

Evaluating the new ICD-10 categories of depressive episode and recurrent depressive disorder

Wolfgang Hiller ^{*}, Gabriele Dichtl ¹, Heidemarie Hecht ², Wolfgang Hundt, Werner Mombour, Detlev von Zerssen

Max-Planck-Institute of Psychiatry, Munich, FRG

(Received 12 July 1993; revision received 5 November 1993; accepted 17 November 1993)

Abstract

The new ICD-10 categories of 'depressive episode' (F32) and 'recurrent depressive disorder' (F33) were evaluated on the levels of diagnoses, subtypes, syndromes and symptoms. Interrater reliability was analyzed from findings of four independent diagnosticians who rated 100 written psychiatric case records with the help of criteria-related checklists. Agreement was sufficient if patients with and without the diagnosis of a depressive disorder were included ($\kappa = 0.82$), but low in the depressed subsample ($\kappa = 0.40$). Unclear boundaries became apparent for the classification of depressive syndromes as moderate or severe. The results suggest that the ICD-10 concept of depressive disorder is very similar to the well-known international concept of 'major depression'.

Key words: Depressive episode; Recurrent depressive disorder; ICD-10; Interrater reliability; Case record

1. Introduction

The diagnostic concept of depressive disorder has been modified repeatedly through numerous revisions of psychiatric classification systems. One of the currently most important innovations is the

tenth International Classification of Diseases (ICD-10), representing the official diagnostic system of the World Health Organization (WHO). The two new categories of 'depressive episode' and 'recurrent depressive disorder' have been introduced by ICD-10 within a separate section for affective disorders. These diagnoses are very similar to the concept of major depression which has become popular as part of the worldwide used American DSM-III and DSM-III-R systems (American Psychiatric Association, 1980, 1987). Disorders with a milder depressive symptomatology are classified separately in all systems as dysthymic disorder or adjustment disorder with depressed mood.

* Corresponding author (Present address). Clinic Roseneck, Center for Behavioural Medicine, Am Roseneck 6, D-83209 Prien, FRG.

¹ Dr. Dichtl is now at the Consultant Service for Neurology and Psychiatry, Municipal Hospital, Rosenheim, FRG.

² Dr. Hecht is now at the Department of Psychiatry, University of Freiburg, FRG.

The new ICD-10 system has approached DSM-III-R in many aspects. Both classifications are theoretically neutral, the definition of disorders is predominantly descriptive, and explicit criteria are specified which can be used for a more standardized diagnostic assessment. The new categories of depressive episode and recurrent depressive disorder are operationalized by the so-called ICD-10 diagnostic criteria for research (WHO, 1993). The general concept of both diagnoses is identical. A depressive episode is to be diagnosed only for a single (or the first) depressive episode. If two or more episodes have occurred and were separated by at least one depression-free interval of at least two months, the diagnosis of recurrent depressive disorder must be made.

The traditional but controversial distinction between psychotic and neurotic depression has been given up by ICD-10, and the term 'endogenous depression' is no longer used. However, ICD-10 provides an analogous diagnostic subtype which can be evaluated *in addition* to the diagnoses of single or recurrent depressive disorder. This so-called 'somatic syndrome' includes characteristic symptoms of the 'endogenous' symptom pattern, e.g., depression worse in the morning, waking up in the morning 2 hours or more before the usual time, lack of emotional reaction to events or activities that normally produce an emotional response. Nevertheless, it must be emphasized that 'endogenous depression' and 'somatic syndrome' should *not* be used as synonyms because ICD-10 defines the concept of somatic syndrome by purely descriptive features and does not include any etiological or pathogenetic assumptions.

First results from ICD-10 field trials (Sartorius et al., 1993; Dilling et al., 1990; Stieglitz et al., 1990) have shown that the new categories of depressive disorder seem to find good acceptance among clinicians in different psychiatric institutions. In these investigations, the interrater reliability of recurrent depressive disorder was acceptable even though the less strictly defined ICD-10 clinical descriptions and diagnostic guidelines were employed instead of diagnostic criteria. Another study was conducted by us (Hiller et al.,

1993a), referring to 100 clinical case records which were rated by four independent diagnosticians. We found that the observed interrater agreement for depressive disorders was 82% higher than agreement expected by chance alone. This represents an excellent result if compared with the reliabilities obtained according to traditional classifications (e.g., Spitzer and Fleiss, 1974).

To our knowledge, subtypes of depressive disorder and additional diagnostic sub-classifications as provided by ICD-10 were not evaluated until now. In the study reported here, we intended to analyze the reliability of the new categories more closely by referring to all levels of sub-classification. Interrater agreement and the nature of disagreements will be examined for (a) the distinction between single vs. recurrent depressive episodes; (b) the assessment of grades of severity; and (c) the presence or absence of a somatic syndrome. We will also consider the reliability on the level of single symptoms which are relevant for the general depressive syndrome as well as for the additional somatic syndrome. Diagnostic congruence will be evaluated not only for the differentiation between depressive and other psychiatric disorders, but also *within* the specific group of depressive patients.

