Cross-National Study of Diagnosis of the Mental Disorders: Methodology and Planning

BY JOSEPH ZUBIN, PH.D.

The author describes the instruments that have been developed to provide an objective basis for making reliable diagnoses in the cross-national study. Armed with these tools, the investigators designed an experimental approach to examining patients admitted to a London and a New York mental hospital. The author also outlines the next steps in the study.

The seed from which this study grew was planted during the Work Conference on Field Studies in the Mental Disorders, held under the joint auspices of the American Psychopathological Association and Biometrics Research in 1959 under a grant from the National Institute of Mental Health and reported in book form in 1961 (19).

The group of outstanding research workers who attended came to the conclusion that the diagnostic and classification problems in the mental disorders could not be resolved through conferences. Only an actual field study, in which patients from different cultures or nationalities were examined by indigenous clinicians and by clinicians from other cultures, could help clarify the complex issues involved. With the help of the late Dr. Paul H. Hoch, and of Drs. Benjamin Pasamanick and Morton Kramer, a grant was obtained from the NIMH to begin a field study which we are now reporting.

The differences uncovered in the conference and the variety of factors that might have given rise to them were so numerous that some limitation had to be made on the scope of the study. I had suggested that we ought to begin with mental retardation rather than with mental disease because of the availability of criterion tests for mental retardation and the lack of such tests for mental disease. This did not evoke great enthusiasm, and finally the area of choice fell on the functional psychoses, in which some of the greatest differences of opinion arise. In the beginning we decided that the countries to be chosen for comparison should have as much in common as possible and yet present at least apparent differences in the incidence and prevalence of the functional psychoses.

The discrepancy in the national statistics of the U. S. and U. K. with regard to the relative frequency of affective and schizophrenic disorders offered a ready preliminary target for investigation. By limiting ourselves to these two countries we avoided the baffling linguistic barrier that would be involved if a non-English-speaking country were chosen. Second, the rather highly developed state of British and American psychiatry would ensure that the final result would take into consideration the rich cultural heritage in the field of psychopathology in both countries and would benefit from the contrast in points of view.

In our initial study we wanted to determine if possible whether the discrepancy in national statistics on hospitalized patients represented a real or only an apparent difference. In the process of answering this rather limited question we hoped to be able to work out the techniques and methods that would enable us to study the more
basic question of the distribution of mental disorders in both countries, the possible differentials that might exist, and the reasons for their existence.

In an attempt to answer the questions of the initial study regarding the relative frequencies of affective and schizophrenic disorders we had to find, first of all, a method for arriving at a diagnosis in both countries that would be objective, reliable, and valid—that is, free as far as possible of local prejudices. Second, we had to be prepared to trace the sources of the apparent or real discrepancy wherever they might be found.

Review of Diagnosis and Classification

With regard to developing an objective, reliable, and valid method for diagnosis in both countries, a review was made of the current state of diagnosis in the world literature(18). The results indicated that despite the shortcomings of the present diagnostic system, the degree of agreement on diagnosis was quite good with regard to the broad categories of organic and functional psychoses, neuroses, and personality disorders, although it was rather low for some of the specific conditions within these categories. With regard to schizophrenia, the agreement reached up to 70-80 percent. For the affective disorders it was a little more variable, but the same general degree of agreement was attained.

It should be remembered, however, that in most of these studies the agreement refers to diagnosis of patients hospitalized for unspecified durations. Whether similar degrees of agreement could be reached on new admissions is debatable. To eliminate diagnosis completely because of imperfect agreement and to accept instead the view that mental disorders are nothing more than the response of an essentially normal individual to severe stress and strain seems unwarranted at this point. Somewhere between the full acceptance of current nosology and its complete rejection lies the middle view—that the current diagnostic system represents a good starting point from which to improve approaches to classification.

To improve current diagnostic procedures it is necessary to find out the sources of disagreement. Three basic sources of variability have been described by Ward and associates(15) in an experimental investigation of this problem: inconstancy in the patient's behavior accounted for five percent of the variability; inconsistent behavior on the part of the diagnostician, 32.5 percent; and inadequate guidelines in the nosological system, 62.5 percent.

