

JAD 00762

Reliability of checklist-guided diagnoses for DSM-III-R affective and anxiety disorders

Wolfgang Hiller, Michael von Bose, Gabriele Dichtl and Dorothee Agerer

Max-Planck-Institute of Psychiatry, Psychiatric Outpatient Department, Munich, F.R.G.

(Received 7 June 1990)

(Revision received 8 August 1990)

(Accepted 15 August 1990)

Summary

The test–retest reliability of DSM-III-R diagnoses for affective and anxiety disorders was determined under clinical routine conditions in a psychiatric outpatient department. The sample consisted of 60 patients, and the Munich Diagnostic Checklists (MDCL) were administered for diagnostic evaluation and classification. Each subject was independently examined by two of four participating diagnosticians (two psychiatrists, two psychologists). Acceptably high levels of agreement were indicated by several statistics (including κ) for most disorders. Reliability was analyzed for diagnoses, subclassifications, and symptoms. Reduced agreement was found only for dysthymia, agoraphobia, and social phobia. Major causes were information variance and weaknesses of operationalization. Overall results were satisfactory when compared with other reliability studies.

Key words: Reliability; Affective disorders; Anxiety disorders; DSM-III-R; Classification; Munich Diagnostic Checklists (MDCL)

Introduction

The reliability and reproducibility of psychiatric diagnoses are influenced substantially by the clearness and precision of diagnostic criteria, and by the accuracy of clinical investigation (Ward et al., 1962; Helzer et al., 1977; Spitzer et al., 1978; Grove et al., 1981; Blashfield, 1984). DSM-III-R,

published by the American Psychiatric Association (1987), and the forthcoming ICD-10 represent major advancements and a new generation of classification systems, since diagnoses are operationalized and consistently focused on objective and observable (descriptive) signs and symptoms.

More basically, these systems require a new diagnostic process for clinical routine as well as for scientific work (Robins et al., 1981; Wittchen et al., 1985; Robins and Helzer, 1986). The clinician (or scientist) has to explore the patient's psychopathology in a systematic and comprehensive manner, and he must refer to the definitions

Address for correspondence: Dr. Wolfgang Hiller, Max-Planck-Institute of Psychiatry, Psychiatric Outpatient Department, Kraepelinstr. 10, D-8000 Munich 40, F.R.G.

of the classification system when evaluating whether a specific diagnosis can be given or not. All relevant diagnostic information must be obtained for an adequate classification, and this can be enhanced if diagnostic criteria are explicitly used as guidelines during verbal and/or non-verbal examination.

DSM-III-R and the earlier DSM-III system have stimulated the development of comprehensive interviews (Robins et al., 1981, 1988; Di Nardo et al., 1983; Spitzer et al., 1987) in order to structure and standardize diagnostic procedures. With such instruments, considerable improvements of reliability were found in empirical studies (Burnam et al., 1983; Di Nardo et al., 1983; Semler et al., 1987, 1989). However, the practical use of interviews does not seem to be unrestricted, since it can often be difficult to integrate highly structured protocols into clinical routine work. First, interviews tend to be time-consuming (frequently more than 1 h for a single patient), and second, they are considered to be somehow inflexible in application (since the diagnostician must adhere to a fixed sequence which often deviates from the usual course of free clinical explorations).

Thus, the question arises to which degree systematic diagnostic evaluations can be conducted under routine clinical conditions with strict time limits (e.g., 15–45 min in outpatient clinics, or private practices), and whether such investigations can be supported (and guided) by instruments. The approach reported here was derived from earlier methods used to assess psychopathology. For example, checklists or rating scales such as those developed by Wittenborn (1955) or Lorr et al. (1963) have proved useful for both clinical routine and scientific work. Most rating scales have been designed for an empirical grouping of symptoms (syndromes) on the basis of statistical analyses, but it seems to be possible to adapt this method for the purpose of classification (i.e., to assign patients to diagnostic categories).

This report describes an investigation of the usefulness of diagnosis-related checklists. We administered the Munich Diagnostic Checklists (MDCL) which have recently been developed by us (Hiller et al., 1989). The goals of our study were (1) to evaluate the test–retest reliability of checklist-derived DSM-III-R diagnoses under routine

outpatient conditions for affective and anxiety disorders, (2) to analyze the nature and sources of diagnostic disagreement, and (3) to compare the reliability of MDCL diagnoses with those established by use of structured and standardized interviews.

Method

Data were obtained from 60 psychiatric outpatients (30 women, 30 men) who were consecutively examined at the Psychiatric Outpatient Department of the Max-Planck-Institute of Psychiatry in Munich during a 10-week period. They were referred from a general hospital, private psychiatrists, and non-psychiatric physicians for diagnostics and treatment proposals. All patients were examined and treated within the frame of our usual routine work.

Diagnostic assessment

The Munich Diagnostic Checklists (MDCL) were administered to serve as a guideline and tool for the assessment of DSM-III-R diagnoses. This instrument consists of 30 pocket-size lists, each allowing to rate the criteria of a specific diagnostic category, and to combine signs and symptoms for a diagnostic decision. Thus, each checklist can be used by the diagnostician in order to accept or reject a diagnosis during or immediately after exploration and/or clinical examination. The MDCL can easily be integrated into usual modes of verbal exploration, and standardized questioning, probing, or a fixed order of progression is not required. During the examination, however, the diagnostician must make sure that *all* diagnostically relevant criteria are evaluated.

To illustrate the design and structure of the MDCL, the first page of the MDCL 'Manic or Hypomanic Episode' is displayed in Fig. 1 (giving all symptoms to be evaluated and coded for a manic or hypomanic syndrome). The MDCL exist in English and German. A more detailed description of the instrument is given elsewhere (Hiller et al., 1990).

