Soc Psychiatry Psychiatr Epidemiol (1998) 33: 140–144

ORIGINAL PAPER

R. Heun · H. Müller · H.J. Freyberger · W. Maier

Reliability of interview information in a family study in the elderly

Accepted: 24 January 1997

Abstract The aim of the present study was to evaluate the combined test-retest and interrater reliability of different psychiatric lifetime diagnoses yielded in the course of a family study in elderly patients and controls. The following interviews and questionnaires were used in combination: the Composite International Diagnostic Interview (CIDI), the Structured Interview for the Diagnosis of Dementia of the Alzheimer Type, Multi-infarct Dementia and Dementias of Other Aetiology (SIDAM), the General Health Questionnaire (GHQ-12) and questionnaires for neurasthenia and recurrent brief depression (RBD). Depressive and dementia disorders can be diagnosed with good reliability in a family study setting with the use of these instruments. The diagnoses of phobic disorders, neurasthenia, RBD, subthreshold RBD and psychiatric caseness as indicated by GHQ-12 scores were less reliable in this setting and are therefore less suitable for use in family studies.

Introduction

Family studies are performed to assess the role of familial, possibly genetic, factors in the occurrence of psychiatric disorders. For this purpose frequencies of different psychiatric lifetime diagnoses in first-degree relatives of different patient groups are compared with the figures for first-degree relatives of a control group (Gershon et al. 1982, Kendler 1990). Lifetime diagnoses of relatives of psychiatric patients and controls cannot be based on clinical observation alone, but also on results from interviews or questionnaires, or a combination thereof (Weissman et al. 1986). Different interview modules and cognitive tests are necessarily combined to cover different psychiatric disorders, including subthreshold disorders, and to account for possible comorbidity. To make an accurate lifetime diagnosis, interviews may therefore range from one to several hours in length.

In the present family study we compared the lifetime risks of several psychiatric disorders in first-degree relatives of elderly patients with depression or dementia of the Alzheimer type with the equivalent risks in the control families from the general population.

The presence of a possible selection bias during the recruitment of participants and relatives and the validity of family history information in this setting have been described elsewhere (Heun et al. 1995, 1996). However, assessments of the reliability of interview modules and resulting diagnoses as have thus far been carried out are not applicable to a family study setting:

1. For a comparison of psychiatric disorders in family studies it is essential that disorders are reliably detected, before they can be reliably labelled. However, the first aspect cannot be evaluated by comparing the accuracy of diagnoses in patient samples.
2. The reliability of interviews and questionnaires has primarily been evaluated in young cooperative or in clinical samples, but rarely in the general population or in first-degree relatives. Even the Composite International Diagnostic Interview (CIDI, WHO 1990), intended for the examination of general population samples, has rarely been evaluated for its reliability in non-patient samples (see Wittchen 1994).
3. The answers obtained in psychiatric interviews generally depend on the educational and social level, literacy and the perception of abnormality of the sample of interest (Shrout 1995). The perception of normality might be influenced by the presence of a diseased relative in the family. Consequently, reliability data obtained in the general population may not be useful in a family study.

R. Heun (✉) · H. Müller · H.J. Freyberger · W. Maier
Department of Psychiatry and Psychotherapy,
University of Bonn, Sigmund-Freud-Strasse 25,
D-53105 Bonn, Germany

4. The reliability of any test or interview may be influenced by the application of other diagnostic instruments during an interview session. Thus, reliability assessments for most instruments have been performed separately, but not in combination with other instruments as in family studies.

It was therefore the aim of the present study to assess the combined test-retest and interrater reliability of interview modules used for the assessment of psychiatric disorders in a family study in elderly patients and controls under normal conditions according to the following study design:

1. The investigated reliability study sample was comparable with the family study sample and consisted of patients, control subjects and their relatives.
2. The reliability study was performed during the family study.
3. All raters performing the family study interviews also participated in the reliability analysis.
4. The selected time interval between tests and retests was comparable with the usual delays between first contacts and interviews as well as with the average delay due to missed interviews that had to be performed at a later date during the family study.

It should be emphasized that it was not the goal of the present reliability study to assess the performance of interviews and of tests under optimal conditions, but to determine the diagnostic accuracy of psychiatric interviews in a family study setting.

