The Reliability of Psychiatric Diagnosis

WILLIAM M. GROVE

INTRODUCTION

In science, one needs to be able to understand and repeat another's work. In psychopathology this is difficult, since our descriptions of patients are often ambiguous and idiosyncratic (Elkin, 1947). Measurement research in psychopathology corresponds to the development of laboratory tests in clinical medicine; it is often tedious but indispensable. This chapter first surveys some methodological and statistical problems that arise in diagnostic reliability research. Recent developments in the statistics of diagnostic reliability are reviewed. The outcomes of studies of diagnostic reliability are then surveyed. Ongoing studies related to diagnosis are touched on. The reader is cautioned against overoptimism about the attainable certainty of psychiatric diagnosis.

At the outset, one must realize that reliability of diagnoses is a joint function of the diagnostic scheme and the diagnosticians. For example, Helzer and Coryell (1983) have recently told how one of them was unable to confirm findings of another investigator using the dexamethasone suppression test as an indicator of melancholia. After some detective work, the failure to replicate was attributed to variance in how the agreed-on DSM-III (American Psychiatric Association, 1980) diagnostic rules were being interpreted. Diagnostic reliability studies correspond to the calibration of a new laboratory's assay procedures and reagent preparations.

WILLIAM M. GROVE • Department of Psychiatry University of Iowa, Iowa City, IA 52242.
nel are developing a WHO-ADAMHA Comprehensive International Diagnostic Interview. Wittchen has recently completed a physician re-examination study of both DIS and CIDI using his German translations. Spitzer and colleagues are working on a clinician interview for the making of DSM-III diagnoses, the Structured Clinical Interview for DSM-III (Spitzer, 1983). While these developments pertain most directly to the information-gathering phase of diagnosis, the development of structured interviews has already enlarged our knowledge of diagnostic reliability. Future developments will clarify whether diagnostic concepts which have proved reliable in Anglo-American psychiatry travel well to other cultures, and to what degree skilled clinical judgment is necessary for reliable diagnosis.

Readers might feel rather comfortable about psychiatric diagnosis at this point. They should not. Many people become overconfident of diagnostic accuracy when they read of high reliability. One has to remind oneself that while unreliability constrains validity in a general way, raising reliability is no panacea. High reliability can always be bought by gerrymandering diagnostic definitions, or by local agreements on minutiae of interpretation of specified criteria. Therefore, reliability studies can never be more than a necessary first step in diagnostic research. Given the amorphous and overlapping nature of many psychiatric syndromes, one must be suspicious of highly reliable diagnoses until they are proven to be highly valid, too.

REFERENCES


HOW UNRELIABILITY CONSTRAINS VALIDITY

Three independent quantities are sufficient to describe any procedure that yields a dichotomous diagnosis (Galen & Gambino, 1975). The first quantity, \( P \), is the prevalence of the disorder in question. The second and third parameters, sensitivity and specificity, denoted by \( Se \) and \( Sp \), respectively, capture the accuracy of the diagnostic process. Sensitivity is the probability that a truly ill person is correctly labeled by the diagnostic process, and specificity is the probability that a truly well person is correctly labeled. If we denote the complement of the prevalence, \( 1 - P \), by \( Q \), it is easy to prove that

\[
\hat{P} = P \, Se + Q (1 - Sp)
\]

which shows that the prevalence of a disorder is not even unbiasedly estimated by the apparent prevalence \( \hat{P} \) (the frequency with which the disorder is diagnosed by an imperfect rater). If two imperfect interchangeable raters with stable diagnostic habits, making diagnostic errors independently, jointly diagnose patients, their expected percentage agreement is

\[
p_a = P \, Se^2 + Q (1 - Sp)^2 + P (1 - Se)^2 + Q \, Sp^2
\]

This forbidding-looking expression says that the probability of two raters agreeing on the presence or absence of an illness is composed of two parts, agreement on presence of illness and agreement on absence. Each of these, in turn, comprises cases on which both raters judge correctly and cases on which both err. All reliability coefficients in common use depend on \( p_a \) and thus on \( P \) as well. It is due to the complexity of Equation (2) that the interpretation of reliability coefficients as constraints on validity is so tortuous.

