Preliminary Report of the Reliability of Research Diagnostic Criteria (RDC)

Applied to Psychiatric Case Records

Robert L. Spitzer, M.D., Jean Endicott, Ph.D., Eli Robins, M.D.
Judith Kuriansky, Ed. M., and Barry Gurland, M.D.

Robert L. Spitzer, M.D.
Director, Evaluation Section, Biometrics Research, New York State Department of Mental Hygiene, New York State Psychiatric Institute; Associate Professor of Clinical Psychiatry, Department of Psychiatry, College of Physicians and Surgeons, Columbia University.

Jean Endicott, Ph.D.
Co-Director, Evaluation Section, Biometrics Research, New York State Department of Mental Hygiene, New York State Psychiatric Institute; Research Associate, Department of Psychiatry, College of Physicians and Surgeons, Columbia University.

Eli Robins, M.D.
Wallace Renard Professor of Psychiatry and Head of the Department, Washington University School of Medicine, St. Louis, Missouri 63110.

Judith Kuriansky, Ed. M.
Research Scientist, Section on Diagnosis and Psychopathology, Biometrics Research, New York State Department of Mental Hygiene, New York State Psychiatric Institute.

Barry Gurland, M.D.
Head, Section on Diagnosis and Psychopathology, Biometrics Research, New York State Department of Mental Hygiene, New York State Psychiatric Institute; Associate Professor of Clinical Psychiatry, College of Physicians and Surgeons, Columbia University.
Introduction

A crucial problem in psychiatry, that effects both clinical work as well as the research study of methods to improve psychiatric treatment, is the generally low reliability of current psychiatric diagnostic procedures. This sad state of affairs is well recognized in our field, has been amply demonstrated in numerous studies, and is responsible in part, for the low regard with which psychiatric diagnosis is often held both within psychiatry, and the general field of medicine. Furthermore, inability to agree on the names of the conditions we treat and study, introduces a source of noise into any attempt to study the etiology, course, and treatment of these conditions.

The sources of diagnostic unreliability have been studied and can be divided into problems in eliciting data from the patient and problems in synthesizing this data according to the diagnostic criteria (Beck, et al, 1962). Most studies have show that the primary source of unreliability is the variability among clinicians in the diagnostic criteria that they use. While the Diagnostic and Statistical Manual of the American Psychiatric Association (DSM-11) provides some general guidelines for differential diagnosis by describing the major features of each of the conditions, the paucity of specific criteria in this manual force the clinician to rely heavily on his own concepts of the diagnostic categories that he has formed, based upon his particular training, experience and interests.

An alternative approach towards the diagnostic procedure, is the development of specific inclusion and exclusion criteria for each diagnosis that the clinician is required to use, regardless of his own personal concept of the disorder. With this approach, the clinician's task is twofold: to determine the presence or absence of specific clinical phenomena and then to apply the comprehensive rules provided for making the diagnoses. A consequence of this approach
Research Diagnostic Criteria

-2-

Is the need for an 'Undiagnosed psychiatric disorder' category for those patients who do not fulfill the criteria of any of the conditions. This approach, unlike the standard approach to diagnosis, seeks to identify relatively homogeneous subgroups of patients by minimizing false positives and tolerating a small number of false negatives.

This paper is a preliminary report of the reliability of a set of specific diagnostic criteria for a selected group of functional psychiatric disorders, called the Research Diagnostic Criteria. This report presents the reliability of these criteria as applied to psychiatric case records. A future paper will present a more extensive exposition of the rationale and development of the RDC.

The Research Diagnostic Criteria (RDC)

The RDC were developed by Spitzer, Endicott and Robins with the assistance of the other participants in a collaborative project on the psychology of depressive disorders. This project is sponsored and coordinated by the Clinical Research Branch of the NIMH. The latest version of the RDC appears as an appendix following this paper. The RDC is an elaboration and expansion of the specified criteria developed by a group of psychiatrists at Washington University School of Medicine in St. Louis for a smaller number of diagnoses (Feigner, et al, 1972). They have offered considerable evidence for the validity of these conditions as separate diagnostic entities. In developing the RDC, additional diagnoses were added, such as schizo-affective disorders, some personality disorders, and a category called borderline or atypical psychosis. In addition, many different non-mutually exclusive ways of categorizing major depressive illness were provided for studies focusing on depressive illness.

