Journal of Consulting and Clinical Psychology
1996, Vol. 64, No. 6, 1285-1289

0022-006X/96/$3.00

METHODOLOGICAL DEVELOPMENTS

Measuring Diagnostic Agreement

James Langenbucher, Erich Labouvie, and Jon Morgenstern
Rutgers—The State University of New Jersey

Diagnostic agreement tests the reliability and concordance of diagnostic systems. The introduction of measures of agreement with reputations for baserate independence (e.g., Yule's *Y* and *Q*), and new studies occasioned by the publication of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV;* American Psychiatric Association, 1994) and the International Classification of Diseases—10 (ICD-10; World Health Organization, 1992) make it necessary to study the relationship of illness baserates to measures of agreement. Testing diagnostic concordance for diagnoses of drug dependence from the third edition of the *DSM* (American Psychiatric Association, 1980) versus *DSM-IV* diagnoses of drug dependence under 3 baserate conditions, it was found that Yule's *Y* and *Q* proved as vulnerable to differences in baserates as kappa or percent agreement and that specificity covaried with baserate rather than being fixed, as most theoretical discussions assume. The uncritical use of *Y* and *Q*, therefore, is likely to lead to optimistic interpretations of agreement. Kappa should be preferred for most purposes, although an adjustment to the computational formulas for *Y* and *Q* is presented that can diminish their positive bias.

Quantitative studies of diagnostic reliability (the agreement of different clinicians applying common diagnostic rules) and diagnostic concordance (the agreement of different diagnostic systems about an illness) have been refined gradually over the past 2 decades, beginning with the field trials from the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III;* American Psychiatric Association, 1980; see also Spitzer, Forman, & Nee, 1979). Now, new tests of reliability and concordance have become necessary as the fourth edition of the *DSM* (*DSM-IV;* American Psychiatric Association, 1994) and the International Classification of Diseases (ICD-10; World Health Organization, 1992) enter practice, replacing with revised rules those of earlier systems.

The simplest measure of reliability and concordance is *percent agreement.* This measure does not, however, control for chance agreement and, in most research, has been supplanted by Cohen's (1960) kappa (κ) and other coefficients. Cohen's kappa is affected by low sensitivity (the proportion of ill patients who satisfy a diagnosis) and specificity (the proportion of well patients who do not satisfy a diagnosis). Following important theoretical notes by Kraemer (1979) and Grove, Andreasen, McDonald-Scott, Keller, and Shapiro (1981), it was concluded that κ also is related to diagnostic baserate (illness prevalence); assuming that sensitivity and specificity remain constant, κ is attenuated as baserates approach the extreme values of 0 or 1.0. Grove et al. (1981) recommended that κ not be used when the baserate is 0.05 or less, although this precluded tests of diagnostic reliability or concordance in most population samples and some clinical ones.

This so-called "baserate problem" led Spitznagel and Helzer (1985) to recommend the use of Yule's (1912) *Y* coefficient when baserates are extreme. They claimed that *Y* is ". . . effectively independent of prevalence within the range of prevalence rates typical in psychiatry. As such, it can be used to compare concordances across studies in which the prevalence of disorder differs" (Spitznagel & Helzer, 1985, p. 727). This recommendation was surprising, because Spitznagel and Helzer had shown that, for fixed levels of sensitivity and specificity, *Y*, like κ decreases as baserates approach 0 or 1.0. The Spitznagel and Helzer article was immediately criticized by Kraemer (1987), Shrout, Spitzer, and Fleiss (1987), and Uebersax (1987). Shrout et al. (1987) noted that *Y* is not only baserate dependent, it is also based on a nonlinear (square root) transformation of the odds ratio and tends to become large if the cells in a 2×2 contingency table representing diagnostic disagreement are unbalanced. Spitznagel and Helzer (1985) had proposed an adjustment to compensate for this problem, but their critics were unconvinced. An alternative measure, Yule's *Q* (Clogg & Shihadeh, 1994), avoids a square root transformation of the odds ratio but shares with *Y* the tendency to be positively biased when the cells for disagreement are unequal.