The dictum that “reliability constrains validity” originally formulated for continuous mental test scores holds in a certain way for dichotomous diagnoses as well. For example, imagine an instance of a common kind of study in psychopathology, in which a diagnostic group is compared with other patients or normal controls, on some characteristic related to etiology, outcome, or treatment response. Now, some members of the diagnostic group and some controls will be misclassified when study subjects are selected. In fact, ignoring sampling error, a proportion equal to

\[
PPV = \frac{P \, Se}{P \, Se = Q (1 - Sp)}
\]

of the subjects in the diagnostic group actually have the illness, and similarly a proportion equal to

\[
NPV = \frac{Q \, Sp}{P (1 - Se) + Q \, Sp}
\]

of the controls actually do not have the illness. These two quantities, \( PPV \) and \( NPV \), are called the positive predictive value and negative predictive value, respectively.

Consider an example. Suppose that the true prevalence of an illness in a hospital is 30%, the sensitivity of the diagnostic process is 50% and the specificity is 75%. One is interested in studying suicide risk in patients versus controls. If the rate of suicide in the truly ill group is \( r_i = 15\% \) and in controls, the true relative risk is 15. However, in our example \( PPV = .4615 \) and \( NPV = .7778 \), which means that just 46% of all study cases truly have the disorder and only about three quarters of the controls are really controls. With such contamination of samples, the estimated relative suicide risk is \( \frac{r_i \cdot PPV + r_c(1 - PPV)}{r_i(1 - NPV) + r_c \cdot NPV} = 7.5\% / 4\% = 1.8 \), approximately. This difference could easily escape detection, and we might conclude that the rates of suicide were identical in the two groups, in the face of a true fifteen-fold difference. On the other hand, if another study were conducted in the same hospital with better diagnoses (say \( Se = .8; \, Sp = .9 \)), then the relative suicide risk would be about 12%/2% or 5.3. This figure, while still a drastic underestimate, is much more accurate and is more likely to be discovered in a small-sample study. Please note that these examples are optimistic because they ignore sampling error; the situation will be worse, on the average, in any real study.

Carey and Gottesman (1978) have pointed out that, because prevalences vary from clinic to clinic, two raters with stable diagnostic habits will experience changes in diagnostic reliability as they move from clinic to clinic. If errors of commission matter and errors of omission do not, it can happen that the less reliable diagnostic is more valid. Real studies often proceed by choosing just the definitely ill and comparing them to certain noncases, such as medical-surgical controls. In such a situation, kappa may not index the decrement in validity caused by diagnostic errors since only specificity matters: incorrect labeling of some patients as noncases does not decrease validity.

COLLECTING DATA FOR RELIABILITY STUDIES

Determining the reliability of diagnoses involves the study of rater agreement or consistency. If we make the testable assumptions that our
raters are equally skilled and that their habits do not change, we may make certain predictions. First, if two (or more) raters each examine separate series of patients and if specific patients are assigned to specific raters only by chance factors, then the raters should agree, within sampling error, in their frequencies for specific diagnoses. For example, if patients are sent to one of several wards in a hospital solely by the availability of an empty bed, and if doctors are rotated among wards, then the frequency of chart-diagnosed schizophrenia should be the same across wards. Also, in the absence of a change in catchment populations or referral patterns, the frequency of diagnosed schizophrenia should be stable over time. Clearly, the latter assumptions are stronger than the former, making studies of diagnostic frequency over time riskier than the comparison of rates between wards or between hospitals. Both these procedures were once used to test diagnostic reliability, but such methods have been superseded.

Agreement in diagnostic frequencies for independently diagnosed patient series is necessary but not sufficient for proving diagnoses reliable. Two raters could see patients together, each call 20% of them schizophrenic, and never agree on a solitary diagnosis of schizophrenia. Therefore, one would prefer to examine agreement between diagnosticians on a single patient series.

