Reliability Studies of Psychiatric Diagnosis

Theory and Practice

William M. Grove, MA; Nancy C. Andreasen, MD, PhD; Patricia McDonald-Scott, MA; Martin B. Keller, MD; Robert W. Shapiro, MD

- The existing literature on the reliability of psychiatric diagnosis falls into two periods, the earlier reporting low reliability and the latter reporting much higher figures. The reasons for this trend are examined in the context of a discussion of the design of diagnostic reliability studies. The problems of research design and execution in studies of diagnostic reliability are reviewed, and statistical problems are examined. Solutions to many of these problems are suggested, including recommendations of appropriate reliability coefficients and data analyses.

(Arch Gen Psychiatry 1981;38:408-413)

For years, achieving adequate diagnostic reliability in psychiatry was considered to be a hopeless undertaking. A number of landmark studies suggested that psychiatrists looking at the same patients frequently disagreed about the appropriate diagnoses. As a consequence, the importance of diagnosis was minimized in both research and clinical work. More recently, several studies have shown considerably higher reliability of major diagnoses and of psychiatric symptom measurements than did the older studies. Several major studies have shown that good diagnostic reliability can be achieved. Other studies have used the Present State Examination (PSE) or the Schedule for Affective Disorders and Schizophrenia (SADS) to examine the reliability of symptom measurements with equally favorable results. The reversal of nihilistic attitudes about psychiatric diagnosis has led to a rigorous (and successful) attempt to rework the entire American diagnostic system used by clinicians, DSM-III, which demonstrated in field trials that good agreement could be achieved even in routine psychiatric practice.

The sanguine view of psychiatric diagnosis suggested by these recent studies has arisen for several reasons. First, with the exception of the DSM-III field trials, the recent studies were designed to assess the reliability of diagnosticians especially trained for diagnostic agreement. The diagnosticians were project raters in one or another large-scale research project, not clinicians diagnosing patients in day-to-day practice.

A second difference between studies showing low and high reliability is the use of standardized interview schedules, such as the PSE or the SADS. Since Ward et al found that 5% of diagnostic disagreements resulted from differences among diagnosticians in interview technique, it seems that a modest improvement in reliability might result from the use of standardized interviews.

The third and final difference between older and newer diagnostic reliability studies lies in the use of diagnostic criteria. There is wide agreement that the use of such criteria mitigates or removes the major source of diagnostic unreliability. Ward et al reported that 80% of all diagnostic disagreements stemmed from inadequate nosology: 25% from unclear criteria, 17.5% from disagreements on weighing of symptoms in arriving at diagnoses, 30% from forced choice (only one diagnosis could be made under DSM-I for a present episode of illness), and 7.5% from "impractically fine distinctions." All four of these problems are practically eliminated in DSM-III, the Research Diagnostic Criteria (RDC), and the Feighner criteria.

The success of these more recent studies suggests a number of questions. Was higher reliability achieved because structured interview schedules were used, because diagnostic criteria were used, or both? Was the intensive training often used to "calibrate" diagnosticians responsible? Can less experienced raters also achieve good reliability? Is the good reliability an artifact of less demanding designs for measuring interrater agreement? Would a longer test-retest interval between two interviews in the test-retest studies cause a drastic decrease in agreement? What aspects of the evaluation and diagnostic process can

See also pp 381, 393, 400, and 417.

Accepted for publication Oct 31, 1980.
From the Department of Psychiatry, University of Iowa, Iowa City (Mr Grove and Dr Andreasen); New York State Psychiatric Institute, New York (Ms McDonald-Scott); and the Department of Psychiatry, Harvard University, Cambridge, Mass (Dr Keller and Shapiro).

Deceased.

Reprint requests to Department of Psychiatry, University of Iowa, 500 Newton Rd, Iowa City, IA 52242 (Dr Andreasen).
most economically be improved? How can reliability studies be made a part of ongoing, large-scale research studies?

We became concerned with these issues as they emerged in the design and execution of the National Institute of Mental Health (NIMH) Collaborative Studies of the Psychobiology of Depression. The present article will discuss a number of enduring problems in the design, execution, and analysis of diagnostic reliability studies. Identical problems occur with symptom measurement procedures, so the conclusions presented here apply to such measurements as well.

