Eur Arch Psychiatry Clin Neurosci (1991) 241:165–169

European Archives of Psychiatry and Clinical Neuroscience

© Springer-Verlag 1991

The Reliability of the SADS-LA in a Family Study Setting

Marion Leboyer[1,2], Wolfgang Maier[3], Mardjane Teherani[2], Dirk Lichtermann[3], Thierry D'Amato[1], Petra Franke[3], Jean-Pierre Lépine[2], Jürgen Minges[3], and Peter McGuffin[4]

Workgroup of the European Science Foundation Network on "Molecular Biology of Mental Diseases"
[1]INSERM, Unité 155, Château de Longchamp, Paris 75016, France
[2]Service de Psychiatrie, Pr. Widlöcher, Hôpital Salpetrière, Paris 75013, France
[3]Department of Psychiatry, University of Mainz, W-6500 Mainz, Federal Republic of Germany
[4]Department of Psychological Medicine, University of Wales, Cardiff, England

Received February 13, 1991

Summary. The joint-rater and test-retest reliability study of two translated versions of the SADS-LA (Schedule for Affective Disorders and Schizophrenia – Lifetime version – modified for the study of anxiety disorders), one in French and the other in German, have been tested in family study settings, in a sample of patients and first-degree relatives. The test-retest reliability study demonstrated that identification of major affective disorders and schizophrenia was performed with sufficient reliability; however, diagnoses of subtypes of major disorders (e.g. bipolar II disorder) and identification of minor disorders was less reliable. The implications of these findings in phenotype identification during family studies in psychiatry are discussed.

Key words: Reliability – Diagnoses – Structured interviews – Family studies

Introduction

The identification of phenotypes is of crucial importance in the expanding field of linkage studies in psychiatric disorders. Diagnostic misclassification in linkage studies may change the lod score substantially, thus reducing the power of linkage analyses (Martinez et al. 1989). It is a well-known fact that cases identification and discrimination between different diagnoses in psychiatry is far from perfect. However, in the absence of a known pathophysiological mechanism, measuring misclassification in family studies is impossible and can only rely on reproducibility of diagnoses. It is generally accepted that standardized interviews enhance the reliability of psychiatric classifications both for the identification and diagnostic subgrouping of cases. Therefore, it is important to measure reliability of instruments used in family studies before starting any genetic project.

Offprint requests to: M. Leboyer

To date, most family studies in Europe and in the United States have used the Schedule for Affective Disorders and Schizophrenia – Lifetime Version (SADS-L; Endicott and Spitzer 1978) or a modification thereof. The SADS-L has also been preferred for linkage studies of affective disorders or schizophrenia (Merikangas et al. 1989, Weissman et al. 1986). Historically, the SADS was developed to ensure systematic collection of signs and symptoms for making Research Diagnostic Criteria (RDC) diagnoses (Spitzer et al. 1978). The lifetime version of the SADS was developed to record information concerning past as well as current episodes of illness. The SADS-L has recently been updated (Fyer et al. 1985, Manuzza et al. 1986) to a new version, now called SADS-LA. The SADS-LA allows RDC, DSM-III, and DSM-III-R diagnoses and includes more details for studying anxiety disorders.

In order to identify a sufficient number of large multiply affected families, a collaborative project in linkage studies of major psychiatric disorders was initiated by the European Science Foundation (ESF; "network on molecular neurobiology for mental illness"). The SADS-LA has been chosen by the scientific committee of this network as the basis for the identification of phenotypes. After an initial training of the investigators, a reliability study of the SADS-LA in a multicenter setting was required by the ESF committee before starting the process of recruiting informative families.

The aims of this study were thus

1. to make sure that diagnostic assessment can be handled with sufficient reproducibility in the scheduled collaborative linkage project,
2. to identify diagnostic categories with insufficient reliability,
3. to provide a basis for acceptable definition of cases (affected or non-affected),
4. to test the reliability of two translated versions of the SADS-LA, one in German and the other in French.

The results of this reliability study may also be useful for psychiatric geneticists working in other linkage study programs.