One can employ any of several methods to obtain multiple diagnoses on a series of patients. Several raters can read a series of patient charts or prepared clinical vignettes. This may not be a stringent test of reliability, but it allows one to identify problems in interpreting chart material and points up ambiguities in diagnostic criteria. Alternatively, raters can diagnose recorded interviews. In videotape or audiotape studies, the original interviewer may be counted as a rater for reliability purposes or not, depending on whether he or she could have noticed diagnostically important patient behavior not captured on tape. This design is well suited to intercenter or international reliability studies.

A rigorous method is to conduct joint reliability interviews. Two or more raters can sit in, each interviewing some of the patients (to average out interviewer effects). With live interviews, subtle behavior can be noticed by all, perhaps making diagnoses more reliable. However, it is difficult for a taped or observed interviewer to follow any but the most rigidly structured interview schedule without giving coraters diagnostic cues, biasing agreement upwards. Interviews with optional questions which may be skipped are especially problematic.

Test–retest reliability studies offer a stringent test of agreement. Here one avoids accidental cueing of fellow raters, offering a high methodological hurdle. However, such studies are no panacea. Only two or three raters can see each patient. A problem with a short- or medium-term retest study is that the patient may recall the previous interview, perhaps reciting remembered answers and inflating interrater agreement. This can only be avoided by using a longer follow-up, but then the patient's condition may change. Cloninger, Miller, Wette, Martin, and Guze (1979) have treated the analysis of diagnostic stability over time, under a restrictive assumption that various causal factors linearly influence diagnostic status over the long term. The best (but expensive) solution is to combine test–retest and simultaneous rating designs, preferably with short-term and long-term retest intervals. In this way the various reliabilities, interrater and test–retest, can be computed and compared.

One innovation in test–retest studies deserves special mention. It may be impossible to have each of several raters interview every patient. If one can have each possible pair (or triplet) of raters in a study occur with an equal number of patients, one has a balanced incomplete blocks design (Fleiss, 1981). This was first employed by Rosenzweig, Vandenburg, Moore, and Dukay (1961), but they used the wrong analysis for their design. More recently, the Iowa 500 Study and the National Institute of Mental Health Clinical Research Branch Collaborative Depression Study (Katz, Secunda, Hirschfeld, & Koslow, 1979) have used this design. It requires many fewer interviews per patient than full rater-by-subject designs in order to get reliability estimates.

**MEASURING DIAGNOSTIC AGREEMENT**

Given a series of patients, each diagnosed at least twice, one wishes to quantify diagnostic agreement. The simplest statistic is the observed percentage agreement, but this has undesirable statistical properties. When the prevalence of a given disorder is low, agreement could be made high by calling each patient well. One would prefer an index not subject to inflation by such extraneous factors, but all indices of rater agreement are affected by the true prevalence. One can, however, use statistics that correct for the apparent prevalence, which is as close as one can ordinarily come to knowing the prevalence anyway. An important class of such statistics is exemplified by kappa (Cohen, 1960), defined as

\[
K = \frac{p_e - p_c}{1 - p_c}
\]

where \(p_e\) is the observed percentage agreement and \(p_c\) is the agreement expected if raters assigned diagnoses to patients randomly, each rater having...
his or her own frequency of making the diagnosis. Cohen's original proposal was that chance agreement would occur as follows. If rater A made \( p_1 = 40\% \) positive diagnoses while rater B made a \( p_2 = 30\% \) and if each rater assigned diagnoses at random, then they would agree by chance \((.4)(.3) + (.1)(.7) = .34\) of the time. Cohen therefore defined chance agreement for dichotomous diagnoses as

\[
p_c = p_1 p_2 + (1 - p_1)(1 - p_2)
\]

(6)

This makes \( K \) a chance-corrected index of interrater agreement, and it also makes \( K \) independent of the observed rater "prevalences." However, it cannot make \( K \) independent of the true prevalence (Kraemer, 1978).

