Arch. Psychiat. Nervenkr. 227, 3—15 (1979)

Archiv für Psychiatrie und Nervenkrankheiten
Archives of Psychiatry and Neurological Sciences

Interrater-Reliabilität bei Diagnosen, AMP-Syndromen und AMP-Symptomen

U. Baumann[1] und B. Woggon[2]

[1] Institut für Psychologie der Christian-Albrechts-Universität, Kiel, Bundesrepublik Deutschland
[2] Psychiatrische Universitätsklinik Zürich, Forschungsdirektion (Direktor: Prof. Dr. J. Angst), Zürich, Schweiz

Interrater Reliability of Diagnosis, AMP Syndromes and AMP Symptoms

Summary. Psychic disorders can be classified into three levels: symptom, syndrome, and diagnosis. For each of these levels of reference, interrater reliability has been calculated. For this purpose, 48 patients (25 depressives and 23 schizophrenics) were interviewed by two raters each and the diagnoses were registered on the AMP system (forms 3 and 4). Additionally, each rather made an ICD diagnosis. With ICD numbers of three digits interrater reliability amounted to Kappa = 0.84; with ICD numbers of four digits it amounted to Kappa = 0.65. According to the degree of accuracy, numerical agreement with the AMP syndromes lies between Kappa = 0.61 and 0.85. Single symptoms had the lowest reliability (median: Kappa 0.45 and 0.53). Reasons for these differences and possibilities for improvement are discussed.

Key words: Diagnosis – AMP system – Interrater reliability.

Zusammenfassung. Psychische Störungen lassen sich auf drei Ebenen klassifizieren: Symptom, Syndrom, Diagnose. Für jede dieser Bezugsebenen wurde in dieser Studie die Beurteilerübereinstimmung berechnet. Dazu wurden 48 Patienten (25 depressive and 23 schizophrene Patienten) von jeweils 2 Untersuchern interviewt und die Befunde auf dem AMP-System (Bogen 3 und 4) festgehalten. Zusätzlich gaben die Untersucher eine ICD-Diagnose. Die Interrater-Reliabilität betrug bei dreistelligen ICD-Ziffern Kappa = 0.84, bei vierstelligen ICD-Ziffern Kappa = 0.65. Je nach Genauigkeitsgrad liegt die numerische Übereinstimmung bei den AMP-Syndromen zwischen Kappa = 0.61 und 0.85. Von geringster Reliabilität waren die einzelnen Symptome (Median: Kappa = 0.45 resp. 0.53). Es werden Gründe für die Unterschiede und Verbesserungsmöglichkeiten diskutiert.

Schlüsselwörter: Diagnose – AMP-System – Interrater-Reliabilität.

Sonderdruckanforderungen an: Prof. Dr. U. Baumann, Psychologisches Institut der Universität Kiel, Olshausenstraße 40—60, D-2300 Kiel, Bundesrepublik Deutschland

0003-9373/79/0227/0003/$ 02.60

1. Einleitung

Nach Mombour (1977) werden zur Systematik psychischer Störungen drei Ebenen unterschieden: Symptomatik, Syndromatik und Nosologie. Heimann (1977) postuliert ein hierarchisches Klassifizierungssystem mit den Ebenen Diagnose, Syndrom, Symptom, Psychophysiologie und Biochemie. Während sich die Symptomatik und Syndromatik meist auf den psychopathologischen Querschnittsbefund bezieht, kommen bei der Nosologie resp. Diagnose zusätzliche Gesichtspunkte hinzu, wie z. B. Verlauf und Ätiologie. Symptome und Syndrome stehen in hierarchischem Bezug, indem Syndrome als Symptomkomplexe zu sehen sind (von Zerssen, 1973). So definiert auch Mombour (1977, S. 120) Syndrom als „das überzufällig häufige, gemeinsame Auftreten einer Reihe von Einzelsymptomen". Während der empirische Syndrombegriff dem dimensionalen Ansatz der Persönlichkeitspsychologie Cattells und Eysencks entspricht, stellen Nosologien kategoriale Klassifikationssysteme dar. In der Persönlichkeitspsychologie ist der kategoriale Ansatz weitgehend durch den dimensionalen ersetzt worden (vgl. Eysenck und Eysenck, 1969), in der Psychiatrie dagegen finden wir beide Betrachtungsweisen nebeneinander (Kendell, 1975; Frank, 1975). Unabhängig von den inhaltlichen Argumenten, die mehr für die eine oder die andere Betrachtungsweise sprechen, ist bei allen Klassifikationssystemen das Kriterium der formalen Genauigkeit — die Reliabilität — von Wichtigkeit. Von den klassischen Reliabilitätsarten (vgl. Lienert, 1969) Retest-, Paralleltest- und Halbierungsmethode (verallgemeinert in der Konsistenz) kommen für die Syndromerfassung im Rahmen der Selbstbeurteilung (z. B. von Zerssen, 1976) alle angeführten Reliabilitätsarten in Frage. Basiert die Syndromerfassung auf Fremdbeurteilung (vgl. Pichot, 1974; CIPS, 1977), so kommt zum Patienten als weitere Variationsquelle der Beurteiler (Rater) hinzu, so daß unterschieden werden muß zwischen der formalen Genauigkeit innerhalb *eines* Raters und zwischen *verschiedenen* Ratern. Für letztere Fragestellung, die meist im Vordergrund steht, wird die *Interrater-Reliabilität* (Beurteiler-Übereinstimmung) berechnet. Auch für nosologische Klassifikationssysteme stellt die Interrater-Reliabilität das Hauptkriterium dar, an der die Güte des Systems gemessen wird. Daneben liegen auch vereinzelte Reteststudien vor (z. B. Kendell, 1974; Duckworth und Kedward, 1978). Als weitere Reliabilitätsart für kategoriale Systeme wird auch der Vergleich von Häufigkeitsverteilungen von Diagnosen benutzt; bekannt ist vor allem der Vergleich USA/England (Kendell, 1975; Leff, 1977).