Often the criteria in the RDC take the form of listing specific symptoms, and/or clinical characteristics, a number of which are necessary for the criteria to be fulfilled. Many conditions have exclusion criteria as well as inclusion criteria. A consequence of this approach is the recognition of the need for an 'undiagnosed psychiatric disorder' category for the patients who do not meet the criteria
for any of the specific disorders, yet exhibit significant psychopathology.

Each diagnosis is judged either as absent, probable or definitely present. When two or more diagnoses are mutually exclusive (for example, depressive personality and minor depressive illness) they are so noted in the criteria. Otherwise, more than one diagnosis may be given to the same patient for the same period of illness, or for different periods of illness (for example, schizophrenia and alcoholism).

Some of the diagnostic terms in the RDC are different from those used in traditional nomenclatures. This is done to help avoid confusion when the category is defined in such a way that it has only partial correspondence with a traditional category. For example, the traditional category of neurotic depression includes patients who would be categorized in the RDC as either major or minor depressive illness. For several of the diagnostic categories, the criteria differ little, if at all, from those generally used for example, the diagnosis of obsessive-compulsive neurosis or phobic neurosis. For other categories where a consensus as to the diagnostic criteria does not exist in the field, it was necessary to adopt what seemed to us as reasonable, although admittedly arbitrary, criteria. For example, the RDC schizo-affective schizophrenia includes the symptom picture of both schizophrenia and affective illness but none of the criteria relate to acuteness of onset, although many clinicians include this variable in their use of the term. Of course, it will be possible to examine the data for all schizo-affectives and compare them with those whose onset of illness is acute to determine if use of this additional criterion increases diagnostic validity.
Study

An initial version of the RDC was used in a study of 120 case records of patients from the New York State Psychiatric Institute. The purpose of this study was to determine the reliability of the RDC as applied by research assistants to a large sample of heterogeneous patients, and to compare the reliability of these diagnoses using the RDC with the reliability of diagnoses made by experienced psychiatrists using the official nomenclature of the American Psychiatric Association, DSM-11, under conditions approximating those usually present in most research studies in which clinical diagnoses are made using the DSM II categories.

This case record study was possible because Ms. Judith Kuriansky and Dr. Barry Gurland of the Psychopathology and Diagnosis Section of Biometrics Research, had a set of case records used in a previous study of diagnostic practice conducted at the New York State Psychiatric Institute (Kuriansky, et al, 1974). That study involved an examination of the case records from two decades in which the hospital diagnosis of schizophrenia had more than doubled (from 28% during the decade 1932-1941 to 77% during the decade 1947-1956).

That study used 129 records of patients age 20-59; 64 cases from each decade selected to reproduce the original proportions of the hospital diagnoses. The case records were Xeroxed and all references to diagnosis and year of admission were obliterated. Since the research diagnostic criteria used in this study are only for functional conditions, we excluded eight cases in which the diagnosis of an organic brain syndrome was noted as possibly present in the case record. This left a total of 120 cases.

Three independent diagnoses on each case were available: those made by the hospital psychiatrist; by Dr. Gurland, a British trained psychiatrist; and by one of a group of 16 experienced North American psychiatrists, who acted as re-diagnosticians of the cases. Both Dr. Gurland and the re-diagnosticians used the DSM-11 when making their diagnoses.
Because the earlier study focused on diagnostic practice, no attempt was made to alter the diagnostic practices of the re-diagnosticians in an effort to achieve high interrater reliability. Therefore, they did not meet together or participate in any special training sessions and the degree to which they referred to the DSM-II manual is unknown. The design of the earlier study was such that any lack of reliability among the re-diagnosticians would have no systematic biasing effect on their comparison of the patients from the two decades.

All of the cases were read and diagnosed independently by Dr. Spitzer and six of his research assistants who were paired in varying combinations for each case. Each research assistant read from 30 to 60 cases. Only one of the research assistants had professional training in psychopathology (psychiatric social work). However, all of the research assistants had considerable on-the-job training in making psychiatric ratings and in the use of the RDC and most had attended a ten session seminar on descriptive psychiatry given by Dr. Spitzer to the first year psychiatric residents at the Psychiatric Institute.