Methods

Recruitment of patients, controls and relatives

The family study sample was consecutively recruited from inpatients aged over 60 years with depression (age of onset >40 years) or dementia of the Alzheimer type (DAT) at the Department of Psychiatry, University of Mainz. Control subjects and their families were recruited from the general population. The family study design has recently been described elsewhere (Heun et al. 1995, 1996). During 4 months of the family study, 36 consecutively recruited subjects were asked for a second interview with another rater. The second interview was performed in 31 subjects between 14 and 42 days after the initial interview. The sample consisted of eight patients and six control subjects from the general population and 17 first-degree relatives (17 women, mean age 62.5 ± 16.7 years, range 25–85 years). The eight interviewers were medical students in the 6th year of medical school, with several months of experience in psychiatry and extensive training in interview modules. The raters conducted between six and nine interviews (i.e. three to five first and three to five second interviews) each. The order of interviews followed a predetermined pattern: the same two raters were allowed to perform interviews only twice to prevent an influence on reliability results caused by fixed rater combinations. For practical reasons, i.e. availability of raters and interviewees, a full stratification or, alternatively, complete randomization was not feasible. The raters were blind to the proband status of the interviewed subjects (i.e. patient, control subject or relative of either group).

The combined test-retest and interrater reliability of psychiatric diagnoses and of test scores obtained using the following instruments was evaluated:

1. The Composite International Diagnostic Interview (CIDI) providing psychiatric lifetime diagnoses according to ICD-10 (WHO 1991), including an interview module assessing the criteria for Recurrent Brief Depressive episodes (RBD, Angst et al. 1990), and subthreshold RBD (Maier et al. 1994).
2. The Structured Interview for the Diagnosis of Dementia of the Alzheimer type, Multi-infarct Dementia and Dementias of Other Aetiology (SIDAM, Zaudig et al. 1991), which includes the Mini-Mental State examination (MMS, Folstein et al. 1975), and provides a dementia diagnosis, a global cognitive SIDAM-score (SISCO) plus subscores for the assessment of memory, orientation, intellectual ability and higher cortical functions.
3. A new module for the assessment of neurasthenia according to ICD-10 criteria (WHO 1991; Maier, unpublished manuscript.
4. The General Health Questionnaire (GHQ 12-item version, Goldberg 1972), which was assessed for its possible usefulness as a screening test for psychiatric morbidity in a family study setting. This GHQ version was selected because it offers the advantage of being the shortest and most simple of the available versions. The GHQ symptoms were classified according to the binary code (0-0-1-1).

All instruments were administered in the described sequence, all questions were asked orally without omissions and thus allowing some repetitions. The strict order to be followed by the interviewers appeared to be best suited to prevent the occurrence of individual variations in test applications.

The sample size was determined to allow comparability with other reliability studies and, consequently, should be sufficiently large to provide useful estimates of reliability for different disorders (e.g. Gureje and Obikoya 1990, Wittchen 1994). However, reliability studies should be performed under the realistic conditions prevailing during family studies, and should not interfere with the performance of the entire family study by reducing the willingness of future participants who might fear being asked for two instead of one interview. The reliability assessment was therefore limited to a period of 4 months.

The reliability evaluation used intraclass correlation coefficients (ICC) for metric data (Shrout 1995) and Kappa-coefficients for categorical data, i.e. individual lifetime diagnoses (Landis and Koch 1977). Several diagnoses were possible per subject and were individually compared according to the categories indicated in Table 1.

Results

Fourteen out of 31 interviewed subjects had a minimum of one psychiatric lifetime diagnosis in at least one of the interviews (see Table 1). Complete identity of the five-character ICD-10 codes for all diagnoses was established in only 3 out of 14 diseased individuals. Consequently, further reliability evaluations were restricted to the major categories (i.e. the first three ICD-10 characters). Table 1 shows the frequency of psychiatric disorders in both interviews and the reliability of psychiatric diagnoses, i.e. Kappa values and 95% confidence intervals. Table 2 depicts the test results obtained on both interviews as well as reliability values (ICC) for metric data.