2. Method

This investigation is part of a larger research program conducted to evaluate the reliability of major psychiatric disorders with the help of criteria-related diagnostic checklists. The design of the study has been described in detail elsewhere (Hiller et al., 1993a). In short, we employed the International Diagnostic Checklists (IDCL) to rate psychopathology and to assess ICD-10 diagnoses from 100 written case records. This material referred to a selected sample of psychiatric inpatients from the Munich Follow-Up Study (MFS; Wittchen et al., 1992). All of these patients had received the clinical ICD-8 diagnosis of an endogenous psychosis at the end of the index investigation and at follow-up 5 to 8 years later. The IDCL have first been introduced under the label 'MDCL' (Munich Diagnostic Checklists; cf

Hiller et al., 1990a,b; Bronisch et al., 1992). They are now associated to the family of instruments provided by the WHO for diagnostics according to ICD-10. We have described the IDCL in a separate article (Hiller et al., 1993b). All investigations were performed at the Max-Planck-Institute of Psychiatry in Munich (FRG).

2.1. Characteristics of case records and patients

The case records comprised the psychiatric symptomatology of the patients at admission and during hospitalization (mental status examination), mental disorders of family members, personal and social history, former psychiatric disturbances and other medical diseases. Further biographical data and any mention of a present or past psychiatric diagnosis were omitted before the case records were distributed to the participating diagnosticians. This was done because diagnoses were to be determined directly from the psychopathological descriptions and without influence from previous nosological considerations. The sample consisted of 47 male and 53 female patients with a mean age (at first admission) of 39.2 years ($SD = 9.44$ years). The original selection criteria in the MFS had been: (a) definite or probable clinical diagnosis according to the former ICD-8 system; (b) age between 20 and 65 years; (c) IQ of 85 or above; (d) length of inpatient treatment at least 10 days. The clinical ICD-8 diagnoses were 45 schizophrenic psychoses (295), 33 affective psychoses (296), and 22 other psychoses (295.7).

2.2. Diagnostic assessment

All case records were rated by four independent clinicians, two psychiatrists and two clinical psychologists (each one female and one male). They had clinical experience between 2 and 3 years and their clinical training had taken place in other psychiatric institutions or within various departments of the Max-Planck-Institute of Psychiatry. All patients described in the case reports were personally unknown to the four clinicians. The IDCL were used for all diagnostic assessments and the raters were instructed not to sim-

ply confirm their diagnostic impressions; rather, they were asked to check diagnostic criteria carefully and to make a specific diagnosis only if all relevant criteria were judged to be fulfilled. It was clear that none of the clinicians knew the diagnoses given by his colleagues when evaluating a specific case record. They also agreed to avoid any communication about individual cases until the collection of all data was completed.

The ICD-10 lists employed in this study referred to the 1990 draft of the ICD-10 diagnostic criteria for research. In this paper, we will restrict our analyses to the ratings for depressive disorders which can be made using the IDCL 'Depressive episode'. The four pages of this pocket-sized list are displayed in Fig. 1. On the first page, the ten single depressive symptoms as defined by ICD-10 are listed. For an individual patient, each symptom can be rated as present, probably present or absent, and the symptomatology can be documented as current or past. The second page allows for a specification of the depressive syndrome into the categories mild, moderate or severe. The diagnostician can further evaluate the criteria G1 to G3 which define the minimum duration of the depressive syndrome (2 weeks) and rule out hypomanic or manic episodes and organic factors. A decision between the distinct diagnoses of depressive episode, recurrent depressive disorder and other/unspecified depressive disorder can be made on the third page. The diagnostician may determine the diagnostic code by specifying the type of disorder (3rd character), the current degree of severity (4th character) and the presence or absence of a somatic syndrome (5th character). The last page of the checklist provides the explicit criteria of the additional somatic syndrome.

2.3. Distribution of diagnoses

A total of 400 ICD-10 diagnoses were assessed, i.e., one for each patient (100) by each diagnostician ($4 \times 100 = 400$). We obtained 88 diagnoses of depressive disorder (i.e., depressive episode *or* recurrent depressive disorder), 34 of bipolar disorder, 174 of schizophrenia, 16 of acute and transient psychotic disorder, 47 of schizoaf-

IDCL International Diagnostic Checklist for ICD-10

Depressive episode Name: _____ Age: _____ Date: _____

- (1) **Depressed mood** to a degree that is definitely abnormal for the individual, present for most of the day and almost every day, largely uninfluenced by circumstances, and sustained for at least 2 weeks.
- Probably
 No Yes
- Define the pattern of depressive symptomatology
 • Relate all symptoms to the period coded in (1)
- (2) Loss of interest or pleasure in activities that are normally pleasurable
- Probably
 No Yes
- (3) Decreased energy or increased fatigability
- No Yes
- (4) Loss of confidence or self-esteem
- No Yes
- (5) Unreasonable feelings of self-reproach or excessive and inappropriate guilt
- No Yes
- (6) Recurrent thoughts of death or suicide, or any suicidal behaviour
- No Yes
- (7) Complaints or evidence of diminished ability to think or concentrate, such as indecisiveness or vacillation
- No Yes
- (8) Change in psychomotor activity, with agitation or retardation (either subjective or objective)
- No Yes
- (9) Sleep disturbance of any type
- No Yes
- (10) Change in appetite (decrease or increase) with corresponding weight change
- No Yes

Specify if symptomatology is current or previous:

Yes Probably Yes Probably Yes Probably

Current: Symptomatology exists currently for the first time.