For the present study, the MDCL for major depressive episode, manic or hypomanic episode, dysthymia, cyclothymia, adjustment disorder, panic disorder, agoraphobia, simple phobia, social

<i>MDCL</i>		<i>Munich Diagnostic Checklist for DSM-III-R</i>								
Manic or Hypomanic Episode		Name: _____								
		Age: _____ Date: _____								
A	Distinct period of <i>abnormally</i> and persistently <i>elevated, expansive, or irritable</i> mood	No Stop ← <input type="checkbox"/>	Probably <input type="checkbox"/>	Yes <input type="checkbox"/>						
B	<div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> <ul style="list-style-type: none"> • Define the pattern of symptomatology • Relate all symptoms to the <i>same</i> period • Consider only <i>non-organic</i> symptomatology </div>	<table style="margin-left: auto; margin-right: auto;"> <tr> <td style="padding: 0 10px;">No</td> <td style="padding: 0 10px;">Probably</td> <td style="padding: 0 10px;">Yes</td> </tr> <tr> <td style="padding: 0 10px;"> </td> <td></td> <td></td> </tr> </table>			No	Probably	Yes			
No	Probably	Yes								
(1)	<i>Inflated self-esteem</i> or <i>grandiosity</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
(2)	<i>Decreased need for sleep</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
(3)	<i>More talkative</i> or <i>pressure to keep talking</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
(4)	<i>Flight of ideas</i> or subjective experience that <i>thoughts are racing</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
(5)	<i>Distractibility</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
(6)	<i>Increased activity</i> (socially, occupationally, sexually) or <i>psychomotor agitation</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
(7)	<i>Activities</i> which have a high potential of painful consequences	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
<div style="border: 1px solid black; padding: 5px;"> At least 3 items from (1) to (7) (4 items if mood is only irritable) </div>		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
		Stop ←								

Fig. 1. Section of the MDCL 'Manic or Hypomanic Episode' used to evaluate manic signs and symptoms (criteria A and B in DSM-III-R).

phobia, obsessive-compulsive disorder and generalized anxiety disorder were employed.

Sample characteristics

Subjects were accepted into the MDCL test-retest study if there had been no evidence of an acute or chronic organic mental disorder, or of a psychosis (i.e., schizophrenia and related dis-

orders). Other inclusion criteria were a minimum age of 18 years, and the patient's agreement to participate in the study (i.e., allowing his data to be used for scientific purposes, and returning for a second interview with a different diagnostician). All subjects were informed about the reasons for being interviewed twice, and they were instructed to consider both explorations as independent (i.e.,

TABLE 1
BASE RATES OF MDCL LIFETIME DIAGNOSES (DSM-III-R)^a

Diagnosis	No. of patients (%) ^b
Mood (affective) disorders	
Major depression	29 (48.3)
Bipolar disorder	4 (6.7)
Dysthymia	8 (13.3)
Adjustment disorder/uncomplicated bereavement	5 (8.3)
Total	46 (76.7)
Anxiety disorders	
Panic disorder	5 (8.3)
Agoraphobia (with or without panic disorder)	10 (16.7)
Simple phobia	3 (5.0)
Social phobia	3 (5.0)
Obsessive-compulsive disorder	4 (6.7)
Total	25 (41.7)
Overall total	71 (118.4)

^a Total number of positive diagnoses; cases with disagreement were counted as positive.

^b Percentages were computed by dividing the number of patients with the particular diagnosis by the total number of patients (since multiple diagnoses could be given, the total percentage exceeds 100%).

to give complete information in both occasions, and not to regard the second interview as a continuation of the first one).

The sociodemographic data of the sample at the time of investigation were as follows: (1) age 41.2 ± 12.3 years (mean \pm SD) with a range of 20–74 years; (2) 21 married, 25 single, 13 divorced or separated, 1 widowed; (3) educational level: 24 primary school with or without graduation, 17 college, 7 vocational schools, 11 university or comparable institution, 1 other schools.

Table 1 shows the base rates of MDCL diagnoses for the 60 patients of our sample. A total of 71 diagnoses was obtained for mood (affective) and anxiety disorders (representing a mean of 1.2 diagnoses for each patient), since multiple diagnoses could be made according to the DSM-III-R concept of comorbidity. We diagnosed mood (affective) disorders in 46 cases (76.7%) and anxiety disorders in 25 cases (41.7%).

Test-retest procedure

The MDCL were administered to each of the 60 patients on two separate occasions by one of four independent diagnosticians (initial diagnostic evaluation before treatment or other interventions). Each subject was carefully screened for disorders that may have occurred at any time in his life (resulting in *lifetime diagnoses*). The approximate investigation length was 30–75 min. Time intervals between the two examinations were kept to a minimum (1–4 days) in order to reduce variance caused by possible changes in the patients' acute psychopathology.

We further intended to minimize any bias due to the distribution and assignment of individual interviewers, and systematically rotated pairings of interviewers so that each of the six pairs (AB/AC/AD/BC/BD/CD) examined an equal number of patients. Each member of a pair was the first interviewer in an equal number of cases. Consequently, each diagnostician was test-retest partner for each of his colleagues across 10 subjects, for five as the test interviewer and for the remaining five as the retest interviewer.

Two psychiatrists (male and female) and two clinical psychologists (male and female) participated as diagnosticians. Both physicians and one psychologist had 2–3 years experience in clinical psychiatry. The other psychologist had worked in a psychiatric hospital for less than a year, and had counseling experience with psychiatric patients for many years. Whenever a specific patient was examined for the second time (retest), the interviewer did not know the result of the first investigation. The MDCL had been introduced into our clinical routine work 4 months before the beginning of the test-retest study. During this period, each of the four diagnosticians became familiar with the content and clinical use of the lists, and their application was supervised daily by the department staff in routine case conferences.