Problems in reliability studies may be grouped into design and execution problems and statistical problems. Design and execution problems include the choice of an experimental design and the advantages and disadvantages of competing designs. Statistical problems center on the choice of a reliability coefficient, its computation, and its interpretation.

**DESIGNING AND EXECUTING THE RELIABILITY STUDY**

There are several serviceable designs in reliability studies. They range from simpler, less stringent designs that will give higher estimates of reliability to others that are more stringent and tend to give lower estimates. Available reliability study designs include the following: ratings of written case vignettes, ratings of videotaped interviews, ratings completed by an interviewer and an observer working with a "live" patient, short-interval (hours to days) and long-interval (months) test-retest studies of live patients. No doubt other paradigms could be devised. The study sample may also vary from hospitalized patients to random samples from the "normal" population. In the assessment of reliability, it is desirable to use more than one design whenever possible because different designs have different strengths and weaknesses and complement one another. For example, a videotape study may involve hundreds of raters, allowing stratification of diagnosticians on experience and background and giving a very precise estimate of rater variability. A test-retest study may involve more patients and fewer raters, leading to a more precise estimation of subject variability and more representative patient sampling than would a videotape study. When both studies are done using the same diagnosticians, interviews, and diagnostic criteria, the knowledge gained is greater than the sum of the separate studies because one design answers criticisms of the other.

A rarely addressed factor in choosing a reliability study design is economic cost. It is unfortunate that the more informative studies cost more than the simpler ones. As research costs go up, the cost-benefit ratio for various designs will become even more important in choosing among them. Of course, one study is much better than none. It is unfortunate that "none" is what is done in much published psychiatric research.

**Case Vignettes**

The least expensive and least stringent study has multiple raters make ratings and/or diagnoses from written case vignettes. The vignettes may be selected to present certain difficulties to the raters. The vignettes are easily presented to the raters, and they tell the same story to every rater. This type of study quickly establishes whether raters know the criteria and interpret case records in the same way.

This design is least stringent because it tends to reduce to a minimum sources of variability in rating and diagnosing. The patients’ accounts cannot vary, interpreting style cannot vary, and differing interpretations of patients' nonverbal behavior cannot occur. The only sources of variability that influence agreement in the vignette design are variances between diagnosticians in interpretation of criteria and in interpretation of vignettes. Furthermore, no idiiosyncratic ratings in ratings for a single subject by a single rater may occur, since individual raters rate each vignette only once.

Case vignette studies have many useful features. The difficulty of the cases can be varied to examine the reliability of both "easy" and "hard" cases. The method is relatively efficient. A large number of cases can be studied relatively quickly. If the obtained reliability in such a study is low, there is little point in doing more expensive and stringent studies until more groundwork (ie, scale or criteria development, rater training) has been laid. In large studies involving many raters, the vignette method can be used to identify outliers. More training can then be given to the less reliable raters, while more reliable raters can receive positive feedback. In this way, the reliability study may be integrated into the ongoing training of the raters.

**Video Tape Studies**

The next design uses videotaped and filmed interviews. The videotapes can be shown to raters in groups, and if raters are not allowed to stop the tape, the situation is similar to live interviews. This paradigm is very useful for interrater reliability work in collaborative studies because it is much cheaper to send a tape around than to send raters from center to center to conduct live interviews. Many more raters can view a videotape than could ever separately interview a live patient, and all without the patient’s gaining practice at being interviewed.

The videotape study, because it allows many raters to observe the same patient, lends itself to the study of observer bias. Raters can be classified by professional background (physician vs psychologist vs social worker), by degrees of experience, or by any other potentially biasing factor. If enough raters in each category can be obtained, the one-way analysis of variance that may be applied to the ratings provides a precise assessment of differences in symptom ratings or of diagnostic assignments. Such a study, using all 36 active raters in the Collaborative Depression Study (CDS), showed no effect of professional background or experience level on diagnostic agreement (N. C. Andreasen, MD, PhD; P. McDonald-Scott, MA; W. M. Grove, MA, et al, unpublished data, June 1980).