Die Ergebnisse der Reliabilitätsstudien für nosologische Systeme sind in verschiedenen Überblicksarbeiten dargestellt (Kreitman, 1961; Zubin, 1967; Spitzer und Fleiss, 1974; Zubin et al., 1975). In einer Re-Analyse älterer Studien kommen Spitzer und Fleiss (1974) zum Schluß, daß die Reliabilität zufriedenstellend sei bei den Diagnosegruppen Organische Psychosen, Alkoholismus und Schwachsinn; mittlere Reliabilitäten resultieren bei der Diagnose Schizophrenie, während sie bei den übrigen Diagnosen nicht ausreichend hoch ist. Die Ergebnisse sind mit denen von Zubin (1967) nicht direkt vergleichbar, da Spitzer und Fleiss in ihrer Arbeit den Kappakoeffizienten verwenden, ältere Arbeiten aber das methodisch problematische Übereinstimmungsprozent, bei dem die Auftretenswahrscheinlichkeiten der Diagnosen nicht miteingehen.

In den Überblicksarbeiten wird meistens stillschweigend die Vergleichbarkeit der Studien angenommen; damit postuliert man Übereinstimmung in den relevanten Komponenten der Versuchsplanung. Daß sich aber unter dem Titel „Reliabilität von Diagnosen" verschiedenste Designs verbergen, läßt sich leicht nachweisen (vgl. auch Helzer et al., 1977). So können die Interviews auf Videoband aufgenommen werden, wobei die Übereinstimmung der Rater bei der Beurteilung des vorgespielten Videobandes berechnet wird. Eine andere Variante besteht in dem Interview durch Rater A und einem weiteren Interview durch Rater B oder in dem Interview durch Rater A unter Dabeisein von B usw. Je nach Versuchsplan stehen den Beurteilern verschiedene Informationsmengen zur Verfügung. Ein weiteres Problem stellt die Vorgabe unterschiedlicher Diagnose-Spektren dar. So ist bekannt, daß die Übereinstimmung bezüglich dreistelliger ICD-Ziffern (ICD, 1975) höher ist als bei vierstelligen.

Studien können sich auch in der Interview-Gestaltung unterscheiden: Länge des Interviews, Interviewer-Stil, Strukturiertheit des Interviews usw. Diese wenigen Beispiele zeigen bereits deutlich, daß die Vergleichbarkeit von Studien nicht ohne weiteres gegeben ist und damit durchschnittliche Übereinstimmungswerte problematisch sind.

Gründe für die mangelnde Reliabilität von Diagnosen werden in der bekannten Arbeit von Ward et al. (1962; vgl. auch Wittenborn, 1972) diskutiert: Inkonstanz beim Patienten und Interviewer, Inadäquatheit der Nosologie. Die formale Güte sucht man durch Standardisierungen zu verbessern: durch halb- oder vollstrukturierte Interviews soll der Einfluß des Untersuchers reduziert werden, durch formalisierte Informationsverarbeitung wird die klinische Urteilsbildung ersetzt (Bsp. System DIAGNO, CATEGO). Daß damit die Reliabilität verbessert werden kann, wird z. B. durch die internationale Studie zum Diagnosevergleich gezeigt (Kendell, 1975), aber auch durch die von Zubin et al. (1975) berichteten Kappa-Werte.

Im Gegensatz zu den Reliabilitätswerten von Diagnosen sind die — ebenfalls durch Fremdbeurteilung — im Rahmen von Ratings gewonnenen Syndromwerte meistens in ihrer formalen Güte befriedigend und liegen vielfach in der Größenordnung von 0.8—0.9 (vgl. CIPS-Manual, 1977); vereinzelt werden aber auch tiefere Werte bis $r = 0.5$ berichtet.

2. Fragestellung

Stellt man die drei eingangs erwähnten Klassifikationsebenen (Symptom, Syndrom, Diagnose) einander gegenüber, so interessiert u. a. auch der Genauigkeitsunterschied, der mit diesen Bezugsebenen verbunden ist. Kreitman et al. (1961) haben in einer vielschichtigen Studie die Beurteilerübereinstimmung bezüglich verschiedener Merkmalsbereiche verglichen (Dauer der Krankheit, Symptome, Diagnosen usw.), wobei die Daten von *einer* Stichprobe stammten. Die Übereinstimmung bezüglich der Symptome bezog sich aber nicht auf das Vorhandensein resp. Fehlen, sondern auf den Stellenwert des Symptoms bezüglich der zu stellenden Diagnose.

Untersuchungen in ähnlicher Art für die drei Klassifikationsebenen liegen aber bisher nicht vor, so daß meistens die Kennwerte, die man an *unterschiedlichen* Stichproben gewonnen hat, miteinander in Bezug gesetzt werden. In unserer Arbeit sollen aber an *derselben* Stichprobe die entsprechenden statistischen Werte gewonnen werden, womit ein exakter Vergleich möglich wird.

Nach der Testtheorie (Lienert, 1969) sollten die aus Symptomen zusammengesetzten Syndromskalen in ihrer Genauigkeit höher sein als die einzelnen Symptome. Die Diagnosen wiederum sind erfahrungsgemäß infolge ihrer höheren Komplexität unreliabler als Syndrome.

3. Versuchsplan

Zur durchgeführten Studie sind detaillierte Angaben in Woggon et al. (1978) dargestellt. Einige wichtige Punkte seien hier nochmals angeführt:

Es wurden 48 Patienten (25 depressive und 23 schizophrene Zustandsbilder) von jeweils 2 Beurteilern (1 Interviewer, 1 Nebeninterviewer) untersucht (mittlere Interviewdauer $M = 45$ min).