In practice, few nosologists pay attention to these issues, using their preferred measure of diagnostic agreement—κ, *Y*, *Q* or percent agreement—on the basis of familiarity and custom.

James Langenbucher, Erich Labouvie, and Jon Morgenstern, Research Diagnostic Project, Center of Alcohol Studies, Rutgers—The State University of New Jersey.

This research was supported by Grant 2RO1-DA05688-04 from the National Institute on Drug Abuse.

We thank Dennis Gorman and Barbara S. McCrady for their comments on earlier drafts of this article. Peter E. Nathan, Kevin Miller, Wendy Boswell, Mary Jo Breiner, and more than 30 diagnosticians and site subdirectors, were helpful in conducting the study.

Correspondence concerning this article should be addressed to James Langenbucher, Research Diagnostic Project, Center of Alcohol Studies, Rutgers—The State University of New Jersey, Piscataway, New Jersey 08855-0969.

Most assume, based on the prior theoretical work introduced earlier, that high levels of diagnostic agreement are more, not less, difficult to achieve in normal samples because of their low illness baserates; population-based nosologic studies should, therefore, constitute conservative tests, allowing clinical researchers to proceed in apparent safety if their nosologies are proven reliable or concordant in normal samples.

It is our concern that this security is unfounded and may have developed because theoretical demonstrations of the relationship between baserates and diagnostic agreement such as those offered by Spitznagel and Helzer (1985) or Shrout et al. (1987) made the critical assumption that sensitivity and specificity remain constant across populations with different baserates. This assumption may not be tenable. It is reasonable instead to expect a shift in specificity under different baserate conditions because the proportion of cases that are true negatives (well patients who do not satisfy a diagnosis) may be much higher in normal samples than in clinical ones. Because specificity is determined by the proportion of true negatives in the sample, diagnoses may take in normal samples higher specificities, and lower specificities in clinical samples, than previous demonstrations suggest. Therefore, because measures of diagnostic agreement are dependent on specificity, the relationship between baserates and measures of diagnostic agreement may, in some instances, be the reverse of those expected by nosologists, with higher reliability and concordance in normal samples than in clinical ones.

In view of the possible covariation between baserate and specificity, it may be useful to complement with systematic empirical comparisons the theoretical discussions of coefficients of agreement to which we are accustomed. Therefore, in the study reported here, we used actual field data from a multisite study of substance users to examine empirically the behavior of various coefficients of agreement. We studied four *DSM-III* versus *DSM-IV* drug dependence diagnoses (alcohol, amphetamine, cannabis, and opiate dependence) under three different baserate conditions: (a) a highly concentrated (severe) clinical sample, in which all cases met at least one *DSM-IV* drug dependence criterion; (b) a "gated" clinical sample in which all cases reported at least regular use of the drug; and (c) an unrestricted clinical sample, in which nothing was assumed about the use of a particular drug. The *DSM-III* and *DSM-IV* differ in their definitions of drug dependence, so this study is properly viewed as a concordance study, but the findings speak to the paradigm of diagnostic agreement, reliability, and concordance, generally.

Method

Participants

Three-hundred seventy mixed substance users were recruited from eight clinical sites in five states in the northeastern United States. Most participants were male (81%), White (83%), and had at least a high school education (87%). Alcohol dependence was the diagnosis made most frequently, although most participants met criteria for more than one substance use disorder. Substance users were paid for participating, and informed consent was obtained before assessment. Main findings of this study have been reported elsewhere (e.g., Langenbucher, 1995; Langenbucher & Chung, 1995; Langenbucher, Morgenstern, Labouvie, Miller & Nathan, 1996; Langenbucher, Morgenstern, Labouvie, & Nathan, 1994a, 1994b; Langenbucher, Morgenstern & Miller, 1995; Langenbucher, Sulesund, Chung & Morgenstern, 1996; Morgenstern, Langenbucher & Labouvie, 1994).

