Brit. J. Psychiat. (1974), 125, 341–7

A Re-analysis of the Reliability of Psychiatric Diagnosis

By ROBERT L. SPITZER and JOSEPH L. FLEISS

Introduction

Classification systems such as diagnosis have two primary properties, reliability and validity. Reliability refers to the consistency with which subjects are classified; validity, to the utility of the system for its various purposes. In the case of psychiatric diagnosis, the purposes of the classification system are communication about clinical features, aetiology, course of illness and treatment. A necessary constraint on the validity of a system is its reliability. There is no guarantee that a reliable system is valid, but assuredly an unreliable system must be invalid.

Studies of the reliability of psychiatric diagnosis provide information on the upper limits of its validity. This paper discusses some of the difficulties in appraising diagnostic reliability, offers a re-analysis of available data from the literature, and suggests a possible course of action to improve psychiatric diagnosis.

Zubin (1967) reviewed the major studies of the reliability of psychiatric diagnosis performed until 1966. He noted that diagnostic reliability is referred to in three different ways: agreement between independent diagnosticians examining the same patients, stability in diagnosis over time, and similarity in diagnostic frequencies for comparable samples. It is the first sense—interjudge agreement—that is fundamental.

There are inherent limitations to the interpretation of the other two uses of the term. For agreement between initial and subsequent diagnosis, one must consider the possibility that some of the disagreement may be due to changes in the patient's condition and not just to unreliability. The difficulty with interpreting differences in distributions between populations is that one is forced to assume, often without evidence, that the populations do not differ in psychopathology, when in fact they may.

Measuring Diagnostic Reliability

More studies of diagnostic reliability have been of the interjudge type than of either of the other two types. There are two features of the data reported in these studies, however, which limit an assimilation of their results. One is the choice of an index of agreement and the other is a failure to take into account the base rates of the various diagnoses. The hypothetical data of Table I will illustrate some of the complexities involved in measuring diagnostic agreement.

Table I

Hypothetical data (in proportions) for agreement on three diagnoses by two diagnosticians

Diagnostician A	Diagnostician B			
	Psychosis	Neurosis	Organic	Total
Psychosis ..	·75	·01	·04	·80
Neurosis ..	·05	·04	·01	·10
Organic ..	0	0	·10	·10
Total	·80	·05	·15	1·00

To measure the degree of agreement on a single diagnosis (e.g. neurosis), one may collapse the original table into a 2×2 table as shown in Table II. Some studies (Schmidt and Fonda, 1956; Kreitman, 1961) report the proportion of overall agreement, i.e., the proportion of all patients on whom there is agreement as to the

Table II

Hypothetical data (in proportions) for agreement on neurosis by two diagnosticians, from Table I

Diagnostician A	Diagnostician B		
	Neurosis	Other	Total
Neurosis ..	·04	·06	·10
Other ..	·01	·89	·90
Total	·05	·95	1·00

https://doi.org/10.1192/bjp.125.4.341 Published online by Cambridge University Press

presence or the absence of the diagnosis. For the data of Table II, the proportion of overall agreement is ·04+·89 = ·93.

Other studies (Beck *et al.*, 1962; Sandifer *et al.*, 1964) report the proportion of specific agreement, which is an index obtained by ignoring all subjects agreed upon as not having the given diagnosis (in Table II, ignoring the 89 per cent of patients agreed upon as not having a neurosis). One first determines the average proportion of all subjects given the specified diagnosis by either diagnostician (for the data of Table II, this proportion is ½ (·10+·05) = ·075), and then finds the proportion agreed upon as having that diagnosis (for the present example, this proportion is ·04). The proportion of specific agreement is the ratio of these two proportions. For the data of Table II the resulting value is ·04/·075 = ·53. This index can be interpreted as the probability that one diagnostician will make the specified diagnosis given that the other has done so.