Methods

The reliability of the SADS-LA has been tested in a setting very similar to a realistic family study situation. Consequently, subjects were either patients or relatives. We thus undertook an inter-rater and test-retest reliability study at two different sites where clinicians were participating in the ESF project (Paris/France, Mainz/Germany) with two translated versions of the SADS-LA. No back-translation was performed, but appropriateness and accuracy were checked by the authors of the SADS-LA and their native function and between co-workers (Abby Fyer, personal communication).

Assessment

At both sites, one pair of raters performed interviews for the inter-rater reliability study in a balanced way, exchanging roles between interviewer and observer randomly. In addition, a third clinician re-interviewed patients for the test-retest study. Thus, in each center, three highly experienced clinicians carried out the interviews. They had been introduced to the SADS-LA by an officially authorized SADS-LA trainer.

In- and out-patients were systematically asked if they were willing to participate. Subjects were only included in the study if they agreed to being interviewed 1–5 months after the first interview. Subjects (patients or relatives) were interviewed first by a group of two interviewers, one who performed the interview and did the rating and the other who observed and rated (joint rater reliability). At the end of this *first* interview, the interviewers and the observer filled a diagnostic coding sheet, on which they could list all the diagnoses whose criteria were met by the patient during the interview. The raters also had to fill a discrepancy sheet, on which they could list all the questions answered by the observer and the interviewer in a different way. They were asked to list the discrepant questions and to describe the type of mistakes which explained best the discrepancy between the two raters. This procedure was originally used in the CIDI-WHO field trial (Wittchen et al. 1991).

A few months later (mean 3.2 months; range: 1.5–4.9) the proband was invited to be re-interviewed by a third interviewer, different from the first two raters and blind to the first. The third interviewer was also asked to fill out a diagnostic coding sheet. A best estimate diagnosis combining information obtained through both interviews and case notes was used to compare diagnoses obtained at the end of each interview. The reliability was assessed by comparing these diagnoses.

Samples

One center is an in- and out-patient clinic mainly for anxiety and affective disorders (Hôpital Bichat/Paris), the other a university department serving a medium-size city for the treatment of acute cases (Mainz/Germany). Consequently, mainly affective disorders were recruited in the French department ($n = 19$) and a sample of mainly psychotic patients was recruited in the German center ($n = 19$). Patients were recruited consecutively. At the German center, the sample of patients was derived from a family study in a consecutive sample of in- and out-patients. Consequently, at the

Table 1. Test-retest-reliability diagnoses of current and/or previous episodes (DSM-III/DSM-III-R) (no diagnostic hierarchy) by the SADS-LA in patients

	DSM-III			DSM-III-R		
	Relative frequency by either rater[a]	Yule[b]	Kappa mean[b]	Relative frequency by either rater[a]	Yule[b]	Kappa mean[b]
Schizophrenia/schizophreniform disorder	7/36	0.77	0.71	7/36	0.75	0.71
Schizoaffective disorder	./.	./.	./.	5/36	0.45	0.27
Psychotic disorders[c]	13/36	0.84	0.80	13/36	0.56	0.55
Affective disorders:						
Unipolar	21/36	0.49	0.49	19/36	0.51	0.46
Bipolar	11/36	0.70	0.64	10/36	0.49	0.49
W/psychotic features	11/36	0.96	0.92	7/36	0.83	0.78
W/o psychotic features	13/36	0.86	0.86	12/36	0.95	0.72
Major depression or bipolar disorder	23/36	0.90	0.86	22/36	0.84	0.75
Any affective disorder (incl. dysthymia/ cyclothymia)	25/36	0.60	0.60	24/36	0.53	0.50
Anxiety disorders:						
Panic disorder/agoraphobia	7/36	0.32	0.22	9/36	0.43	0.33
Generalized anxiety disorder	4/36	0.16	0.07	6/36	0.61	0.47
Phobic disorders (simple, social, phobia)	7/36	0.56	0.55	9/36	0.66	0.65
Any anxiety disorder (incl. obsessive compulsive disorder, phobic disorders)	10/36	0.35	0.35	13/36	0.49	0.49
Alcoholism/drug abuse	4/36	0.73	0.69	5/36	0.94	0.77

[a] Number of patients receiving a particular diagnosis by any of the two raters (++ or −+ or +−)

[b] Yule and Kappa coefficients were calculated by using the pseudobayesian approach, in order to match the zero cells in the contingency tables. Based on the pseudobayesian approach means and variances of Kappa coefficients were estimated by the jackknife procedure

[c] Psychotic disorders are defined as schizophrenia or schizoaffective disorders or "other psychotic disorders" or affective disorders with psychotic features (DSM-III/DSM-III-R)

Mainz center, additional relatives were able to be included in this reliability study.