Den psychopathologischen Befund hielten die Rater mit dem AMP-System (Angst et al., 1969; Scharfetter, 1972) Bogen 3 (psychischer Befund) und Bogen 4 (somatischer Befund, dabei nur 1. Spalte) fest. Zusätzlich hatten sie auch Angaben zur Diagnose, Explorierbarkeit des Patienten, Anschaulichkeit der Symptomatik und Sicherheit der Diagnosen zu machen; dazu lag eine sechsstufige Skala vor (6 = ausgezeichnet, 1 = ungenügend). Die 4 Projektbeurteiler wurden untereinander in Zweiergruppen aufgeteilt unter Berücksichtigung sämtlicher Kombinationen (4 Rater: 6 Kombinationen). Jeweils 1 Rater führte das Interview durch, der zweite war als „Nebeninterviewer" dabei und konnte Zusatzfragen stellen.

Für die Festlegung der Diagnosen wurde den Beurteilern eine Liste mit 13 Diagnosen vorgelegt. Es handelt sich um 6 Unterdiagnosen von ICD 295 (Schizophrenie: 295.0 Schizophrenia simplex, 295.1 hebephrene Form, 295.2 katatone Form, 295.3 paranoide Form, 295.6 schizophrene Rest- und Defektzustände, 295.7 schizoaffektive Psychosen), 5 Unterdiagnosen von ICD 296 (affektive Psychosen: 296.0 Involutionsdepression, 296.1 Manie, 296.2 endogene Depression, 296.3 zirkuläre Verlaufsform, 296.8 andere affektive Psychosen), ICD 298.0 (reaktive depressive Psychose) und ICD 300.4 (depressive Neurose).

Das Spektrum war also auf die Diagnosen Schizophrenie und Depression eingeengt. Dies war dadurch bedingt, daß das in unserer Studie verwendete AMP-System für den Bereich der endogenen und der körperlich begründbaren Psychosen konzipiert worden ist und seine Anwendung in Psychopharmakastudien vorwiegend bei schizophrenen und depressiven Patienten erfolgt. Die Überprüfung der Beurteilerübereinstimmung führt bei den dreistelligen ICD-Ziffern infolge des eingeengten Diagnosespektrums sicher zu oberen Schätzungen der Reliabilitätskoeffizienten, da eine Zunahme der möglichen Kategorien die Höhe der Koeffizienten beeinflußt.

4. Statistische Auswertung

Zur statistischen Auswertung wurde in Woggon et al. (1978) ausführlich Stellung bezogen, insbesondere zur Verwendung von Kappa (Cohen, 1960; Fleiss, 1971). Bei den nosologischen Diagnosen stellt sich — im Gegensatz zu den Symptomen — das Problem der extremen Randverteilungen nicht; dagegen haben wir bei der Analyse der vierstelligen ICD-Diagnosen im Verhältnis zur Raterzahl sehr viele Kategorien (theoretisch 13, benützt 12), so daß entsprechende Kennwerte mit Vorsicht zu interpretieren sind (Kappa nach Fleiss, 1971).

Für die Syndromauswertung ist nach Tinsley und Weiss (1975) bei Intervallskalenniveau zwischen „Agreement" und „Reliability" zu unterscheiden. Im ersten Fall geht es um die numerische Übereinstimmung zwischen 2 Ratern, im zweiten Fall um die gleiche Reihenfolge.

Beispiel: Einstufung von 5 Patienten durch 3 Rater in Skalen mit Werten zwischen 1 und 11.

	Patient				
	1	2	3	4	5
Rater 1	5	3	8	1	9
Rater 2	5	2	9	3	8
Rater 3	7	5	10	3	11

Rater 1/2 stimmen numerisch besser überein als Rater 1/3, wenn man die Summe der absoluten Differenzen als Maß nimmt (1/2: Differenzsumme = 5; 1/3: Differenzsumme = 10). Geht man aber von der Rangfolge der Werte aus, so ist diese bei den Ratern 1/3 identisch, bei 1/2 nicht.

Will man kategoriale Daten (Symptome, Diagnosen) in ihrer formalen Güte mit Intervalldaten (Syndromskalen) vergleichen, so muß man einen Koeffizienten haben, der in beiden Skalenebenen anwendbar ist. Bei dichotomen Daten und gleicher Randverteilung bei beiden Ratern ist Kappa mit dem Phi-Koeffizienten (und damit Produkt-Moment-Korrelation) identisch (Cohen, 1960; Bartko, Carpenter, 1975), in den übrigen Fällen ist Kappa < Phi. Ein Vergleich unter Berücksichtigung der Übereinstimmung in der Rangfolge ist daher bei unseren Daten (Diagnosen: mehrere Kategorien) nicht möglich. Nimmt man aber die numerische Übereinstimmung zum Kriterium, so läßt sich Kappa auch für Intervalldaten formulieren (Tinsley, Weiss, 1975):

$$\text{Kappa} = (N_b - N_e) \,/\, (N - N_e)$$

N_b Zahl der Patienten, bei denen beide Rater numerisch übereinstimmen
N_e Zahl der zufällig zu erwartenden Übereinstimmungen
N Gesamtzahl der untersuchten Patienten

Zwei Probleme sind dabei zu klären: numerische Übereinstimmung und zufällig zu erwartende Übereinstimmung. Von Übereinstimmung ist sicher nicht nur dann zu sprechen, wenn zwischen beiden Ratern keine Differenz besteht ($D = 0$), sondern auch bei kleineren Divergenzen. Als Grenze für die zu zählenden Divergenzen schlagen wir daher den bei Testprofilen üblichen kritischen Wert vor, der beim intraindividuellen Testvergleich Anwendung findet (Lienert, 1969):

$$D_{\text{crit}\,a} = (X_1 - X_2) = z_{\text{crit}\,a} \cdot s_X \cdot \sqrt{2(1 - r_{tt})}$$