Current and previous: Symptomatology exists currently, and it had also been present in past history.

Previous: Symptomatology existed in past history (specify _____)

Depressive episode

Specify from symptoms (1) to (10): are the below criteria met for either mild, moderate or severe depressive episode?

mild	moderate	severe
<ul style="list-style-type: none"> A total of at least 4 of the symptoms (1) to (10), and at least 2 of them must be from (1) to (3) 	<ul style="list-style-type: none"> A total of at least 6 of the symptoms (1) to (10), and at least 2 of them must be from (1) to (3) 	<ul style="list-style-type: none"> A total of at least 8 of the symptoms (1) to (10), including all 3 of the symptoms (1), (2) and (3)
<input type="checkbox"/> met probably <input type="checkbox"/> not met	<input type="checkbox"/> met probably <input type="checkbox"/> not met	<input type="checkbox"/> met probably <input type="checkbox"/> not met

G1 The depressive episode lasts for at least 2 weeks. Stop ← No Probably Yes

G2 Rule out: Hypomania and mania
 Presence of hypomanic or manic symptoms sufficient to meet the criteria for hypomanic or manic episode at any time in the individual's life. (consider the diagnoses of bipolar affective disorder, mania or hypomania using the corresponding IDCL)

Yes Probably No

G3 Rule out: Organic aetiology
 The episode is attributable to psychoactive substance use or to any organic mental disorder.

Yes Probably No

Depressive episode

If G1 to G3 and criteria for mild, moderate or severe episode are met:

met probably not met

Check other diagnoses for mood (affective) disorders (listed at the bottom of page 4).
 • In the absence of a specific mood (affective) disorder, consider the residual categories for other or unspecified depressive episode or recurrent depressive disorder (bottom of page 4).
 • Determine the appropriate diagnosis (page 3).

fective disorder and 41 of other disorders. Of the 88 diagnoses of depressive disorder, 24 were made by the male psychologist, 22 by the male psychiatrist and 21 by each the female psychologist and the female psychiatrist. 29 patients of the total sample received a diagnosis of depressive disorder from *at least* one of the four diagnosticians.

2.4. Statistical analyses

Interrater reliability will be expressed mainly by κ (kappa), a chance-corrected statistic for categorical data. This measure was first introduced by Cohen (1960) and later extended by other authors to the case of multiple ratings with equal or unequal numbers of ratings per subject (e.g., Fleiss, 1981). Perfect interrater agreement is indicated by a κ value of 1 (which is the maximum possible value), while a κ of 0 results if the observed agreement is equal to the agreement expected by chance alone. We did not perform tests of significance but rather referred to a conventional interpretation of the magnitude of κ with values of 0.75 or above representing excellent agreement, 0.40–0.75 representing fair to good agreement and below 0.40 representing poor agreement (Fleiss, 1981). The frequencies (base rates) of diagnoses, subtypes, syndromes and symptoms will be reported in addition since κ may decrease rapidly if base rates are very low or very high (κ should then be interpreted only with caution).

A second measure of reliability will be the overall *percentage agreement* among all raters. We refer to the proportion of observed agreement which is used by Fleiss (1971) for the calculation of chance-corrected agreement. The interpretation of this measure can be demonstrated as follows: An agreement of, for example, 86% is equivalent to the conditional probability of 0.86 that a subject receives the diagnosis under consideration from a second rater, given that also the first rater had chosen this diagnosis.

All reliability analyses will be done separately for two different samples: (a) the total sample of 100 patients including $4 \times 100 = 400$ diagnoses; and (b) the more specific sample of the 29 pa-

tients who had received at least one diagnosis of a depressive disorder, including $4 \times 29 = 116$ diagnoses. We decided to use both samples as different and alternative frames of reference because each approach has important advantages as well as disadvantages. These can be summarized as follows:

(a) The total sample is necessary to determine the overall reliability for depressive disorder. This analysis must include patients with *and* without a depressive disorder, which means that diagnostic agreement about *inclusion* (of the 29 patients with a depressive disorder) as well as *exclusion* (of the 71 patients with other disorders) is to be considered. However, reliability may be overestimated by referring to the total sample when sub-classifications *within* the group of depressive patients are analyzed. By definition, disagreement about subtypes can occur only for the 29 depressive patients but not for the remaining 71 patients. Inadequately high reliabilities may therefore result if complete agreement about the absence of sub-classifications is counted for the 71 patients with other diagnoses although these sub-classifications had, in fact, not been evaluated for this group.

