Reprinted from the Archives of General Psychiatry
July 1967, Volume 17

262

Quantification of Agreement in Psychiatric Diagnosis

A New Approach

Robert L. Spitzer, MD; Jacob Cohen, MD; Joseph L. Fleiss, MS; and Jean Endicott, PhD, New York

THE LAST few years have seen a number of studies dealing with the problem of the reliability of psychiatric diagnoses. What makes it difficult to assimilate the various findings is the lack of uniform methods for quantifying the salient features of the data. Thus, one study will report an overall rate of perfect agreement of 54%,[1] while another will report an overall contingency coefficient of 0.714.[2] Still another will report that, given that one diagnostician has made a particular diagnosis, the probability that another diagnostician will make the same diagnosis is 0.57.[3]

Furthermore, as generally used, all of these methods suffer from one or more deficiencies which are illustrated using the hypothetical data of Table 1. (1) Chance agreement is not taken into account, or equivalently, the base rates at which the various diagnoses are made are not used to qualify the agreement measure. Thus, to merely report that the percent agreement of the data in Table 1 is 59% (summing the observed agreement in the diagonal cells) is entirely misleading. Consider the following hypothetical circumstance: diagnostician A makes the diagnosis of psychosis 70% of the time, neurosis 20%, and personality disorder 10%, but does so randomly (ie, without even examining the patient). Assume, furthermore, that B proceeds similarly, but at baseline rates of 80%, 10%, and 10%. The expected values under this hypothesis, based on elementary probability theory, are those given in parentheses in Table 1. For example, the probability that both will call a patient psychotic is 0.70 times 0.80, or 0.56. Thus, if the two diagnosticians merely operated randomly as above, the numbers in parentheses in Table 1 would be obtained and the agreement would likewise by 59%! (2) The method which evaluates an entire table of joint judgments by means of a χ^2 test and the associated contingency coefficient[4] in fact credits departures from chance as heavily in the disagreement as in the agreement direction. For example, Table 1 yields a contingency test χ^2 value of 7.89, which with four degrees of freedom is significant at the 10% level, and a contingency coefficient of 0.270. The fact that χ^2 and the contingency coefficient are not 0 only reflects the fact that A's assignments are *associated* with B's assignments more than chance, but not that they are in *agreement* more than chance. Note that in Table 1 all the departures from expected values are in the cells off the diagonal, that is, in cells representing disagreement. For example, the diagnosis of a patient as psychotic by A and as personality disorder by B occurs in nine cases, while chance alone would lead to an expectation of only seven such disagreements. (3) The most common measure used in reporting agreement, percent agreement, is virtually never accompanied by a significance test. Although χ^2 does provide significance tests, as shown above, it is ambiguous in assessing agreement as such. (4) All of these methods fail to distinguish degree of diagnostic disagreement. Thus, the disagreement between neurosis and personality disorder is treated as if it were as much a disagreement as that

Submitted for publication Jan 24, 1967.

From the Research Division, Washington Heights Community Service, New York State Psychiatric Institute, and Biometrics Research, New York State Department of Mental Hygiene (Dr. Spitzer); New York University (Dr. Cohen); Department of Biostatistics, Columbia University School of Public Health (Mr. Fleiss); and Biometrics Research, New York State Department of Mental Hygiene (Dr. Endicott).

Reprint requests to 722 W 168th St, New York 10032 (Dr. Spitzer).

between neurosis and psychosis, despite the obvious clinical difference in the gravity of the two discrepancies.

This paper presents a procedure for examining a series of pairs of psychiatric diagnoses which suffers from none of the above deficiencies. It takes chance agreement into account, it has a statistical rationale, and it differentiates degrees of disagreement.

Degrees of Diagnostic Deviance

In Fould's study[5] of the reliability of psychiatric diagnosis, a diagnostic agreement scale is described whereby all possible pairs from a limited set of diagnoses are assigned agreement scores ranging from 0 to 6. Sandifer[6] has worked out a more elaborate scheme which includes all of the official diagnoses as well as qualifying phrases listed in the American Psychiatric Association (APA) diagnostic manual.[7] For any pair of official diagnoses explicit instructions are provided for determining the level of disagreement on a 7-point scale. The method is as follows: all diagnoses are grouped into 13 major classes, such as acute brain syndromes, chronic brain syndromes, schizophrenia, etc. For each of the major classes the seven possible levels of disagreement provided for are defined for various combinations of the two diagnoses. Level 0 is defined as exact (diagnosis and qualifying phrase) agreement, level 1 represents slight disagreement, etc, and level 6 represents complete disagreement. If the pair is paranoid schizophrenia and catatonic schizophrenia, the level assigned is 1. If the pair is paranoid schizophrenia and paranoid state, the level assigned is 2. If the pair is paranoid schizophrenia and neurosis-depressive type, the level assigned is the maximum, 6.