Kappa is therefore an exceptionally useful statistic, and it has several nice algebraic properties. For a large class of agreement statistics denoted by \( t \), the quantity

\[
T = \frac{t - t_c}{1 - t_c}
\]

(7)

is equal to \( K \), provided only that \( t_c \) is computed under the assumption that raters assign diagnoses at random so as to obtain their observed frequencies of positive diagnosis. Many reliability indices reduce to \( K \) once chance agreement is allowed for. A very useful fact is that \( K \) is almost exactly equal to an intraclass correlation coefficient computed on two raters' dichotomous diagnostic data, coded 1 for a positive diagnosis, 0 otherwise (Fleiss & Cohen, 1973).

Kappa depends on \( P \) (for fixed Se and Sp) and yet \( P \) is conceptually irrelevant to diagnostic agreement. To get around this problem, which is serious with the low base rates seen in epidemiological studies, an agreement index originally proposed by Yule has been studied in some detail by Spitznagel and Helzer (1985). This index is

\[
Y = \frac{P_{00} P_{11} - P_{01} P_{10}}{P_{00} P_{11} + P_{01} P_{10} + 1}
\]

(8)

where \( P_{00} \) is the agreement on negative diagnoses, \( P_{11} \) is the agreement on positive diagnoses, and \( P_{01} \) and \( P_{10} \) are the disagreement cells. This index agrees closely with kappa for base rates above .4 when diagnoses are of reasonable quality, but is much less sensitive to declining prevalence than kappa (i.e., as \( P \) approaches zero). It may be that this is bought at the cost of \( Y \) 's insensitivity to declining agreement as well. However, this index deserves statistical study, and shows promise for use in epidemiological work.

Analysis becomes complicated for studies with multiple raters. One needs to define agreement when more than two raters diagnose a case. One could define it as pairwise agreement (proportion of all possible rater pairs who agree), majority agreement, or total agreement (all raters agree). It turns out to be simplest to use pairwise agreement.

The definition of a kappa-type statistic for pairwise agreement in complex designs has been discussed by Fleiss and Cuzick (1979), who provided a kappa-type statistic useful when different and varying numbers of raters examine each patient; it is assumed that each patient is seen by a separate set of raters. Uebersax (1983) gave a kappa-type statistic for the general case in which the number of raters per patient varies, but raters may overlap in the patients they see.

I prefer the general approach of Landis and Koch (1977) to that of defining special kappas. They proposed to bypass the computation of chance agreement, instead computing an intraclass correlation coefficient for the diagnostic "score" coded 1 if present, 0 otherwise. As mentioned above, when two raters see a series of patients together, \( K \) is essentially identical to an intraclass correlation coefficient. The intraclass correlation is a "signal" to "signal plus noise" ratio, and assesses the proportion of measurement variance accounted for by systematic patient differences. One can use intraclass correlations for any reliability design whatsoever, they do not require that raters be the same for all patients or be different for each patient, and do not require that the number of raters per patient be constant. All one needs to compute them is a good analysis of variance program. Intraclass correlations are the only approach presently available for analyzing the results of balanced incomplete blocks designs. I employed the method of Landis and Koch (1977) to define a kappa-type agreement statistic wherever possible in this review, in reanalyzing old reliability data. The price paid for this flexible approach is that different components of measurement variability must be estimated in different situations, so that no one simply defined statistic indexes agreement in all instances.

The derivation of standard errors for kappas can be difficult. It is usually not very important to test the hypothesis \( K = 0 \), as it is implausible that agreement is no better than chance. One would prefer to obtain a confidence interval for the true intraclass correlation or kappa. Unfortunately, this is an area where often no exact or even approximate results are available. Often one must be satisfied with an estimate of kappa without a standard error. Landis and Koch (1977) gave a method for deriving approximate standard errors which is generally applicable to all sorts
of designs, but it requires complex and expensive computation. Uebersax (1983) gave another method which is likewise general but very expensive and only approximate.