Table III presents the values of the two indices of agreement for the three diagnoses of Table I. The two indices order the diagnoses quite differently. The proportions of overall agreement seem to be similar, with that for organic brain syndrome being best and that for neurosis being second best. The proportions of specific agreement are of different orders of magnitude, and indicate that agreement on psychosis is best and agreement on neurosis poorest.

Table III

Indices of agreement between two diagnosticians on three diagnoses of Table I

Diagnosis	Proportion of overall agreement		Proportion of specific agreement		
	Obtained	Chance expected	Obtained	Chance expected	Kappa
Psychosis ..	·90	·68	·94	·80	·69
Neurosis ..	·93	·86	·53	·07	·50
Organic ..	·95	·78	·80	·12	·77

The two indices are obviously not comparable. A further complication is that neither can be interpreted independently of the rates at which the diagnoses are made. For one thing, the values associated with the poorest possible agreement may be appreciably greater than 0. For example, given that the two diagnosticians diagnose psychosis 80 per cent of the time, the lowest value possible for the proportion of overall agreement on psychosis is ·60 and the lowest value possible for the proportion of specific agreement on psychosis is ·75.

Secondly, some degree of agreement is to be expected solely on the basis of chance. To take an extreme example, suppose that diagnosticians A and B jointly diagnosed a sample of patients without even examining them, but merely kept to their usual base rates. One would then expect that 64 per cent of the time (= ·8 × ·8), they would agree on the diagnosis of psychosis. Given their base rates, only agreement beyond that expected by chance alone would be meaningful.

A statistic for measuring agreement on nominal categories such as diagnosis, which incorporates a correction for chance, was originally proposed by Cohen (1960) and later generalized by Spitzer *et al.* (1967a), Cohen (1968), Fleiss (1971), Light (1971), and Fleiss *et al.* (1972). The statistic, named kappa, contrasts the observed proportion of agreement with the proportion expected by chance alone by means of the formula kappa $=(p_o - p_c)/(1 - p_c)$, where p_o is the observed proportion of agreement and p_c is the proportion expected by chance.

Whether p_o is taken to be the proportion of overall agreement or the proportion of specific agreement, one obtains identically equal values of kappa after correcting for chance. The term p_c is obtained by determining expected cell frequencies (as one does, e.g., in calculating the standard chi square statistic), and then calculating the proportion of agreement on the table with expected frequencies. Kappa varies from negative values for less than chance agreement, though 0 for chance agreement, to +1·0 for perfect agreement. Kappa may be interpreted as an intra-class correlation coefficient (Fleiss and Cohen, 1973).

Table III gives the chance expected values of the two proportions of agreement and the resulting values of kappa. The ordering effected by kappa is different from either of the other two orderings. After correcting for chance, one finds agreement to be best for organic brain syndrome, next best for psychosis, and poorest for neurosis.

Studies of Diagnostic Reliability

The major studies of the reliability of psychiatric diagnosis, fortunately, report both the base rates and the diagnostic agreement values, thus permitting the calculation of chance corrected agreement, kappa.

(I) Schmidt and Fonda (1956) studied 426 patients admitted to a state hospital in Connecticut. Each patient was diagnosed within the first week of admission by one of a group of eight psychiatric residents, and within the third week by one of three senior psychiatrists. The data available to the psychiatric residents were the usual admission reports as well as their own physical and mental status examination. The data available to the senior psychiatrist included all of the data available to the psychiatric residents as well as additional data that had been collected by other staff members and by themselves during their own brief examinations.

(II) Kreitman (1961) studied 90 consecutive new referrals to an out-patient clinic in England. Each patient was interviewed by one of three consulting psychiatrists, and completely independently by one of two research psychiatrists. The only sources of information to both sets of psychiatrists were the patient, a family member and a letter of referral.

(III) Beck *et al.* (1962) studied 153 patients randomly selected from new referrals to two out-patient services in Philadelphia. Each patient was randomly assigned to be interviewed by two of four experienced psychiatrists. Each psychiatrist conducted an independent interview and apparently had no source of information other than the patient himself.