Apparently the major portion of the disagreement is attributable to the diagnostic system itself; next in importance is the variability of the diagnostician; while the variance attributable to the patient is relatively negligible. If we could introduce more systematic structured approaches to interviewing, so that each clinician would cover the same ground with each patient as far as is consistent with clinical needs, we could stand to gain an almost 33 percent increase in accuracy—or at least in agreement. The shortcomings of the nosological system may, of course, be reduced by carefully overhauling the diagnostic schema—something which the American Psychiatric Association is now doing. With the reduction of these two major sources of error, agreement at least could be enhanced.

In order to eliminate or reduce as much as possible the sources of the variability due to the interviewer, a review was undertaken of the various attempts that have been made to eliminate them. One of the methods now in vogue for this purpose is the use of rating scales for evaluating patient behavior, ranging from such techniques as the early Malamud and Sands Psychiatric Rating Scale to the more modern version of Lorr's Inpatient Multi-dimensional Psychiatric Scale, the Overall and Gorham Brief Psychiatric Rating Scale, and similar scales by Wittenborn and others. These scales have demonstrated their efficacy in specific studies for the evaluation of drug therapy and are a great help in the identif-
cation of target symptoms and their elimination by specific treatments.\(^2\)

The most obvious strength of the rating scale approach is that large and amorphous terms (such as "anxiety") are fractionated into smaller units of behavior on which it is easier in principle to reach agreement. The prime source of variability that Ward points to—inadequate guidelines in the nosological system—is thus reduced.

The rating scale approach has one outstanding defect, namely, that it generally depends upon unstructured unsystematic interviewing and observational techniques for its data. In order to overcome this defect, systematic structured interviews had to be developed that would provide the data to be rated and the rules for the objective ratings themselves. Fortunately, both the Biometrics Research group in New York(12, 14) and the Maudsley Hospital group in London(16) were already working on such instruments prior to the beginning of the current project. We will describe the results of their combined efforts later.

Sources of Variability

The reduction in variability in incidence and prevalence of mental disorders due to the method of interviewing would still leave other sources of variability intact. Among these are certain biasing factors: 1) the tolerance of the community to various types of behavioral deviation, 2) the utilization of mental health facilities by the population, and 3) the training of the clinician. Since the initial detection of a mental disorder is made not by a psychiatrist but by a layman—the patient himself, his family, friends, policeman, etc.—the local norms regarding the tolerance for mental disorders become important, and any differences in the attitudes of the communities toward mental disorder would produce a difference in the number detected.

The pattern of utilization of hospital facilities—whether all socioeconomic classes, for example, use the hospital equally freely, or any other selective factors affected hospitalization—would tend to introduce differences in the number and kinds of patients hospitalized.

The training of the psychiatrist and the school to which he belongs might also introduce differences. The method used in determining the diagnosis would play a part.

The variation due to local social-cultural norms in bringing patients to the attention of health facilities cannot practically be eliminated but can be evaluated by means of anthropological and sociological investigations. Furthermore, actual field investigations of the general population with a uniform and sensitive screening instrument, perhaps of the type of structured interview to be described, would yield information on this source of variability. The training of the psychiatrist, the school to which he belongs, and the particular methods he uses in arriving at a diagnosis (the relative weights he attaches to the various symptoms) need also not be altered in order to attain uniform results, but their effects on the resulting diagnostic label can be determined by comparison with the profiles of behavior noted on the objective structured interview. By means of videotapes, the particular style and predilection of the diagnostician can be determined.

This has been demonstrated in a recent study(6) in which the results of viewing the same videotape yielded a difference of opinion regarding diagnosis, one group of viewers diagnosing the same patient as neurotic, the other group as psychotic. On examination of the average profile of ratings based on the observations of the two groups, it was found that those who diagnosed neurosis perceived little apathy, while those who diagnosed psychosis perceived much apathy in the patient's behavior. This is an example of specifying particular aspects of functioning that have a strong weighting in the eventual diagnostic decision.