Indem wir für Alpha ein 5%-Niveau setzen, ist die zulässige Divergenz, die noch als Übereinstimmung zählt, kleiner als beim höheren Signifikanzniveau. Ähnliches gilt für die Reliabilitätsschätzung r_{tt}: je geringer der Wert, um so größer ist die tolerierte kritische Differenz. Eine Differenz zwischen 2 Ratern, die diesen kritischen Wert übersteigt, zählt als Nichtübereinstimmung, andernfalls als Übereinstimmung (N_e). Setzt man in die Gleichung die bei den AMP-Syndromen (vgl. Abschnitt 7) vorliegende Standardabweichung $s_X = 10$ und $z = 1.96$ ein, so ergeben sich folgende kritische Werte:

$D_{\text{crit}} = 8.8$ ($r_{tt} = 0.9$); $D_{\text{crit}} = 12.4$ ($r_{tt} = 0.8$); $D_{\text{crit}} = 15.2$ ($r_{tt} = 0.7$).

Die zufällig zu erwartende Übereinstimmung schätzen wir, indem wir von der Differenz zweier unabhängiger Zufallsvariablen mit $s = 10$ ausgehen (AMP-Syndrome). Die Differenz ist verteilt nach $M_D = 0$ und $s_D = 14.1$. Die durch $\pm D_{\text{crit}}$ eingeschlossene Fläche (normiert durch $z = (D_{\text{crit}} - M_D) \,/\, s_D$) gibt den zufällig zu erwartenden Prozentsatz an, resp. die Anzahl (bezogen auf $N = 48$):

$D_{\text{crit}} = 8.8$ $P = 0.465$ $N_e = 22.3$
$D_{\text{crit}} = 12.4$ $P = 0.621$ $N_e = 29.8$
$D_{\text{crit}} = 15.2$ $P = 0.717$ $N_e = 34.4$

Wir werden bei unseren Berechnungen alle 3 Schätzungen für r_{tt} und die damit verbundenen N_e benutzen, da wir keine exakten Werte für r_{tt} haben.

5. Beschreibung der Beurteiler und Patienten

Die 4 Rater hatten mehrmonatige Erfahrung mit dem AMP-System, in dessen Verwendung sie mittels Interrater-Training geschult worden sind.

Bei den Patienten handelte es sich um 25 depressive und 23 schizophrene Patienten, die bezüglich Lebensalter (17—68 Jahre) und Krankheitsdauer (0—26 Jahre) heterogen waren. Weitere Angaben zur Stichprobe in Woggon et al. (1978).

6. Ergebnisse der nosologischen Diagnosen

6.1. Übereinstimmungen

Bei den 48 Patienten stimmen die *beiden Rater* bezüglich der dreistelligen ICD-Diagnosen in 44 Fällen überein (23mal ICD 295-Schizophrenie; 20mal ICD 296-affektive Psychosen, 1mal ICD 300-Neurose), was einem Kappa = 0.84 mit $P < 0.01$ oder einem Übereinstimmungsprozent von 92% entspricht. In drei Fällen haben wir Diskrepanzen zwischen ICD 295/296, in einem Fall zwischen ICD 296/300. Fassen wir ICD 296 und 300 zusammen, so erhöht sich Kappa auf 0.88, resp. der Prozentsatz auf 94%.

Bei den vierstelligen ICD-Ziffern ist das Übereinstimmungsprozent 71% (34/48). Da wir sehr viele Kategorien (theoretisch 13) im Vergleich zur Raterzahl (pro Patient 2) haben, scheint uns die Berechnung von Kappa problematisch zu sein, es soll daher nur als deskriptives Maß angeführt werden: Kappa = 0.65 (berechnet nach Fleiss, 1971). In Tabelle 1 finden sich die detaillierten Ergebnisse.

Die gefundenen Divergenzen bezüglich der Übereinstimmung vierstelliger ICD-Diagnosen spiegeln die auch im klinischen Alltag immer wieder auftretenden Schwierigkeiten bei der Diagnosestellung von Untergruppen-Diagnosen wider. Gerade die Abgrenzung hebephrener Zustandsbilder von paranoiden Schizophrenien (295.1/295.3) ist häufig auf Grund der Gewichtung vorhandener paranoider Denkinhalte recht subjektiv. Bei der Abgrenzung monopolarer endogener Depressionen (296.2) von bipolaren Formen (296.3) liegt die Schwierigkeit auf einem anderen Gebiet, nämlich der Bewertung anamnestischer Angaben zum Auftreten manischer Phasen in der Vorgeschichte. Da die Diagnosen von den Ratern auf Grund der Exploration des Patienten, d.h. also ohne Kenntnis fremdanamnestischer Angaben, gestellt wurden, konnten nur die Angaben des Patienten selbst für die Untergruppen-Diagnose verwendet werden. Diese wurden offenbar von den Ratern nicht übereinstimmend interpretiert, was die Not-

Tabelle 1. Übereinstimmungen bei vierstelligen ICD-Ziffern (Diagnosenamen s. Abschnitt 3)

Übereinstimmung		Divergenz	
Diagnose	Anzahl	Diagnosepaar	Anzahl
295.1	5	295.1/.3	4
.2	1	295.0/.3	1
.3	9	295.1/.7	1
.6	1	296.0/.3	1
.7	1	296.2/.3	3
296.0	6	295.1/296.1	1
.2	7	295.7/296.0	1
.3	3	295.7/296.3	1
300.4	1	296.8/300.4	1
Total	34	Total	14

wendigkeit fremdanamnestischer Angaben zur Stellung nosologischer Diagnosen unterstreicht.

