

An Empirical Comparison of Diagnoses and Reliabilities in ICD-10 and DSM-III-R

Wolfgang Hiller¹, Gabriele Dichtl², Heidemarie Hecht³, Wolfgang Hundt⁴, and Detlev von Zerssen⁴

¹Clinic Roseneck, Center for Behavioral Medicine, Am Roseneck 6, 8210 Prien, Federal Republic of Germany

²Consultant Service for Neurology and Psychiatry, Municipal Hospital, Pettenkoferstr. 10, 8200 Rosenheim, Federal Republic of Germany

³Department of Psychiatry, University of Freiburg, Hauptstr. 5, 7800 Freiburg, Federal Republic of Germany

⁴Max-Planck-Institute of Psychiatry, Kraepelinstr. 10, 8000 München 40, Federal Republic of Germany

Received December 4, 1992

Summary. The psychiatric classification systems ICD-10 and DSM-III-R were compared by applying both sets of diagnostic criteria to the same sample of patients suffering from affective and psychotic disorders. Four independent raters assessed diagnoses according to both systems to 100 written case records which had been prepared in a traditional, standard format. The International Diagnostic Checklists (IDCL) were employed to rate relevant psychopathological signs and symptoms and to apply diagnostic decision rules. The results showed that ICD-10 yielded a generally higher reliability for all main disorders except for bipolar disorder. Overall reliability was $\kappa = 0.53$ for diagnoses according to DSM-III-R and 0.59 for diagnoses according to ICD-10. Agreement was best for affective disorders, moderate for schizophrenia and unacceptable for schizoaffective disorder. Insufficient boundaries were found in both systems between schizoaffective disorder on one side and schizophrenia and bipolar disorder on the other side. The different duration criteria for schizophrenia of six months in DSM-III-R and one month in ICD-10 tend to have considerable consequences for frequency rates of schizophrenia in a typical clinical setting.

Key words: Classification – ICD-10 – DSM-III-R – Reliability – Case Records – International Diagnostic Checklists (IDCL)

Introduction

The use of fixed diagnostic criteria and the assessment of operationalized diagnoses have become standard methods in psychiatric classification. They have been introduced into clinical psychiatry by DSM-III and its revised form DSM-III-R (APA 1987). This approach is now ex-

tended by chapter V of the tenth revision of the International Classification of Diseases (ICD-10) which has been developed by the World Health Organization (WHO 1990) for the worldwide use in WHO member countries.

The new ICD-10 system resembles DSM-III-R in many ways (Cooper 1988; Sartorius 1988; Maier et al. 1990). It has adopted a number of principles which have proven useful in the clinical and scientific application of DSM-III and DSM-III-R. Above all, diagnostic criteria for research are incorporated for the first time into an ICD system in order to define the symptomatology and other relevant characteristics (e.g., time of onset, course, severity) of mental disorders. During clinical examinations, the diagnostician can refer to these criteria in order to decide whether a specific diagnosis can be given or not. ICD-10 has also given up the traditional etiological distinction between neurosis and psychosis and most disorders are now defined by primarily descriptive features. First results from ICD-10 field trials showed a good degree of acceptance of the new system (Dilling et al. 1990; Freyberger et al. 1990).

The arrangement of diagnostic classes in ICD-10 is very similar to that of DSM-III-R. ICD-10 offers a separate chapter for affective disorders with the distinction between depressive and bipolar disorders and a subclassification of depressive disorders according to the severity of depressive syndromes. Another chapter exists for nonorganic psychotic disorders including schizophrenia and related disorders. ICD-10 and DSM-III-R are equal in delineating schizophrenia from disorders with prominent delusions (i.e., delusional or paranoid disorders) and from schizophrenia-like episodes of shorter duration (i.e., schizophreniform and acute/transient psychotic disorders). The category of schizoaffective disorder has been placed in both systems in the same chapter with schizophrenia and not with affective disorders.

On the other hand, a number of disparities can be found when comparing diagnostic criteria in both clas-

sifications more closely. Examples are the definitions of psychotic and affective syndromes which are not completely identical in their content and in the number of associated signs and symptoms (though very similar). However, more important are differences between the duration criteria for schizophrenia, since DSM-III-R defines a minimum of six months and ICD-10 no more than one month. For the diagnosis of schizoaffective disorder, ICD-10 does not require a distinct psychotic episode without prominent affective symptoms as does DSM-III-R.

It remains unclear from the pure descriptions of disorders which consequences are to be expected from such differences when subjects are classified in typical clinical settings or epidemiological surveys. Studies are therefore needed to investigate the practical degree of similarity between corresponding categories in ICD-10 and DSM-III-R. If substantial differences were found for individual diagnoses, it should further be evaluated which of the competing definitions provided more empirical evidence.

To our knowledge, the study reported here is the first one to investigate the degree of comparability of diagnoses in ICD-10 and DSM-III-R by means of an interrater design. We applied the diagnostic criteria of both systems to a set of 100 case records which were derived from routine clinical case histories of a psychiatric hospital. The same patients had been investigated in the Munich Follow-Up Study (MFS; c.f. Wittchen et al. 1992) where they all had received a definite or probable clinical diagnosis of an endogenous psychosis. For the purpose of the present study, all case records were read and evaluated by four independent diagnosticians. The following questions are addressed: (1) Which impact have differences in the ICD-10 and DSM-III-R definitions of mental disorders on the distribution and frequency of diagnoses in a typical clinical setting? (2) What are the interrater reliabilities using the case records method and how do they differ between both classification systems? (3) How stable are the reliability coefficients themselves? (4) Which diagnostic discrepancies are most common and which reasons account for such inconsistencies?

Method

The present study is part of a larger research program conducted to evaluate the practicability and reliability of a set of new diagnostic checklists. This instrument, the International Diagnostic Checklists (IDCL), is associated to the family of instruments provided by the WHO for diagnostics according to ICD-10. The checklists have first been introduced under the label "MDCL" (Munich Diagnostic Checklists; c.f. Hiller et al. 1990a, 1990b). The IDCL exist in separate versions for ICD-10 and DSM-III-R, and the primary purpose of the instrument is to enable clinicians to evaluate diagnostic criteria and decision rules during usual clinical examinations. A more detailed description of the IDCL is given in a separate article in this volume (see page 218).