The study designs and statistics discussed above have all seen employment by investigators interested in diagnostic reliability. Over the last fifty years, there has been a gradual increase in the methodological sophistication of reliability studies. Early studies compared rates of various diagnoses between wards or hospitals, but in the 1950s investigators became interested in simultaneous rating and test-retest studies of diagnosis. In the 1960s kappa and structured diagnostic interviews were first used in reliability studies, and in the 1970s diagnostic criteria and balanced incomplete blocks designs were introduced. To these studies we now turn. First I will discuss older studies which calculated, or allow one to calculate, chance-corrected agreement.

Studies of Reliability Antedating Specified Diagnostic Criteria

Table 1 gives results of many important older studies giving kappas or allowing a kappa to be computed. There is enormous variation in methodology between these studies, and many had small samples. Only a few studies require comment. Masserman and Carmichael (1938) followed up patients for a year, showing some diagnostic inconsistency for each category studied. They did not distinguish between changes in patients' conditions and discrepancies due to better histories available at follow-up. Norris (1959) studied the agreement between observation unit diagnoses in general hospitals and later diagnoses after transfer to psychiatric hospitals in London. This comparison is hampered by patient changes due to the natural history of disorders and to a longer observation period in hospital. The simultaneous rating and test-retest studies control for these problems. The best of these studies were by Schmidt and Fonda (1956), Beck, Ward, Mendelson, Mock, and Erbaugh (1962), and Sandifer, Pettus, and Quade (1964).

Schmidt and Fonda (1956) examined the agreement between diagnoses made by psychiatric residents and staff physicians, a design which causes two problems. The resident's history was a major source of staff diagnoses; this would push agreement up. On the other hand, the staff doctor had access to ward observations the resident did not; this would tend to lower kappa. These biases may not offset each other.

Ward et al. (1962) asked raters, who interviewed patients in a test-retest design, why they disagreed on diagnoses when they disagreed. It
turned out that a majority of disagreements occurred because raters used different criteria for diagnosis, or because raters found the diagnostic criteria in DSM-I to be impossibly fine distinctions. This suggests that a clearer, minimally inferential set of criteria, agreed on by all raters beforehand, would lead to much higher diagnostic reliability. Sandifer et al. (1964), using a simultaneous-rating design with 10 judges, obtained distinctly higher reliability for schizophrenia and personality disorder than did Beck et al. This may be due to design differences (simultaneous rater vs. test–retest).

Two other studies in Table 1 deserve comment. The investigators of the World Health Organization's International Pilot Study of Schizophrenia (World Health Organization, 1973) conducted reliability exercises within each of their nine world-wide study centers. Both simultaneous-rating and test–retest studies were done, using the Present State Examination (PSE; Wing, Birley, Cooper, Graham, & Isaacs, 1967) as their structured interview, but diagnosing patients without specific rules. The reliabilities for schizophrenic and affective disorder diagnoses were higher than most in Table 3, suggesting that reducing interviewer variability increased reliability. Unfortunately, almost no intercenter reliability data were gathered, and none reported. This was perhaps unavoidable given the vast distances separating many centers. Clearly, intracenter agreement and intercenter agreement can differ.

Spitzer, Endicott, Robins, Kurlansky, and Gurland (1975), involved in designing psychiatric interviews for many years, studied the reliability of diagnosing case records, mostly oriented toward the diagnosis of schizophrenia. The results, based on DSM-II (American Psychiatric Association, 1968) diagnoses, were disappointing. This may be partly due to the deficiencies of case records as the sole source of diagnostic information.

Reliability of Criteria-Based Diagnoses

Not satisfied with this state of affairs, and as part of piloting the NIMH Clinical Research Branch Collaborative Program on the Psychobiology of Depression (CDS), Spitzer, Endicott, and Robins (1978) developed the Research Diagnostic Criteria (RDC). These criteria showed the influences of Eli Robins and colleagues at Washington University, who developed the Feighner (or St. Louis group) criteria, and also of Spitzer and colleagues' experiences in designing interviews. With collaborators, these investigators conducted a 150-patient pilot study for the CDS, allowing the "debugging" of the RDC under field conditions. While the St. Louis group criteria considerably antedate the RDC, reliability studies of the former postdate the RDC, so I will discuss the RDC first.