(b) The smaller sample of the 29 depressive patients was thus employed mainly for the analysis of sub-classifications of depressive disorder. This method represents a more stringent test of reliability because only subjects with a relevant depressive symptomatology are considered. Patients with other diagnoses are ignored and somehow trivial agreements are avoided, e.g., the exclusion of specific features of a depressive symptomatology in a schizophrenic patient who has never experienced a depressive episode. Nevertheless, the restriction to the 29-patients sample is not free of disadvantages. It must be expected that lower reliability values result from this method since the high proportion of (negative) agreement in the 71 remaining patients is disregarded. Reliabilities from the selected smaller sample are not sufficiently comparable with those from the more frequently used procedure of considering total samples. The interpretation of our results will therefore be based on *both* the total and the 29-patients sample.

3. Results

We will first analyze the subtypes of depressive disorder as differentiated by the ICD-10 diagnostic code. The composition of this code is shown on the third page of the IDCL 'Depressive episode' (Fig. 1). While the first two characters are 'F3' for *all* affective disorders, the third character is used to differentiate between single depressive episode (F32) and recurrent depressive disorder (F33). The fourth character specifies the current degree of severity with '0' for mild, '1' for moderate, '2' for severe without psychotic symptoms and '3' for severe with psychotic symptoms. The digit '1' in the fifth character defines the additional specification of a somatic syndrome (or, whenever present, mood incongruent psychotic symptoms).

The interrater reliabilities of the diagnoses and all principle sub-classifications are presented in Table 1. We obtained an excellent κ value of 0.82 for the global category of depressive disorder in the total sample, while the corresponding value among the 29 depressive patients was only 0.40 and thus insufficient. The base rate of 75.9% in the smaller sample refers to the overall number of 116 diagnoses. It should be remembered that not all of the 29 patients had received a diagnosis of depressive disorder from *all* four diagnosticians; rather, seven patients had been classified

as depressive by only one diagnostician, two by two diagnosticians, three by three diagnosticians and 17 by all four diagnosticians.

Our analyses for the global diagnosis of depressive disorder refer to a very strict definition of interrater agreement. In some cases, diagnoses of a residual 'other' or 'unspecified' depressive disorder had been made. This was based on the finding that criteria of a specific depressive disorder from the sections F32 and F33 had not been fulfilled completely. Specific and residual diagnoses were counted as *disagreements* in Table 1. However, if analyzed as *one* group, the reliability of depressive disorder was slightly better in the total sample ($\kappa = 0.85$, agreement 88.2%, base rate 23.3%) while unchanged in the smaller sample ($\kappa = 0.40$; agreement 88.2%, base rate 80.2%).

The further results in Table 1 show a good reliability for the differentiation between single depressive episode (F32) and recurrent depressive disorder with more than one depressive episode (F33). Both sub-classifications were determined with a reliability of $\kappa = 0.70$ or above in the total sample and 0.63 or above in the depressive sample. Percentage agreement was above 70% and even up to 87.2% among the depressed patients.

In contrast, we found generally less congruence for the current degree of severity which is used to specify the fourth character of the diag-

Table 1
Reliability of ICD-10 depressive disorder and sub-classifications

	<i>n</i> = 100 sample			<i>n</i> = 29 sample		
	κ	%	Base rate	κ	%	Base rate
Depressive disorder (F32 and F33)	0.82	85.6	22.0	0.40	85.6	75.9
Type of disorder (third character)						
Depressive episode (F32)	0.70	71.6	6.8	0.71	78.4	29.3
Recurrent depressive disorder (F33)	0.72	76.9	16.3	0.63	87.2	66.3
Current degree of severity (fourth character)						
Mild severity (F3x.0)	0.53	55.6	4.5	0.69	75.0	18.5
Moderate severity (F3x.1)	0.44	47.2	6.0	0.36	53.1	28.4
Severe, without psychotic symptoms (F3x.2)	0.30	33.3	4.5	0.15	35.3	22.2
Severe, with psychotic symptoms (F3x.3)	0.45	48.0	6.3	0.40	58.1	27.2
Specification of somatic syndrome (fifth character)						
With somatic syndrome	0.63	68.7	16.3	0.23	79.3	71.6

κ = kappa, % = percentage agreement.

nostic code. Values were somehow acceptable only for mild depressive disorder with $\kappa = 0.53$ in the total sample and 0.69 in the smaller sample. The remaining κ values were not above 0.45 and agreement rates were well below 60%. Severe depressive disorder without psychotic symptoms was least reliable with $\kappa = 0.15$ and only 35.3% agreement in the depressive sample. The insufficient values in the total sample could probably be attributed to the low base rates of no more than 4.5 to 6.3%, but the corresponding κ values were even smaller in the 29-patients sample where the base rates were medium sized from 22.2 to 28.4%.