Selection and Preparation of Case Records

We selected 100 written case records from the sample of the Munich Follow-Up Study (MFS; Wittchen et al. 1992). The pa-

tients of this study had been examined and treated as psychiatric inpatients in the former Psychiatric Department of the Max-Planck-Institute of Psychiatry in Munich (FRG) between 1973 and 1975. The original selection criteria 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. For the present study, we used the subsample of patients with the clinical diagnosis of an endogenous psychosis according to ICD-8. We further included only those patients who had participated in a six to eight year follow-up investigation and who additionally had received a clinical diagnosis of a specific disorder according to DSM-III (given independently by another diagnostician prior to our study). The clinical diagnoses from the MFS were of course unknown to the raters of our study.

The case records generally consisted of four to nine pages and were written in a traditional, standard format. They included the psychiatric symptomatology at admission and during hospitalization (mental status examination), familial and social development, former psychiatric disturbances and other medical diseases. However, other biographical data and any mention of a present or past psychiatric diagnosis were omitted before the case reports were distributed to the participating diagnosticians. We chose this procedure because diagnoses were to be determined directly from the psychopathological descriptions in the case records and should not be influenced by previous nosological considerations. The identities of the patients were disguised in the records by using code names.

Diagnostic Assessment

All case records were subsequently given to each of four independent clinicians who were asked to read the reports carefully and work out diagnoses according to ICD-10 as well as DSM-III-R. The diagnosticians were two psychiatrists and two clinical psychologists (each one male and one female) with clinical experience in a psychiatric hospital between two and three years. They had received their clinical training in different psychiatric facilities or different departments of the Max-Planck-Institute of Psychiatry. All patients described in the case reports were personally unknown to the four clinicians (time elapsed from first admission to our evaluation procedure had been 15 years or more).

Diagnostic assessment was done throughout with the help of the IDCL. All diagnosticians were sufficiently experienced with the use of this instrument since they had administered the checklists for a longer period of time in routine diagnostic examinations. They were also familiar with the contents and concepts of ICD-10 and DSM-III-R. The ICD-10 lists employed in this study referred to the 1990 draft of the diagnostic criteria for research (WHO 1990a). The raters were instructed not to simply confirm their diagnostic impressions, but to check diagnostic criteria carefully and to make a specific diagnosis only when all relevant criteria were judged to be fulfilled. Further, ICD-10 and DSM-III-R were to be considered as independent classification systems, including the logical possibility to diagnose two different disorders whenever criteria for the same disorders differ between the systems. 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.

Patient Characteristics

47 of the patients were male, 53 female. The mean age (at first admission) was 39.2 years (SD = 9.44 years) with a range from 25 to 62 years. Familial status was 47 single, 38 married, 14 divorced or separated, and 1 widowed. The educational level of the patients varied between primary school and high school. 33 patients had attended school for eight to nine years, 35 for ten years, 5 for 12 years, 26 for 13 years and 1 for less than eight years.

Statistical Analyses

We employed the κ statistic for the evaluation of chance-corrected agreement between raters. This measure has been proposed for the case of two independent raters (Cohen 1960) as well as for the comparison of multiple raters formulating diagnoses for the same group of subjects (Fleiss 1971). The maximum value of κ is 1, indicating perfect diagnostic concordance. A value of 0 or below results whenever the observed interrater agreement does not exceed the amount of agreement expected by chance alone. We performed a test of significance for each κ value (one-tailed at 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 of 0.70 or above indicates excellent congruence (Fleiss 1981).

Alternative measures of interrater agreement are the overall percentage of agreement and Yule's Y . For the case of dichotomous ratings by two diagnosticians, Spitznagel and Helzer (1985) have suggested to use Yule's Y instead of κ because Y is independent from varying base rates of diagnoses under consideration. However, it is still controversial whether a measure independent from base rates is appropriate for reliability studies or not (Shrout et al. 1987). We therefore report Y as an additional statistic. A Pseudo-Bayes estimation was applied for Y whenever a single cell of a fourfold classification table became 0 (otherwise, Y would have reached the endpoint value of 1 despite incomplete congruence; c.f. Bishop et al. 1975).

Results

We will first compare frequency rates of the individual ICD-10 and DSM-III-R diagnoses that were obtained in the sample. In a next step, diagnostic reliability will be analyzed and corresponding categories in both classification systems will be compared in detail by referring to their reliability measures as a critical yardstick. We will then evaluate the nature of disagreements for specific diagnoses.

Frequency of Diagnoses

An important characteristic of disorders is their relative frequency of occurrence (or base rate) in the population

or in specific clinical settings. For the present sample, the distribution of DSM-III-R and ICD-10 diagnoses is given in Table 1. Absolute values are shown for each of the four diagnosticians as well as for the total of 400 diagnoses which were made within each classification system.

Substantial differences between equivalent diagnoses in DSM-III-R and ICD-10 were found for schizophrenia, schizophreniform disorder (the corresponding ICD-10 category is labelled "acute and transient psychotic disorder") and schizoaffective disorder. Schizophrenia according to DSM-III-R was diagnosed in 30.25% of all diagnostic formulations of the four clinicians, but the same diagnosis according to ICD-10 was given in even 43.5%. This represents an increase of more than 40% (from 121 to 174 diagnoses). The reverse tendency was observed for schizophreniform/acute and transient psychotic disorder with a decrease of almost 70% from 50 diagnoses in DSM-III-R (12.5%) to only 16 diagnoses in ICD-10 (4.0%). We found that the main reason for these discrepancies was the different duration criterion for schizophrenia in both systems (six months in DSM-III-R, only one month in ICD-10). Our data show that a relatively large number of incongruently diagnosed cases may result from this difference, and a higher prevalence of schizophrenia must certainly be expected under various clinical and scientific conditions when ICD-10 is used. It does not seem to be justified to transfer research results from one system to the other, e.g. data obtained from a sample of schizophrenic patients according to ICD-10 should not simply be generalized to schizophrenics according to DSM-III-R.