The problems of this design have much in common with those of the vignette design. The patient's story is not allowed to vary, so that discrepancies in self-presentation cannot be measured. Disagreements due to interviewing styles cannot occur. On the other hand, rater interpretations of videotaped interviews are likely to be more heterogeneous than interpretations of written vignettes. This is because the patient's apparent mental state is shown; rater interpretation of quality of patient thought, language, and communication, blunting of affect, nonverbal manifestations of depression, and so on can affect the diagnosis.

The major problem with the videotape design is that the interviewer may inadvertently cue observers as to how he is rating the subject. When he pauses to rate and how he phrases probe questions following a general question can give away the rating and cause spuriously high agreement. With an interview like the SADS, which covers more ground than the PSE by skipping sections when several key questions are answered negatively, the further complication arises that certain questions must be rated identically (ie, blank) since a section was skipped. An additional disadvantage is that this design cannot assess variability in interviewing habits.

**Interrater Agreement Designs**

Another design that is useful in reliability studies has one rater interview a subject and another rater observe and rate simultaneously and independently. This is the standard training experience for raters in the CDS, with item-by-item discussion of the ratings later. The CDS also requires that every tenth interview be an interrater reliability interview so that reliability can be assessed on an ongoing basis as the study progresses. One advantage of such a design over the case vignette and videotape studies is that it is more cost effective; it uses subjects who will be interviewed anyway, it needs no equipment, and it measures the questions following that variation that the videotape study does. The main disadvantage of the design is identical to that of the videotape study, principally that spuriously high agreement may result from observing or receiving cues from the interviewer. In fact, in the live interviewer-observer study, the observer cannot be prevented...
from seeing the rater’s movements, while the video camera could be made to record only the patient’s movements. Unlike the videotape study, the interviewer-observer design cannot be used with many raters per subject since it is usually physically impractical to interview a patient before more than a limited number of observers.

**Test-Retest Designs**

The most stringent test of diagnostic reliability is the test-retest design, where subjects are re-interviewed after a time interval has elapsed. This introduces three very important sources of variance into the design, which unfortunately are interwoven in a way that makes them hard to separate. Two sources are variations in the patient’s story in the two interview occasions and variations in rating style of the diagnosticians. On the simplifying assumption that individual interviewers maintain the same style from one patient to another and from one diagnostic category to another, the two sources can be examined separately. However, this is probably only a crude approximation to the true state of affairs. The third source of variability is true change of the patient’s status over time. The longer the interval, the more likely is such change. The vagaries of memory are presumably greater at longer intervals as well.

This design is expensive to execute within a single clinical center because it disrupts normal office scheduling to set up interviews with pairs of raters at appropriate intervals. A rater can do only two interviews like the SADS per day, and so in a week at most ten interviews per rater can be collected. In a multicenter study like the CDS, the cost is very high since raters must travel to other centers if intercenter reliability is to be assessed.

Though the cost is high, the data from a test-retest study can answer nearly all the questions the other designs can answer and several others in addition. It is the only design that permits a stringent evaluation of variability in rating style of individual diagnosticians. It provides the closest approximation of actual interviewer behavior in the research or clinical situation.

A useful innovation to increase efficiency and decrease cost for this expensive design is to rotate all possible interviewer pairs so that each pair occurs an equal number of times without regard to the order of interviewers within subjects. This makes efficient use of interviewer and patient time; a first patient can be interviewed in the morning by one rater while a second rater interviews a second patient. The first patient is reinterviewed in the afternoon by the second rater while the first rater interviews a third patient, and so on. Statistically, this creates a balanced incomplete block design that has a relatively straightforward analysis for determining reliability coefficient.

A further refinement of the test-retest design used in the CDS was to interview the subject on three occasions, using both short and longer intervals so that the results could be compared (Andreasen et al, p 400). In this study, interviews were conducted at short intervals (morning, afternoon) and long intervals (several months). The objection has been raised concerning short-interval studies that the subjects might simply be reiterating their morning story from memory to the afternoon rater. However, our subjects had been interviewed several months earlier as well. We therefore compared the initial diagnoses from several months previously with the consensus diagnosis of the morning and afternoon raters. The initial vs consensus reliability was, of course, lower than the morning vs afternoon reliability, but not drastically so. Thus, one could show with such a double design that most of the agreement was not artifactual.