The fifth-character specification of presence vs absence of a somatic syndrome was sufficiently reliable only in the complete sample ($\kappa = 0.63$). An unacceptable κ of 0.23 was obtained for the depressive sample, although percentage agreement was higher than in the total sample (79.3% vs 68.7%). In this case, the lower κ reflects a higher base rate and thus a higher amount of chance agreement in the depressive sample where a somatic syndrome was determined in more than 70% of all 116 diagnoses.

In a next step, we analyzed the amount of agreement on the level of single depressive symptoms and syndromes. These results are summa-

rized in Table 2. Most of the symptoms had good reliabilities in the total sample with highest values for depressed mood, unreasonable feelings of self-reproach or guilt, sleep disturbance and disturbance of appetite and weight change ($\kappa = 0.70$ or above). The lowest value in this sample was found for the symptom of diminished ability to think or concentrate with $\kappa = 0.44$. However, all symptoms were clearly less reliable in the depressive sample where eight of the ten symptoms had κ values not above 0.50. Relatively good agreement was obtained only for the symptom of unreasonable feelings of self-reproach or guilt (0.63), followed by sleep disturbance (0.53) and disturbance of appetite with weight change (0.48). Highest base rates were computed for the symptoms of depressed mood and loss of interest or pleasure which were assessed in more than 80% of the depressed patients. The least common symptom with a base rate of only 56.9% was that of unreasonable feelings of self-reproach or guilt.

Table 2 also presents the results for the subtyping of depressive syndromes into mild, moderate and severe (as operationalized by the ICD-10 definition of depressive disorder). This differentiation is based on a symptom count with at least four single depressive symptoms for the mild, six

Table 2

Reliability of depressive symptoms as defined by ICD-10 depressive disorder and of specifications into mild, moderate and severe depressive syndrome

	<i>n</i> = 100 sample			<i>n</i> = 29 sample		
	κ	%	Base rate	κ	%	Base rate
Depressive symptoms						
1. Depressed mood	0.71	81.6	35.8	0.19	91.7	89.7
2. Loss of interest or pleasure	0.67	77.4	32.5	0.28	87.0	81.9
3. Decreased energy or increased fatiguability	0.51	64.9	27.8	0.14	72.6	68.1
4. Loss of confidence or self-esteem	0.62	71.3	25.3	0.43	79.3	63.8
5. Unreasonable feelings of self-reproach/guilt	0.71	77.9	23.8	0.63	83.8	56.9
6. Suicidal thoughts or behaviour	0.59	68.7	24.5	0.41	77.0	61.2
7. Diminished ability to think or concentrate	0.44	58.1	25.3	0.17	69.4	62.9
8. Psychomotor agitation or retardation	0.61	70.1	24.3	0.34	77.2	65.5
9. Sleep disturbance	0.74	80.4	25.5	0.53	86.7	71.6
10. Disturbance of appetite with weight change	0.70	76.3	20.8	0.48	79.8	61.2
Specification of the depressive syndrome						
Mild depressive syndrome	0.58	60.0	5.0	0.66	73.5	17.3
Moderate depressive syndrome	0.38	43.9	9.5	0.15	47.6	38.8
Severe depressive syndrome	0.47	52.7	10.8	0.18	54.8	43.9

κ = kappa, % = percentage agreement.

for the moderate and eight for the severe syndrome. We found an acceptable interrater agreement only for the mild depressive syndrome with $\kappa = 0.58$ in the total sample and 0.66 among depressive patients. The values for moderate and severe depressive syndromes were clearly lower and close to chance agreement in the 29-patients sample.

The nature of discrepancies between the three distinct syndromes is illustrated graphically in Fig. 2. We determined how often each syndrome, diagnosed by each rater, was confirmed or disconfirmed by each of the other raters. Mild depressive syndromes were confirmed by 73.5% of all corresponding ratings whereas discrepancies were found in 12.2% due to a moderate syndrome and in 14.3% due to a severe syndrome. It can be seen from Fig. 2 that only 47.6% of the moderate syndromes and 54.8% of the severe syndromes were confirmed by corresponding ratings, reflecting the lower reliabilities of these sub-classifications. 46.7% of the moderate syndromes were incongruently classified as severe and 39.5% of the severe syndromes were incongruently classified as moderate. Thus, the boundaries between the moderate and severe syndrome were clearly weaker than those between the mild and the other two syndromes.

It was already shown in Table 1 that the additional subtype of a somatic syndrome yielded a good interrater agreement only in the total sample. Table 3 gives the reliabilities of the eight symptoms which are to be evaluated in order to decide whether a somatic syndrome is present or