Table 1 further shows an increase of the frequency rate for schizoaffective disorder of more than 70% from 25 diagnoses in DSM-III-R (6.25%) to 47 diagnoses in ICD-10 (11.75%). A review of individual case records revealed that this difference can mainly be explained by the more restrictive nature of diagnostic criteria within DSM-III-R. Patients presenting with both psychotic and

Table 1. Sample frequency rates of diagnoses for DSM-III-R and ICD-10

	Psychiatrist, female	Psychiatrist, male	Psychologist, female	Psychologist, male	No. of diagnoses	%
<i>DSM-III-R</i>						
Schizophrenia	30	36	26	29	121 / 400	30.25
Schizophreniform disorder	16	9	16	9	50 / 400	12.5
Schizoaffective disorder	2	4	9	10	25 / 400	6.25
Major depression	22	22	24	25	93 / 400	23.25
Bipolar disorder	9	8	9	11	37 / 400	9.25
Other disorders	21	21	16	16	74 / 400	18.5
<i>ICD-10</i>						
Schizophrenia	44	44	45	41	174 / 400	43.5
Acute/transient psychotic disorder	4	4	5	3	16 / 400	4.0
Schizoaffective disorder	7	6	19	15	47 / 400	11.75
Depressive disorder	21	22	21	24	88 / 400	22.0
Bipolar disorder	11	10	4	9	34 / 400	8.5
Other disorders	13	14	6	8	41 / 400	10.25

affective syndromes often failed to fulfil criterion B of schizoaffective disorder in DSM-III-R which requires at least one psychotic episode without prominent mood symptoms for at least two weeks. There is no comparable specification in the ICD-10 definition of schizoaffective disorder. ICD-10 merely requires the presence of psychotic and affective symptoms within the same episode of the disorder.

No dramatic differences were found for the frequency rates of depressive and bipolar disorders. They were diagnosed with rates from 22.0 to 23.25% (depression) and 8.5 to 9.25% (bipolar disorder). There were only small variations in the number of these diagnoses given by each of the four diagnosticians. Table 1 shows that the only exception is bipolar disorder in ICD-10 with only four diagnoses from the female psychologist and 9 to 11 diagnoses from the other three raters. In this case, the small number of diagnoses of bipolar disorder was mainly compensated for by an increased number of diagnoses of schizoaffective disorder. This indicates the somehow insufficient delineation between bipolar and schizoaffective disorder in ICD-10 which may be a source of diagnostic inconsistencies.

Among the class of "other disorders" were ten diagnoses of delusional (paranoid) disorder and three of brief reactive psychosis in DSM-III-R, and two diagnoses of delusional disorder, seven of schizotypal disorder and 13 of simple schizophrenia in ICD-10. The ICD-10 category of schizotypal disorder exists in a similar form as a personality disorder on axis II in the DSM-III-R system, but this axis was not evaluated in our study. There is no equivalent diagnosis in DSM-III-R for the category of simple schizophrenia in ICD-10.

An interesting difference between both classification systems was found for the frequency of residual diagnoses. These categories are labelled "not otherwise specified" in DSM-III-R and "other" or "unspecified" in ICD-10. For example, if a patient presents with a psychotic syndrome but criteria for none of the specific psychotic disorders (like schizophrenia or schizoaffective disorder) are found to be fulfilled, the residual category for psychotic disorder has to be chosen. The high rate of 18.5% for "other disorders" in DSM-III-R (Table 1), as compared to only 10.25% in ICD-10, is due to the generally more narrow diagnostic specifications in the DSM-III-R system. As a consequence, diagnoses of specific disorders had to be rejected more often than in ICD-10. Residual affective disorders were diagnosed in our sample in 3.75% according to DSM-III-R and in only 1.75% according to ICD-10, and the difference was even larger for residual psychotic disorders (10.0% in DSM-III-R vs. 2.0% in ICD-10).

Diagnostic Reliability

In a second step, we analyzed the interrater reliabilities for individual diagnoses in both classification systems. Fig. 1 gives κ values and overall percentage agreements for the main disorders and for the general diagnostic agreement within ICD-10 and DSM-III-R. The κ values in Fig. 1 reflect the amount of agreement between all four diagnosticians. They were computed according to a

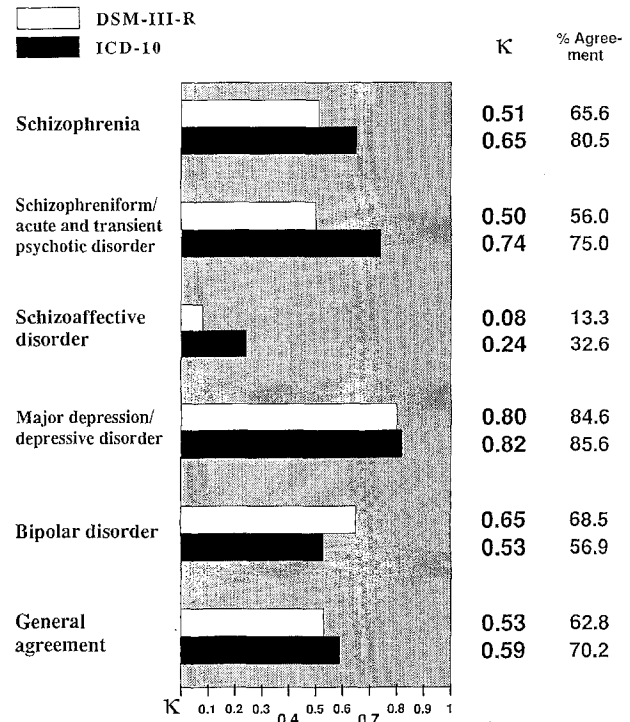


Fig. 1. Reliability of diagnoses according to DSM-III-R and ICD-10 (κ = kappa)

procedure described by Fleiss (1971) in which diagnoses of several raters for a set of cases are considered simultaneously. The κ values are also illustrated graphically by means of bar charts, showing more clearly the different levels of interpretation (κ above 0.40 as acceptable, κ above 0.70 as excellent; c.f. Fleiss 1981).

It can be seen that κ values for all categories except for bipolar disorder were superior in ICD-10 as compared to DSM-III-R. The overall diagnostic agreement was 70.2% with $\kappa = 0.59$ for ICD-10 and 62.8% with $\kappa = 0.53$ for DSM-III-R. Excellent agreement with κ above 0.70 was reached for the ICD-10 diagnoses of depressive and acute/transient psychotic disorder, and for the DSM-III-R diagnosis of major depression. Values of above 0.60 were obtained for schizophrenia in ICD-10 and for bipolar disorder in DSM-III-R. All κ values in Fig. 1 except the ones for schizoaffective disorder were statistically significant at the 5% level.