**Summary**

While these are only some of the possible reliability study designs, they cover the range of stringency required for a thorough examination of diagnostic reliability. One would be very fortunate to be able to use all of them, but any one of them would be better than none. Clearly these different designs call for different statistical analyses. However, before a statistical analysis can be performed, one has to decide what reliability coefficient to calculate. Some problems in choice, computation, and interpretation merit serious study and will now be detailed.

**STATISTICAL PROBLEMS WITH RELIABILITY**

Several problems with reliability of psychiatric diagnosis that may be called statistical arise. The following five such problems should be pointed out: (1) the relationship between reliability and validity, (2) the choice and computation of a reliability coefficient, (3) the problem of low base rates, (4) the interpretation of reliability estimates, and (5) the relationship between reliability and accuracy of diagnosis.

**Reliability vs Validity**

Classical mental test theory states that reliability constrains validity. For variables within the range of convenience of classical mental test theory (intelligence, aptitudes, achievement), this is true by definition of the term "true score." In classical mental test theory, the true score is the subject's average for an indefinitely large number of observed scores. It can be shown that if

\[
\text{Reliability} = \frac{\rho}{\sqrt{\rho^2 + \psi^2}}
\]

and if

\[
\text{Validity} = \frac{\rho^2}{\sqrt{\rho^2 + \psi^2}}
\]

where \( \rho \) denotes the population value of Pearson's r. A criterion variable is a single perfect measure of the construct being measured.

Note that reliability here is a squared correlation, while validity is another squared correlation. The correlation of a measurement with a perfect criterion value cannot exceed the correlation of a measurement with its own true score, else the perfect criterion could replace the measurement's own true score.

Cronbach and diagnostic reliability are now commonly assessed by the coefficient \( k \) or some related agreement statistic. The notion of true score in such situations differs from the classic situation because here the true score is 0 if the subject is not ill and 1 if he is. Fleiss and Cohen have shown that \( k \) is equivalent to an intraclass correlation coefficient, which means that the constraint on validity just mentioned applies when reliability is assessed by

While the distinction between reliability and validity seems clear in the abstract, in practice it is often subtle. The best example of their confounding occurs in the long-interval test-retest reliability study. A six-month test-retest reliability study could justifiably be reported as a follow-up study of diagnostic stability or as a study of course of illness. Yet stability of diagnosis and assessment of outcome are usually considered to be validators of diagnostic categories.

Another convergence of reliability and validity occurs when investigators talk about using a "gold standard" to assess reliability. One often hears the term "true score" to mean agreement, not between two raters, but between a rater and the correct diagnosis. This is, in fact, an assessment of validity, not reliability. However, usually a gold standard or criterion diagnosis is unavailable, though one is sometimes set simply for heuristic purposes.

Finally, it should be noted that although achieving reliability is important, it is not enough. Diagnosticians conceivably may agree perfectly but be wrong all the time. A reliable measure that has no validity is worthless. We could measure the height of all schizophrenics reliably, for example, but the measure would not provide any useful statement about the description, definition, outcome, or etiology of schizophrenia.

**Choice and Computation of a Reliability Coefficient**

Bartko and Carpenter have reviewed all the major reliability coefficients for psychiatric diagnostic data that had been proposed up to the time of that review. Every one had some mathematical or empirical fault. Typically, the assumptions of coefficients with a pleasing mathematical elegance do not precisely match the actual processes involved in psychiatric diagnosis. Indeed, since diagnoses are made in many ways for many purposes, no coefficient could summarize the merits of a diagnostic system in a single number. The task for researchers is to choose the best available coefficient or to devise a better one. Bartko and
Carpenter recommend $\kappa$ for use with dichotomous data like diagnoses. Indeed, $\kappa$ is now the most commonly used coefficient for estimating the reliability of a psychiatric diagnosis. Kappa is a coefficient of interrater agreement, corrected for chance agreement. Its formula is