The scaling of psychiatric diagnostic disagreement is necessarily judgmental. However, Sandifer found that when three experienced clinicians assigned levels of disagreement to pairs of actual diagnoses, the clinicians agreed within one point on the scale.

Two of us (R.L.S., and J.E.) made some minor modifications in Sandifer's procedure: some level assignments were altered, and the categories "nonspecific illness with mild symptomatology" and "no mental illness" were added; in addition, it was judged that the levels were not at equal intervals along a disagreement continuum. To increase the fidelity of the scaling to this judgment, the values 0, 1, 2, 3, 5, 7, and 9 were assigned to the seven levels of disagreement, thus increasing the interval lengths at the disagreement end of the scale.

Table 1.—*Hypothetical Data for Computational Illustration of Different Methods of Reporting Diagnostic Agreement

		Diagnostician B			
		Psychotic	Neurotic	Personality Disorder	Overall %
Diagnostician A	Psychotic	56 (56)	5 (7)	9 (7)	70
	Neurotic	18 (16)	2 (2)	0 (2)	20
	Personality disorder	6 (8)	3 (1)	1 (1)	10
	Overall %	80	10	10	100

In each cell (i), the upper entry is observed frequency, and the lower entry is chance expected frequency.

* Overall p_o agreement = 59%; overall p_c agreement = 59%; overall $\kappa = 0$.

$\chi^2 = 7.89$, df = 4, $P < 0.10$, contingency coefficient = 0.270.

The Measurement of Agreement

To measure agreement for nominal scales such as diagnostic category, an index of agreement,* kappa (K),[8] was developed. Kappa has the following properties: (1) it is the proportion of agreement corrected for, ie, over and above chance agreement; (2) it varies from negative values for less than chance agreement, through 0 for chance agreement, to +1.0 for perfect agreement; and (3) its sampling characteristics are known and therefore can be subjected to statistical significance testing.

To increase the utility of K for use in the study of diagnostic agreement, a modification called "weighted K" was developed.[9] Weighted K, K_w, shares the above noted properties of K; but whereas K does not distinguish among degrees of disagreement, weighted K does. Thus, weighted K provides for partial credit when disagreement is not complete.

*For the purist, it should be noted that it is "shared discrimination" rather than "agreement" that is indexed. In a situation where all of 100 patients were diagnosed schizophrenic by two diagnosticians the "agreement" is 100%, but the shared discrimination, in the absence of discrimination of this diagnosis from others, is indeterminate.

The formulae are:

$$K = \frac{p_o - p_c}{1 - p_c},$$

where p_o is the observed proportion of agreement, and p_c is the chance expected proportion of agreement as described above; and

$$K_w = 1 - \frac{\Sigma\ w_i p_{oi}}{\Sigma\ w_i p_{ci}},$$

where w_i is the disagreement level assigned to a cell (ie, to a given pair of diagnoses), p_{oi} is the observed proportion in that cell, p_{ci} is the chance proportion in the cell computed as described above, and the summation is over all cells. The diagonal cells, which represent complete agreement, are assigned $w_i = 0$, and thus if there is perfect agreement for all pairs, K_w is 1.

If, in using the hypothetical data in Table 2, the question is what is the chance corrected agreement between diagnosticians A and B, taking into account the levels of disagreement for the three diagnoses,† one employs weighted K, which is found to be 0.507. If one asks what the chance corrected agreement is between diagnosticians A and B with all disagreements treated as equal, one employs K. The unweighted K for the same data is 0.429. Since for the hypothetical data in Table 1 chance agreement is equal to observed agreement, K for overall agreement is 0.