Instrument

The Composite International Diagnostic Interview—Expanded Substance Abuse Module (CIDI-SAM; Robins, Cottler, & Babor, 1990), a fully structured, highly reliable interview for substance use problems, was administered to each participant by a masters or doctoral-level clinician experienced in assessing addictive disorders and extensively trained for the study (Langenbucher & Chung, 1995). *DSM-III* and *DSM-IV* diagnoses were made with CIDI-SAM algorithms developed at Washington University.

Procedure

The computational formulas for percent agreement, κ, Y, and Q are given in Figure 1. All are expressed as functions of the proportions in the cells of the contingency table—"a, b, c, and d"—both for the case of concordance (top of Figure 1) and reliability (bottom of Figure 1). According to Carey and Gottesman (1978), the empirical assessment of the reliability of a diagnosis or diagnostic system is not meaningful if two raters or instruments yield different baserates. When significantly different baserates are observed, it is prudent to suspect that different diagnostic constructs are being evaluated and the data can be tested for concordance but not reliability. Therefore, in this table we also show McNemar's test (Hays, 1973) to assess whether the raters or instruments being compared generate equivalent diagnostic baserates. If equality of baserates is found (i.e., McNemar's test is not significant), as it should be if reliability is being tested, the proportions b and c logically can be replaced with $B = (b + c)/2$ in the computational formulas in Figure 1 for κ, Y and Q.

For analyses of diagnostic concordance in the concentrated clinical sample, only participants meeting at least one *DSM-IV* dependence criterion for the drug were used. This restricted the sample to cases with high baserates of drug dependence. For analyses of diagnostic concordance in the gated clinical sample, only participants who reported more than experimental use of a substance (regular drinking, or use of a drug six times or more) were included. This restricted the sample to at least experienced users, with a moderately high diagnostic baserate. For analyses of diagnostic concordance in the unrestricted clinical sample, all 370 participants were used, whether they reported any lifetime use of the drug or not. (There is no unrestricted clinical sample for alcohol because all participants reported at least regular drinking). With these three nested samples, a total of 11 cross-classifications of *DSM-III* versus *DSM-IV* diagnoses were performed for alcohol, amphetamine, cannabis, and opiate dependence. The κ, Y, and Q and proportion (agreement) were calculated for each cross-classification.

Results

Baserates and Measures of Concordance

Differences in baserates and associated variations in the coefficients of agreement are shown in the upper panel of Table 1. It is plain in this table that the relationship between baserate and measure of agreement shares a common trend across drug categories and type of measure: Without exception, increases in baserate—even cases in which baserates move from an extreme low value to a moderate one—are associated with decreases in the measures of agreement. This finding may surprise nosologists who are accustomed to the view that the lowest level of diagnostic agreement will be found in samples with the lowest illness baserates. However, the finding is quite consistent with the theoretical relationships illustrated by Spitznagel and

CONCORDANCE

		System 1: No Dx	System 1: Dx
System 2	No Dx	a	b
	Dx	c	d

Coefficients of Agreement

$$Kappa = \frac{ad-bc}{ad-bc+\frac{1}{2}(b+c)}$$

$$Y = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$$

$$Q = \frac{ad-bc}{ad+bc}$$

$$percent(agree) = \frac{a+d}{a+b+c+d}$$

RELIABILITY

McNemar's Test:

$$\chi^2(df=1) = \frac{N(|b-c|-\frac{1}{N})^2}{(b+c)}$$

If χ^2 *not significant, and* $B = \frac{b+c}{2}$:

$$Kappa = \frac{ad-B^2}{ad-B^2+B}$$

$$Y = \frac{\sqrt{ad}-B}{\sqrt{ad}+B}$$

$$Q = \frac{ad-B^2}{ad+B^2}$$

$$percent(agree) = \frac{a+d}{a+b+c+d}$$

Figure 1. Definitions and computational formulas for coefficients of agreement. a, b, c, and d are proportions, with a + b + c + d = 1.0. Dx = diagnoses.