Validity Studies

We have thus far concerned ourselves primarily with the reliability of the diagnosis. What about validity? Judging current diagnostic validity by the criteria of concurrent, predictive (prognostic), construct, and content validity, there is much to be desired. Concurrent validity based on psychological test diagnosis involves too much interdependence or even collusion to yield any comfort. The prognostic validity of diagnosis seems to be of value in manic-depressive psychosis, since the past history of episodes of this illness tends to be predictive, but prognosis for other diagnostic categories is so dependent on type of follow-up, type of evaluation, etc., that we must conclude that prognostic validation of diagnosis is still in the future. (It is proposed to follow up a sample of the patients interviewed by the bilateral project in subsequent years to throw light on this question.)

Use of diagnosis for choice of therapy presents no better a picture than prognosis. For example, although Bannister and associates(1) found a statistically significant association between diagnosis and type of therapy, the number of "correct" choices (most popular therapy for a given diagnosis) was only 18 percent—that is, in 82 percent of the choices no stable relationship between diagnosis and therapy was noted. Whether the current drug era will produce more "correct" choices of therapy remains to be seen. Content validity would be perhaps the easiest type of validity to evaluate if we had good criteria on which to base the type of behavior expected of each diagnostic group. This type of validity founders because of the lack of general agreement in the delimitation of the type of behavior belonging to each diagnostic group.

The area of construct validity offers the widest field for experimental evaluation, and such constructs as "flatness of affect," communicability of speech, overinclusive thinking, deviation in processing of stimulus input, and encoding in the central nervous system, etc., are being examined for their potential value in validating the diagnosis of schizophrenia, for example.

In some areas, the testing of validity may have to wait until proper research methods are developed, and this may of course require much laborious investigation from a number of specialized fields. In the meantime, however, if one determines that a measure or a method is "reliable," one can at least feel confident that different investigators are talking about the same population. Reliability in general is a matter of agreement, whereas validity is more a matter of discovery. In addition, not all validity studies offer the same kind of benefits. If prognosis, for example, is essentially independent of etiology, as some have suggested, then prognostic validity would tend to have more practical than theoretical significance.

The end goal of diagnosis is etiology, but because etiology is practically unknown in psychopathology we resorted to the construction of potential or "ideal" etiologies in the form of scientific models.

Scientific Models

The following scientific models for etiology were considered: ecological, developmental, conditioning and learning, genetic, internal environment, and neurophysiological. Variations from one culture to another in the importance of factors in each of these models or variations in their interaction could produce differences in the actual distribution of the different types of disorders.

It would be well to take up each of the models in turn, describing its assumptions, the causal agents presumably salient to it, the deviant behaviors associated with these agents, and the techniques provided for measuring them. Unfortunately, space will not permit it; the reader can find this discussion elsewhere(17). Here we shall be concerned with the variety of techniques and methods which have been developed for dealing with the various hypotheses emanating from these models.

Each of these models has given rise to a variety of hypotheses, although only a few examples can be cited. Thus the ecological model has initiated a large number of epidemiological studies to investigate the ecological sources of mental disorder. The developmental model is responsible for the series of studies on the influence of early
environment on subsequent personality adjustment. The internal environment model has initiated investigations of the body fluids of schizophrenics and of the electrolyte metabolism of depressives. The neurophysiological model has initiated the study of the inhibitory role of descending afferents on cortical evoked potentials in hysterical anesthesia (4) and a host of investigations on the cortical arousal level of schizophrenics, including Mednick's assumption of higher arousal leading to quicker conditioning and generalization.

The learning model has been responsible for providing animal analogues for behavior disorders in the form of the conditioned emotional response(5) and for psychosomatic disorders in the operant control of autonomic and visceral functioning(9). It has also given rise to the hypothesis that the schizophrenic is more influenced by the immediately preceding event than by a more distant event(10).

It became clear at the very start that not all the important variables could be investigated in our initial study. A strategic starting point had to be found. The first step, as was already pointed out, was to develop an interview which would be objective, reliable, and valid—or at least have face validity.

The current practices in arriving at a diagnosis in both countries leave much to be desired from the point of view of the above-mentioned criteria. They may be satisfactory enough for clinical purposes but could hardly be useful for any scientific investigation.