If one asks what the chance corrected agreement is between diagnosticians A and B for any given diagnosis, such as psychosis, one collapses the data into a fourfold table to make the dichotomous distinction of psychosis-nonpsychosis, and employs K. (Since the distinction is dichotomous, unweighted kappa and weighted K are the same.) This is illustrated using the hypothetical data in Table 3, obtained by collapsing the data in Table 2 for the diagnosis of psychosis. The distinction of psychosis vs all other diagnoses is made with K of 0.596. When a similar operation is performed for neurosis vs all others, K is 0.450, while for personality disorders vs all others, K is 0.222. Thus, by computing both the weighted (or unweighted) K for all diagnoses and the unweighted K for each of the separate diagnoses, one obtains an overall measure of diagnostic agreement, as well as individual indices for separate diagnoses.

†The weights used for the paired diagnostic *categories* in the hypothetical data of Table 2 were choosen for illustrative purposes only. They should not be confused with the weights for levels of diagnostic disagreement assigned by us to pairs of *specific* diagnoses.

Table 2.—*Hypothetical Data for Computational Illustration of* K *and Weighted* K

		Diagnostician B							
		Psychotic		Neurotic		Personality Disorder		Total	Prop
Diagnostician A	Psychotic	106	0	10	9	4	5	120	(0.60)
		(0.530)	(0.390)	(0.050)	(0.150)	(0.020)	(0.060)		
	Neurotic	22	9	28	0	10	3	60	(0.30)
		(0.110)	(0.195)	(0.140)	(0.075)	(0.050)	(0.030)		
	Personality disorder	2	5	12	3	6	0	20	(0.10)
		(0.010)	(0.065)	(0.060)	(0.025)	(0.030)	(0.010)		
	Total	130		50		20		200	
	Proportion	(0.65)		(0.25)		(0.10)			(1.00)

In each cell (i), upper left is observed frequency, upper right is cell weight = w_i, lower left is cell observed proportion = p_{oi}, and lower right is cell chance proportion = p_{ci}.

* For weighted κ,

$$\Sigma\ w_i p_{oi} = 0\ (0.530) + 9\ (0.050) + 5\ (0.20) + 9\ (0.110) + 0\ (0.140) + 3\ (0.050) + 5\ (0.010) + 3\ (0.060) + 0\ (0.030) = 1.92, \text{ and}$$

$$\Sigma\ w_i p_{ci} = 0\ (0.390) + 9\ (0.150) + 5\ (0.060) + 9\ (0.195) + 0\ (0.075) + 3\ (0.030) + 5\ (0.065) + 3\ (0.25) + 0\ (0.010) = 3.895, \text{ and}$$

$$\kappa_w = 1 - \frac{1.92}{3.895} = 1 - 0.493 = 0.507.$$

For unweighted κ,

$$p_o = 0.530 + 0.140 + 0.030 = 0.700,\ p_c = 0.390 + 0.075 + 0.010 = 0.475, \text{ and}$$

$$\kappa = \frac{p_o - p_c}{1 - p_c} = \frac{0.700 - 0.475}{1 - 0.475} = \frac{0.225}{0.525} = 0.429.$$

Table 3.—*Hypothetical Data for Computational Illustration of Agreement for a Given Diagnosis

		Diagnostician B					
		Psychotic		All Others		Total	Prop
Diagnostician A	Psychotic	106 (0.530)	(0.390)	14 (0.070)	(0.210)	120	(0.60)
	All others	24 (0.120)	(0.260)	56 (0.280)	(0.140)	80	(0.40)
	Total	130		0.70		200	
	Proportion	(0.65)		(0.35)			(1.00)

* $p_o = 0.530 + 0.280 = 0.810$; $p_c = 0.390 + 0.140 = 0.530$.
κ for psychosis $= \frac{0.810 - 0.530}{1 - 0.530} = \frac{0.280}{0.470} = 0.596$.

Table 4.—*Summary Results From Computer Program for Calculating Psychiatric Diagnostic Agreement (N = 100) (Overall Weighted K = 0.60)*

Disagreement Level	No. of Pairs	%
0	37	37
1	17	17
2	21	21
3	10	10
5	2	2
7	7	7
9	6	6
Average:	2.0	

		Frequency		
Diagnostic Category	Kappa	1st Diag	2nd Diag	Agreement
Brain syndromes	0.85	6	8	6
Manic depressive, manic	—	0	0	0
Psychotic depressions	0.29	7	5	2
Schizophrenic reaction	0.73	63	60	55
Paranoid reaction	0	0	2	0
Psychophysiologic disorder	—	0	0	0
Psychoneurotic reaction	0.42	14	14	7
Personality disturbance	0.39	2	3	1
Sociopathic personality	0.88	5	4	4
Situational reaction	0.26	3	4	1
Mental deficiency	—	0	0	0
Slightly ill	—	0	0	0
Not ill	—	0	0	0