Differences between the classification systems were most prominent for schizophrenia ($\kappa = 0.65$ in ICD-10 vs. 0.51 in DSM-III-R) and schizophreniform/acute and transient psychotic disorder (0.74 vs. 0.50). A detailed analysis of individual case records showed that the time specification of DSM-III-R for schizophrenia (minimum duration of six months) was again a main source of disagreement. It remained unclear in a considerable number of cases whether this critical duration threshold was exceeded or not. Clinical judgement was often complicated by less sharply defined prodromal or residual symptoms. When the DSM-III-R categories of schizophrenia and schizophreniform disorder were analyzed as one diagnostic class, we found an agreement of 77.6% with $\kappa = 0.61$. Nevertheless, this agreement is smaller

Table 2. Pairwise reliabilities of DSM-III-R diagnoses with kappa (κ) and Yule's Y ; insignificant κ values (5% level) are indicated by an asterix

	A vs. B	A vs. C	A vs. D	B vs. C	B vs. D	C vs. D	Mean κ
Schizophrenia	59 5 $\kappa = 0.64$ 11 25 $Y = 0.68$	63 11 $\kappa = 0.55$ 7 19 $Y = 0.60$	58 13 $\kappa = 0.40$ 12 17 $Y = 0.43$	58 16 $\kappa = 0.49$ 6 20 $Y = 0.55$	57 14 $\kappa = 0.52$ 7 22 $Y = 0.56$	61 10 $\kappa = 0.42$ 13 16 $Y = 0.47$	0.51
Schizophreniform disorder	84 7 $\kappa = 0.68$ 0 9 $Y = 0.82$	78 6 $\kappa = 0.55$ 6 10 $Y = 0.65$	82 9 $\kappa = 0.50$ 2 7 $Y = 0.70$	79 5 $\kappa = 0.23^*$ 12 4 $Y = 0.39$	88 3 $\kappa = 0.63$ 3 6 $Y = 0.77$	81 10 $\kappa = 0.41$ 3 6 $Y = 0.60$	0.50
Schizoaffective disorder	94 2 $\kappa = -0.03^*$ 4 0 $Y = 0.00$	91 0 $\kappa = 0.34^*$ 7 2 $Y = 0.54$	89 1 $\kappa = 0.14^*$ 9 1 $Y = 0.52$	87 4 $\kappa = -0.06^*$ 9 0 $Y = -0.01$	86 4 $\kappa = -0.06^*$ 10 0 $Y = -0.01$	83 7 $\kappa = 0.13^*$ 8 2 $Y = 0.27$	0.08
Major depression	74 4 $\kappa = 0.77$ 4 18 $Y = 0.80$	74 2 $\kappa = 0.83$ 4 20 $Y = 0.86$	74 1 $\kappa = 0.86$ 4 21 $Y = 0.90$	71 5 $\kappa = 0.66$ 7 17 $Y = 0.71$	73 2 $\kappa = 0.81$ 5 20 $Y = 0.85$	73 2 $\kappa = 0.87$ 3 22 $Y = 0.89$	0.80
Bipolar disorder	88 4 $\kappa = 0.55$ 3 5 $Y = 0.72$	89 2 $\kappa = 0.76$ 2 7 $Y = 0.85$	88 1 $\kappa = 0.78$ 3 8 $Y = 0.88$	87 4 $\kappa = 0.42$ 5 4 $Y = 0.61$	88 1 $\kappa = 0.71$ 4 7 $Y = 0.85$	87 2 $\kappa = 0.67$ 4 7 $Y = 0.79$	0.65
<div style="display: flex; align-items: flex-start;"> <div style="margin-right: 20px;"> A = Psychiatrist, female B = Psychiatrist, male C = Psychologist, female D = Psychologist, male </div> <div style="margin-right: 20px;"> A vs. B 59 5 11 25 </div> <div> <div style="display: flex; align-items: center;"> <div style="font-size: 2em; margin-right: 10px;">↗</div> <div> interpretation of fourfold table: 59 agreements about <i>absence</i> of diagnosis; 25 agreements about <i>presence</i> of diagnosis; 5 disagreements with A rating presence and B rating absence of diagnosis; 11 disagreements with B rating presence and A rating absence of diagnosis. </div> </div> </div> </div>							

than those found for the separate ICD-10 categories of schizophrenia and acute and transient psychotic disorder. Our results thus indicate that it seems to be easier and more objective to distinguish schizophrenia from shorter and less severe psychotic episodes by applying a cut-off criterion of one month as in ICD-10 (instead of six months as in DSM-III-R).

Highest levels of agreement with κ of 0.80 and above were obtained in both systems for depressive disorders. We also had an acceptable reliability for the diagnosis of bipolar disorder in DSM-III-R ($\kappa = 0.65$), but the corresponding value for ICD-10 was somewhat smaller ($\kappa = 0.53$). It should be noted that the diagnosis of a specific affective disorder (e.g., major depression) and the diagnosis of a residual category for affective disorders (e.g., depressive disorder not otherwise specified) were counted as *discrepancies* in the results presented in Fig. 1. However, we often observed only minor differences in the diagnosticians' ratings of clinical syndromes. For example, if one rater found five depressive symptoms during a depressive episode and his colleague assessed only four symptoms, a diagnostic disagreement of major depression vs. depressive disorder not otherwise specified could result (since a minimum of five symptoms is required for major depression). We therefore extended our analyses by treating specific and residual categories for affective disorders as congruent. Under this condition, the following values were derived: major depression/depressive disorder (DSM-III-R): $\kappa = 0.86$, 89.6% agreement; bipolar disorder (DSM-III-R): $\kappa = 0.66$, 69.6%; depressive disorder (ICD-10): $\kappa = 0.85$, 88.2%; bipolar disorder (ICD-10): $\kappa = 0.55$, 59.3%.