As suggested earlier, it was concluded that only systematic structured interviews with objective scoring could serve our purpose. In reviewing the variety of interviews available we noted that they ranged all the way from the unstructured interview commonly used in making diagnoses through partially structured interviews to interviews that were quite rigorously organized.

Development of Instruments

One type of instrument developed in Biometrics Research, under the direction of Drs. Eugene I. Burdock and Anne Hardesty(2), was the Structured Clinical Interview, which is essentially a soft-sell or low-pressure approach to interviewing consisting of neutral questions, serving as stimuli, that elicit information from the subject in an open-ended way, but that nevertheless allow for evaluating the content of the response along a precoded set of 179 items covering a wide spectrum of psychopathology. This structured interview proceeds without probing and is therefore suitable for any type of person, patient or normal, whose state of mental health is to be investigated.

The second level of interviewing, represented by the Psychiatric Status Schedule(13) is more probing in nature. It was developed by Dr. Robert L. Spitzer with the help of other members of the Biometrics Research staff including Drs. Eugene I. Burdock, Anne Hardesty, and Jean Endicott. In this technique there is a more definite attempt at determining psychopathology through direct questioning and probing wherever it is called for. This interview consists of a systematic structured interview and has some 500 items of a specific, dichotomous nature which have to be scored as the interview progresses. It covers a wider spectrum of psychopathology than the Structured Clinical Interview, including alcoholism and drug addiction as well as deviations in role performance in everyday life.

The third technique is that developed at the Institute of Psychiatry of the Maudsley Hospital in London by Dr. John Wing and his associates in the Medical Research Council social psychiatry research unit. The most probing instrument of all, it attempts to determine to the satisfaction of the interviewer himself whether each of the 500 items of psychopathology included in the inventory is present or absent in the patient.

The interview finally hammered out by the two staffs in New York and London consisted of a combination of the two last-named instruments—the Psychiatric Status Schedule and the Maudsley MRC Schedule. The PSS has been in use long enough to permit the development of both clinical scoring scales and factorially derived scales as well as a program for computer diagnoses. The Maudsley part of the interview has not yet been subjected to such statistical manipulation, although it does provide clinically...
based dimensions for evaluation of the patient and is in use by several World Health Organization projects in schizophrenia.

Since the standardized mental state schedule deals only with the current state of the patient, a history schedule was also formulated to capture the important steps in the development of the illness. An instrument for collecting ecological, developmental, and genetic data was also developed so as to provide information on the role that these important factors play in psychopathology.

The reliability of the information recorded during the mental state examination is high. The median intraclass correlation coefficient is .85 for both the scales of the PSS and the sections of the MRC.

With respect to the history data, the "hard" items—education, marital status, etc.—are typically the more reliable, with a median coefficient of approximately .90. The median reliability for the "soft" items—parents' personalities, sexual adjustment, etc.—is approximately .60.

Present Study

In planning the initial study, we came to realize that the best place to begin was with the admission service of a large non-private (state) hospital. The results of such a study could not be definitive in answering the basic question as to incidence or prevalence because they would depend on the representativeness of the hospital selected in each country, the degree of the utilization of the chosen hospital in each country by various segments of the population, and the ecological characteristics of the catchment area from which the patients were drawn. Nevertheless it would provide some idea of the distribution of the diagnoses at least in two contrasted large hospitals and also provide a testing ground for the feasibility of the approach to cross-national studies. Furthermore, discrepancies between local hospital diagnoses and our own standardized project staff diagnoses might reveal the local deviations in diagnostic tendencies.

Before we can answer the basic question regarding the comparison of the distribution of mental disorders in the U. K. and the U. S., a sampling study of hospitals in Greater London and Greater New York will have to be undertaken.

Secondly, it is proposed to conduct a study of the individuals coming for help in whatever facility available to them (including physicians in private practice and other resource people), and eventually a field study of suitably selected catchment areas will be made in order to cover the general population.

Testing Hypotheses Emanating from Etiological Models

With regard to the six etiological models, again the task of tackling all of them proved too great. The problem that they present is two-fold: 1) what are the parameters of each model that may be related to deviant behavior, and 2) what type of deviation would each of these factors tend to produce? Furthermore, as shown in figure 1, the models do not operate independently but interact with each other.