(IV) Sandifer *et al.* (1964) studied 91 patients from three hospitals in North Carolina. A psychiatric resident presented material about each patient to a group of ten experienced psychiatrists. Following each presentation the patient was interviewed by one of the 10 diagnosticians. After jointly observing the patient, each diagnostician made his own diagnosis.

(V) The U.S.–U.K. Diagnostic Project (Cooper *et al.*, 1972) conducted a series of studies comparing diagnostic practice in the United States and the United Kingdom. In one study, 250 consecutive admissions to a single New York State mental hospital and 250 consecutive admissions to a London area mental hospital were diagnosed by the hospital physician according to his usual practices, and independently by members of the project, who used a structured interview schedule. In a second study, 192 consecutive admissions to nine New York State mental hospitals and 174 consecutive admissions to nine London area mental hospitals were studied similarly. Most of the project members had received their psychiatric training in London. Because the results of the two studies .within for each city were similar, only mean agreement values for the New York and the London samples are reported. The agreement measured is between the project's and the hospitals' psychiatrists.

(VI) Spitzer *et al.* (in preparation) studied 100 consecutive admissions to the Washington Heights Community Service of the New York State Psychiatric Institute. Each patient was diagnosed by one of 15 admitting residents within the first few days of admission. Each patient was also diagnosed up to three months after admission by one of two supervising psychiatrists after reviewing the case record prepared by the admitting resident. No attempt was made to prevent the admitting therapist from discussing his diagnostic formulation with the supervising psychiatrist. It is assumed that such discussions often took place, though not invariably.

Results

Table IV presents the values of kappa calculated from the data presented in the original reports. Values are reported only for those categories for which original data were provided. Although the different studies used slightly different classification schemes (American Psychiatric Association, 1952 and 1968; H.M. Stationery Office, 1968), the results are reported for broad categories whose definitions

https://doi.org/10.1192/bjp.125.4.341 Published online by Cambridge University Press

TABLE IV

Kappa coefficients of agreement on broad and specific diagnostic categories from six studies

	Study							
Category	I	II	III	IV	V		VI	Mean
					New York	London		
Mental deficiency				·72				·72
Organic brain syndrome ..	·82	·90					·59	·77
Acute brain syndrome ..				·44				·44
Chronic brain syndrome ..				·64				·64
Alcoholism					·74	·68		·71
Psychosis	·73	·62		·56	·42	·43	·54	·55
Schizophrenia	·77		·42	·68	·32	·60	·65	·57
Affective disorder					·19	·44	·59	·41
Neurotic depression			·47		·20	·10		·26
Psychotic depression				·19	·24	·30		·24
Manic-depressive				·33				·33
Involutional depression ..			·38	·21				·30
Personality disorder or Neurosis..	·63			·51	·24	·36		·44
Personality disorder			·33	·56	·19	·22	·29	·32
Sociopathic			·53					·53
Neurosis		·52		·42	·26	·30	·48	·40
Anxiety reaction			·45					·45
Psychophysiological reaction ..				·38				·38

are similar in all of the classification systems used.

There are no diagnostic categories for which reliability is uniformly high. Reliability appears to be only satisfactory for three categories: mental deficiency, organic brain syndrome (but not its subtypes), and alcoholism. The level of reliability is no better than fair for psychosis and schizophrenia and is poor for the remaining categories. Using uncorrected agreement values, Zubin (1967) found agreement on the combined category of personality disorder and neurosis to be almost as high as for psychosis. It is clear from Table IV that after correction for chance, agreement on the combined category is poorer than on psychosis.