The reliability of schizoaffective disorder was higher in ICD-10 than in DSM-III-R ($\kappa = 0.24$ vs. 0.08), but both values were unacceptable. This indicates that the assessment of schizoaffective disorders is highly problematic and unsatisfactory regardless of the specific definitions in both classification systems.

Pairwise Comparisons of Reliability

Interrater studies with more than two diagnosticians have the advantage that a set of several independent reliability measures can be determined for the same sample. In the present study, each of the four diagnosticians made his diagnostic decisions independently from each of his colleagues. It is thus possible to assess reliability by contrasting agreement and disagreement for each of the six pairs of diagnosticians (A/B, A/C, A/D, B/C, B/D, C/D). The six resulting reliability measures reflect some of the interrater variability which must be expected when judgements from clinicians with different backgrounds of training, experience and diagnostic habits are compared.

Table 2 presents the pairwise reliabilities for the five principal DSM-III-R disorders of our study. The exact binary data distribution (diagnosis given vs. not given) is shown for each pair of diagnosticians as well as the resulting κ and Y values. For example, the clinicians A (male psychiatrist) and B (female psychiatrist) had agreed that 25 patients had schizophrenia, but they disagreed for the remaining 16 patients. This distribution yields of κ of 0.64, which means that the agreement of A and B is

Table 3. Pairwise reliabilities of ICD-10 diagnoses with kappa (κ) and Yule's Y ; insignificant κ values (5% level) are indicated by an asterix

	A vs. B	A vs. C	A vs. D	B vs. C	B vs. D	C vs. D	Mean κ
Schizophrenia	$\frac{50}{6} \mid \frac{6}{38} \kappa = 0.76$ $Y = 0.76$	$\frac{49}{7} \mid \frac{6}{38} \kappa = 0.74$ $Y = 0.74$	$\frac{49}{7} \mid \frac{10}{34} \kappa = 0.65$ $Y = 0.66$	$\frac{47}{9} \mid \frac{8}{36} \kappa = 0.66$ $Y = 0.66$	$\frac{45}{11} \mid \frac{14}{30} \kappa = 0.49$ $Y = 0.50$	$\frac{48}{7} \mid \frac{11}{34} \kappa = 0.63$ $Y = 0.64$	0.65
Acute and transient psychotic disorder	$\frac{96}{0} \mid \frac{0}{4} \kappa = 1.00$ $Y = 1.00$	$\frac{95}{1} \mid \frac{0}{4} \kappa = 0.88$ $Y = 0.90$	$\frac{95}{1} \mid \frac{2}{2} \kappa = 0.56$ $Y = 0.81$	$\frac{95}{1} \mid \frac{0}{4} \kappa = 0.88$ $Y = 0.90$	$\frac{95}{1} \mid \frac{2}{2} \kappa = 0.56$ $Y = 0.81$	$\frac{94}{1} \mid \frac{3}{2} \kappa = 0.48$ $Y = 0.78$	0.73
Schizoaffective disorder	$\frac{89}{4} \mid \frac{5}{2} \kappa = 0.26^*$ $Y = 0.50$	$\frac{79}{14} \mid \frac{2}{5} \kappa = 0.31$ $Y = 0.58$	$\frac{82}{11} \mid \frac{3}{4} \kappa = 0.30$ $Y = 0.52$	$\frac{78}{16} \mid \frac{3}{3} \kappa = 0.16^*$ $Y = 0.38$	$\frac{80}{14} \mid \frac{5}{1} \kappa = 0.01^*$ $Y = 0.03$	$\frac{74}{7} \mid \frac{11}{8} \kappa = 0.36$ $Y = 0.47$	0.23
Depressive disorder	$\frac{76}{3} \mid \frac{2}{19} \kappa = 0.85$ $Y = 0.88$	$\frac{77}{2} \mid \frac{2}{19} \kappa = 0.88$ $Y = 0.90$	$\frac{75}{4} \mid \frac{1}{20} \kappa = 0.86$ $Y = 0.90$	$\frac{75}{3} \mid \frac{4}{18} \kappa = 0.79$ $Y = 0.83$	$\frac{72}{6} \mid \frac{4}{18} \kappa = 0.72$ $Y = 0.76$	$\frac{74}{5} \mid \frac{2}{19} \kappa = 0.80$ $Y = 0.84$	0.82
Bipolar disorder	$\frac{86}{3} \mid \frac{4}{7} \kappa = 0.63$ $Y = 0.75$	$\frac{89}{0} \mid \frac{7}{4} \kappa = 0.50$ $Y = 0.70$	$\frac{87}{2} \mid \frac{4}{7} \kappa = 0.67$ $Y = 0.79$	$\frac{89}{1} \mid \frac{7}{3} \kappa = 0.39$ $Y = 0.72$	$\frac{86}{4} \mid \frac{5}{5} \kappa = 0.48$ $Y = 0.65$	$\frac{90}{6} \mid \frac{1}{3} \kappa = 0.43$ $Y = 0.74$	0.52
A = Psychiatrist, female B = Psychiatrist, male C = Psychologist, female D = Psychologist, male							
A vs. B $\frac{49}{7} \mid \frac{6}{38}$ <div style="display: inline-block; vertical-align: middle; margin-left: 10px;">  <p>interpretation of fourfold table: 49 agreements about <i>absence</i> of diagnosis; 38 agreements about <i>presence</i> of diagnosis; 6 disagreements with A rating presence and B rating absence of diagnosis; 7 disagreements with B rating presence and A rating absence of diagnosis.</p> </div>							

64% higher than expected by chance alone. All six κ values for each of the specific disorders were summarized in Table 2 by their mean values which are identical to the corresponding overall κ values in Fig. 1.

It can be seen that there is considerable variability of diagnostic reliability for all disorders due to the individual pairwise comparisons. For example, the lowest reliability for schizophrenia was $\kappa = 0.40$ between A and D and the highest was $\kappa = 0.64$ between A and B. κ was 0.50 or above only for three of the six reliabilities for schizophrenia. For schizophreniform disorder, four of the six κ values were 0.50 or above, but only 0.23 was reached when comparing B and C. Excellent values up to 0.87 were computed for major depression and up to 0.78 for bipolar disorder. In contrast, four of the six values for schizoaffective disorder were below 0 and the best agreement was not higher than 0.34.