The diagnoses made by CIDI for major depression and dementia showed a good reliability (Kappa >0.6); however, the diagnoses of phobic disorders appeared less reliable. The diagnoses for neurasthenia, RBD, subthreshold RBD, and psychiatric caseness were not sufficiently reliable, despite the fact that the number of

Table 1 Frequencies of psychiatric lifetime diagnoses in first, and second interviews, and reliabilities of diagnostic decisions (Kappa values for categorial data) in 31 subjects – 14 subjects with a minimum of one psychiatric lifetime diagnosis in at least one of the interviews and 17 subjects with no lifetime diagnosis in either interview (*CIDI* Composite International Diagnostic Interview, WHO 1990; *MMS* Mini-Mental State, Folstein et al. 1975; *SIDAM* Structured Interview of the Diagnosis of Dementia of the Alzhelmer type, Multi-infarct Dementia and Dementias of other Aetiology, Zaudig et al. 1991; *SISCO* SIDAM total score, *GHQ* General Health Questionnaire, Goldberg 1972)

Diagnosis	Criterion/ instrument	Frequency in first interview (%)	Frequency in re-interview (%)	Kappa (95% confidence interval)
Major depression	ICD-10/CIDI	12.9	22.9	0.67 (0.34–1)
Dysthymia	ICD-10/CIDI	6.5	3.2	–
Any depressive disorder	ICD-10/CIDI	16.1	25.8	0.71 (0.41–1)
Phobic disorder	ICD-10/CIDI	16.1	9.7	0.20 (0–0.67)
Panic disorder	ICD-10/CIDI	0	3.2	–
Any anxiety disorder	ICD-10/CIDI	12.9	12.9	0.14 (0–0.57)
Nicotine dependence	ICD-10/CIDI	3.2	6.5	0.65 (0.02–1)
Dementia	MMS $\leqslant$ 23/CIDI and SIDAM	9.7	13.0	0.78 (0.43–1)
Any psychiatric disorder	ICD-10/CIDI	25.8	39.7	0.57 (0.23–0.91)
Dementia of Alzheimer type	ICD-10/SIDAM	9.7	9.7	1.0 (0.65–1)
Dementia	SISCO $\leqslant$ 33/SIDAM	9.7	9.7	1.0 (0.65–1)
Recurrent Brief Depression	Angst et al. 1990, i.e. monthly episodes during 1 year/questionnaire	3.7	14.8	0.37 (0.10–0.64)
Subthreshold recurrent Brief Depression	Angst et al. 1990, but monthly episodes for 6–11 months/Questionnaire	14.8	29.6	0.19 (0–0.51)
Neurasthenia	ICD-10/questionnaire	3.3	0	–[a]
Psychiatric caseness	$\geqslant$4/12 symptoms in GHQ	25.9	18.5	0.18 (0–0.53)
Psychiatric caseness	$\geqslant$3/12 symptoms/GHQ	25.9	20.5	0.32 (0–0.67)

[a] Neurasthenia was part of a major depression or anxiety disorder in six cases, and was consequently excluded by ICD-10 criteria

Table 2 Reliability of metric data in 31 subjects interviewed two times: Scores and numbers of relevant symptoms (mean ± SD; ranges in parentheses)

	First interview	Re-interview	Intraclass correlation coefficient
MMS	27.4 ± 3.6 (16–30)	26.7 ± 3.8 (16–30)	0.78
SIDAM			
SISCO (total score)	47.5 ± 8.2 (24–55)	47.4 ± 8.7 (22–55)	0.91
Orientation	9.2 ± 1.8 (4–10)	9.3 ± 1.9 (3–10)	0.94
Memory	16.6 ± 3.5 (8–20)	16.9 ± 3.7 (7–20)	0.80
Intellectual ability	4.4 ± 1.1 (1–5)	4.4 ± 1.3 (0–5)	0.55
Higher cortical functions	17.4 ± 2.8 (11–20)	17.0 ± 3.1 (8–20)	0.75
No. of symptoms in GHQ	2.1 ± 3.0 (0–10)	1.9 ± 3.3 (0–12)	0.45
No. of symptoms of neurasthenia	2.0 ± 2.8 (0–9)	2.0 ± 3.0 (0–10)	0.55
No. of depressive symptoms in RBD episodes	3.6 ± 4.8 (0–15)	4.2 ± 5.1 (0–18)	0.40

positive symptoms showed moderate intraclass correlations between the two interviews. In contrast, the cognitive test scores and subscores measured by SIDAM provided excellent reliability (ICC >0.8, see Table 2).

Discussion

Family studies face problems in obtaining adequate information on psychiatric disorders in patients, control subjects from the general population and, most importantly, in their relatives, because, for instance, of the limited availability, time resources and compliance of non-patient samples. These circumstances may reduce the quality of diagnostic interviews compared to interviews performed in less problematic settings.