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \tag{1}$$

where $p_o$ is the observed proportion of agreement, and $p_e$ is the agreement expected by chance. "Chance" is assumed to operate as follows. When two raters make diagnoses, a proportion of them will agree because the diagnosis given by rater 1 just happens to coincide with rater 2's diagnosis. If rater 1 assigns a diagnosis with probability $p_1$, and rater 2 does the same with probability $p_2$, then chance agreement will occur with probability

$$p_e = p_1 p_2 + (1 - p_1)(1 - p_2) \tag{2}$$

The first term represents agreement on a positive diagnosis by chance, and the second term represents agreement on a negative diagnosis by chance. Then the quantity $p_o - p_e$ is the excess agreement not attributable to chance. The denominator simply makes the maximum value of $\kappa$ 1, indicating perfect agreement.

Objections have been raised to the use of $\kappa$. The main objection runs as follows. Kappa is an agreement index corrected for chance agreement. "Chance agreement" means the agreement that would be observed if two raters assigned diagnoses to cases at random. Now this is not what diagnosticians do. They assign the easy cases, or "textbook" cases, to diagnoses with little or no error; they may guess or diagnose randomly on the others. If one knew which cases were textbook cases, one could treat them separately; but that is a difficult matter. If one could reliably assess "textbookiness," one could treat the certain cases separately from the uncertain cases. To our knowledge, this has never been done successfully.

The concept of textbook vs difficult cases does point up an interesting problem. If a study uses predominantly easy cases to estimate reliability, estimates will, of course, be higher than if the sample is one of difficult cases. Similarly, if the threshold for diagnosis is set in between two groups of cases widely separated in difficulty (Figure), reliability will be higher than if the threshold is located in the middle of a hump in the difficulty distribution. This problem, which is related to what the diagnosticians in the Ward et al. study called "impractically fine distinctions," has not received systematic study, to our knowledge.

Janes and Maxwell, who do not try to separate easy from difficult cases, suggest assuming that the chance in chance agreement operates as follows. Each rater assigns half of the nontextbook or questionable cases correctly and to negative diagnoses. Each rater also assigns half of those textbook cases, the ones on which the raters disagreed, to the rater $1 +$ rater $2$-cell and half to the rater $1 -$ rater $2$+ cell. Textbook cases are assumed to be diagnosed without error.

The flaw in this scheme is that it assumes a rating process model that does not depend on the rater's observed behavior (ie, diagnostic base rates). It suggests that when in doubt the rater mentally flips a fair coin to make the diagnosis. We hope nobody really does this. Certainly the SADS, RDC, and PSE instructions, as well as the nature of the Feighner diagnostic criteria, make it clear that when in doubt one is to rate no symptom, diagnose no illness, and call no case "undiagnosed illness." For these diagnostic criteria, the Maxwell model of chance agreement ought not to hold and probably does not hold. Kappa, on the other hand, can be visualized as embodying the following model of chance agreement. When in doubt on a nontextbook case, each rater mentally flips a biased coin, with the probability of getting "heads" (giving the diagnosis) equal to his own base rate. For items with low base rates, this approximates the instruction to rate no symptom or no illness. For this reason, $\kappa$ seems a sounder choice than the Maxwell coefficient.

Recently, a new reliability coefficient, called the $\pi$ coefficient, has been recommended. The equation relating observed agreement to the reliability coefficient $\pi$ is

$$p_o = \pi^2 + (1 - \pi)^2 \tag{3}$$

where $p_o$ is the observed percentage agreement between two raters. The model that suggests this equation is as follows. Raters can agree by both being correct (with probability $\pi \times \pi$, or $\pi^2$) or by both being incorrect (with probability $1 - \pi^2$). It is unfortunate that the derivation of this coefficient assumes that the probability of making a correct diagnosis if the subject is ill (sensitivity) equals the probability of being correct if the subject is well (specificity), each being equal to $\pi$. This assumption is not generally true, and the coefficient does not seem to quantify any aspect of a real rating process. Because $\pi$ does not correct for chance agreement, $\pi$ always exceeds $\kappa$, often by an amount sufficient to make an incautious researcher conclude that an unreliable item or diagnosis is very reliable. For example, for a $\pi$ of .9 and an observed base rate of .1 for both raters 1 and 2, $\kappa$ would be 0. Therefore, we cannot recommend $\pi$ as a useful reliability coefficient.
There are other advantages to using $\kappa$ as an agreement statistic. It is almost identical to an interclass correlation coefficient of the form