Helzer (1985) and Shrout et al. (1987): When baserates range between 0.2 and 1.0, κ decreases as the baserate increases, and Y decreases as the baserate increases from 0.1 to 1.0, even if specificity and sensitivity are presumed to remain constant. It is our view that this effect may be further exaggerated because of the systematic covariation of specificity with baserate.

Covariation of Baserate With Sensitivity/Specificity

To determine the extent to which specificity and sensitivity systematically covary with baserate, we computed specificities and sensitivities for both diagnostic systems relative to each other. Results are given in Table 2. As seen there, sensitivity was generally stable across different baserate conditions. In contrast, specificity showed a marked tendency to decrease as the baserate increased, as we had expected. In other words, as the diagnostic baserate increased, decreases in specificity were not offset by corresponding increases in sensitivity, as would be required to maintain stable coefficients of agreement (e.g., Grove et al., 1981). Thus, low baserates in unrestricted samples may, because of the covariation between baserate and specificity, be related to higher estimates of diagnostic agreement than will be found in more restricted clinical samples.

Computational Adjustments for Reliability Studies

Results of McNemar's tests confirmed that *DSM-III* and *DSM-IV* yielded significantly different baserates in all cases except for opiates. Because equivalent diagnostic baserates between *DSM-III* and *DSM-IV* were found for opiates, it was possible for illustrative purposes only to treat coefficients of agreement for opiate diagnoses as though they were tests of diagnostic reliability and to apply to them the computational adjustments—replacing the proportions b and c with $B = (b + c)/2$, as in the lower panel of Table 1—that we suggested should be used when testing diagnostic reliability. These additional coefficients of agreement are given in the bottom panel of Table 1. As shown there, replacing the proportions b and c with $B = (b + c)/2$ in the computational formulas for κ, Y and Q resulted in slightly more conservative estimates of agreement for Y and Q.

Discussion

This study showed that low baserates in broad research samples are related to higher measures of diagnostic agreement than are found in studies of more ill clinical groups. Coefficients of agreement may be high in unrestricted samples such as population studies merely on the basis of low baserates and high diagnostic specificity. Yule's Y and Q coefficients appear to offer no relief from this and are no more stable than κ under different baserate conditions. In addition, in demonstrating the substitution of $B = (b + c)/2$ for the proportions b and c in the calculation of reliability coefficients, we found no change in the size of κ but substantial shrinkage in Y and Q when the substitution was made. This suggests that, in reliability research, positive biases in the estimates of Y and Q that are due to differences in diagnostic baserate between the raters or diagnosticians being compared are likely to be more severe than biases in the estimates of κ. Therefore, κ should be the default measure in most situations, and the adjusted computational routines in Figure 1 must be used whenever Y and Q are preferred in studies of diagnostic reliability.

This study has a number of limitations that bear remark. *DSM-III* and *DSM-IV* are only two of a larger number of di-

Table 1
Base Rates and Coefficients of Agreement by Substance and Sample

Substance and sample	Base rate		Measure of agreement				% chance agreement
	DSM-III	*DSM-IV*	κ	*Y*	*Q*	% agreement	
Alcohol							
Unrestricted							
Gated	0.68	0.79	0.65	0.78	0.97	86	60
Concentrated	0.77	0.89	0.44	0.67	0.91	85	71
Amphetamines							
Unrestricted	0.13	0.02	0.18	0.65	0.87	88	85
Gated	0.34	0.06	0.13	0.44	0.74	69	64
Concentrated	0.40	0.07	0.11	0.38	0.66	64	59
Cannabis							
Unrestricted	0.09	0.15	0.61	0.79	0.97	92	78
Gated	0.14	0.22	0.59	0.74	0.96	88	71
Concentrated	0.21	0.37	0.53	0.70	0.95	80	58
Opiates							
Unrestricted	0.17	0.16	0.90	0.94	1.00	97	72
Gated	0.51	0.46	0.84	0.87	0.99	92	50
Concentrated	0.73	0.66	0.73	0.79	0.98	89	57
Opiates[a]							
Unrestricted			0.90	0.93	1.00		
Gated			0.84	0.84	0.98		
Concentrated			0.73	0.75	0.96		