With the exception of the U.S.–U.K. study (number V) of the New York hospitals, all the studies summarized here involved diagnosticians of similar background and training. In addition, special efforts were made in some of the studies to have the participant diagnosticians come to some agreement on diagnostic principles prior to the beginning of the study. One would have expected these features of similar background and prior consensus on principles to contribute to good reliability. One can only assume, therefore, that agreement between heterogeneous diagnosticians of different orientations and backgrounds, as they act in routine clinical settings, is even poorer than is indicated in Table IV. Further, there appears to have been no essential change in diagnostic reliability over time (the studies summarized in Table IV were arrayed in chronological order).

DISCUSSIONS AND CONCLUSIONS

In spite of the obvious unreliability of psychiatric diagnosis, there exists evidence for sensitivity to and agreement on the major psychiatric problems experienced by a patient. Gurland *et al.* (1972), in a detailed analysis of data on the patients in the U.S.–U.K. diagnostic study, found that hospital psychiatrists were sensitive to patient's psychopathology. A number of patients in the New York sample were identified by the project psychiatrists as suffering from severe depression but not from any signs of schizophrenia. The hospital psychiatrists diagnosed most of these severe depressives as schizophrenic, but treated the majority of them with anti-depressant medication or with ECT. The hospital staffs obviously recognized the depression in their patients, when it was present, but failed to incorporate that recognition into their diagnoses.

As one of its studies of diagnostic practice, the U.S.–U.K. diagnostic project showed videotape recordings of a small number of psychiatric interviews to large numbers of American, British, and Canadian psychiatrists (Copeland *et al.*, 1971; Kendell *et al.*, 1971; Sharpe *et al.*, in press). Some of the interviews gave rise to strikingly large diagnostic differences between the three countries; in one case the percentage of psychiatrists diagnosing schizophrenia ranged from 2 per cent in the British Isles to 69 per cent in the United States, the proportion for Canadian psychiatrists being intermediate. In another study, Sandifer *et al.* (1968) reported substantial diagnostic differences between American, English and Scottish psychiatrists.

The participant psychiatrists in the videotape studies also judged the presence or absence of technically described psychiatric signs and symptoms, and made ratings on the Inpatient Multidimensional Psychiatric Scales (IMPS) of Lorr *et al.* (1962), a series of 89 rating scales defined in non-technical language. As Katz *et al.* (1969) found in an earlier study, the U.S.–U.K. study found poor agreement between psychiatrists in judging the presence or absence of symptoms described in technical terms. With respect to ratings on the IMPS, however, there were striking similarities in the psychiatrists' perceptions of psychopathology. Although American psychiatrists tended to rate the presence of more severe pathology than their British and Canadian colleagues, all psychiatrists were in excellent agreement as to the most serious and the least serious problem areas. In other words, mean profiles across the factors of the IMPS were at different mean levels, but were effectively parallel. This parallelism obtained for each of the tapes shown, even though the profile for each tape highlighted different aspects of psychopathology.

The reliability of psychiatric diagnosis as it has been practised since at least the late 1950's is not good. It is likely that the reasons for diagnostic unreliability are the same now as when Beck *et al.* (1962) studied them. They found that a significant amount of the variability among diagnosticians was due to differences in how they elicited and evaluated the necessary information, and that an even larger amount was due to inherent weakness and ambiguities in the nomenclature. Since that time there have been two major innovations which may provide solutions to these problems.

Several investigators have developed structured interview schedules which an interviewer uses in his examination of the patient (Spitzer *et al.*, 1967b and 1970; Wing *et al.*, 1967). These techniques provide for a standardized sequence of topics, and ensure that variability among clinicians in how they conduct their interviews and in what topics they cover is kept to a minimum. For rating the pathology observed, these schedules contain precoded items which explicitly define the behaviours to be rated rather than relying on technical terms which have different meanings to different clinicians.