Statistical analysis

Agreement between diagnosticians on diagnoses and symptoms was calculated using κ , a chance-corrected measure of congruence for binary ratings (Cohen, 1960; Fleiss, 1981). Values of κ range from -1.0 to 1.0 with higher values representing higher levels of agreement. We per-

formed a test of significance for each κ value (one-tailed on the 5% level of error), but an interpretation of the magnitude of κ was considered to be more important (since κ may easily become statistically significant despite unsatisfactory agreement). It has been suggested that κ values of 0.70 and greater indicate excellent congruence (Fleiss, 1981).

The κ statistic currently represents the standard method to assess diagnostic agreement in psychiatry (Shrout et al., 1987), but it has been criticized because it is strongly influenced by the base rate of the diagnosis under study (Carey and Gottesman, 1978; Grove et al., 1981; Spitznagel and Helzer, 1985). That is, κ tends to decrease rapidly under low base rate conditions, and there is considerable controversy whether this is an adequate feature of κ or not (Spitznagel and Helzer, 1985; Shrout et al., 1987). However, Spitznagel and Helzer (1985) have proposed Yule's Y as a measure independent of base rates. Y is closely related to κ since it can be regarded as an approximation to maximum κ across all possible base rates, and we additionally employed this measure to analyze our data. A pseudo-Bayes estimation was applied whenever a single cell of the fourfold classification table became 0 (otherwise, Y would have reached the endpoint value of 1.0 despite incomplete congruence; cf. Bishop et al., 1975).

The congruence measures presented here refer to the reliability of probable or certain diagnoses (or symptoms). This corresponds to the usual clinical situation where an intervention seems to be justified even if the diagnosis is still provisional, or if it has to be complemented by additional information. However, the MDCL also makes it possible to evaluate agreement for certain diagnoses only (i.e., by contrasting certain and probable diagnoses). We have conducted this analysis, but the results are almost identical with the data presented here, and no systematic bias could be found (these analyses can be requested from the authors).

Analysis of disagreement

We further analyzed major reasons for diagnostic disagreement which have been discussed extensively in the literature (Ward et al., 1962; Helzer et al., 1977; Grove et al., 1981; Shrout et al., 1987).

Referring to detailed protocols that had been made for each examination, we differentiated the following sources of variance.

I = Information variance. This disagreement is due to inconsistencies, between the two sittings, in the patient's verbal report or his willingness to give more detailed information (e.g., the patient reports depressed mood in the first exploration, but denies it in the second sitting).

E = Exploration style. This reflects discrepant findings due to differences in the clinicians' styles or techniques of investigation, including variations in questioning and interacting with the patient (e.g., exploring precisely and insistently, or in a rather global manner for specific symptoms).

S = Subject variance. This source of disagreement refers to changes in the patient's condition between the two occasions (e.g., the patient is severely anxious in the first sitting, but not during the second one).

C = Criterion variance. This kind of disagreement stems from insufficient specification, or ambiguous (vague) definition of DSM-III-R criteria (e.g., unclear description).

W = Clinical weighting. This refers to variations due to different clinical interpretations of aspects of the clinical picture. Diagnosticians may agree on specific signs and symptoms, but nevertheless differ in their judgement of importance (e.g., compulsive behavior is found by both clinicians, but it is considered to be clinically relevant by only one of them).

Results

The results of our investigation will be presented separately for mood (affective) and anxiety disorders. Then they will be compared to data of other DSM-III reliability studies.

Mood (affective) disorders

The reliability of MDCL diagnoses for mood disorders (including specific depressive and bipolar disorders) and other conditions with affective symptomatology (adjustment disorder and uncomplicated bereavement) is displayed in Fig. 2. Each row gives, for a single category, a fourfold table with the exact distribution of congruently and incongruently assigned patients, κ , Yule's Y ,

MDCL Diagnosis	Retest		κ^*	Y	% Agree-ment	Source(s) of Disagreement
	Test	+				
Major Depression	31 4 4 21		.73	.73	86	I, S, W, E, C
Bipolar Disorder	56 1 0 3		.85	.85	98	W
Dysthymia	52 4 1 3		.50	.72	92	I, S, W, E
Mood Disorders	25 4 1 30		.83	.86	92	I, S, W, E, C
Adjustment Disorder / Uncompl. Bereavement	55 0 2 3		.73	.79	97	I

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 κ

* all κ values are significant at the 5% level.

Fig. 2. Test-retest reliability of MDCL diagnoses for mood (affective) disorders (DSM-III-R).

the overall percentage agreement, and a summary of source(s) of disagreement. Values of κ are illustrated graphically, showing more clearly the different levels of interpretation (κ above 0.40 as acceptable, κ above 0.70 as excellent; cf. Fleiss, 1981).

All κ values in Fig. 2 were significant at the 5% level. High congruence was obtained for major depression (κ and $Y = 0.73$) and bipolar disorder (κ and $Y = 0.85$). Major depression included cases with single or recurrent major depressive episodes. One of the patients with bipolar disorders was judged to have a hypomanic episode (often referred to as 'bipolar II'), and the appropriate diagnosis was bipolar disorder not otherwise specified.

A somewhat lower level of agreement ($\kappa = 0.50$) was found for dysthymia. However, we computed a Y coefficient of 0.72 for this diagnosis, indicating that κ is considerably influenced by the relatively low base rate of dysthymia in our study. Overall percentage agreement for dysthymia was 92, thus even higher than the corresponding value for major depression (86%).

The three categories of major depression, bipolar disorder, and dysthymia were additionally combined into the more global categorization of mood disorders (i.e., diagnosing *at least* one of these three disorders), and it can be seen from Fig. 2 that this combination yielded an excellent rate of $\kappa = 0.83$ with 92% overall agreement.