Two cases dropped out in Paris (final sample size $n = 17$) for test-retest, one case dropped out in the Mainz patient sample (final sample size $n = 18$ for test-retest) and two cases dropped out in the Mainz relatives' sample (final sample size $n = 20$ for test-retest).

The sample of patients consisted of 22 females and 13 male patients and the sample of relatives consisted of 12 female and 8 male subjects; mean age was 42.6 years in the patient sample and 36.9 years in the relatives' sample.

Analysis

Diagnoses may refer to a particular episode or to a lifetime (across all the previous episodes). Diagnostic data have been considered hierarchically (using the diagnosis highest in a pre-established diagnostic hierarchy only) and non-hierarchically (counting primary as well as secondary). Among the functional psychiatric disorders, the diagnostic hierarchy places maximal weight on schizophrenia followed by schizoaffective disorders, major affective and minor affective disorders, anxiety disorders and, subsequently, alcoholism and drug abuse.

Here, the evaluation was based on:

a) non-hierarchical diagnoses of any previous or current episode and
b) hierarchical lifetime diagnoses.

The Kappa and the Yule coefficients were both used for the assessment of agreement; both coefficients had been corrected for random agreement. The Kappa coefficient was used most frequently; however, it is strongly dependent on the base rate of assessments. Therefore it has to be supplemented by the Yule coefficient for matching the dependency on the base rate (Spitznagel and Helzer 1985). Both coefficients were calculated by using the jackknife method for receiving an unbiased estimate. If a cell in a contingency table was empty, the coefficients were calculated by using the pseudobayesian approach (Bishop et al. 1975); in spite of being the most acceptable method, this contains some inherent arbitrariness. Kappa values greater than 0.75 suggest excellent agreement, Kappas between 0.74 and 0.60 suggest good agreement, Kappas between 0.59 and 0.40 suggest fair agreement, and Kappas below 0.40 suggest poor agreement. A Kappa of 0.0 indicates chance agreement.

Results

Joint-rater Reliability

Schizophreniform symptomatology and major affective episodes were diagnosed by the two raters with perfect agreement both for patients and relatives. There were minor discrepancies (in one case) with respect to the extent of the temporal overlap between affective and schizophreniform symptomatology, resulting in a disagreement for schizoaffective disorder (DSM-III) (Kappa = 0.72). The main differences between the two raters were observed for phobic disorders documented by the discrepancy sheet. Especially, the agreement on the severity of psychosocial impairment was poor, resulting in the relatively low joint-rater reliability when using DSM-III (Kappa = 0.75 with SD = 0.04) for phobic disorders; as for DSM-III-R phobic disorders could be diagnosed more reliably (Kappa = 0.89 with SD = 0.03), as no judgement is required if phobic symptoms are due to other conditions.

Test-Retest Reliability in Patients

As shown in Tables 1 and 2, the rate of test-retest agreement with respect to all psychotic disorders was quite high, both by using a hierarchical lifetime diagnoses or a non-hierarchical diagnosis of any episode (Kappa > 0.70). The agreement for schizophrenia is also satisfactory (Kappa > 0.70). However, schizoaffective disorders (DSM-III-R) can hardly be diagnosed reliably (Kappa = 0.27); this is mainly due to the fact that a very unspecific temporal overlap between affective and schizophreniform symptomatology is required for the DSM-III-R diagnosis of schizoaffective disorder. As psychotic patients with affective symptomatology are either diagnosed as affective or schizoaffective disorders,

Table 2. Test-retest-reliability diagnoses of hierarchical[a] lifetime diagnoses (DSM-III/DSM-III-R) by the SADS-LA in patients