Neben der Übereinstimmung zwischen den Ratern ist auch die Übereinstimmung zwischen den Diagnosen der *Rater* und der *Krankengeschichten-Diagnose* von Interesse. Die Krankengeschichten-Diagnose wird vom klinischen Oberarzt im Rahmen einer gemeinsamen Untersuchung des Patienten zusammen mit dem behandelnden Assistenzarzt unter Berücksichtigung möglichst detaillierter Angaben zum gesamten Krankheitsverlauf (verfügbare Akten, Krankengeschichten, Befragung von Bezugspersonen) gestellt. Die drei Diagnosesteller (2 Rater, 1 Krankengeschichten-Diagnose) stimmen in 36 von 48 (75%) der Fälle überein, wenn sie die dreistelligen ICD-Ziffern benützen. Damit gelangen wir zu einem Kappa = 0.71 mit $P < 0.01$. Engen wir das Spektrum ein auf ICD 295 versus ICD 296/300 (Zusammenfassung depressiver Zustandsbilder), so steigt Kappa auf 0.79 mit einem Prozentsatz an Übereinstimmung von 90%. Die 12 Divergenzen beziehen sich 3mal auf Unterschiede zwischen ICD 298/296, 4mal auf ICD 300/296, 5mal auf ICD 296/295 (davon 3mal 296./295.7).

Vergleicht man die drei Diagnosesteller bezüglich der vierstelligen ICD-Ziffern, so ist das Übereinstimmungsprozent 42% (20/48) und Kappa = 0.56 (deskriptives Maß). Von den Diskrepanzen spielt sich der größte Teil innerhalb der Hauptgruppe Schizophrenie und affektive Psychosen ab, zum Teil innerhalb des depressiven Bereichs:

Innerhalb ICD 295. (Schizophrenie):	10
Innerhalb ICD 296. (affektive Psychosen):	6
Innerhalb ICD 296./300. (depressives Zustandsbild):	4
Innerhalb ICD 296./298. (depressives Zustandsbild):	3
	23

Die 5 restlichen Diskrepanzen lauten:

295.7 (KG) — 296.3 (beide Rater)
296.2 (KG) — 295.3 (beide Rater)
296.3 (KG) — 296.3 (Rater A) 295.7 (Rater B)
296.2 (KG) — 295.7 (Rater A) 296.0 (Rater B)
296.3 (KG) — 296.1 (Rater A) 295.1 (Rater B)

Die Ursachen für die Diskrepanzen liegen zum einen an den unterschiedlichen Informationsquellen, die für die Diagnosefindung zur Verfügung standen (s. oben). Eine weitere Ursache ist die spontane Fluktuation psychopathologischer Symptome, die man oft innerhalb kurzer Zeiträume auch ohne Einfluß verschiedener Behandlungsverfahren beobachten kann. Da Krankengeschichten-Diagnose und Projekt-Diagnose nicht am gleichen Tag gestellt wurde, können Diskrepanzen dadurch erklärt werden. Eine Zusammenfassung der Ergebnisse finden wir in Tabelle 4 (S. 13).

Erwartungsgemäß nimmt die Übereinstimmung mit zunehmender Auffächerung der Klassifikationseinheiten ab (Kappa 0.84 zu 0.65 und 0.71 zu 0.56), ebenso aber auch mit Zunahme der Raterzahl (0.84 zu 0.71 und 0.65 zu 0.56).

Ein Vergleich unserer Ergebnisse mit der Literatur ist — wie bereits in der Einleitung begründet — wegen der Verschiedenheit der Versuchspläne schwierig. Dazu kommt, daß die Arbeiten von Spitzer und Fleiss (1974) und Zubin et al. (1975) meist für jede Diagnoseeinheit ein Kappa rechnen (Diagnose X versus

Rest), was keine globale Beurteilung des Klassifikationssystems zuläßt. Bei Spitzer und Fleiss (1974) werden für „Schizophrenia" Kappa = 0.57 und „affective disorder" Kappa = 0.41 als Durchschnittswert über 6 Studien angegeben. In einer neueren Untersuchung mit standardisierten Kriterien kommen Spitzer und Mitarbeiter (nach Zubin et al., 1975) zu Kappa = 0.70 („affective") und Kappa = 0.84 („schizophrenia"). Mittelt man die Kappa der vorgegebenen Untergruppen des schizophrenen und depressiven Bereichs, so gelangt man zu einem Kappa von 0.58 (Tabelle 2 in Zubin et al., 1975, S. 647: Gruppen „nonaffective schizophrenia" bis und mit „minor depressive illness"). Die Koeffizienten liegen also in der gleichen Größenordnung wie unsere Werte.

6.2. Nosologische Diagnosen und Charakteristika des Interviews

Vergleicht man die 23 übereinstimmend als schizophren diagnostizierten Patienten mit den 21 übereinstimmend als depressiv erklärten Patienten (ICD dreistellig), so findet man keinen Unterschied bezüglich der Sicherheit, mit der die Diagnose gestellt wurde, und bezüglich der Anschaulichkeit der Symptomatik. Dagegen gelten die depressiven Patienten eher als gut explorierbar im Gegensatz zu den schizophrenen Patienten, was der klinischen Erfahrung entspricht.

Die Wertverteilung der Einstufung der Interviewer (mittlerer Wert über beide Rater) (1—6) wurde in vier gleiche Bereiche unterteilt (Erweiterung des Mediantests) und über die Tafel ein Chiquadrat gerechnet. Bei der Güte der Explorierbarkeit ist die aufgeführte Chiquadrattafel auf dem 10%-Niveau signifikant, wir haben also einen leichten Trend ($\chi^2 = 7.4$, df = 3, $P < 0.1$).