Note. *DSM-III* = *Diagnostic and Statistical Manual of Mental Disorders*, third edition; *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition.
[a] Agreement treated as reliability problem, with computational adjustments as in Figure 1, lower panel.

agnostic systems that support interrater reliability and concordance studies, and substance use disorders constitute only one of many diagnostic categories that require this kind of research. Research on other psychiatric categories, diagnosed by other systems, may produce a different pattern of results (McGorry et al., 1995). Also, variations in baserate are in this study generated from samples that are not independent but are instead nested subsets of each other. The latter issue, however, may not be relevant if it is acknowledged that clinical populations, at least at a conceptual level, represent themselves nested subsets

Table 2
Sensitivity and Specificity of Each System Relative to the Other by Substance and Sample

Substance and sample	*DSM-III* relative to *DSM-IV*		*DSM-IV* relative to *DSM-III*	
	Sensitivity	Specificity	Sensitivity	Specificity
Alcohol				
Unrestricted				
Gated	0.84	0.92	0.98	0.61
Concentrated	0.84	0.83	0.98	0.38
Amphetamines				
Unrestricted	0.75	0.88	0.12	0.99
Gated	0.75	0.69	0.13	0.98
Concentrated	0.75	0.63	0.13	0.97
Cannabis				
Unrestricted	0.54	0.98	0.86	0.92
Gated	0.54	0.98	0.86	0.88
Concentrated	0.51	0.97	0.91	0.77
Opiates				
Unrestricted	0.97	0.97	0.88	0.99
Gated	0.97	0.88	0.88	0.97
Concentrated	0.97	0.73	0.88	0.92

Note. *DSM-III* = *Diagnostic and Statistical Manual of Mental Disorders*, third edition; *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition.

of normal populations. Thus, we think none of the aforementioned limitations affects the technical facts about the effects of baserates on measures of agreement that the present study makes clear.

Do our findings dispute Spitznagel and Helzer (1985), who reintroduced *Y* into diagnostic research more than a decade ago? No. Our clarification may, in fact, have been anticipated by them when they remarked, perhaps too ambiguously.

> We have found that *in the pure case of validity* (italics added) it is possible to define a measure of agreement [*Y*] that is closely related to κ and perfectly independent of the base rate. *This perfect independence of the baserate is lost when the situation shifts from pure validity,* (italics added) but a practical degree of base rate independence remains. (p. 726)

We found *Y* to have no such "practical degree" of independence but to be related to baserates in a manner similar to κ in an actual field test of diagnostic concordance.

A principal implication of this study is that satisfactory measures of diagnostic agreement found in normal samples should not be taken as proof that different diagnostic algorithms will select symptomatically similar cases from a clinical sample. Different systems or interviewers may, in broadly sampled studies, (a) agree about the diagnostic status of most of the very large number of cases that are true negatives (these are cases that are unaffected by illness and report no symptoms, causing high diagnostic specificity); (b) disagree entirely about which few cases in particular are ill, or warrant a diagnosis; and (c) still show good measures of agreement. Therefore, clinicians who assume consistency of results on the basis of diagnostic agreement proven in population-based studies or even in unrestricted clinical samples, in which coefficients of agreement are presumed to be conservatively tested, are naive. Estimates of diagnostic agreement with extremely ill participant groups may be only half as strong as those found in samples with lower illness baserates, because diagnostic specificity is not a constant, as theoretical demonstrations have assumed, but rather covaries systematically with baserate, as these data make clear. These facts suggest that, in new studies of diagnostic agreement driven by the publication of the *DSM–IV* and ICD-10, no good substitute will be found for the systematic, empirical comparison of diagnosticians (reliability research) and diagnostic algorithms (concordance research) as they do their work under a broad range of sampling conditions encountered in the field.