The last row of Fig. 2 shows satisfactory reliability for adjustment disorder and uncomplicated bereavement ($\kappa = 0.73$). These categories were analyzed as *one* diagnostic unit, since the clinical picture of both conditions was highly similar, and all of our subjects with uncomplicated bereavement fulfilled the criteria of adjustment disorder except criterion E (exclusion of uncomplicated bereavement). They reported a significant depressive syndrome (as a reaction to the death of a loved person), but the diagnosis of a major depressive episode had to be rejected (since criterion B of a major depressive episode excludes cases with uncomplicated bereavement).

The important category of major depression was analyzed in more detail, and Fig. 3 gives agreement rates for *syndrome* and *episode* of major

MDCL Subclassification	Retest		κ^*	Y	% Agreement
	Test				
Major Depressive Syndrome	24	4	.70	.70	85
	5	27			
Major Depressive Episode	28	4	.77	.77	88
	3	25			
Melancholic Type	46	4	.67	.74	90
	2	8			
Chronic course	54	1	.64	.80	95
	2	3			

* all κ values are significant at the 5% level.

Fig. 3. Test-retest reliability of MDCL subclassification of major depression (DSM-III-R).

depression, for melancholic type, and for the specification of chronic course. We found good diagnostic congruence for all subclassifications with κ between 0.64 and 0.77, and overall percentage agreement between 85 and 95. There is a reduced number of positive cases for melancholic type and chronic course, since these specifications are evaluated only for patients with a *current* disorder.

The reliability of major depression was additionally analyzed at the symptom level. At least five symptoms are required for a major depressive syndrome, and it should be considered that diagnosticians may agree on syndrome level *despite* discrepancies for individual symptoms. Reliability rates for symptoms are therefore expected to be lower than for syndromes or diagnoses.

An analysis of the nine depressive symptoms (as given by criterion A of major depressive episode) is displayed in Fig. 4. All κ values reached statistical significance, and two thirds were above 0.60. Best rates were obtained for fatigue or loss of energy, and suicidal thoughts or suicidal attempt. In contrast, it was relatively difficult (and less reliable) to decide if psychomotor agitation or retardation was present, since patients often re-

ported the subjective impression of being agitated or retarded, but they were uncertain if this was observable to others (as required by DSM-III-R).

Reasons for the obtained discrepancies were analyzed in detail. For major depression, Fig. 2 shows that all sources of disagreement, as differentiated in our study, were found. They frequently influenced the diagnostic process in combination with each other, and a single *main* reason of disagreement could not be identified in a number of cases.

We often observed that disagreement about a depressive syndrome resulted when patients' reports about the nature and duration of their complaints were vague, ambiguous, inexact, and sometimes even inconsistent *within* one exploration. Diagnostic decisions could then be influenced by even subtle variations in the subject's current state (*S*), by precision of inquiry (*E*), and by clinical judgement (*W*).

For example, a 54-year-old patient reported a recurrent symptomatology with predominant lack of energy, fatigue, hypersomnia, and he characterized his mood using words like 'bad', 'tired', 'worn out', and 'weak'. However, he hesitated to accept expressions like 'depressed' or 'sad'. His

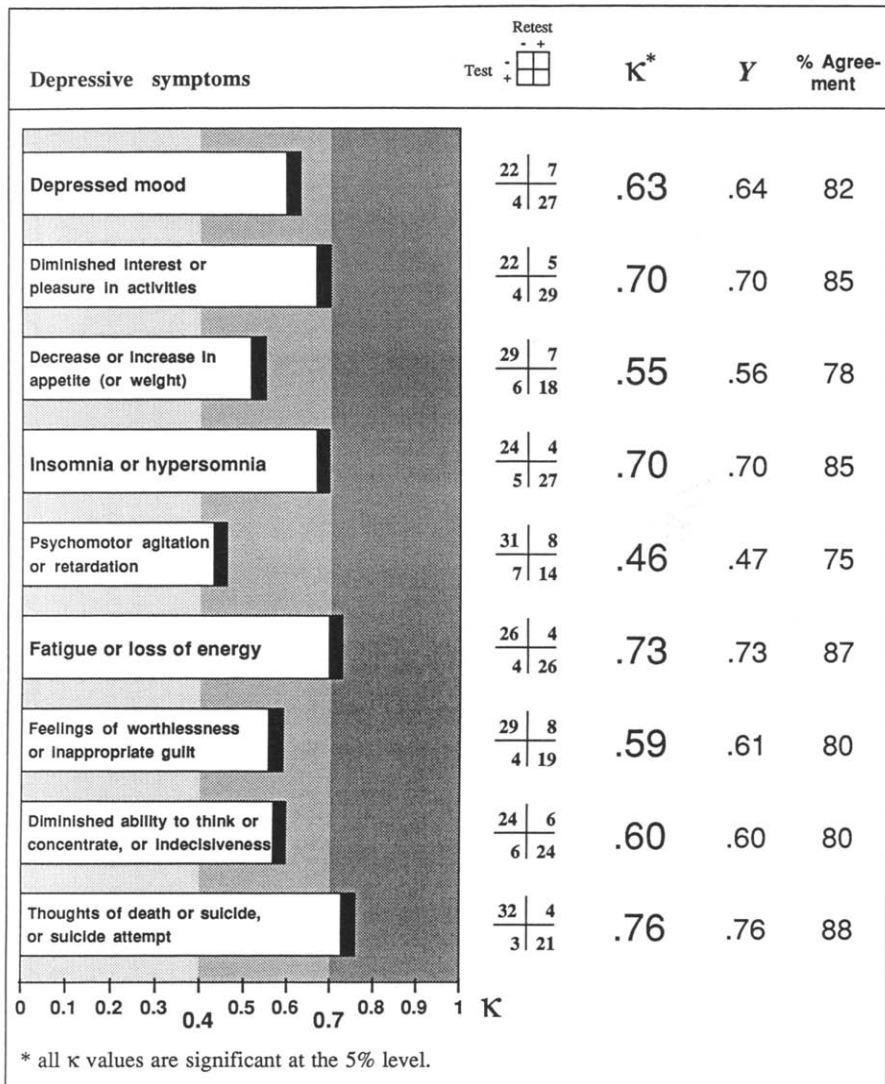


Fig. 4. Test-retest reliability of depressive symptoms (MDCL) (from DSM-III-R major depressive syndrome).