The Spitzer et al. (1975) study cited above, which showed poor reliability of DSM-II diagnoses, was also a study of the reliability of an early draft of the RDC. Using the same case records as for the test of DSM-II reliability, but comparing Spitzer to a research assistant when both used RDC definitions of illnesses, much higher reliabilities for all RDC diagnoses studied were obtained. However, this could be due to their developing a private consensus about how to diagnose difficult cases. In that case other investigators would not replicate these results.

Table 2 shows that this did not happen. To obtain pooled figures that would allow generalization, I made weighted averages of kappas in the table (by Fisher z-transformation). The mean reliability for schizophrenia is .79, for mania or bipolar I .92, for major depression .85, and for alcoholism .94 (drug abuse, essentially I.0). Several facts are noteworthy.

First, investigators not identified with developing the RDC obtained results as good as those of Spitzer et al. Second, there is a modest positive correlation between sample size and the reliability obtained; this argues against a hypothesis that small studies relied on handpicked patients who are easy to diagnose. Similarly, it is not just handpicked raters who gave good reliability. Analyses of data for all 36 raters active in the CDS study (Andreasen, McDonald-Scott, Grove, Keller, Shapiro, & Hirschfeld, 1982) showed that within broad limits rater experience made no observable difference to reliability. Test–retest studies show reliabilities equal to, or little worse than, simultaneous-rating studies. Lifetime or past diagnoses are as reliable as current diagnoses (almost all studies looking at other than current diagnoses used lifetime diagnoses). This latter finding is hard to fathom, since subjects have to rely on their memories to answer questions about past illnesses. A last conclusion: schizophrenia may have benefited least from the advent of RDC and structured interviews.

Table 3 shows studies of St. Louis group criteria. One smaller study used lay judges and a highly structured interview (Coryell, Cloninger, & Reich, 1978), and is not comparable in aim to the other studies. Helzer, Clayton, Pambakian, Reich, Woodruff, and Revelly (1977) looked at the test–retest stability of several diagnoses using a checklist developed at Washington University to record interview data. Helzer and colleagues have been very interested in substituting lay interviewers for psychiatrists, and have made comparisons between them. They used a new interview schedule, the Renard Diagnostic Interview (RDI; Helzer, Robins, Croughan, & Welner, 1981). The results are complex, and their interpretation depends heavily on the assumption that the psychiatrist is a highly accurate judge. Helzer et al. (1981) found that lay judges were more sensitive
### Table 2. Interrater Reliability of RDC Diagnoses