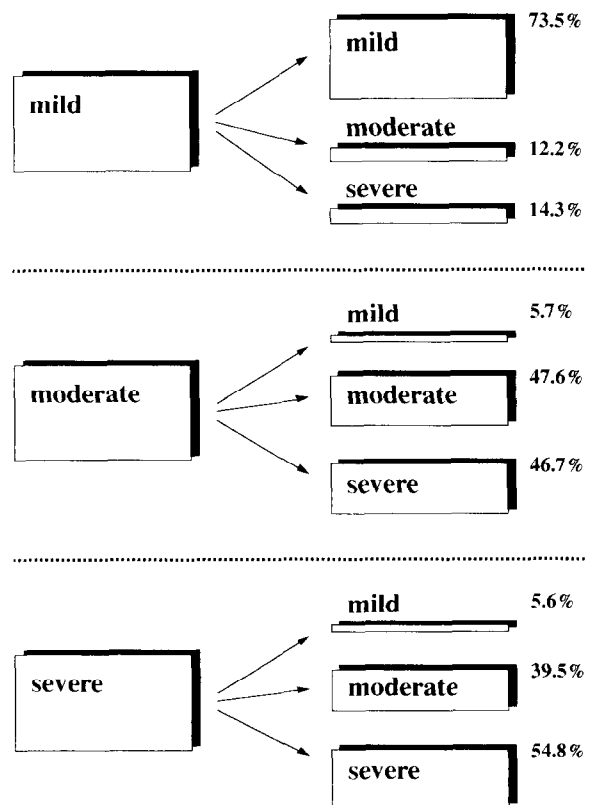


Fig. 2. Interrater discrepancies between different syndromes of depression.

not. The κ values of these symptoms were between 0.43 and 0.72 in the total sample and between 0.18 and 0.63 in the depressive sample. A sufficient reliability with κ above 0.60 in both samples was found only for the symptom of de-

Table 3
Reliability of symptoms as defined for the ICD-10 diagnostic specification of somatic syndrome

	n = 100 sample			n = 29 sample		
	κ	%	Base rate	κ	%	Base rate
1. Marked loss of interest or pleasure	0.62	70.5	21.5	0.19	75.0	69.0
2. Lack of emotional reactions to events or activities	0.43	51.5	14.3	0.18	55.3	45.7
3. Waking in the morning 2 h or more than usual	0.52	56.3	8.0	0.43	57.8	25.9
4. Depression worse in the morning	0.72	76.2	15.8	0.63	81.6	50.0
5. Objective psychomotor retardation or agitation	0.60	66.7	17.5	0.30	69.7	56.9
6. Marked loss of appetite	0.65	69.9	15.3	0.45	72.4	50.0
7. Weight loss of 5% or more	0.57	60.8	8.5	0.48	62.6	28.4
8. Marked loss of libido	0.57	62.1	11.0	0.45	65.1	36.2

κ = kappa, % = percentage agreement.

pression worse in the morning. Agreement for the first two symptoms, i.e., loss of interest or pleasure and lack of reactivity to events or activities, was close to chance agreement with κ of 0.18 and 0.19 in the 29-patients sample.

4. Discussion

Empirical standards such as the objectivity and reproducibility of diagnostic findings have become increasingly important to test the quality of new psychiatric classifications. ICD-10 has introduced a new concept of depressive disorders which defines explicit diagnostic criteria and subdivides depressive syndromes according to the number of depressive episodes and their degree of severity. In the present study, we conducted a first evaluation of these diagnoses by referring to the reliability of subtypes, syndromes and symptoms. We analyzed the interrater agreement between four independent diagnosticians, each of whom rated 100 case records describing the psychopathology and course characteristics of psychiatric inpatients.

Our results showed that the global diagnosis of depressive disorder was reliable only when patients with *and* without this disorder were included. If the sample was restricted to depressed patients alone, interrater agreement was poor and κ values decreased from 0.82 (total sample) to 0.40 (depressive sample). Analyses of sub-classifications according to the different digits of the diagnostic code showed that single vs. recurrent depressive episodes were well discriminated with κ of 0.60 or above, while the fourth-character reliability of currently moderate or severe depressions was well below 0.50 and thus insufficient. Agreement for single signs and symptoms was generally good to excellent with most κ values above 0.70 in the total sample. However, our data were discouraging for the differentiation of mild, moderate and severe depressive syndromes. Of all severe syndromes 39.5% were incongruently classified as moderate by competing ratings of other diagnosticians, and the corresponding disagreement rate for moderate syndromes was even 46.7%.

We wish to emphasize that not all aspects of differential diagnosis in depressed patients were represented in our investigation. We studied a sample of severely impaired inpatients who had received the (traditional) clinical diagnosis of an endogenous psychosis. For these patients, the most crucial delineations are those between unipolar or bipolar affective disorders, schizoaffective disorder and schizophrenia. The question of diagnostic differentiation is considerably different if patients with milder forms of psychiatric disorders are examined. For example, the differential diagnosis of depressive vs. dysthymic disorder will be likely to cause ambiguous diagnostic decisions if depressive states are fluctuating and less clearly separated by episodes. We have demonstrated these difficulties in a previous study with psychiatric outpatients, where the test–retest reliabilities were only 0.73 for the DSM-III-R category of major depression and no more than 0.50 for dysthymic disorder (Hiller et al., 1990b).