descriptions could not be clarified even by intense questioning, and it somehow remained unclear if the diagnosis of a major depression was adequate in this case.

Typical diagnostic problems also resulted when patients with severe patterns of alcoholism reported feelings of guilt, despair, and 'bad mood' after drinking. Depressive symptoms like fatigue, loss of appetite, and poor concentration fre-

quently co-existed. A diagnostic decision about the presence of a clinically relevant depressive syndrome often depended upon the patient's ability to give precise information, since an organic etiology of single symptoms had to be ruled out, and it had to be investigated if the depressive syndrome also existed in periods without alcohol consumption. We frequently observed differences between the two explorations in the patient's

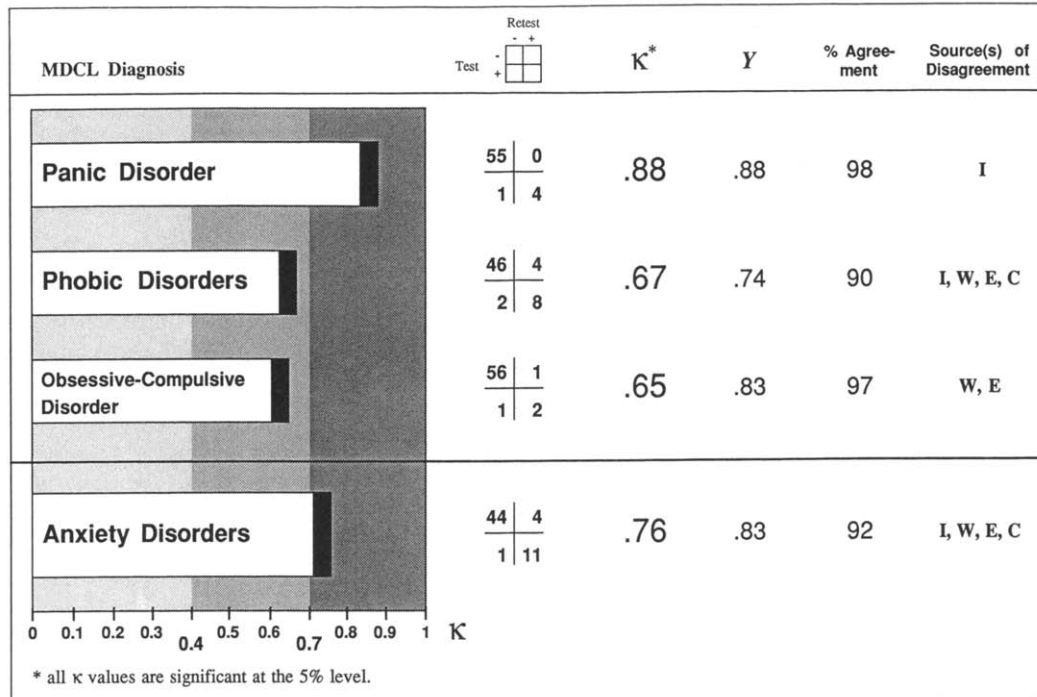


Fig. 5. Test–retest reliability of MDCL anxiety diagnoses (DSM-III-R).

manner to describe and ‘explain’ his complaints, and diagnostic discrepancies were likely to be found in such cases.

An important impact on reliability stemmed from the fact that *lifetime* diagnoses were to be assessed in our study. Of the eight discrepant

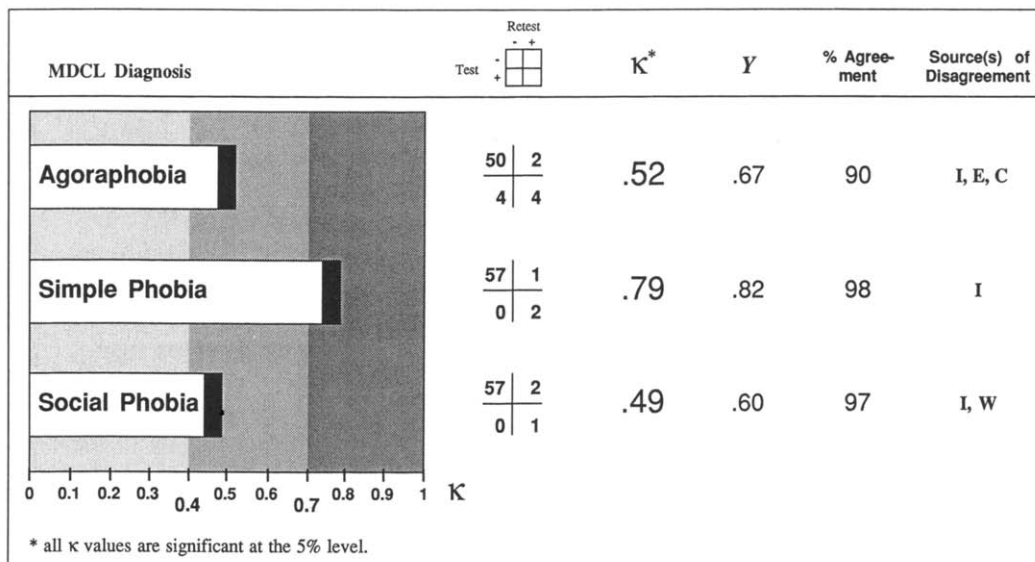


Fig. 6. Test–retest reliability of MDCL diagnoses of phobias (DSM-III-R).

major depression cases (cf. Fig. 2), three patients were not acutely depressed, but reported relatively short depressive episodes (maximum 6 weeks) in their past history (5-20 years ago). These episodes had been related to environmental stress, and they were in no case part of a longstanding depressive disorder.