Güte der Explorierbarkeit

	<4	=4	4.5	≧5
ICD 295	9	2	7	5
ICD 296/300	2	6	5	8

Daß die drei Merkmale (Güte der Explorierbarkeit, Anschaulichkeit der Diagnose und Sicherheit der Diagnosestellung) zusammenhängen, wird durch die Rangkorrelation nach Spearman belegt:

Güte der Explorierbarkeit/Anschaulichkeit der Symptomatik: $r = 0.62$

Güte der Explorierbarkeit/Sicherheit bei Diagnose: $r = 0.62$

Sicherheit bei Diagnose/Anschaulichkeit der Symptomatik: $r = 0.70$

Anschaulichkeit der Symptomatik beeinflußt vermutlich direkt die Sicherheit bei der Diagnose, gleichzeitig sind zwischen Anschaulichkeit und Explorierbarkeit Wechselbeziehungen anzunehmen, die wiederum die Sicherheit bei der Diagnose erhöhen.

Die Konvergenz bzw. Divergenz zwischen den beiden Ratern in den vierstelligen ICD-Diagnosen ist aber unabhängig von den drei das Interview charakterisierenden Variablen. Divergenz kommt also nicht zustande, wenn die Rater den Patienten als schlecht explorierbar erleben oder wenn sie in der Diagnose unsicher sind.

7. Ergebnisse der AMP-Syndrome

7.1. Vergleich mit anderer Stichprobe

Für das AMP-System wurden von Baumann (1974) mittels Faktorenanalyse 9 AMP-Syndrome aufgestellt (vgl. auch Baumann und Angst, 1975).

Die 9 Syndrome lauten: 1 Apathisches S., 2 Halluzinatorisch-desintegratives S., 3 Hostilitäts-S., 4 Manisches S., 5 Somatisch-depressives S., 6 Paranoides S., 7 Katatones S., 8 Gehemmt-depressives S., 9 Hypochondrisches S. Die 9 Faktoren erklären 42.3% der Gesamtvarianz (PC-Analyse mit 1 in der Diagonale, Varimax-Rotation).

Für die Generalisierbarkeit der Ergebnisse ist von Interesse, wieweit die Werte (vgl. Tabelle 2) der vorliegenden Stichprobe vergleichbar sind mit den Daten aus Baumann (1974). Korreliert man jeweils die Profile der schizophrenen und der depressiven Patienten mit den entsprechenden Profilen aus Baumann (1974), so erhalten wir ein rho = 0.67 ($P<0.05$) bei der Gruppe Schizophrenie, resp. rho = 0.85 ($P<0.01$) bei der Gruppe Depression. Betrachtet man die Relation der beiden Profile untereinander, indem man die t-Werte (Unterschied zwischen schizophrener und depressiver Gruppe) aus unserem Material mit den t-Werten aus Baumann (1974) vergleicht, so korrelieren diese mit rho = 0.81 ($P<0.01$). Die vorgefundenen Profilähnlichkeiten deuten also darauf hin, daß die 48 Patienten ähnliche Charakteristika wie die 299 Patienten aus Baumann (1974) aufweisen, damit also die Ergebnisse zumindest für die in Zürich hospitalisierten Patienten generalisierbar sind.

Tabelle 2. AMP-Syndrome: Mittelwerte und Standardabweichungen der untersuchten Stichprobe

AMP-Syndrome	Gruppenmittelwerte der 9 AMP-Psychopathologie-Skalen für die 48 untersuchten Patienten		Gruppenmittelwerte der 9 AMP-Psychopathologie-Skalen für schizophrene Patienten ($N=23$)		Gruppenmittelwerte der 9 AMP-Psychopathologie-Skalen für depressive Patienten ($N=25$)		Unterschied
	M	s	M	s	M	s	
1. Apathisches S.	54,2	8,2	51,4	8,0	56,7	7,7	$P<0.05$
2. Halluzinatorisch-desintegratives S.	48,8	9,6	53,8	10,5	44,3	5,8	$P<0.01$
3. Hostilitäts-S.	51,8	7,3	52,7	7,3	51,0	7,3	n.s.
4. Manisches S.	51,4	8,7	53,4	8,6	49,6	8,7	n.s.
5. Somatisch-depressives S.	49,3	8,6	44,5	7,3	53,7	7,3	$P<0.001$
6. Paranoides S.	46,4	8,5	51,3	8,5	41,8	5,6	$P<0.001$
7. Katatones S.	53,7	6,8	53,7	7,5	53,7	6,2	n.s.
8. Gehemmt-depressives S.	52,7	8,2	47,2	6,1	57,8	6,5	$P<0.001$
9. Hypochondrisches S.	55,7	7,9	52,4	7,8	58,7	6,9	$P<0.01$

7.2. Übereinstimmung

Wie in Abschnitt 4 dargestellt, ist zu unterscheiden zwischen Übereinstimmung in der Rangfolge und numerischer Übereinstimmung. Für die Übereinstimmung in der Rangfolge haben wir jeweils die Rangkorrelation nach Spearman zwischen einem bestimmten Rater und den übrigen 3 Ratern gerechnet und den mittleren Wert über die 4 Rater als Kennwert herangezogen (vgl. auch Woggon et al., 1978; Abschnitt 4). Außer für Skala 5 haben wir Vergleichsdaten von Maurer-Groeli (1976); Skala 5 entfällt, da in der erwähnten Studie keine somatischen Symptome erhoben worden sind, diese aber in Skala 5 miteingehen (vgl. Tabelle 3).