Other limitations of the study reported here come from the specific methodology. We worked with the case record method which is generally considered as a standard procedure to assess diagnostic reliability (e.g., Grove et al., 1981). This method differs from other commonly used procedures such as to conduct live interviews or to rate video-taped explorations, but it is known that each approach is prone to different sources of diagnostic disagreement. The use of written material guarantees that diagnostic decisions are based on the same information for each rater. Another advantage is the inclusion of several raters for the same cases (thus differing from usual test–retest interviews where comparisons are restricted to only two diagnosticians). However, it should be considered that all methods are sensitive to other modifying variables of reliability which may differ from study to study, such as the quality of the patients' reports, the level of the raters' diagnostic expertise or the duration of the specific psychopathological training. Results from one reliability study should therefore *not* be generalized to other settings and conditions.

Despite some methodological differences, similar overall results were found in the international ICD-10 field trial with $\kappa = 0.66$ for depressive

episode and 0.69 for recurrent depressive disorder (Sartorius et al., 1993). As in our study, weaknesses of sub-classification became apparent since κ values of not above 0.42 resulted for the reliability of *mild* depressive disorders. Somewhat lower κ values were obtained in the national samples in German-speaking countries (0.49 for recurrent depressive disorder and 0.17 for depressive episode; cf Zaudig et al., 1990; Freyberger et al., 1990) and in the Western Pacific region (0.56 for depressive episode and 0.44/0.56 for the differentiation between mild and severe depressive episodes, cf Ellis et al., 1990). It should be considered, however, that diagnoses in these field trials were made according to the less stringent ICD-10 clinical descriptions and diagnostic guidelines rather than according to explicit diagnostic criteria.

Although the classification of depression has been a complex and controversial topic for many decades (cf Andreasen, 1982; Paykel, 1983; Ravana and Paykel, 1992), it becomes apparent now that some continuity arose from the concept of operationalized diagnoses. The historical roots of the ICD-10 definition of depressive disorders lie in the work of Feighner and associates (1972) who were the first to construct criteria-based diagnoses for psychiatric disorders. This concept was later taken over by the 'Research Diagnostic Criteria' (RDC; Spitzer et al., 1978), where it was labelled with the term 'major depression'.

Since then, only minor modifications were included into the corresponding concepts of official classification systems such as DSM-III/DSM-III-R and ICD-10. Philipp et al. (1991a) criticized that ICD-10 does not use the term 'major depression' even though the concept of depressive disorders in ICD-10 is practically identical with its predecessors. These authors stated: 'The similarity especially of the item pools being used to define the depressive syndrome is so striking that it could be questioned whether new classification developments should take over identical definitions rather than making new mini-changes and creating new terms for old contents'. (Philipp et al., 1991a, p. 264). A similar opinion was expressed by Mombour et al. (1990) who analyzed qualitative criticisms made during the ICD-10

field trials in German-speaking countries. According to them, an important reason for the acceptance of the new ICD-10 concept of depressive disorders is its similarity to the well accepted DSM-III/DSM-III-R forms of depression.

A comprehensive comparison of diagnoses of depression in competing classification systems has been published by Philipp et al. (1991b). They found that the 1989 criteria of ICD-10 showed a good level of congruence with five other operational definitions and a κ of 0.69 was computed as a corresponding measure of general agreement. The inter-system overlap was somewhat higher for DSM-III and DSM-III-R ($\kappa = 0.74$ and 0.75), similar for RDC (0.70) and lower for the 1987 criteria of ICD-10 (0.66) and the Feighner criteria (0.62). Philipp et al. (1991b) argued that the possible international acceptance of the individual systems could be related to these κ values (with high values indicating a higher degree of acceptance).

To summarize, these results and those presented by us suggest that the new ICD-10 definition of depressive disorder, if considered as a whole, seems to be in accordance with requirements which were formulated for modern and empirically based classifications for mental disorders (e.g., Sartorius, 1988; Cooper, 1988). Improvements may be necessary for sub-classifications such as the severity of depressive syndromes or the additional subtype of somatic syndrome. These aspects, as well as the choice of appropriate assessment methods, should be considered in future diagnostic investigations with depressed patients.

Acknowledgement

Parts of this work were supported by grant Mo 439/1-3 of the German Research Foundation (Deutsche Forschungsgemeinschaft).

References

- American Psychiatric Association (1980) Diagnostic and Statistical Manual of Mental Disorders, 3rd Edn. (1987) 3rd