However, these patients had difficulties remembering the exact symptoms and their duration, or they simply did not report the episodes in one of the interviews because there was no direct relevance to the present situation. We re-calculated agreement measures by considering only present states of depression (i.e., the three patients were eliminated from the group with discrepant results), and found an increased κ of 0.83 for major depression ($Y = 0.83$; 92% overall agreement).

Anxiety disorders

The results for MDCL anxiety diagnoses are summarized in Fig. 5. Panic disorder was analyzed independently of co-existing agoraphobia (i.e., panic disorder with and without agoraphobia was grouped together), and the three categories of agoraphobia (with or without a history of panic disorder), simple phobia, and social phobia were combined into one global category of phobic disorders.

We obtained excellent agreement for panic disorder ($\kappa = 0.88$), and high values for phobic disorders ($\kappa = 0.67$) and obsessive-compulsive disorder ($\kappa = 0.65$). Furthermore, a high level of congruence ($\kappa = 0.76$) was obtained in an overall analysis for anxiety disorders (i.e., agreement in diagnosing *no* or *at least one* specific anxiety disorder). All κ values were significant at the 5% level. Overall percentage agreement varied between 90 and 98.

In a next step, the individual categories of phobias were analyzed separately (Fig. 6). Satisfactory agreement was reached only for simple phobia ($\kappa = 0.79$), whereas a somewhat reduced congruence was found for agoraphobia ($\kappa = 0.52$) and social phobia ($\kappa = 0.49$). These values, however, should be interpreted with caution, since base rates for both categories were low in our sample (5%), and a comparison of κ and Y indicates that higher agreement rates could be expected under more favorable base rate conditions.

Figs. 5 and 6 additionally give sources of disagreement for each of the anxiety categories. Information differences in the patients' reports (I) were found to be an important reason for incongruence in all disorders with only one exception (obsessive-compulsive disorder).

We observed that even minor inconsistencies in the patient's information could cause marked diagnostic discrepancies. For example, a 38-year-old woman with panic attacks reported three anxiety symptoms in the first, and four symptoms in the second interview. A discrepancy in diagnostic level resulted, since a minimum number of four symptoms (as required for the diagnosis of panic disorder) was just missed in the first and barely reached in the second exploration.

It should be stressed that information variance can easily influence *more* than one diagnostic decision. A 24-year-old woman was judged by both diagnosticians to have panic attacks and a severe pattern of sociophobic anxiety (eating in the presence of others). During the first exploration, she reported that the onset of the sociophobic symptoms was related to the fear of getting a panic attack (in a restaurant). The clinician consequently diagnosed panic disorder with agoraphobia, and ruled out social phobia (because criterion B of social phobia excludes cases with fear of having panic attacks). In the retest interview, the patient said that the sociophobic symptoms had started 6 months *before* the first panic attack. Thus, a social phobia was diagnosed in addition to a panic disorder, but this time the panic disorder was classified as *without* agoraphobia (since the phobic symptoms were considered to be part of a social phobia). Thus, both clinicians broadly agreed on symptomatology, but the different information caused *two* diagnostic discrepancies (social phobia and agoraphobia) and only *one* concordance (panic disorder).

Clinical weighting (W) was found to be crucial in obsessive-compulsive disorder and social phobia, especially when the differential diagnosis of a depressive disorder had to be considered (e.g., a patient's report about social fears of being criticized or embarrassed at work was considered to be part of a severe depressive syndrome by one clinician, but not by the other one).

A surprisingly strong influence of insufficiently

defined diagnostic criteria (*C*) was found for agoraphobia, where diagnosticians often disagreed about the severity of the symptomatology. Criterion A of agoraphobia (without history of panic disorder) contains a description of agoraphobic fears as well as of negative consequences (which should be used to estimate severity). In our study, a number of patients reported relatively moderate agoraphobic fears, and it was open to subjective interpretation whether negative consequences of *clinical relevance* had emerged. For instance, when no avoidance behavior was reported, the clinician had to decide whether the agoraphobic situations were endured despite *intense* anxiety. Unfortunately, DSM-III-R does not define intense anxiety either quantitatively or qualitatively.

Comparison with other studies

The relevance of our results was evaluated by comparisons with other reliability studies. We refer to three investigations where test-retest agreement was assessed for DSM-III disorders with the help of standardized diagnostic interviews: (1) the *Diagnostic Interview Schedule* (DIS; Robins et al., 1981), Spanish version, administered by Burnam et al. (1983) to 61 monolingual hispanic outpatients; (2) the *Anxiety Disorders Interview Schedule* (ADIS), a structured instrument to diagnose DSM-III anxiety disorders, developed and tested by Di Nardo et al. (1983) in a sample of 60 patients; and (3) the *Composite International Diagnostic Interview* (CIDI; Robins et al., 1988), basically a version of the DIS to which items of the Present State Examination (PSE; Wing et al., 1974) have been added for a more comprehensive diagnostic assessment. A reliability study was conducted by Semler et al. (1987) with 60 inpatients.