Table 3 gives the equivalent data for ICD-10 diagnoses. The mean κ values are again practically equal to the global κ given in Fig. 1 (minimal deviations are presumably due to rounding errors). All pairwise reliabilities for depressive disorder in ICD-10 were above $\kappa = 0.70$ and all but one of the values for schizophrenia were above 0.60. For acute and transient psychotic disorder, only one κ was below 0.50 and even complete congruence was reached in one case (A/B). Again, we found an unsatisfactory situation for schizoaffective disorder since the maximum value obtained was only 0.36.

Analysis of Diagnostic Disagreements

We analyzed in more detail in which way diagnosticians disagreed on specific diagnoses. Fig. 2 gives a graphical

illustration of these results for schizophrenia in both classification systems. We determined how often each single schizophrenia diagnosis (made by any diagnostician) was confirmed by another rater. Fig. 2 shows that this was the case in 66% within DSM-III-R and in 80% within ICD-10. The remaining 34%, respectively 20%, were split up into the specific categories which had been chosen by the second diagnostician when he was in disagreement with the schizophrenia diagnosis of the first diagnostician.

Most discrepancies for schizophrenia in DSM-III-R were related to schizophreniform disorder (10%), schizoaffective disorder (7%) and the residual diagnosis of psychotic disorder not otherwise specified (7%). These three categories accounted for 24% of disagreements whenever one diagnostician diagnosed schizophrenia. That means, every fourth diagnosis of schizophrenia was unreliably delineated from shorter psychotic episodes (schizophreniform disorder), mixed psychotic and affective syndromes (schizoaffective disorder) and atypical psychotic states (residual psychotic disorder). The remaining 10% of discrepancies were shared by delusional (paranoid), affective and other disorders.

For schizophrenia in ICD-10, the main source of disagreement was the alternative diagnosis of schizoaffective disorder with a rate of 10% (which made up half of all discrepant diagnostic formulations for schizophrenia in this system). Fig. 2 shows that the boundaries between schizophrenia and acute and transient psychotic disorder (2% discrepancies) and the residual category of "other psychotic disorder" (2%) seem clearer in comparison with the corresponding categories of DSM-III-R. Discrepancies between schizophrenia in ICD-10 and oth-

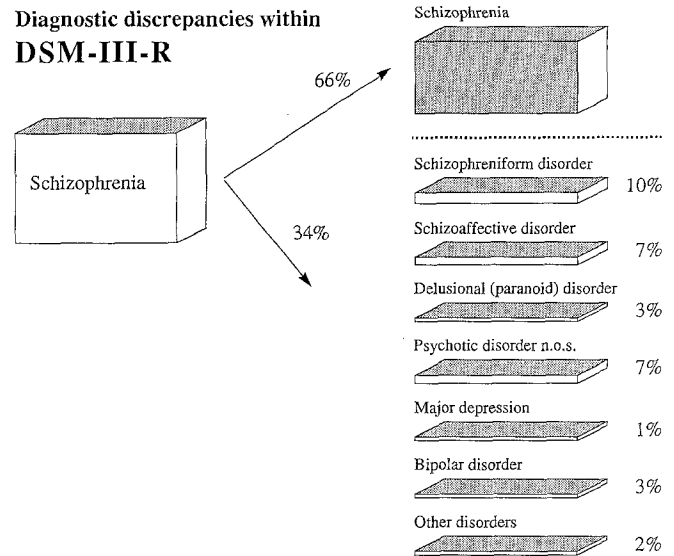
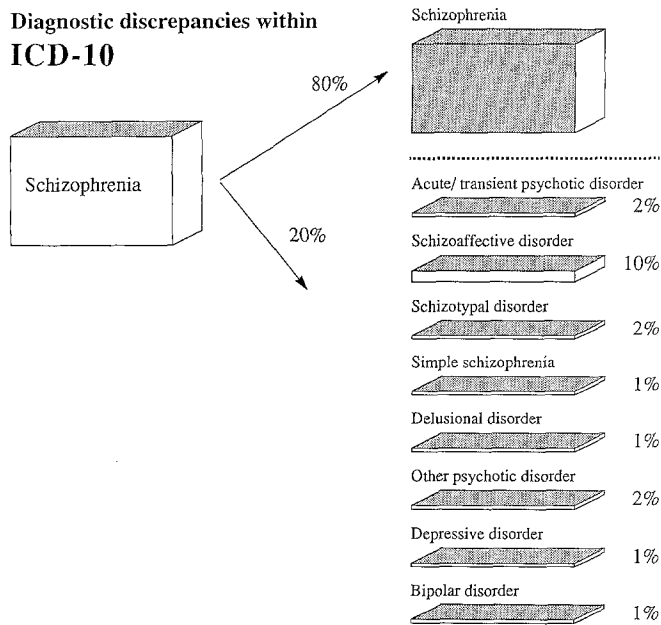


Fig. 2. Diagnostic discrepancies for schizophrenia in DSM-III-R and ICD-10

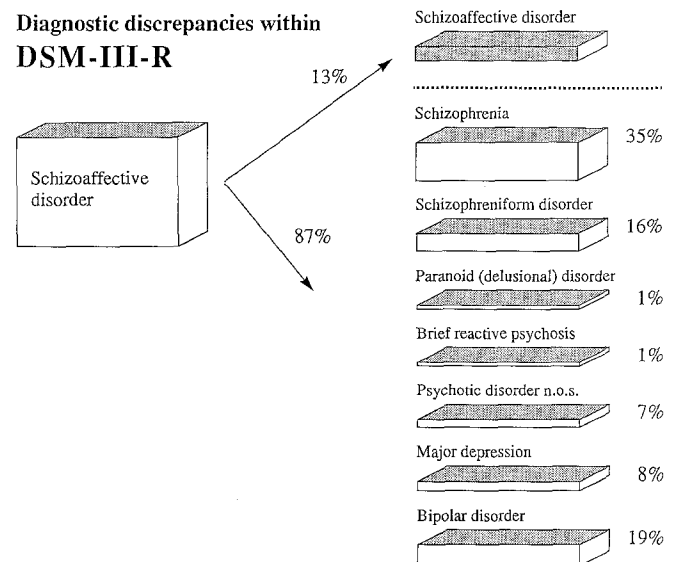
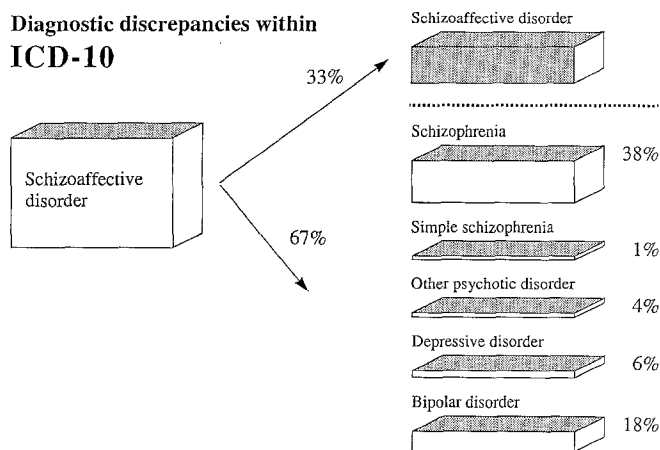