Despite the described problems we demonstrated that the reliability for diagnoses of major psychiatric categories such as dementia and major depression by structured interviews is acceptable in this setting.

The SIDAM was shown to be an adequate tool for the investigation of dementia disorders in family studies, both with regard to the reliability of test scores and the definition of caseness. Zaudig et al. (1991) also observed Kappa values for interrater reliability ranging from 0.60 to 0.95 for different types of dementia including DAT using this instrument. In support of these results, Kukull et al. (1990) reported substantial interrater agreement (Kappa) regarding the diagnosis of DAT in a clinical setting.

The CIDI provided adequate reliability for the major diagnostic categories (first three characters of the ICD-10 code), e.g. major depression, but less so for the subcategories defined by the complete five-character ICD-10 codes. The CIDI is therefore suitable for use in the assessment of major depression disorders in family studies. Kappa values for major categories comparable with those obtained by our study have been reported by Lopez (1994). Andrews et al. (1995) and Wittchen (1994) found excellent joint-rater reliability (Kappa $\geqslant$ 0.9) for all diagnostic categories in the CIDI. Robins et al. (1988) and Gureje and Obikoya (1990) reported excellent, respectively good test-retest reliability without, however, providing complete data in support of their results. Nevertheless, joint-rater reliability and the evaluation of patient samples might inflate Kappa values in comparison with combined test-retest and interrater reliability in mixed samples. Semler et al. (1987) also reported acceptable test-retest reliabilities (Kappa values above 0.5) for most of the CIDI diagnoses made in 60 inpatients. In contrast, reliability for phobic disorders was low in the present family study. A higher reliability for phobic disorders was reported by Lopez (1994), which might be explained by that study's selection of a highly cooperative group of subjects and the low mean age of the sample; younger patients might have had fewer recall problems regarding earlier symptoms than elderly subjects.

Reliabilities for RBD, subthreshold RBD, and neurasthenia were again less satisfactory than those for major depression and dementia: in the present sample the number of RBD symptoms and the diagnosis of either RBD or subthreshold RBD showed only a low reliability (ICC or Kappa) which therefore limits their suitability for family studies – at least where small samples are involved. Reliability data for the diagnosis of RBD have not yet been reported.

Interrater reliability for the number of neurasthenia symptoms as well as for the diagnosis of neurasthenia was again poor in the present study. To our knowledge, reliability regarding the assessment of neurasthenic symptoms has rarely been examined. Mindus et al. (1978) reported on a scale designed for the evaluation of neurasthenic symptoms in workers exposed to jet fuel and in healthy controls; the authors found a high correlation between the symptom scores of two audio tape-based raters. However, in contrast to our investigation, their study excludes patient variation between interviews and thus increases the correlations between test and retest scores.

In agreement with our data, Winefield et al. (1989) observed a moderate correlation of GHQ-12 scores in young Australians when retests were conducted 1 year after the first interviews (Pearson correlation coefficient $P = 0.43$). In contrast, Piccinelli et al. (1993) reported an excellent test-retest reliability in 83 subjects with mean age of approximately 40 years (ICC > 0.80). However, in elderly subjects, who might have memory problems and possibly experience relevant health changes between two assessments (including changes induced by therapy), a moderate test-retest reliability might be expected. The reliability for caseness defined by different GHQ scores was low. A search of the relevant literature revealed that Kappa values for caseness defined by the GHQ-12 scores have not previously been reported for elderly subjects. The results of the present study may, however, be relevant only for the applied GHQ version, other versions of the GHQ have been shown to have a higher reliability (Burvill and Knuiman 1983, Naugton and Wiklund 1993).

Conclusions and limitations

From the present study it can be concluded that depression and dementia can be reliably diagnosed in family studies. This applies to a lesser extent to the diagnoses of phobic disorders, neurasthenia, RBD, subthreshold RBD and to psychiatric caseness defined by GHQ-12 thresholds. It might be argued that this study underestimates the reliability of diagnoses made on the basis of interviews and questionnaires, because the instruments were used during a lengthy interview procedure lasting up to several hours. However, the described combination of diagnostic instruments is required in a family study setting, where different diagnoses are to be compared in many relatives, thus requiring reliability assessment to be performed under realistic conditions. Most importantly, the reliability sample of the present study was representative for the family study sample as a whole and did not represent a selected subsample (31 out of 36 contacted subjects participated in the re-interviews).