	DSM-III			DSM-III-R		
	Relative frequency by either rater[b]	Yule[c]	Kappa mean[c]	Relative frequency by either rater[b]	Yule[c]	Kappa mean[c]
Schizophrenia/schizophreniform disorder	7/36	0.77	0.71	7/36	0.75	0.71
Schizoaffective disorder	./.	./.	./.	3/36	0.59	0.32
Affective disorders:						
Unipolar	16/36	0.63	0.64	15/36	0.59	0.55
Bipolar	8/36	0.53	0.43	7/36	0.45	0.32
W/psychotic features	7/36	0.86	0.82	6/36	0.73	0.73
W/o psychotic features	17/36	0.88	0.83	16/36	0.86	0.87
Major depression or bipolar disorder	23/36	0.86	0.82	22/36	0.78	0.75
Any affective disorder (incl. dysthymia/ cyclothymia)	25/36	0.79	0.79	24/36	0.73	0.71

[a] For diagnostic hierarchy we used Jasper's rule: schizophrenia/schizophreniform > schizoaffective disorder > affective disorder psychotic > affective disorder non-psychotic > anxiety disorder > alcoholism/drug abuse

[b] Number of patients receiving a particular diagnosis by any of the two raters (++ or −+ or +−)

[c] Yule and Kappa coefficients were calculated by using the pseudobayesian approach, in order to match the zero cells in the contingency tables. Based on the pseudobayesian approach means and variances of Kappa coefficients were estimated by the jackknife procedure

the diagnoses of affective disorders with psychotic features according to DSM-III-R is also less reliable (Kappa = 0.78 with SD = 0.05 and 0.73 with SD = 0.04) as compared with the corresponding coefficients for DSM-III-diagnoses (Kappa = 0.92 with SD = 0.02 and Kappa = 0.82 with SD = 0.02).

Diagnoses of affective disorder without any further subtyping were satisfactory for DSM-III and DSM-III-R when performing lifetime diagnoses or diagnoses of the previous episode (Table 1, Table 2). But the discrimination between unipolar and bipolar depression could only be conducted with a low degree of reliability if bipolar II disorder (DSM-III-R) or bipolar disorder not otherwise specified are counted as bipolar disorder; the coefficients of reliability (Kappa) are 0.49 with SD = 0.03 (Table 1), and 0.32 with SD = 0.04 (Table 2) for DSM-III-R and 0.64 with SD = 0.04 (Table 1) and 0.53 with SD = 0.02 (Table 2) for DSM-III. These differences are mainly explained by hypomanic status defining bipolar II being expressed by the patient in the first session but not in the second one. In these cases, discrepancies were never due to the patient having experienced a new symptomatology between the two interviews but to the fact that both patients denied having ever experienced manic or hypomanic symptomatology.

Test-retest reliability was very low for all kinds of anxiety disorders (Table 1). One source of disagreement has already been discussed in the joint-rater reliability session; another source of disagreement can be exemplified by comparing the agreement between DSM-III and DSM-III-R (Table 1). The coefficients of agreement (Kappa, Yule) for anxiety disorders in DSM-III-R are generally higher because the rater does not have to judge whether or not the anxiety symptomatology is due to major disorders. In addition, there are series of cases where panic disorder was only diagnosed in the first session and generalized anxiety disorder was diagnosed only in the second session; this observation explains that the reliability of the global "anxiety disorders" category is higher than the reliability of specific anxiety disorders. The interrater reliability for alcohol and drug abuse is satisfactory (Kappa > 0.64).

Table 2 presents lifetime diagnoses using a strict hierarchical procedure allocating a single diagnosis to a patient as is the rule in linkage studies. The coefficients of reliability obtained are comparable to those gained when not using a hierarchical procedure and allowing for comorbidity. With regard to anxiety disorders, reliability is increased when considering a broad category formed of all anxiety disorders. The reason for this is that anxiety disorders will only be taken into consideration if they are not comorbid with major affective or psychotic disorders.