<table>
<thead>
<tr>
<th>Study</th>
<th>M</th>
<th>N</th>
<th>Sc</th>
<th>Mn</th>
<th>MD</th>
<th>ASPD</th>
<th>Nr*</th>
<th>Alc (Drug)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andreasen et al. (1981)</td>
<td>T-RT</td>
<td>50</td>
<td>1.00*</td>
<td>.87</td>
<td></td>
<td>.94</td>
<td></td>
<td></td>
<td>5 centers; lifetime diagnosis</td>
</tr>
<tr>
<td></td>
<td>T-RT</td>
<td>Same</td>
<td></td>
<td>.88</td>
<td>.75</td>
<td>.72</td>
<td></td>
<td></td>
<td>Same day</td>
</tr>
<tr>
<td></td>
<td>SR</td>
<td>8</td>
<td>.68</td>
<td>.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6 month</td>
</tr>
<tr>
<td>Hesselbrock et al. (1982)</td>
<td>T-RT</td>
<td>42</td>
<td>.72</td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td>5 centers; many raters</td>
</tr>
<tr>
<td></td>
<td>T-RT</td>
<td>Same</td>
<td>.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Videotape</td>
</tr>
<tr>
<td>Hostetter et al. (1983)</td>
<td>SR</td>
<td>120</td>
<td>.86</td>
<td>.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SADS-L vs. DIS</td>
</tr>
<tr>
<td>Keller et al. (1981a)</td>
<td>T-RT</td>
<td>25</td>
<td>.52</td>
<td>.72</td>
<td>.85</td>
<td>.67</td>
<td></td>
<td></td>
<td>3-4 days; present diagnosis</td>
</tr>
<tr>
<td>Keller et al. (1981b)</td>
<td>T-RT</td>
<td>25</td>
<td>.60</td>
<td>.52</td>
<td>.82</td>
<td></td>
<td></td>
<td></td>
<td>Same day</td>
</tr>
<tr>
<td>Mazure &amp; Gershon (1979)</td>
<td>T-RT</td>
<td>47</td>
<td>1.00*</td>
<td>.88</td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td>Past diagnosis</td>
</tr>
<tr>
<td>Rounsaville et al. (1980)</td>
<td>SR</td>
<td>15</td>
<td>6 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Drug abusers</td>
</tr>
<tr>
<td></td>
<td>T-RT</td>
<td>11</td>
<td>Specific category K’s = .64-1.00</td>
<td></td>
<td></td>
<td></td>
<td>3-10 days</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>T-RT</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rounsaville et al. (1982)</td>
<td>T-RT</td>
<td>117</td>
<td>.47</td>
<td>.52</td>
<td>.71</td>
<td></td>
<td></td>
<td></td>
<td>6 months</td>
</tr>
<tr>
<td>Spitzer et al. (1974)</td>
<td>T-RT</td>
<td>100</td>
<td>.78</td>
<td>.59*</td>
<td>&lt; 1 week</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T-RT</td>
<td>Same</td>
<td>.65</td>
<td>.45*</td>
<td>&lt; 1-3 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spitzer et al. (1978)</td>
<td>SR</td>
<td>68</td>
<td>.80</td>
<td>.82</td>
<td>.88</td>
<td>.86(0.76)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study A</td>
<td>SR</td>
<td>Same</td>
<td>.75</td>
<td>.89</td>
<td>.97</td>
<td>.88(0.89)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SR</td>
<td>150</td>
<td></td>
<td>.98</td>
<td>.90</td>
<td>.97(0.95)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study B</td>
<td>SR</td>
<td>Same</td>
<td>.91</td>
<td>.93</td>
<td>.91</td>
<td>.98(1.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study C</td>
<td>T-RT</td>
<td>60</td>
<td>.65</td>
<td>.82</td>
<td>.90</td>
<td>1.00(1.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T-RT</td>
<td>Same</td>
<td>.73</td>
<td>.77</td>
<td>.71</td>
<td>.95(0.73)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study D</td>
<td>SR</td>
<td>49</td>
<td></td>
<td>.85</td>
<td></td>
<td>1.00(1.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study E</td>
<td>T-RT</td>
<td>29</td>
<td>.85</td>
<td>.76</td>
<td>.78−</td>
<td>1.00(1.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spitzer et al. (1975)</td>
<td>SR</td>
<td>120</td>
<td>.78-84</td>
<td>.76-92</td>
<td>.66-74</td>
<td>.66-66</td>
<td>.39-66</td>
<td>.66-1.00</td>
<td>Case records</td>
</tr>
</tbody>
</table>

*Headings for this and following tables: MD = Major Depression; AD = Affective Disorder (MD or Mn); Alc = Alcoholism; Drug = Abuse.

*Bipolar

*Clinical

*Major Depression or Mania.
than psychiatrists, who in turn tended to show higher specificity. But these trends were not strong and the two classes of interviewer usually agreed quite closely.

Table 4 gives results of reliability studies using DSM-III criteria. Originally there was no interview associated with the DSM-III, and studies used case summaries (Hyler, Williams, & Spitzer, 1982), structured interviews created for other purposes, or unstructured interviews (Spitzer, Forman, & Nee, 1979). In the reliability field trials of DSM-III, special efforts were made to involve many psychiatrists, so that these studies are certainly among the most representative of any done using specified diagnostic criteria: the raters are far from handpicked, though many were expert diagnosticians. The results of the field trials were analyzed in two separate phases: an initial report (Spitzer, et al., 1979), and a separate study comparing the reliability of diagnoses based on case vignettes to those based on interviews (Hyler, et al., 1982). The former report suggests that DSM-III, when used by clinicians, may be about as reliable as the RDC in the hands of research interviewers. Case record diagnoses of DSM-III disorders always tended to be a little less reliable than interview diagnoses, making chart review studies suspect.