Computer Program

In connection with a study of computerized diagnosis which requires the comparison of diagnoses from a computer and from clinicians,[10] a computer program (KAPPA) was written to calculate the weighted and unweighted K both for overall agreement and agreement per diagnostic group. The program accepts 332 discrete, four-digit diagnoses: 66 diagnoses times five qualifying phrases from the APA diagnostic manual[7] plus two additional diagnoses, nonspecific illness with mild symptomatology and no mental illness. The printout lists the following: (1) for each patient, the two diagnoses given him and their disagreement level; (2) the frequency distribution for each diagnostician on an diagnoses given by either; (3) the overall weighted K; (4) the number of pairs of diagnoses at each disagreement level and the mean disagreement level; and (5) for each of 14 major diagnostic categories the unweighted K, the number of times each each of 14 major diagnostic categories the unweighted K, the number of times each diagnostician used the category, and the number of times they agreed on it. The program is written in FORTRAN IV for the computer (IBM 7094) and is easily modified for other diagnostic schemes and weights. The summary results obtained using this computer program on pairs of diagnoses assigned to 100 psychiatric patients by two psychiatrists are shown in Table 4. (These diagnoses were made by the treating resident and the attending psychiatrist on new admissions to the Washington Heights Community Service. Since they were not made completely independently, the values of K are probably inflated.)

If agreement were reported in terms of percent perfect agreement, the results in Table 4 would be only 37% (disagreement level 0). However, the overall weighted K indicates that taking into account levels of agreement, (ie, "partial credit"), the chance corrected agreement is 60%. The results in-

dicate that there is considerable variability in the level of agreement for different diagnostic categories: the greatest agreement was for the categories of brain syndromes and sociopathic personality, whereas the lowest agreement was for the psychotic depressions, situational reaction, and paranoid reaction. Although there is a small number of cases in each of these categories, so that the figures reported are not very stable, these findings are in keeping with those generally reported.

Summary

Several defects in current procedures used to assess diagnostic agreement are described. A new measure, weighted kappa (K_w), is described which allows for differences in the gravity of disagreement, ie, gives partial credit for less than complete disagreement; includes a correction for chance agreement; and has known statistical properties allowing for significance testing.

A computer program is described which accepts as input pairs of any ox 332 discrete psychiatric diagnoses on a series of cases, and computes an overall weighted K, Ks for individual diagnostic categories, and related statistics.

This work was supported in part by grant No. MH 08534 from the National Institute of Mental Health.

References

1. Beck, A.T., et al: Reliability of Psychiatric Diagnoses: 2. A Study of Consistency of Clinical Judgment and Rating, *Amer J Psychiat* **119**:351-357, 1962.

2. Schmidt, H.O., and Fonda, C.P.: The Reliability of Psychiatric Prognosis: A New Look, *J Abnorm Soc Psychol* **52**:262-267, 1956.

3. Sandifer, M.G.; Pettus, C.; and Quade, D.: A Study of Psychiatric Diagnosis, *J Nerv Ment Dis* **139**:350-356, 1964.

4. Peters, C.C., and Van Voorhis, W.R.: *Statistical Procedures and Their Mathematical Bases*, New York: McGraw-Hill Book Co. Inc., 1940.

5. Foulds, G.A.: The Reliability of Psychiatric, and the Validity of Psychological, Diagnosis, *J Ment Sci* **101**:851-862, 1955.

6. Sandifer, M.G.: Degrees of Diagnostic Deviance, Jan 1966.

7. Committee on Nomenclature and Statistics: *Diagnostic and Statistical Manual of Mental Disorders*, Washington, DC: American Psychiatric Association, 1952.

8. Cohen, J.: A Coefficient of Agreement for Nominal Scales, *Educ Psychol Measures* **20**:37-46, 1960.

9. Cohen, J.: Weighted Kappa: Nominal Scale Agreement With Provision for Degrees of Disagreement, *Amer Psychol*, to be published.

10. Spitzer, R.L., and Endicott, J.: A Computer Program for Psychiatric Diagnosis Utilizing the Differental Diagnostic Procedure, Scientific Proceedings of the American Psychiatric Association annual meeting Detroit, May 8-12, 1967.