Results

Reliability was determined using the statistic kappa (Spitzer, Fleiss, Cohen, Endicott, 1967). Kappa is the proportion of agreement corrected for chance agreement as a function of the base rates. It varies from negative values for less than chance-expected agreement through 0 for exactly chance-expected agreement to plus 1 for perfect agreement.

The reliability of the RDC for the entire set of 120 cases is shown in the first three columns of Table 1. The first column gives the agreement between Dr. Spitzer (RLS) and the first Biometrics research assistant who read the case. The second column gives the agreement between RLS and the second Biometrics research assistant who read the case. The third column gives the agreement between the two Biometrics research assistants. The results indicate very good agreement for all of the major
categories. To our knowledge, reliabilities for schizophrenia, and affective illness are higher than have been reported in diagnostic studies of this kind. (In a review of such studies (Spitzer and Fleiss, 1974) the range of kappa values for schizophrenia over six studies where the data was available for such calculations was .32 to .77, and for affective illness over five studies it was .19 to .59.)

The reliabilities for the specific diagnostic categories are generally very high with the exception of schizo-affective schizophrenia and borderline or atypical psychosis, two categories that are by their very nature difficult to define with precision. We believe that an additional reason for the low reliabilities obtained here is that the criteria in these two categories were modified to a greater degree than were other categories while the study was in progress. In the case of schizo-affective schizophrenia, one of the first Biometric raters seemed to over-diagnose the condition and, by chance, was given many more cases than the other raters when this was the major differential diagnostic problem. We believe that the kappa of .54 between RLS and the second Biometrics rater is a better estimate of what can be achieved with the criteria in their present form. In the case of borderline or atypical psychosis, this category is more often given as a subsidiary diagnosis than as the major diagnosis, The average agreement for this category among the three pairs of raters as either major or subsidiary diagnosis is .39, which although low, is higher than when the category is used as the major diagnosis.

Interrater agreement using the RDC with specially trained research assistants was considerably higher than that between Dr. Barry Gurland (BG) and the 16 re-diagnosticians, using the DSM-II categories, for the major categories of schizophrenia and affective illness. This was also the case for all of the specific diagnoses with the exception of schizo-affective schizophrenia.

How does the reliability obtained by BG and the re-diagnosticians using
the DSM-III categories compare with other studies of diagnostic reliability? The kappa values between BG and the re-diagnosticians for schizophrenia and affective illness (.48 for both) is approximately in the middle of the range of values previously noted in this paper as having been obtained in studies of the reliability of psychiatric diagnoses. Thus the reliability of diagnostic judgement obtained by BG and the re-diagnosticians is approximately that obtained in the usual research study employing psychiatric diagnoses.

Examination of the agreement between RLS and BG (fourth column) yields kappa values that are generally intermediate between those obtained when both raters use the RDC and both raters use the DSM-III categories. The most likely explanation for these results is that BG, although using DSM-III categories, was trained in England, has a special interest in diagnosis, and has a diagnostic orientation that is more similar to the orientation made explicit in RDC than the approach characteristic of the 16 North American psychiatrists. The good agreement between RLS using the RDC system and BG using the DSM-III categories was also reflected in very similar distributions across all diagnoses for the total sample of cases. This indicates that the RDC is a refinement of the standard nomenclature rather than a different system yielding high reliability at the expense of being at variance with expert diagnostic practice.

Several issues need to be considered in appraising these results. First of all, this study dealt with case records rather than actual patient interviews. Is it possible that the high values obtained here are inflated because case records present stereotypes which are easier to categorize than live patient evaluations? This is unlikely for several reasons. The case records usually contained voluminous progress notes that often presented conflicting diagnostic cues which actually increased the difficulty in arriving at a diagnosis. In addition, the actual agreement values for the case records are in the same range as were obtained
Research Diagnostic Criteria

In a previous study using diagnoses based on interviews of newly admitted inpatients. Since the agreement between the re-diagnosticians and BG was not particularly good, this indicates that the use of case records did not guarantee a high reliability.