A comparison of the different findings is graphically displayed in Fig. 7 (giving κ values). MDCL reliability was similar or moderately superior for most disorders. Exceptions are agoraphobia and social phobia. The ADIS reached higher values for these disorders, but it should be taken into consideration that this instrument was constructed primarily for anxiety disorders (thus being more differentiated in this field).

It can be seen from Fig. 7 that low values for dysthymia were found in all studies. Thus, the obtainable reliability of this diagnosis seems to be

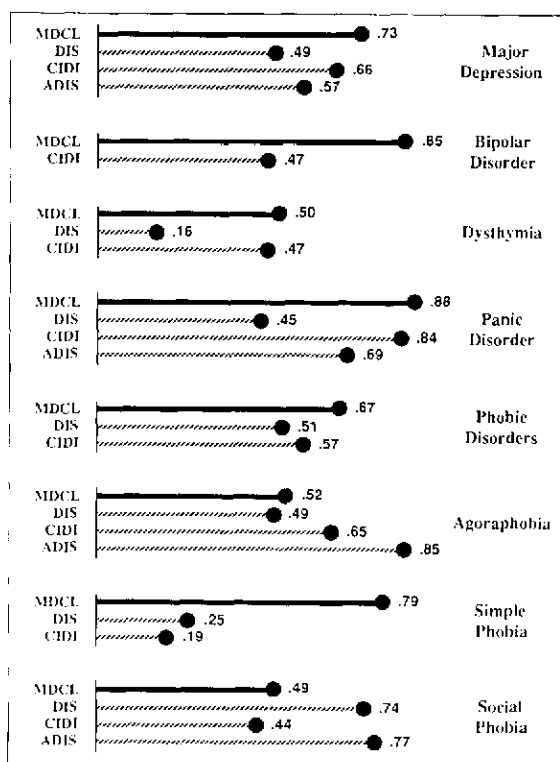


Fig. 7. Comparison of reliability studies (κ values).

unsatisfactory, independently of specific instruments. We frequently observed that it was difficult for patients to accurately respond to the lengthy time specifications of this disorder (i.e., depressed mood most of the time for 2 years, and never without this mood for more than 2 months at a time).

Discussion

This study evaluated a new methodological approach of assessing psychiatric diagnoses, which had been stimulated by the specific conditions of a psychiatric outpatient department. The classification system DSM-III-R and its criteria served as a basic frame of reference, and checklists were integrated into the usual clinical examinations. Thus, the diagnostic procedure was standardized to some degree (as compared to completely free evaluations). Our checklists, the MDCL, were subjected to a stringent test of reliability.

Good to excellent concordance was obtained for most categories of affective and anxiety disorders. κ values of 0.80 or higher were found for panic disorder and bipolar disorder, of above 0.70 for major depression and adjustment disorder/uncomplicated bereavement, and of above 0.60 for phobic disorders and obsessive-compulsive disorder. Somewhat reduced agreement resulted for dysthymia, agoraphobia, and social phobia, but a comparison with other studies suggests that problems in diagnosing these disorders may be due to weaknesses in their definition and operationalization. It was further demonstrated for most symptoms of major depression that κ values of above 0.60 could be reached with the checklist approach.

In all, our findings bear comparison with the reliability obtained for standardized interviews, though each method should be considered to have limitations of its own. Checklist-guided interviews are less structured, allowing for a more idiosyncratic and subjective style of investigation, but they integrate clinical judgement and thus reduce errors due to information variance (which must be expected in instruments such as the Diagnostic Interview Schedule, where interviewers are instructed to code patients' answers without any clinical weighting).

Qualitative analyses (sources of disagreement) showed that clinical disagreement only partly depends on specific exploration styles if diagnostic criteria are accepted *and* accurately referred to as guidelines. In our opinion, precision and unambiguity of diagnostic criteria are important prerequisites for reliable diagnoses. This is supported by results of the DSM-III field trials, Phase Two (American Psychiatric Association, 1980), where a surprisingly high interrater reliability was obtained by purely free clinical evaluations. For example, κ was 0.80 for major affective disorders and psychoactive substance use disorders, and 0.72 for anxiety disorders.

However, caution should be used in drawing general conclusions about the reliability and validity of particular methods until findings are replicated by independent investigations. Some of the above-mentioned issues could certainly be clarified in further research by directly comparing the different procedures (checklists vs. standardized interviews). Furthermore, it seems important to

investigate to which degree diagnoses are reliable over longer periods of time.

From a clinical perspective, the interpretation of reliability rates should involve a distinction between primary and additional diagnoses. We found that disagreement can indirectly arise from the DSM-III-R concept of comorbidity (allowing for multiple diagnoses in a single patient in order to comprehensively characterize the symptomatology). Patients often present with one dominant disorder (e.g., depression), accompanied by one or more *additional* complaints (e.g., anxiety, alcoholism) which are less severe and of questionable diagnostic relevance.

For example, it is generally not difficult to reliably diagnose agoraphobia in a patient who suffers from serious and clearly restricting agoraphobic fears. However, if only intermittent or relatively mild agoraphobic symptoms are present (often merely found by intense questioning) in a predominantly depressed patient, it may remain unclear if an additional diagnosis is justified (i.e., if the degree of severity is sufficient for a diagnosis of agoraphobia). Concordance rates should generally be expected to increase if only principal diagnoses are considered.

Diagnostic instruments and procedures cannot be characterized by one *true* reliability, since the obtainable reliability always depends upon specific conditions (e.g., in- vs. outpatients, severe vs. mild disorders, experienced vs. non-experienced diagnosticians; cf. Klerman, 1985). However, dependable diagnoses are needed for adequate clinical communication as well as for therapeutic considerations, and we feel confident that checklists, if conceptualized relatively simply and transparently, can enhance diagnostic evaluation and classification even under clinical routine conditions.