Von den drei Skalen mit der geringsten Übereinstimmung (Skalen 3, 4, 5) haben in der vorliegenden Stichprobe die Skalen 3 und 4 wenig diagnostische Differenzierungsfähigkeit zwischen schizophrenen und depressiven Patienten (vgl. Tabelle 2); damit ist die Variabilität dieses Merkmals eingeengt, was die Beurteilerübereinstimmung beeinträchtigt. Die eher geringe Beurteilerübereinstimmung in Skala 5 (somatisch-depressives Syndrom) scheint uns schwierig erklärbar zu sein. Insgesamt ist die Beurteilerübereinstimmung bezüglich der Rangfolge mit einem durchschnittlichen Wert von 0.70 noch befriedigend, liegt aber doch unter den Werten der IMPS von Lorr und Klett (0.81—0.95, vgl. Pichot, 1974) oder der Hamilton-Skala mit 0.90 (CIPS, 1977). Ein Vergleich mit den Daten von Maurer-Groeli (1976) ist problematisch, weil es sich um ein anderes Design handelt (nur schizophrene Patienten; 2 Rater); die Ergebnisse hängen vom Rating ab (Erst- vs. Zweitrating) und weisen beträchtliche Unterschiede auf.

Die numerische Übereinstimmung (vgl. Abschnitt 4) wurde mittels Kappa bei Berücksichtigung verschiedener Schätzungen für r_{tt} (0.7—0.9) gerechnet. Je

Tabelle 3. Übereinstimmung bei AMP-Syndromen

AMP-Skala	Übereinstimmung in der Rangordnung				Numerische Übereinstimmung (Kappa)		
	Eigene Daten		Maurer-Groeli		$r_{tt}=0.9$	$r_{tt}=0.8$	$r_{tt}=0.7$
			Rating				
	rho	Range	1	2			
1. Apathisches S.	.78	(.71—.87)	.41	.72	.73	.89	.93
2. Halluzinatorisch-desintegratives S.	.68	(.34—.89)	.79	.84	.69	.73	.71
3. Hostilitäts-S.	.55	(.47—.65)	.73	.78	.49	.78	.85
4. Manisches S.	.64	(.56—.72)	.81	.62	.61	.84	.86
5. Somatisch-depressives S.	.64	(.50—.75)	—	—	.53	.56	.56
6. Paranoides S.	.72	(.65—.78)	.69	.79	.61	.89	1.00
7. Katatones S.	.77	(.48—.95)	.60	.82	.49	.67	.85
8. Gehemmt-depressives S.	.68	(.60—.86)	.67	.71	.65	.73	.78
9. Hypochondrisches S.	.87	(.77—.91)	.59	.53	.49	.78	.71

niedriger r_{tt} ist, um so größere Divergenzen werden als Übereinstimmung gezählt und um so höher fällt Kappa aus (Tabelle 3). Gehen wir von der strengsten Schätzung aus ($r_{tt}=0.9$), so bewegen sich die Kappa-Koeffizienten in der gleichen Größenordnung wie die durchschnittlichen Rangkorrelationen; eine Ausnahme bilden die Skalen 7 (Katatones Syndrom) und 9 (Hypochondrisches Syndrom), die ein eher geringes Kappa aufweisen. Da in der Literatur bisher für Ratingskalen nur Angaben zur Übereinstimmung der Rangreihe vorliegen, kann die Güte des Ergebnisses nicht im Vergleich zu anderen Skalen beurteilt werden. Absolut gesehen ist aber auch in dieser Berechnung die Übereinstimmung mittel (Kappa 0.4—0.6, vgl. Woggon et al., 1978). Gründe für die eher mäßigen Übereinstimmungskoeffizienten sind vor allem in der z. T. geringen Reliabilität der Symptome zu sehen.

8. Diskussion

Wenn wir die Ergebnisse der verschiedenen Klassifikationsebenen — Symptom, Syndrom, Diagnose — einander gegenüberstellen, so berücksichtigen wir bei den Syndromen nur die numerische Übereinstimmung (vgl. Abschnitt 4). Für die Symptome sind Detailangaben in Woggon et al. (1978) zu entnehmen. Tabelle 4 gibt eine Zusammenfassung der Ergebnisse.

Für die Interpretation sind primär die Kappa-Koeffizienten von Interesse. Symptome haben die geringste Beurteilerübereinstimmung. Wie in Woggon et al. (1978) diskutiert, können aber einzelne Symptome durchaus in ihrer Genauigkeit verbessert werden, wenn z. B. die Definitionen präzisiert werden. Durch Addition der Symptome kommt es zu einer Verbesserung der Übereinstimmung, wie aus

Tabelle 4. Beurteiler-Übereinstimmung

Klassifikationseinheit	Kappa	Übereinstimmungsprozent
Symptom		
Median der 110 berechenbaren Symptome	.45	90%
Median der 70 beurteilbaren Symptome	.53	80%
Syndrom		
Median der 9 Skalen: $r_{tt}=0.9$	.61	—
$r_{tt}=0.8$	.78	—
$r_{tt}=0.7$	.85	—
Diagnosen		
2 Beurteiler ICD-dreistellig	.84	92%
ICD-vierstellig	.65	71%
3 Beurteiler ICD-dreistellig	.71	75%
ICD-vierstellig	.56	42%

der Testtheorie bekannt ist (Verbesserung der Reliabilität durch Testverlängerung). Die Übereinstimmung in den Diagnosen ist besser, als man es auf Grund der Literatur erwarten würde. Durch die Begrenzung des Diagnosespektrums dürfte es sich aber um eine obere Schätzung handeln, insbesondere was die dreistellige ICD-Ziffer betrifft. Berücksichtigt man bei den Syndromen, daß sie z. T. aus unpräzis definierten Symptomen zusammengesetzt sind, und bei den Diagnosen, daß sie in dieser Studie auf klinisch-intuitivem Wege und nicht formalisiert zustande gekommen sind, so zeichnen sich für Syndrome und Diagnosen Verbesserungsmöglichkeiten ab, die Übereinstimmungen erreichen lassen, die die gleiche Größenordnung wie gutkonstruierte Tests haben. Bemühungen im Zusammenhang mit dem System DIAGNO (vgl. Zubin et al., 1975) belegen dies deutlich, ebenso einzelne klinische Ratingskalen, die präziser definierte Items umfassen (Bsp. iMPS von Lorr und Klett). Für das AMP-System (Klassifikationsebene Symptom und Syndrom) ist mit den im Gange befindlichen Überarbeitungen eine Verbesserung der Beurteilerübereinstimmung zu erwarten.