$$\text{Intraclass } R = \frac{(\sigma^2_{\text{subjects}})}{(\sigma^2_{\text{subjects}} + \sigma^2_{\text{raters}} + \sigma^2_{\text{error}})}$$

(4)

for the usual situation where a number of raters all rate each of a number of subjects. This means that dichotomous reliability data can be analyzed with any good analysis of variance computer program if the data are coded 0 for disease absent and 1 for disease present (or any two other distinct values). The equivalence also makes it possible to compute a reliability coefficient for designs where not every rater rates every subject, as in the balanced incomplete block design used in the CDS test-retest reliability work. For computational details, see Shrut and Fleiss and Fleiss.

Reliability and Base Rates

The third problem is the problem of base rates. To examine this problem, one needs the concepts of sensitivity (rate of true-positive diagnoses) and specificity (rate of true-negative diagnoses). These two quantities summarize the predictive merit of a diagnostic procedure. The maximum possible value of a reliability coefficient like $\kappa$ is always $1$, indicating perfect agreement on all cases. However, if the sensitivity and specificity of a diagnostic procedure are not perfect (some false-positive and some false-negative errors occur), then the value of $\kappa$ and estimates of base rates will both be affected. For psychiatric diagnoses and rare symptoms such as first-rank symptoms of schizophrenia, the effect is to lower the reliability considerably below the sensitivity and specificity. The relationship, which was derived by Knaus, is given by

$$\kappa = \left(1 - \frac{Q^2}{P^2} \right) \left( S_+ + S_- - 1 \right)$$

(5)

where $\kappa$ is the true population value of $\kappa$ (not an estimate of it), $P^2$ is the true base rate ($P^2 = 1 - P$), $P$ is the apparent base rate ($Q = 1 - P$), $S_+$ is the sensitivity, and $S_-$ is the specificity.

The formula has two noteworthy features. First, $\kappa$ is modified both by base rates and by sensitivity and specificity. Since the proportion $P^2/Q^2$ may be either more or less than 1, it may either increase or decrease $\kappa$. Second, the quantity $(S_+ + S_- - 1)$ acts like an average of sensitivity and specificity, allowing high sensitivity to make up for low specificity (or vice versa) in determining $\kappa$.

Numerical examples bring home the practical importance of these relationships. For example, if the sensitivity and specificity are both .85 and the true base rate is .25, then the maximum $\kappa$ value will be only .76. Other examples are given in the Table. This relationship between base rates and reliability may lead the incautious researcher to reject as unreliable what is in fact a rather good diagnostic procedure. Investigators reporting reliability data should include information about base rates. Reliability estimates for symptoms or diagnoses with very low base rates should probably not be reported at all. In reliability studies conducted for the CDS, it has been decided not to calculate $\kappa$ coefficients for diagnoses with observed base rates of .5% or less.

Interpreting the Value of the Reliability Coefficient

One would hope that after choosing and computing a reliability coefficient, its interpretation would be straightforward: low values are cause for alarm, high values are cause for celebration. However, the interpretation of reliability coefficients is also complicated by the same statistical problems that plague the choice of a coefficient. Kappa (or $\kappa$), the best available choice, varies in magnitude as base rate varies (assuming a fixed sensitivity and specificity) and is lowest when the base rate nears 0 or 1. Therefore, there is not a single reliability but a whole series of reliabilities, one for each base rate. Generalization to base rates other than those observed in the reliability study may or may not be valid.