References

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (3rd ed.)*. Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington DC: Author.

Carey, G., & Gottesman, I. I. (1978). Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. *Archives of General Psychiatry, 35,* 1454–1459.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables.* Thousand Oaks, CA: Sage.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry, 38,* 408–413.

Hays, W. L. (1973). *Statistics for the social sciences (2nd ed.)*. New York: Holt, Rinehart & Winston.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44,* 461–472.

Kraemer, H. C. (1987). Letter. *Archives of General Psychiatry, 44,* 192–193.

Langenbucher, J. W. (1995). Studies of diagnosis and severity of substance use disorders. In J. D. Blaine, A. H., Horton, & L. H. Towle (Eds.), *Diagnosis and severity of drug abuse and drug dependence* (NIH Publication No. 95-3884, pp. 44–70). Rockville, MD: National Institute on Drug Abuse.

Langenbucher, J. W., & Chung, T. (1995). Onset and staging of DSM–IV alcohol dependence using mean age and survival/hazard methods. *Journal of Abnormal Psychology, 104,* 346–354.

Langenbucher, J. W., Morgenstern, J., Labouvie, E., Miller, K. & Nathan, P. E. (1996). On criterion weighting in the *DSM–IV. Journal of Consulting and Clinical Psychology, 64,* 343–356.

Langenbucher, J. W., Morgenstern, J., Labouvie, E., & Nathan, P. E. (1994a). Diagnostic concordance of substance use disorders in *DSM–III, DSM–IV* and ICD-10. *Drug and Alcohol Dependence, 36,* 193–203.

Langenbucher, J. W., Morgenstern, J., Labouvie, E., & Nathan, P. E. (1994b). Lifetime *DSM–IV* diagnosis of alcohol, cannabis, cocaine and opiate dependence: Six-month reliability in a multi-site clinical sample. *Addiction, 89,* 1115–1127.

Langenbucher, J. W., Morgenstern, J., & Miller, J. (1995). *DSM–III, DSM–IV* and ICD-10 as severity scales for drug dependence. *Drug and Alcohol Dependence, 39,* 139–150.

Langenbucher, J. W., Sulesund, D., Chung, T., & Morgenstern, J. (1996). Illness severity and self-efficacy as course predictors of *DSM–IV* alcohol dependence in a multisite clinical sample. *Addictive Behaviors, 21,* 543–553.

McGorry, P. D., Mihalopoulos, C., Henry, L., Dakis, J., Jackson, H. J., Flaum, M., Harrigan, S., McKenzie, D., Kulkarni, J., & Karoly, R. (1995). Spurious precision: Procedural validity of diagnostic assessment in psychotic disorders. *American Journal of Psychiatry, 152,* 220–223.

Morgenstern, J., Langenbucher, J. W., & Labouvie, E. (1994). The generalizability of the dependence syndrome across substances: An examination of some properties of the proposed *DSM–IV* dependence criteria. *Addiction, 89,* 1105–1113.

Robins, L. N., Cottler, L. B., & Babor, T. (1990). *Composite International Diagnostic Interview—Expanded Substance Abuse Module (CIDI-SAM).* St. Louis, MO: Authors.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44,* 172–177.

Spitzer, R. L., Forman, J. B., & Nee, J. (1979). *DSM–III* field trials: initial interrater diagnostic reliability. *American Journal of Psychiatry, 136,* 818–820.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the baserate problem in the kappa statistic. *Archives of General Psychiatry, 42,* 725–728.

Uebersax, J. S. (1987). Letter: *Archives of General Psychiatry, 44,* 193–194.

World Health Organization. (1992). *Draft international classification of diseases and related health problems (10th ed.)*. Geneva, Switzerland: Author.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75,* 581–642.

Received October 3, 1995
Revision received February 15, 1996
Accepted March 6, 1996 ■