Fig. 3. Diagnostic discrepancies for schizoaffective disorder in DSM-III-R and ICD-10

er categories were negligible with rates of 2% or below.

The insufficient situation for schizoaffective disorder in both classification systems is illustrated in Fig. 3. The diagnosis of schizoaffective disorder of any of the four raters was confirmed in only 13% in DSM-III-R and 33% in ICD-10. Most discrepancies were due to the alternative diagnoses of schizophrenia (in about one of three cases) and bipolar disorder (in about one of five cases). It becomes apparent from Fig. 3 that the reliability of schizoaffective disorder in DSM-III-R is lower than in ICD-10 mainly because of the additionally unclear distinction between schizoaffective and schizophreniform disorder. No discrepant cases between schizoaffective and acute/transient psychotic disorder in ICD-10 were observed in our sample.

Discussion

Psychiatric classification systems play a central role for the communication about clinical features, etiology, course and treatment of mental disorders. They are expected to define diagnoses and provide rules which delineate different disorders from each other. It is necessary to revise such classifications from time to time in order to reflect the changing knowledge about diagnostic characteristics and the nosological status of single disorders.

In the study described here, we investigated similarities and differences between the currently most important classification systems in psychiatry, ICD-10 and DSM-III-R. Both systems are very similar in their general structure, but differences exist concerning details in

the description and content of specific disorders. The practical relevance of such incongruencies was evaluated empirically by comparing the distributions and reliabilities of corresponding diagnoses in both systems.

We found that the frequencies for depressive and bipolar disorders did not differ substantially between both systems, but the rate for schizophrenia in ICD-10 was more than 40% higher than in DSM-III-R. In contrast, shorter and less severe schizophrenia-like disorders were diagnosed more often in DSM-III-R. This demonstrates that the different duration criteria for schizophrenia of six months in DSM-III-R and one month in ICD-10 tend to have considerable consequences for typical clinical assessments (as in a psychiatric hospital) and that the comparability of schizophrenia diagnoses in both systems is clearly limited.

We further evaluated the usefulness of main diagnoses, as offered in both classification systems, by comparing their interrater reliabilities. All ICD-10 categories except bipolar disorder were found to have higher reliability values with an overall mean κ of 0.59 for ICD-10 diagnoses and 0.53 for DSM-III-R diagnoses. Reliability was generally excellent for affective disorders (only bipolar disorder in ICD-10 had a κ below 0.60) and moderate for schizophrenia. However, we frequently failed to reach diagnostic agreement for schizoaffective disorder in both systems, resulting in unacceptably low κ values of 0.24 (ICD-10) and 0.08 (DSM-III-R). A detailed analysis of the nature of discrepancies associated with schizoaffective disorder revealed that no sufficient boundaries existed between schizoaffective disorder on one side and schizophrenia and bipolar disorder on the other side. Similar evidence has been reported by Sprock (1988).

The present study represents one of the first attempts to compare ICD-10 and DSM-III-R by applying criteria from both systems to the same sample of patients. We used the case records method which is one of the standard procedures to estimate reliability (Helzer et al. 1977; Grove et al. 1981). This approach has advantages and disadvantages as compared to live explorations. For example, the use of written material guarantees that diagnostic decisions are based on the same information for each of the participating diagnosticians and for each set of diagnostic criteria applied. It is easier to involve a larger group of raters in order to estimate the consistency of reliability values by a number of independent pairwise comparisons. On the other hand, it is not secured that all relevant information is included in case summaries and live interviews may be superior in providing a more complete clinical picture of the patient's present and past symptomatology.

A direct comparison between reliabilities from case summaries and live interviews was reported by Hyler et al. (1982). Both methods were applied to the same group of 46 psychiatric patients (from which case records existed). Reliability was higher when based on the interview method for all diagnostic classes except for affective disorders. However, some of the live interviews analyzed by Hyler et al. (1982) had been done jointly and some in separate test-retest trials. It is known that

reliability tends to be overestimated when the different ratings stem from the same interview situation (c.f. Grove et al. 1981), and the high interview reliabilities obtained by Hyler et al. (1982) may therefore have resulted from this specific method.

We have reasons to assume that the reliabilities found in the present study are well comparable to the results from usual test-retest studies with two diagnosticians. Burnam et al. (1983), using Spanish and English versions of the Diagnostic Interview Schedule (DIS), found κ values of 0.48 and 0.66 for schizophrenia and of 0.49 and 0.61 for major depression. Semler et al. (1987) employed a German version of the Composite International Diagnostic Interview (CIDI) and obtained reliabilities of $\kappa = 0.60$ for schizophrenia, 0.54 for schizophreniform disorder, 0.66 for major depression and 0.47 for bipolar disorder. These values are in the same range as those found by us, for example 0.40 to 0.64 for schizophrenia, 0.66 to 0.87 for major depression and 0.42 to 0.78 for bipolar disorder.

General conclusions about the adequacy and usefulness of ICD-10 and DSM-III-R should be drawn from the data presented here only with caution. Diagnostic reliability is only *one* important property of classification systems, but it does in no way guarantee the validity of specific diagnostic categories. Other important aspects are the clearness and precision of psychopathological descriptions and the logic of decision rules for diagnoses which exclude each other. We, the authors, personally believe that DSM-III-R is superior in some of these formal aspects and its didactic value, and we generally agree with Mombour et al. (1990) who have described some considerable shortcomings of ICD-10 in detail.