Acknowledgements

This work was supported by Grant Mo 439/1-3 of the German Research Foundation (DFG). The authors greatly appreciate the helpful comments made by Werner Mombour, M.D.

References

- American Psychiatric Association (1980) *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. American Psychiatric Association, Washington, DC.
- American Psychiatric Association (1987) *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. revised. American Psychiatric Association, Washington, DC.
- Bishop, Y.M.M., Feinberg, S.E. and Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.
- Blashfield, R.K. (1984) *The Classification of Psychopathology. Neo-Kraepelinian and Quantitative Approaches*. Plenum Press, New York, NY.
- Burnam, N.A., Karno, M., Hough, R.L., Escobar, J.I. and Forsythe, A.B. (1983) The Spanish Diagnostic Interview Schedule. Reliability and comparison with clinical diagnoses. *Arch. Gen. Psychiatry* 40, 1189–1196.
- Carey, C. and Gottesman, I.I. (1978) Reliability and validity in binary ratings. Areas of common misunderstanding in diagnosis and symptom ratings. *Arch. Gen. Psychiatry* 35, 1454–1459.
- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.
- Di Nardo, P.A., O'Brien, G.T., Barlow, D.H., Waddell, M.T. and Blanchard, E.B. (1983) Reliability of DSM-III anxiety disorder categories using a new structured interview. *Arch. Gen. Psychiatry* 40, 1070–1074.
- Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions*. 2nd edn. Wiley, New York, NY.
- Grove, W.M., Andreasen, N.C., McDonald-Scott, P., Keller, M.B. and Shapiro, R.W. (1981) Reliability studies of psychiatric diagnosis. Theory and practice. *Arch. Gen. Psychiatry* 38, 408–413.
- Helzer, J.E., Robins, L.N., Taibleson, M., Woodruff, R.A., Reich, T. and Wish, E.D. (1977) Reliability of psychiatric diagnoses: I. A methodological review. *Arch. Gen. Psychiatry* 34, 129–133.
- Hiller, W., Zaudig, M. and Mombour, W. (1989) *The Munich Diagnostic Checklists (MDCL)*. Logomed Hoepker, Munich.
- Hiller, W., Zaudig, M. and Mombour, W. (1990) Development of diagnostic checklists for use in routine clinical care. *Arch. Gen. Psychiatry* (in press).
- Klerman, G. (1985) Diagnosis of psychiatric disorders in epidemiologic field studies. *Arch. Gen. Psychiatry* 42, 723–724.
- Lorr, M., Klett, C.J. and McNair, D.M. (1963) *Syndromes of Psychosis*. Pergamon Press, Oxford.
- Robins, L.E. and Helzer, J.E. (1986) Diagnosis and clinical assessment: the current state of psychiatric diagnosis. *Annu. Rev. Psychol.* 37, 409–432.
- Robins, L.E., Helzer, J.E., Croughan, J. and Ratcliff, K.S. (1981) The NIMH Diagnostic Interview Schedule. Its history, characteristics, and validity. *Arch. Gen. Psychiatry* 38, 381–389.
- Robins, L.N., Wing, J., Wittchen, H.-U., Helzer, J.E., Babor, T.F., Burke, J., Farmer, A., Jablenski, A., Pickens, R., Regier, D.A., Sartorius, N. and Towle, L.H. (1988) The Composite International Diagnostic Interview. *Arch. Gen. Psychiatry* 45, 1069–1077.
- Semler, G., Wittchen, H.-U., Joschke, K., Zaudig, M., von Geiso, T., Kaiser, S., von Cranach, M. and Pfister, H. (1987) Test-retest reliability of a standardized psychiatric interview (DIS/CIDI). *Eur. Arch. Psychiatr. Neurol. Sci.* 236, 214–222.
- Semler, G., Wittchen, H.-U. and Zaudig, M. (1989) The test-retest reliability of the German version of the Composite International Diagnostic Interview on RDC diagnoses and symptom level. In: J.E. Mezzich and M. von Cranach (Eds.), *International Classification in Psychiatry*. Cambridge University Press, Cambridge.
- Shrout, P.E., Spitzer, R.L. and Fleiss, J.L. (1987) Quantification of agreement in psychiatric diagnosis revisited. *Arch. Gen. Psychiatry* 44, 172–177.
- Spitzer, R.L., Endicott, J. and Robins, E. (1978) Research Diagnostic Criteria: rationale and reliability. *Arch. Gen. Psychiatry* 35, 773–782.
- Spitzer, R.L., Williams, J.B. and Gibbon, M. (1987) *Structured Clinical Interview for DSM-III-R (SCID)*. Biometrics Research Department, NYS Psychiatric Institute, New York, NY.
- Spitznagel, E.L. and Helzer, J.E. (1985) A proposed solution to the base rate problem in the kappa statistic. *Arch. Gen. Psychiatry* 42, 725–728.
- Ward, C.H., Beck, A.T., Mendelson, M., Mock, J.E. and Erbaugh, J.K. (1962) The psychiatric nomenclature: reasons for diagnostic disagreement. *Arch. Gen. Psychiatry* 7, 198–203.
- Wing, J.K., Cooper, J.E. and Sartorius, N. (1974) *The Description and Classification of Psychiatric Symptoms: An Instruction Manual for the PSF and CATEGO System*. Cambridge University Press, London.
- Wittenborn, J.R. (1955) *Wittenborn Psychiatric Rating Scales*. Psychological Corp., New York, NY.
- Wittchen, H.-U., Semler, G. and von Zerssen, D. (1985) A comparison of two diagnostic methods. *Arch. Gen. Psychiatry* 42, 677–684.