Could the higher agreement values obtained with the RDC primarily be a function of the time spent by the Biometrics raters in training and working together and developing a shared concept of the various categories, as contrasted with the absence of such shared experiences on the part of the psychiatrists using the DSM-11 categories? No doubt, if BG and the re-diagnosticians had worked together they would have improved their reliability. However, previous studies of diagnostic reliability using the standard nomenclature have employed psychiatrists who worked together in an attempt to improve agreement yet the reliabilities obtained in these studies are well below those obtained here with the RDC. Furthermore, it is likely that a large part of "working together" actually consists of supplementing the general criteria which are provided in DSM-11 with specific criteria to aid the differential diagnostic process. Unfortunately, such working rules are of no value to the profession as a whole unless they are made available for general use. This is precisely what the RDC attempts to do.

The results of this study indicate that the RDC may be a technique for greatly increasing the reliability of psychiatric diagnosis. We recognize that several of the categories need further work before they too can be reliably diagnosed. In addition, it is recognized that the demonstration of reliability is only a partial requirement of a diagnostic system. The fundamental requirement, of course is validity. It is possible for a classification system to be reliable but not valid. Therefore, validity studies will be required before this approach can be shown to be generally useful. However, we believe that these initial results are promising and justify further work on the RDC.
Supported by NIMH Grant 23864, 'NIMH Clinical Research Branch Collaborative Depression Studies, II: SADS' and Biometrics Research, New York State Department of Mental Hygiene.
References


Table 1. Kappa Coefficients of Agreement for Major Diagnosis by Various Pairs of
Raters on 120 Case Records.

<table>
<thead>
<tr>
<th>Diagnostic System Being Used</th>
<th>Both RDC</th>
<th>RDC and DSM-11</th>
<th>Both DSM-11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RLS vs. Biom.1</td>
<td>RLS vs. Biom.2</td>
<td>Biom.1 vs. Biom.2</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>.78</td>
<td>.84</td>
<td>.84</td>
</tr>
<tr>
<td>Other Schizophrenia</td>
<td>.68</td>
<td>.82</td>
<td>.70</td>
</tr>
<tr>
<td>Schizo-affective</td>
<td>.22</td>
<td>.54</td>
<td>.24</td>
</tr>
<tr>
<td>Affective Illness</td>
<td>.70</td>
<td>.74</td>
<td>.70</td>
</tr>
<tr>
<td>Manic Illness</td>
<td>.82</td>
<td>.92</td>
<td>.76</td>
</tr>
<tr>
<td>Major Depressive Illness*</td>
<td>.67</td>
<td>.74</td>
<td>.66</td>
</tr>
<tr>
<td>Minor Depressive Illness**</td>
<td>.56</td>
<td>.66</td>
<td>.56</td>
</tr>
<tr>
<td>Other</td>
<td>.69</td>
<td>.70</td>
<td>.74</td>
</tr>
<tr>
<td>Briquet's Syndrome</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
</tr>
<tr>
<td>Antisocial Personality</td>
<td>-</td>
<td>-</td>
<td>.66</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>1.00</td>
<td>.66</td>
<td>.66</td>
</tr>
<tr>
<td>Drug Dependence</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Obsessive-Compulsive Neur.</td>
<td>.64</td>
<td>.52</td>
<td>.65</td>
</tr>
<tr>
<td>Anxiety Neurosis</td>
<td>.65</td>
<td>.59</td>
<td>.52</td>
</tr>
<tr>
<td>Phobic Neurosis</td>
<td>.66</td>
<td>.39</td>
<td>.66</td>
</tr>
<tr>
<td>Borderline or Atypical Psy.</td>
<td>.24</td>
<td>.06</td>
<td>.20</td>
</tr>
<tr>
<td>Undiagnosed Illness</td>
<td>.45</td>
<td>.30</td>
<td>.59</td>
</tr>
</tbody>
</table>

* In this study considered equivalent to any psychotic depressive illness in DSM-11.

** In this study considered equivalent to neurotic depression in DSM-11.

*** A dash indicates that either the diagnostic category was not available to one
or both raters, or that none or only one of the raters ever used it.