Test-Retest Reliability in Relatives

The reliability of the identification of caseness strongly depends on the stringency of the definition of caseness. Considering only relatives with either psychotic or major affective disorders as cases ($n = 6$ among 20) provides perfect test-retest agreement. Six cases among relatives received a diagnosis of anxiety disorders or other neurotic disorders (DSM-III) by either rater. Including these cases, too, 8 relatives were assessed as cases by both raters, 8 were assessed as non-cases by both, 2 were considered as cases the first time but not in the second interview, and 2 were considered cases only in the second interview (Kappa = 0.6 with SD = 0.03, Yule = 0.06).

In our sample, this drop of reliability is mainly due to the discrepancies with respect to dysthymia and anxiety disorders. Two of the four discrepant cases were due to dysthymia. Dysthymia was diagnosed twice in a discrepant manner (Kappa = −0.05 with SD = 0.06), calculated with the pseudobaysian technique; anxiety disorders were diagnosed twice in consonant manner in both sessions and three times in a discrepant manner (Kappa = 0.5 with SD = 0.08, Yule = 0.6); alcoholism was not diagnosed in the relatives.

Discussion

The reliability study was able to demonstrate that the identification of major affective disorder and schizophrenia could be done with sufficient reliability with two translated versions of the SADS-LA. The negative findings are: first, the subtypes of major disorders are less reliable, and second, the identification of minor disorders, dysthymia and anxiety disorders, reduces the reliability of classification.

Comparison with Other Studies

Two reports on the reliability of the SADS may be used for comparison. Endicott and Spitzer (1978), using the SADS-L, reported nearly perfect agreement in the joint-rater setting for all qualities; this result is similar to our results (Endicott and Spitzer did not report on phobias). The test-retest reliability for features of depression was between 0.8 and 0.9 (Kappa), whereas the anxiety assessment (without applying diagnostic hierarchies) were much lower (0.68 for Kappa); we obtained similarly a drop in the reliability of anxiety disorders. Andreasen et al. (1981) reported good to excellent reliability for all RDC-major disorders, with the exception of bipolar II disorder; the results of our reliability study is in accord with this previous finding.

Manuzza et al. (1989) reported test-retest reliabilities for anxiety disorders (using the SADS-LA); they received very low reliability coefficients for phobias; the diagnosis of simple phobia (past episodes) agreed between two sessions only by chance; the reliabilities for other anxiety disorders were moderately higher; the lifetime reliabilities for panic disorder were 0.86 (Kappa), and for generalized anxiety disorder 0.60 (Kappa), and substantially lower rates for previous episodes. These figures are higher than ours, which may be explained by the differences between the samples; Manuzza et al. (1989) analyzed the ratings of patients with anxiety disorders and the interviewers knew that this is the major problem of the patients to be rated. Our sample was much more

heterogeneous and anxiety was usually not the main complaint of the patient.

Test-retest studies in non-patient samples have been reported in a wide range of interrater-reliability: Andreasen et al. (1981) obtained sufficient reliability for major depression (Kappa 0.75) , bipolar I disorder (Kappa 0.88) and alcoholism (0.72) but not for bipolar II (0.06) and certain subtypes of major depression. Mazure and Gershon (1979) described excellent reliability of caseness in a mixed sample (subjects, relatives, controls). On the other hand, a very low reliability was found by Bromet et al. (1986) for major depression (Kappa 0.41). A general tendency appears in these studies: the longer the intervals between the first and the second assessment, the lower the reliability.

Sources of Disagreement

Insufficient reliability of subtyping major disorders may be highlighted by two observations: first, the discrepancies in assessing the temporal overlap between affective symptoms and schizophrenic symptoms blur the discrimination between the DSM-III-R categories schizoaffective disorder and affective disorders with psychotic features which is left to the qualitative interpretation of the clinician; second, no strict criteria help the clinician in the definition of the degree of impairment which is required to diagnose mania/hypomania in DSM-III-R (but not in DSM-III). This critical point in the DSM-III-R does not imply that the DSM-III would form a better classification system, for neither the category of schizoaffective disorder nor the category of bipolar II disorder have been used within the framework of DSM-III, as no criteria at all have been given for these categories.