With respect to improving the nomenclature, the St. Louis group (Feighner *et al.*, 1972) has offered a system limited to 16 diagnoses for which they believe strong validity evidence exists, and for which specified requirements are provided. Whereas in the standard system the clinician determines to which of the various diagnostic stereotypes his patient is closest, in the St. Louis system the clinician determines whether his patient satisfies explicit criteria. For example, for a diagnosis of the depressive form of primary affective disorder the three requirements are dysphoric mood, a psychiatric illness lasting at least one month with no other pre-existing psychiatric condition, and at least

https://doi.org/10.1192/bjp.125.4.341 Published online by Cambridge University Press

five of the following eight symptoms: poor appetite or weight loss; sleep difficulty; loss of energy; agitation or retardation; loss of interest in usual activities or decrease in sexual drive; feelings of self-reproach or guilt; complaints of or actually diminished ability to think or concentrate; and thoughts of death or suicide.

A consequence of the St. Louis approach is the necessity for an 'undiagnosed psychiatric disorder' category for those patients who do not meet any of the criteria for the specified diagnoses. In actual use, this category is applied to 20–30 per cent of newly-admitted in-patients.

These two approaches, structuring the interview and specifying all diagnostic criteria, are being merged in a series of collaborative studies on the psychobiology of the depressive disorders sponsored by the N.I.M.H. Clinical Research Branch. We are confident that this merging will result not only in improved reliability but in improved validity which is, after all our ultimate goal.

ACKNOWLEDGEMENTS

This research was supported in parts by grants MH 23864 and MH 23964 from the National Institute of Mental Health. Professor Jacob Cohen of New York University made many valuable contributions to this report.

REFERENCES

AMERICAN PSYCHIATRIC ASSOCIATION (1952) *Diagnostic and Statistical Manual of the Mental Disorders.*

—— (1968) *Diagnostic and Statistical Manual of the Mental Disorders. 2nd Edition.*

BECK, A. T., WARD, C. H., MENDELSON, M., MOCK, J. E., & ERBAUGH, J. K. (1962). Reliability of psychiatric diagnoses: 2. A study of consistency of clinical judgments and ratings. *Amer. J. Psychiat.*, **119**, 351–7.

COHEN, J. (1960) A coefficient of agreement for nominal scales. *Educ. psychol. Measmt.*, **20**, 37–46.

—— (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, **70**, 213–20.

COOPER, J. E., KENDELL, R. E., GURLAND, B. J., SHARPE, L., COPELAND, J. R. M. & SIMON, R. (1972) *Psychiatric Diagnosis in New York and London.* (U.S.–U.K. Diagnostic Project.) London: Oxford University Press.

COPELAND, J. R. M., COOPER, J. E., KENDELL, R. E., & GOURLAY, A. J. (1971) Differences in usage of diagnostic labels amongst psychiatrists in the British Isles. *Brit. J. Psychiat.*, **118**, 629–40.

FEIGHNER, J. P., ROBINS, E., GUZE, S. B., WOODRUFF, R. A., WINOKUR, G. & MUNOZ, R. (1972) Diagnostic criteria for use in psychiatric research. *Arch. gen. Psychiat.*, **26**, 57–63.

FLEISS, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychol. Bull.*, **76**, 378–82.

—— & COHEN, J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. psychol. Measmt.*, **33**, 613–19.

—— SPITZER, R. L., ENDICOTT, J. & COHEN, J. (1972) Quantification of agreement in multiple psychiatric diagnosis. *Arch. gen. Psychiat.*, **26**, 168–71.

GURLAND, B. J., FLEISS, J. L., SHARPE, L., SIMON, R. & BARRETT, J. E. (1972) The mislabeling of depressed patients in New York State hospitals. *Disorders of Mood* (eds. J. Zubin & F. A. Freyhan), pp. 17–28. Baltimore: Johns Hopkins Press.

HER MAJESTY'S STATIONERY OFFICE (1968) *A Glossary of Mental Disorders.* General Register Office Studies on Medical and Population Subjects, no. 22.