What are the possible reasons for the lower reliabilities of DSM-III-R? – Of importance is perhaps that a very precise and specific definition of a disorder (as often given by DSM-III-R) may be a source of *disagreement* between different diagnosticians rather than a help to reach congruence. An example is the DSM-III-R criterion of a two week psychotic period without affective symptoms, which delineates schizoaffective from affective disorders more clearly than does ICD-10. When applied, however, disagreement about this specific criterion often caused disagreement about the entire diagnosis.

Similar conclusions can be derived from a comparison of our results with those from ICD-10 field trials in German-speaking countries, conducted in 1987 and 1988 with 134 clinicians in ten psychiatric centers (Dilling et al. 1990). Diagnoses were *not* based on explicit criteria, but a draft of the ICD-10 clinical descriptions and diagnostic guidelines (WHO 1990b) had been employed for all diagnostic decisions. Nevertheless, most κ values obtained in this study were acceptable or high such as 0.68 for schizophrenia, 0.82 for acute and transient psychotic disorder, 0.54 for schizoaffective disorder, 0.49 for recurrent depressive disorder and 0.78 for bipolar disorder (Albus et al. 1990; Zaudig et al. 1990; Freyberger et al. 1990). This underlines our hypothesis that less strictly assessed diagnoses do not necessarily imply a decrease of reliability. However, further investigations contrasting

ICD-10 and DSM-III-R are needed to clarify the relationship between important criteria (and their degree of specificity) and diagnostic reliability.

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

References

- Albus M, Strauß A, Stieglitz R-D (1990) Schizophrenia, schizotypal and delusional disorders (Section F2): results of the ICD-10 field trial. *Pharmacopsychiat* 23 [Suppl IV]:155-159
- American Psychiatric Association, APA (1987) Diagnostic and statistical manual of mental disorders, 3rd edn (revised). APA, Washington
- Bishop YMM, Feinberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. Cambridge MA: MIT Press
- Burnam NA, Karno M, Hough RL, Escobar JI, Forsythe AB (1983) The Spanish Diagnostic Interview Schedule. Reliability and comparison with clinical diagnoses. *Arch Gen Psychiatry* 40:1189-1196
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educational Psychol Meas* 20:37-46
- Cooper JE (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 HJ, Malchow P (1990) Design of the ICD-10 field trial in German-speaking countries. *Pharmacopsychiat* 23 [Suppl IV]:142-145
- Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychol Bull* 76:378-382
- Fleiss JL (1981) Statistical methods for rates and proportions, 2nd ed. New York, Wiley
- Freyberger HJ, Albus M, Stieglitz R-D (1990) ICD-10 field trial in German-speaking countries - summary of the quantitative empirical results. *Pharmacopsychiat* 23 [Suppl IV]:192-196
- Grove WM, Andreasen NC, McDonald-Scott P, Keller MB, Shapiro RW (1981) Reliability studies of psychiatric diagnosis. Theory and practice. *Arch Gen Psychiatry* 38:408-413
- Helzer JE, Robins LN, Taibleson M, Woodruff RA, Reich T, Wish ED (1977) Reliability of psychiatric diagnosis, I. A methodological review. *Arch Gen Psychiatry* 34:129-133
- Hiller W, Zaudig M, Mombour W (1990a) Development of diagnostic checklists for use in routine clinical care. *Arch Gen Psychiatry* 47:782-784
- Hiller W, Bose M von, Dichtl G, Agerer D (1990b) Reliability of checklist-guided diagnoses for DSM-III-R affective and anxiety disorders. *J Affect Disord* 20:235-247
- Hyler SE, Williams JBW, Spitzer RL (1982) Reliability in the DSM-III field trials. Interview v case summary. *Arch Gen Psychiatry* 39:1275-1278
- Maier W, Philipp M, Zaudig M (1990) Comparison of the ICD-10 classification system with the ICD-9- and the DSM-III-R classification of mental disorders. *Pharmacopsychiat* 23 [Suppl IV]:183-187
- Mombour W, Spitzner S, Reger KH, Cranach M von, Dilling H, Helmchen H (1990) Summary of the qualitative criticisms made during the ICD-10 field trial and remarks on the German translation of ICD-10. *Pharmacopsychiat* 23 [Suppl IV]:197-201
- Sartorius N (1988) International perspectives of psychiatric classification. *Br J Psychiat* 152 [Suppl 1]:9-14
- Semler G, Wittchen H-U, Joschke K, Zaudig M, Geiso T von, Kaiser S, Cranach M von, Pfister H (1987) Test-retest reliability of a standardized psychiatric interview (DIS/CIDI). *Eur Arch Psychiatr Neurol Sci* 236:214-222
- Shrout PE, Spitzer RL, Fleiss JL (1987) Quantification of agreement in psychiatric diagnosis revisited. *Arch Gen Psychiatry* 44:172-177
- Spitznagel EL, Helzer JE (1985) A proposed solution to the base rate problem in the kappa statistic. *Arch Gen Psychiatry* 42:725-728
- Sprock J (1988) Classification of schizoaffective disorder. *Compr Psychiat* 29:55-71
- Wittchen H-U, Essau CA, Zerssen D von, Krieg J-C, 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
- WHO, World Health Organization (1990a) ICD-10, Chapter V (F), Mental and behavioural disorders (including disorders of psychological development), diagnostic criteria for research (1990 draft). Geneva, WHO
- WHO, World Health Organization (1990b) ICD-10, Chapter V (F), Mental and behavioural disorders (including disorders of psychological development), clinical descriptions and diagnostic guidelines (1990 draft). Geneva, WHO; German ed: Dilling H, Mombour W, Schmidt MH (eds) (1991) Weltgesundheitsorganisation: Internationale Klassifikation psychischer Störungen, ICD-10, Kapitel V (F), Klinisch-diagnostische Leitlinien. Bern, Huber
- Zaudig M, Stieglitz R-D, Gastpar M, Rösinger C (1990) Mood (affective) and schizoaffective disorders (Section F3): results of the ICD-10 field trial. *Pharmacopsychiat* 23 [Suppl IV]:160-164