Interpretation of reliability data may also depend on the restrictiveness of the diagnostic or measuring system being used. Carey and Gottesman point out that restrictive diagnostic systems like the RDC are built on the assumption that false-positive diagnostic errors are much more serious than false-negative ones. In this situation, the most relevant quality index is the specificity, which measures false-positive diagnostic errors. It is easy to see that the sensitivity of the RDC is less relevant to its quality than the specificity. Kappa, and by its identity with $\kappa$ the intraclass $R$, gives sensitivity and specificity equal weight in determining a reliability index. It is unfortunate that their impact is usually difficult to measure since sensitivity and specificity cannot be determined accurately without a criterion better than the ratings being assessed; such a criterion is usually not available. Nonetheless, one can conclude that a low $\kappa$ or $R$ value is not necessarily a bad thing for restrictive diagnostic systems.

A numerical example may illustrate this fact. If an investigator wants to select a sample of schizophrenics for a biological study, he probably wants a homogeneous sample. He may use the RDC for this purpose. Suppose only half of all true schizophrenics are picked up by these criteria ($S_+ = .50$), but there is only a 1% false-positive error rate ($S_- = .99$). Suppose further that the true base rate of schizophrenia in the investigator's hospital is 25%. Then, $\kappa = .16$ (Table), and yet it can easily be shown that 94% of the investigator's sample truly will be schizophrenic. The interpretation of reliability coefficients should take into account these pitfalls. A further complication is the fact that no useful statistical test is available to determine when reliability is acceptable. It can be determined that it is significantly better than chance, but this information has very little meaning in clinical or research settings. While investigators often make the simple assumption that a $\kappa$ value of more than .5 or .6 is acceptable, in fact the meaning of all estimates of reliability will depend on all the circumstances that have been discussed. One such circumstance is the stringency of design, base rate, sensitivity and specificity, and the coefficient chosen. The wise investigator should be aware of all these factors when interpreting the reliability studies presented in scientific journals.

Reliability and Case Detection

A fifth problem occurs in epidemiological field studies. Ordinarily, nonclinical populations are less variable in diagnosis than are diagnostically mixed clinical populations. This makes reliability of diagnosis lower in community surveys than in hospital studies. Shrut and Fleiss have shown that this can lower the accuracy of the case-finding procedure until the population prevalence of a disorder is overestimated by a factor of two or more. Thus, the reliability among noncases of indicators of diagnosis must be very good for accurate prevalence estimation. Often it is only fair, poor, or even unknown.

COMMENT

It no longer seems necessary to apologize for poor diagnostic reliability in psychiatry. Carefully constructed interview schedules and lists of diagnostic criteria, together with rigorous training of raters, have caused a quantum jump in the magnitude of psychiatric reliability in the last decade.

Much work remains undone. Subtype diagnoses; diag-
noses of slippery concepts like schizophrenia spectrum (schizotypal features or personality) and milder illnesses like personality disorders, minor depression, and hypomania all show poorer reliability than do major diagnoses. Case detection in nonclinical populations is fraught with hazards. Existing coefficients of agreement all seem unrealistic in one or another of their assumptions.

Provisional solutions to many of these problems can be achieved by careful design and execution of several different types of reliability studies and by analyzing reliability data in more than one way. The general results of recent reliability studies have been encouraging, and it is to be hoped that they will be replicated in the future with larger sample sizes and more sensitive measures of reliability than are now available.

This work is from the National Institute of Mental Health-Clinical Research Branch Collaborative Program on the Psychobiology of Depression: Clinical Studies. This study was completed with the cooperation and participation of Collaborative Program investigators Gerald L. Klerman, MD (chairman); Robert M. A. Hirschfeld, MD (project director and co-chairman); Martin M. Katz, MD (National Institute of Mental Health); Martin B. Keller, MD, and Robert W. Shapiro, MD (Boston); Jan A. Fawcett, MD, and William A. Scheffner, MD (Chicago); Nancy C. Andreasen, MD, William Coryell, MD, and George Winokur, MD (Iowa City); Jean Endicott, MD, and Robert J. Spitzer, MD (New York); and Paulo J. Clayton, MD, Theodore Reich, MD, and Eli Robins, MD (St. Louis). Joseph Fleiss, PhD, Philip Lavori, PhD, and John Rice, PhD, also provided assistance.

References