KATZ, M. M., COLE, J. O. & LOWRY, H. A. (1969) Studies of the diagnostic process: The influence of symptom perception, past experience, and ethnic background on diagnostic decisions. *Amer. J. Psychiat.*, **125**, 937–47.

KENDELL, R. E., COOPER, J. E., GOURLAY, A. J., COPELAND, J. R. M., SHARPE, L. & GURLAND, B. J. (1971) The diagnostic criteria of American and British psychiatrists. *Arch. gen. Psychiat.*, **25**, 123–30.

KREITMAN, N. (1961) The reliability of psychiatric diagnosis. *J. ment. Sci.*, **107**, 876–86.

LIGHT, R. J. (1971) Measures of agreement for qualitative data: Some generalizations and alternatives. *Psychol. Bull.*, **76**, 365–77.

LORR, M., MCNAIR, D. M., KLETT, C. J. & LASKY, J. J. (1962) Evidence of ten psychiatric syndromes. *J. consult. Psychol.*, **26**, 185–9.

SANDIFER, M. G., HORDERN, A., TIMBURY, G. C. & GREEN, L. M. (1968) Psychiatric diagnosis: A comparative study in North Carolina, London and Glasgow. *Brit. J. Psychiat.*, **114**, 1–9.

—— PETTUS, C. & QUADE, D. (1964) A study of psychiatric diagnosis. *J. nerv. ment. Dis.*, **139**, 350–6.

SCHMIDT, H. O. & FONDA, C. P. (1956) The reliability of psychiatric diagnosis: A new look. *J. abnor. soc., Psychol.*, **52**, 262–7.

SHARPE, L., GURLAND, B. J., FLEISS, J. L., KENDELL, R. E., COOPER, J. E. & COPELAND, J. R. M. Some comparisons of American, Canadian and British psychiatrists in their diagnostic concepts. *Canad. J. Psychiat.* In press.

SPITZER, R. L., COHEN, J., FLEISS, J. L. & ENDICOTT, J. (1967a) Quantification of agreement in psychiatric diagnosis: A new approach. *Arch. gen. Psychiat.*, **17**, 83–7.

—— ENDICOTT, J., COHEN, J. & FLEISS, J. L. Constraints on the validity of computer diagnosis. (In preparation).

—— —— FLEISS, J. L. & COHEN, J. (1970) Psychiatric Status Schedule: A technique for evaluating psychopathology and impairment in role functioning. *Arch. gen. Psychiat.*, **23**, 41–55.

—— FLEISS, J. L., ENDICOTT, J. & COHEN, J. (1967b) Mental Status Schedule: Properties of factor analytically derived scales. *Arch. gen. Psychiat.*, **16**, 479–93.

https://doi.org/10.1192/bjp.125.4.341 Published online by Cambridge University Press

WING, J. K., BIRLEY, J. L. T., COOPER, J. E., GRAHAM, P. & ISAACS, A. D. (1967) Reliability of a procedure for measuring and classifying 'present psychiatric state'. *Brit. J. Psychiat.*, **113**, 499–515.

ZUBIN, J. (1967) Classification of the behavior disorders. In *Annual Review of Psychology* (eds. P. R. Farnsworth & O. McNemar). Palo Alto, California, *Annual Reviews*, pp. 373–406.

A synopsis of this paper was published in the June 1974 *Journal*.

Robert L. Spitzer, M.D., *Director of Evaluation Section, Biometrics Research, New York State Department of Mental Hygiene at the Psychiatric Institute, 722 West 168 Street, New York, New York 10032; and Associate Professor of Clinical Psychiatry, Columbia University, New York*

Joseph L. Fleiss, Ph.D., *Head of Biostatistics Section, Biometrics Research, New York State Department of Mental Hygiene at the Psychiatric Institute; and Associate Professor of Biostatistics, Columbia University, New York*

(*Received 17 January 1974*)

https://doi.org/10.1192/bjp.125.4.341 Published online by Cambridge University Press