- Edn. revised. American Psychiatric Association, Washington, DC.
- Andreasen, N. (1982) Concepts, diagnosis, and classification. In: E.S. Paykel (Ed.), *Handbook of Affective Disorders*. Churchill Livingstone, Edinburgh.
- Bronisch, T., Garcia-Borreguero, D., Flett, S., Wolf, R. and Hiller, W. (1992) The Munich Diagnostic Checklist for the assessment of DSM-III-R personality disorders for use in routine clinical care and research. *Eur. Arch. Psychiatry Clin. Neurosci.* 242, 77–81.
- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.
- Cooper, J.E. (1988) The structure and presentation of contemporary psychiatric classifications with special reference to ICD-9 and 10. *Br. J. Psychiat.* 152 (Suppl.) 1, 21–28.
- Dilling, H., Freyberger, H.J. and Malchow, P. (1990) Design of the ICD-10 field trial in German-speaking countries. *Pharmacopsychiatry* 23 (Suppl. IV), 142–145.
- Ellis, P.M., Welch, G., Purdie, G.L. and Mellso G.W. (1990) Australasian field trials of the mental and behavioural disorders section of the draft ICD-10. *Aust. NZ J. Psychiatry* 24, 313–321.
- Feighner, J.P., Robins, E., Guze, S.B., Woodruff, R.A., Winokur, G. and Munoz, R. (1972) Diagnostic criteria for use in psychiatric research. *Arch. Gen. Psychiatry* 26, 57–63.
- Fleiss, J.L. (1971) Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382.
- Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions*, 2nd Edn. Wiley, New York, NY.
- Freyberger, H.J., Albus, M. and Stieglitz, R.-D. (1990) ICD-10 field trial in German-speaking countries – summary of the quantitative empirical results. *Pharmacopsychiatry* 23 (Suppl. IV), 192–196.
- 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.
- Hiller, W., Zaudig, M. and Mombour, W. (1990a) Development of diagnostic checklists for use in routine clinical care. *Arch. Gen. Psychiatry* 47, 782–784.
- Hiller, W., von Bose, M., Dichtl, G. and Agerer, D. (1990b) Reliability of checklist-guided diagnoses for DSM-III-R affective and anxiety disorders. *J. Affect. Disord.* 20, 235–247.
- Hiller, W., Dichtl, G., Hecht, H., Hundt, W. and von Zerssen, D. (1993a) An empirical comparison of diagnoses and reliabilities in ICD-10 and DSM-III-R. *Eur. Arch. Psychiatry Clin. Neurosci.* 242, 209–217.
- Hiller, W., Zaudig, M., Mombour, W., and Bronisch, T. (1993b) Routine psychiatric examinations guided by ICD-10 diagnostic checklists (International Diagnostic Checklists). *Eur. Arch. Psychiatry Clin. Neurosci.* 242, 218–223.
- Mombour, W., Spitzner, S., Reger, K.H., von Cranach, M., Dilling, H. and Helmchen, H. (1990) Summary of the qualitative criticisms made during the ICD-10 field trial and remarks on the German translation of ICD-10. *Pharmacopsychiatry* 23 (Suppl. IV), 197–201.
- Paykel, E.S. (1983) The classification of depression. *Br. J. Clin. Pharmacol.* 15, 155S–159S.
- Philipp, M., Maier, W. and Delmo C.D. (1991a) The concept of major depression. I. Descriptive comparison of six competing operational definitions including ICD-10 and DSM-III-R. *Eur. Arch. Psychiatry Clin. Neurosci.* 240, 258–265.
- Philipp, M., Maier, W. and Delmo C.D. (1991b) The concept of major depression. II. Agreement between six competing operational definitions in 600 psychiatric inpatients. *Eur. Arch. Psychiatry Clin. Neurosci.* 240, 266–271.
- Ramana, R. and Paykel, E.S. (1992) Classification of affective disorders. *Br. J. Hosp. Med.* 47, 831–835.
- Sartorius, N. (1988) International perspectives of psychiatric classification. *Br. J. Psychiatry* 152 (Suppl. 1), 9–14.
- Sartorius, N., Kaelber, C.T., Cooper, J.E., Roper, M.T., Rae, D.S., Gulbinat, W., Üstün, T.B. and Regier, D.A. (1993) Progress toward achieving a common language in psychiatry. *Arch. Gen. Psychiatry* 50, 115–124.
- Spitzer, R.L. and Fleiss, J.L. (1974) A re-analysis of the reliability of psychiatric diagnosis. *Br. J. Psychiatry* 125, 341–347.
- Spitzer, R.L., Endicott, J. and Robins, E. (1978) Research Diagnostic Criteria: rationale and reliability. *Arch. Gen. Psychiatry* 35, 773–782.
- Stieglitz, R.-D., Zaudig, M., Freyberger, H.J. and Dittmann, V. (1990) Feasibility, suitability, and interrater reliability of ICD-10 during different stages of the ICD-10 field trial. *Pharmacopsychiatry* 23 (Suppl. IV), 188–191.
- Wittchen, H.-U., Essau, C.A., von Zerssen, D., Krieg, J.-C. and Zaudig, M. (1992) Lifetime and six-month prevalence of mental disorders in the Munich follow-up study. *Eur. Arch. Psychiatry Clin. Neurosci.* 241, 247–258.
- World Health Organization, WHO (1993) ICD-10, Chapter V (F), Mental and Behavioural Disorders (including Disorders of Psychological Development), Diagnostic Criteria for Research. WHO, Geneva.
- Zaudig, M., Stieglitz, R.-D., Gastpar, M. and Rösinger, C. (1990) Mood (affective) and schizoaffective disorders (Section F3): results of the ICD-10 field trial. *Pharmacopsychiatry* 23 (Suppl. IV), 160–164.