Literatur

Angst, J., Battegay, R., Bente, D., Berner, P., Broeren, W., Cornu, F., Dick, P., Engelmeier, M.-P., Heimann, H., Heinrich, K., Helmchen, H., Hippius, H., Pöldinger, W., Schmidlin, P., Schmitt, W., Weis, P.: Das Dokumentationssystem der Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie (AMP). Arzneim. Forsch. (Drug Res.) **19**, 339—405 (1969)

Bartko, J. J., Carpenter, W. T.: On the methods and theory of reliability. J. Nerv. Ment. Dis. **163**, 307—317 (1976)

Baumann, U.: Diagnostische Differenzierungsfähigkeit von Psychopathologie-Skalen. Arch. Psychiat. Nervenkr. **219**, 89—103 (1974)

Baumann, U., Angst, J.: Methodological development of the AMP-System. In: Neuropsychopharmacology, J. R. Boissier, H. Hippius, P. Pichot, eds. Proceedings of the IX. Congress of the CINP (Paris, 1974). Amsterdam: Excerpta Medica 1975

CIPS: Internationale Skalen für Psychiatrie. Berlin: CIPS 1977

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measurement **20**, 37—46 (1960)

Duckworth, G., Kedward, H.: Man or machine in psychiatric diagnosis. Am. J. Psychiat. **135**, 64—68 (1978)

Eysenck, H. J., Eysenck, S. B. G.: Personality structure and measurement. London: Routledge & Kegan 1969

Fleiss, J. L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378—382 (1971)

Frank, G.: Psychiatric diagnosis: a review of research. London: Pergamon 1975

Heimann, H.: Wirkung von Psychopharmaka und zugrundeliegende theoretische Vorstellungen. Pharmakopsychiatrie Neuro-Psychopharmakologie **10**, 119—129 (1977)

Helzer, J. E., Robins, L. N., Taibleson, M., Woodruff, R. A., Reich, T., Wish, E.: Reliability of psychiatric diagnosis. Arch. Gen. Psychiat. **34**, 129—133 (1977)

ICD: Diagnoseschlüssel und Glossar psychiatrischer Krankheiten. Berlin-Heidelberg-New York: Springer 1975

Kendell, R. E.: The stability of psychiatric diagnosis. Br. J. Psychiat. **124**, 352—356 (1974)

Kendell, R. E.: The role of diagnosis in psychiatry. London: Blackwell 1975

Kreitman, N.: The reliability of psychiatric diagnosis. J. Ment. Sci. **107**, 876—886 (1961)

Kreitman, N., Sainsbury, P., Morrisey, J., Towers, J., Scrivener, J.: The reliability of psychiatric assessment: an analysis. J. Ment. Sci. **107**, 887—908 (1961)

Leff, J.: International variations in the diagnosis of psychiatric illness. Br. J. Psychiat. **131**, 329—338 (1977)

Lienert, G. A.: Testaufbau und Testanalyse (3. Auflage). Weinheim: Beltz 1969

Maurer-Groeli, Y.: Untersuchungen zur Interraterreliabilität des AMP-Systems. Arch. Psychiat. Nervenkr. **221**, 321—329 (1976)

Mombour, W.: Systematik psychischer Störungen. In: Klinische Psychologie, L. J. Pongratz (Hrsg.). Handbuch der Psychologie, Bd. 8, 1. Halbband. Göttingen: Hogrefe 1977

Pichot, P. (Hrsg.): Psychological measurements in psychopharmacology. Basel: Karger 1974

Scharfetter, C.: Das AMP-System. Manual, 2. Auflage. Berlin-Heidelberg-New York: Springer 1972

Spitzer, R. L., Fleiss, J. L.: A re-analysis of the reliability of psychiatric diagnosis. Br. J. Psychiat. **125**, 341—347 (1974)

Tinsley, H., Weiss, D.: Interrater reliability and agreement of subjective judgments. J. Counseling Psychol. **22**, 358—376 (1975)

Ward, C. H., Beck, A. D., Mendelson, M., Mock, J. E., Erbaugh, J. K.: The psychiatric nomenclature. Arch. Gen. Psychiat. **7**, 198—205 (1962)

Wittenborn, J. R.: Reliability, validity and objectivity of symptom-rating scales. J. Nerv. Ment. Dis. **154**, 79—87 (1972)

Woggon, B., Baumann, U., Angst, J.: Interrater-Reliabilität von AMP-Symptomen. Arch. Psychiat. Nervenkr. **225**, 73—85 (1978)

Zerssen, D. v.: Syndrom. In: Lexikon der Psychiatrie, C. Müller (Hrsg.). Berlin-Heidelberg-New York: Springer 1973

Zerssen, D. v.: Klinische Selbstbeurteilungsskalen (KSb-S). Weinheim: Beltz 1976

Zubin, J.: Classification of the behavior disorders. Ann. Rev. Psychol. **18**, 373—406 (1967)

Zubin, J., Salzinger, K., Fleiss, J. L., Gurland, B., Spitzer, R. L., Endicott, J., Sutton, S.: Biometric approach to psychopathology. Ann. Rev. Psychol. **26**, 621—671 (1975)

Eingegangen am 26. Juli 1978