The ecological model and the learning model refer primarily to exogenous factors impinging on the individual. The developmental model is partly exogenous, influenced by ecological and learning factors, and partly endogenous, reflecting matura- tion. The genetic, internal environment, and neurophysiological models operate entirely within the skin, but they are mutually inter-related as well as influenced by the ecological forces via learning and development.

In order to test the hypotheses emanating from the six models described previously, we must provide techniques and methods for measuring the hypothesized deviant behavior accounted for by each of the models. We have tentatively divided the models
themselves into three groups: 1) the ecological model; 2) the developmental and learning models; and 3) the genetic, internal environment, and brain-function models.

For the first, culture-dependent interviewing techniques seem currently to be the most satisfactory way of assessing deviant behavior. For the second group, culture-fair techniques need to be provided. For the third group, culture-free techniques are being developed.

We have already discussed the development of the techniques for dealing with the ecological model. Here the interview technique in its modern dress—the structured interview—is apparently of most help. Thus up until now, the project has tackled only one aspect of psychopathology—viz., what becomes apparent through interviewing methods that are based on the assumption that psychopathology reveals itself in deviations from social-cultural norms of behavior (including thought and feeling).

We have, however, taken a complementary step in the direction of discovering culture-fair and culture-free objective indicators of psychopathology. To this end a workshop was organized with some of the outstanding workers in this field to determine what direction the search for such techniques should take. The participants consisted of proponents of two of the models: the learning-theory model, for which culture-fair techniques are being developed, and the neurophysiological model, for which culture-free techniques are being sought. The results of this workshop are not yet available, but a brief summary of some of the outstanding conclusions follows:

1. Although there are no truly culture-free indicators, it is possible to reduce the influence of past history of reinforcement, attitudinal differences, and motivational differences by utilizing responses during the first 1000 milliseconds following stimulation to simple sensory laboratory types of stimuli testing the responsiveness of the individual.

2. As the role of interfering variables such as reinforcement history, attitude, and motivation is lessened, differences between patients and normals usually decrease and sometimes decline to nothing; but those differences that persist after the controlling of irrelevant environmental variables may prove to reflect psychopathology in a more basic sense.

3. It is necessary to specify the attributes of the stimulus to which the patient responds in the laboratory tests in order to make certain that the difference between patient and normal is a valid indicator of psychopathology and not a difference due to attribute preference, stimulation history, sensitization, etc. One way of reducing the number of possible stimulus attributes to which a subject may respond is to design the test task in such a way that it can be performed only by responding to one attribute of the stimulus. Furthermore, if there is more than one way to perform the task, this too may introduce troublesome variability. It is therefore incumbent upon the experimenter to devise tasks that provide only one solution or one avenue of response.

4. While present-day diagnosis is often unreliable and sometimes of doubtful validity it nevertheless provides a good starting point in selecting a sample of certain types of patients. Beginning, for example, with a clinically diagnosed group of schizophrenics and matched normals, we can proceed to apply refined measuring techniques which are hypothesized to differentiate the two groups. In one study of this type we also included the results based on our structured interview utilizing computer diagnosis. When the "pure" groups (consisting of those who were identified as schizophrenic or as normal by both clinician and computer) were compared in a cross-modal reaction time experiment(7), it was found that the variability in reaction time in the pure groups was considerably reduced and the difference between the two pure groups considerably enhanced.

This iterative method of progressively purifying groups by successive approximations can help considerably in improving the discriminating power of the diagnostic interview itself as well as in the search for

METHODOLOGY AND PLANNING

---

3 The Workshop on Objective Indicators of Psychopathology was held at Sterling Forest, Tuxedo, N. Y., in February 1968. It was organized by Drs. Mitchell Kietzman, Samuel Sutton, Peter Venables, and Joseph Zubin. The proceedings will be published at a later date.

new objective indicators in the psychophysiological or sensory laboratory. Shagass' demonstration that slower recovery in the amplitude of evoked potentials to rapidly succeeding somatosensory stimuli characterizes depressives may also be used in this iterative procedure.

