, still indicated that the majority of the diagnoses are not reliable in the community. More than one half of the x values are less than 0.60 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 x statistic will provide a valuable index for working on this problem, with a valid quantification of change-corrected diagnostic agreement in the general population.

This study was supported by the National Institute of Mental Health grants MH-1873 and MH-2550.

References


Agreement in Diagnosis—Strodt et al. 177

Arch Gen Psychiatry—Vol 41, Feb 1987