Poor reliability of minor disorders was explained by the lack of a precise definition of the relevant threshold for psychosocial impairment. This issue already becomes apparent by the joint-rater reliability in phobic disorders and is further stressed in the test-retest setting showing very poor reliability for dysthymia and phobic disorders. An additional source of disagreement in diagnosing anxiety disorders in the patient sample is due to the problem of comorbidity: as there is no precise definition of "disorder x being due to disorder y", it is not surprising that anxiety disorders are especially unreliable when using the DSM-III classification.

The lack of reliability in subdividing affective disorders might partly explain the heterogeneity of clinical phenomenology observed in families of affective subjects as has been evidenced by a series of family studies. It demonstrates that the heterogeneity observed in families will become even more complex if fringe cases are taken into account.

Conclusion

This study illuminates methodological difficulties in identifying phenotypes in psychiatric family studies. Researchers in linkage studies including fringe cases for defining caseness should be aware that reliability of classification ranges with its stringency. There is a general tendency that by lowering the threshold, the reliability decreases too. This fact is not taken into account in any publication of linkage study using a variety of thresholds for identifying cases.

A strategy which may help to overcome the reliability problem is provided by the sib pair method, using only siblings fulfilling criteria for major disorders. The results of this reliability study support a more frequent application of this strategy.

As the reliability and the validity of the diagnostic instruments available is apparently not perfect, the application of structured interviews, however, needs supplementation by follow-up studies in the phenotypes of family members. Temporal stability of diagnostic allocations is a major cue for the identification of "true" cases. Therefore, follow-up studies of families identified for genetic research should become an integral part of linkage studies programs.

References

Andreasen NC, Grove WM, Shapiro RW, Keller MB, Hirschfeld RMA, McDonald-Scott P (1981) Reliability of Lifetime Diagnosis. Arch Gen Psychiatry 38:400–405

Bishop YYM, Feinberg SE, Holland PW (1975) Discrete multivarite analysis: Theory and practice. MIT Press, Cambridge

Bromet EJ, Dunn LO, Connell MM, Dew MA, Schulberg HC (1986) Long term Reliability of Diagnosing Lifetime Major Depression in a Community Sample. Arch Gen Psychiatry 43: 435:440

Endicott J, Spitzer RL (1978) A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. Arch Gen Psychiatry 35:837–844

Fyer AJ, Endicott J, Manuzza S, Klein DF (1985) Schedule for Affective Disorders and Schizophrenia – Lifetime version (modified for the study of anxiety disorders): New York, NY: Anxiety Disorder Clinic, NY State Psychiatric Institute

Manuzza S, Fyer AJ, Klein DF, Endicott J (1986) Schedule for Affective Disorders and Schizophrenia – Lifetime version (modified for the study of anxiety disorders): rationale and conceptual development. J Psychiatr Res 20:317–325

Manuzza S, Fyer AJ, Martin LY, Gallops MS, Endicott J, Gorman J, Liebowitz MR, Klein DF (1989) Reliability of Anxiety Assessment. I. Diagnostic Agreement. Arch Gen Psychiatry 46:1093–1101

Martinez M, Khlat M, Leboyer M, Clerget-Darpoux F (1989) Performance of linkage analysis under misclassification error when the genetic model is unknown. Genet Epidemiol 6, 1:247–253

Mazure C, Gershon ES (1979) Blindness and reliability in lifetime psychiatric diagnosis. Arch Gen Psychiatry 36:521–525

Merikangas KR, Spence MA, Kupfer DJ (1989) Linkage Studies of Bipolar Disorder: Methodologic and Analytic Issues. Arch Gen Psychiatry 46:1137–1141

Spitzer RL, Endicott J, Robins E (1978) Research Diagnostic Criteria: Rationale and Reliability. Arch Gen Psychiatry 35: 773–782

Spitznagel EL, Helzer JE (1985) A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 42: 725–728

Weissman MM, Merikangas KR, John K, Wickramaratne P, Prusoff BA, Kidd KK (1986) Family-genetic studies of psychiatric disorders: developing technologies. Arch Gen Psychiatry 43: 1104–1116

Wittchen HU et al. (1991) Cross cultural feasibility, reliability and sources of variance of the composite international diagnostic interview (CIDI). (in press)