5. The selective readiness with which some organs in certain individuals tend to be more readily conditioned—heart rate, for example—as Miller(9) has shown, can serve not only to clarify psychosomatic conditions but may also be of use in the functional psychoses. If it can be established that offspring of schizophrenics even before they develop clinical schizophrenia seem on the average to have greater unconditioned responsiveness to noise as measured by GSR, as Mednick(8) seems to find, or that adult schizophrenics show a higher intensity gradient in their evoked potentials to energy stimuli as Callaway and Jones(3) and Shagass(11) have shown, we may be able to utilize such differential sensitivity to identify directly or through conditioning methods those individuals who show proneness to psychopathology of one or another kind.

Summary

As a result of the combined efforts of the U. K. and U. S. staffs of our bilateral study of diagnosis of the mental disorders, instruments have been developed that provide an objective basis for making reliable and, we hope, valid diagnoses. Armed with these tools, we designed an experimental approach to examining the newly admitted patients to one hospital in Greater London and to one in Greater New York, limiting ourselves to the age group 35-59 so as to ensure sufficient representation of the two major categories, affective disorder and schizophrenia, which have been found to vary considerably in prevalence between the two countries.

The next step in this study is to get a representative sample of the patients in the hospitals of Greater London and Greater New York to see whether the differences observed between the two individual hospitals will hold true of the two metropolitan areas. The third step is to extend this procedure not only to inpatients but to outpatients and private patients—in fact, all those individuals who come for help to some mental health facility in one of the two countries. The fourth and final step is to do a general population study of a selected catchment area in each country and help clarify the incidence and prevalence of mental disorders against the ecological characteristics of the two catchment areas.

In the meantime, several subdivisions of Biometrics Research are engaged in the attempt to follow up the search for objective indicators initiated by a recent workshop on objective indicators of the mental disorders. We are also engaged in studying the interaction processes going on throughout the interview, with special reference to proper role assumption by the interviewer and the patient and the impact of this on the interview. We are also concerned with the development of objective indicators of the social distance between patients and members of their environment (family, friends, etc.) through a study of linguistic similarities.

Acknowledgments

The original plans for this study were formulated by the U. S. Steering Committee, consisting of the late Dr. Paul Hoch, Drs. Morton Kramer and Benjamin Pasamanick, and this writer, and later Dr. Heinz Lehmann. Once these plans were accepted and a grant was obtained from NIMH, the staffs that were selected in the two countries took over the task of implementing the project.

The psychiatrists in the British team are Drs. John Cooper, R. E. Kendall, and John Copeland. Professor Michael Shepherd and Dr. John Wing are members of the U. K. Steering Committee, and Sir Aubrey Lewis is a part-time psychiatric consultant to the project. Professor A. E. Maxwell's department of biometrics acts as statistical adviser. Dr. Norman Sartorius (psychiatrist), Miss Miriam David (research assistant), and Miss Gillian Stoneham (psychiatric social worker) worked on the project during the study at Netherne Hospital reported here.

The U. S. team consists of Drs. Barry Gurland, Lawrence Sharpe, and Tibor Far-
kas, psychiatrists; Dr. Joseph L. Fleiss, biostatistician; Mr. Robert Simon, social anthropologist; Miss Vivian Hoff, psychologist; and Miss Pamela Roberts, computer operator, among others.

The consulting staff includes Dr. W. Edwards Deming, sampling, and—from Biometrics Research—Dr. Robert Spitzer, psychiatrist; Dr. Ruth Bennett, sociology; Dr. Muriel Hammer, anthropology; Dr. Samuel Sutton, psychophysiology; Dr. Mitchell Kietzman, psychophysics; and Dr. Kurt Salzinger, learning theory.

The project staff is grateful to the staffs of the Netherne Hospital, Dr. R. K. Freudenberg, medical superintendent, and to the Brooklyn State Hospital staff, Dr. Nathan Beckenstein, director, for permission to carry on the project.

The author is indebted to Mr. David Jenness for editorial assistance and to Dr. Samuel Sutton and Dr. Joseph Fleiss for reading the manuscript and making many valuable suggestions.

REFERENCES