The studies by Helzer, Stoltzman, Farmer, Brockington, Plesons, Singerman, and Works (1985), Helzer, Robins, McEvoy, Spitznagel, Stoltzman, Farmer, and Brockington (1985), Robins, Helzer, Ratcliff, and Seyfried (1982) and by Burman, Kanno, Hough, Escobar, and Forsythe (1983) were conducted as part of the NIMH Epidemiological Catchment Area program (ECA: Regier, Myers, Kramer, Robins, Blazer, Hough, Eaton & Locke, 1984). This multicenter collaborative program aims to interview 20,000 individuals in the community to determine prevalence of DSM-III-defined psychiatric disorders and risk factors for these disorders. The Diagnostic Interview Schedule (DIS; Robins, Helzer, Croughan, & Ratcliff, 1981) was developed for the ECA to allow lay interviewers to diagnose DSM-III (as well as St Louis group and RDC) disorders without much clinical judgment; practically every word the interviewer says is spelled out, and follow-up to almost every possible interviewee answer is dictated as well. This is perhaps not ideal for clinical research settings, but that remains to be seen. Reliability of diagnoses in the community based on the DIS will soon be published; the studies in Table 4 are based on lay or psychiatric interviewers examining identified patients. The reliability of the diagnoses appears satisfactory, though perhaps not quite as high as for the RDC. A most important finding is that the Spanish-language DIS reported on by Burman et al. (1983) gives comparable reliability to the English-language version, and yields nearly identical results when used to interview bilingual subjects twice.
A large number of studies have reported reliabilities for specific diagnoses, as opposed to the reliability of sets of criteria for multiple disorders. Some of these were reliability checks, or study patient inclusion for large studies following up patients over time or investigating the genetics of major psychiatric disorders. These are useful for judging how exacting the inclusion process was in various studies, but their bearing on the reliability of diagnoses may be limited by circumstances. This is because almost all studies examined selected patients for whom the diagnostic decision was artificially confined to one or a few possibilities. Since the patients and their diagnoses are often not randomly sampled from any meaningful population, it is hard to decide whether the special circumstances of a given study would bias reliability up (by reducing diagnostic choices) or down. For example, selecting patients who are all deluded makes the diagnosis schizophrenia-not schizophrenia less reliable, by excluding those in whom schizophrenia cannot even be suspected by any rater. Some studies have also recently looked at comparative reliabilities for differing definitions of some disorders, (e.g., borderline or schizotypal personality disorders or schizophrenia). One can collect a large number of such one- or two-diagnosis studies, but since they are not comparable to studies reviewed above, I shall not cover them here.

What Does the Future Hold?

It would appear that many diagnoses have been made more understandable and more reliable by introducing specified diagnostic criteria and structured interviews. Some still-difficult diagnoses can probably be made more reliably. Interviews for personality disorders (Stangl, Pfahl, Zimmerman, Bowers & Corenthal, 1985) and anxiety disorders (DiNardo, O'Brien, Barlow, Waddell & Blanchard, 1983) are being developed. Such efforts have already resulted in interim revisions of a number of DSM-III criteria. A burgeoning area is the diagnosis of psychiatric disorders in the community; the NIMH ECA is producing much data relevant to the reliability of diagnoses in persons not currently in treatment. Such data are critical for interpreting the meaning of “case finding” in psychiatric epidemiology. Last, the emergence of cross-national diagnostic interviews, which started with the PSE, has continued with the Spanish version of the DIS and with subsequent interviews as well. Wittchen and colleagues in Munich are working on a German translation of the DIS (Helzer, 1984), while a collaborative group of World Health Organization and Alcoholism, Drug Abuse, and Mental Health Administration person-