The sample size of the present study was limited, but comparable with other reliability samples. The sample size was sufficient to assess the reliability of different diagnoses, but the confidence intervals of Kappa values were relatively high and thus overlapping. This limits the comparison of reliability results of different diagnoses (i.e. Kappa values). Again, due to the small sample size, we are not able to confirm that the reliability of diagnoses was identical in the different subgroups consisting of patients, controls or relatives of either group, even though an examination of the diagnoses yielded for individual subjects did not provide evidence for such group differences. It is possible that the more detailed information provided by controls and relatives on previous and less severe disorders compensates for the better detectability of the more severe disorders in currently affected patients, who might have provided less accurate information on comorbid conditions and on previous episodes. However, considerably larger samples of patients, controls and relatives are necessary for these more detailed analyses.

References

Andrews G, Peters L, Guzman A, Bird K (1995) A comparison of two structured interviews: CIDI and SCAN. Aust NZ J Psychiatry 29: 124–132

Angst J, Merikangas K, Scheidegger P, Wicki W (1990) Recurrent brief depression: a new subtype of affective disorder. J Affect Dis 19: 87–98

Burvill PW, Knuiman MW (1983) Which version of the General Health Questionnaire should be used in community studies? Aust NZ J Psychiatry 17: 237–242

Folstein MF, Folstein SE, McHugh PR (1975) "Mini-Mental State". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12: 189–198

Gershon ES, Hamovit J, Guroff JJ, Dibble E, Leckman JF, Sceery W, Targum SD, Nurnberger JI Jr, Goldin LR, Bunney WE Jr (1982) A family study of schizoaffective, bipolar I, bipolar II, unipolar, and normal control probands. Arch Gen Psychiatry 39: 1157–1167

Goldberg DP (1972) The detection of psychiatric illness by questionnaire. Maudsley Monographs 21. Oxford University Press, Oxford

Gureje O, Obikoya B (1990) The GHQ-12 as a screening tool in a primary care setting. Soc Psychiatry Psychiatr Epidemiol 25: 276–280

Heun R, Burkart M, Maier W (1995) Selection biases during recruitment of patients and relatives for a family study in the elderly. J Psychiatr Res 29: 491–504

Heun R, Hardt J, Burkart M, Maier W (1996) Validity of the family history method in relatives of gerontopsychiatric patients. Psychiatr Res 62: 227–238

Kendler KS (1990) Toward a scientific psychiatric nosology. Arch Gen Psychiatry 47: 969–973

Kukull WA, Larson EB, Reifler BV, Lampe TH, Yerby M, Hughes J (1990) Interrater reliability of Alzheimer's disease diagnosis. Neurology 40: 257–260

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159–174

Lopez (1994) Reliability of the Brazilian version of the CIDI in a case-control study of risk factors for drug abuse among adults in Rio de Janeiro. Bull Panam Health Org 28: 34–41

Maier W, Herr R, Gänsicke M, Lichtermann D, Houshangpour K, Benkert O (1994) Recurrent brief depression in general practice. Clinical features, comorbidity with other disorders, and need for treatment. Eur Arch Psychiatry Clin Neurosci 244: 196–204

Mindus P, Struwe G, Gullberg B (1978) A CPRS subscale to assess mental symptoms in workers exposed to jet fuel – some methodological considerations. Acta Psychiatr Scand Suppl 271: 53–62

Naugton MJ, Wiklund I (1993) A critical review of dimension-specific measures of health quality of life in cross cultural research. Qual Life Res 2: 397–432

Piccinelli M, Bisoffi G, Bon MG, Cunico L, Tansella M (1993) Validity and test-retest reliability of the Italian version of the 12-Item General Health Questionnaire in general practice: a comparison between three scoring methods. Compr Psychiatry 34: 198–205

Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, Farmer A, Jablenski A, Pickens R, Regier DA, Sartorius N, Towle LH (1988) The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Arch Gen Psychiatry 45: 1069–1077

Semler G, Wittchen HU, Joschke K, Zaudig M, von Geiso T, Kaiser S, von Cranach M, Pfister H (1987). Test-retest reliability of a standardized psychiatric interview (DIS/CIDI). Eur Arch Psychiatr Neurol Sci 236: 214–222

Shrout PE (1995) Reliability. In Tsuang MT, Tohen M, Zahner GEP (eds) Textbook in psychiatric epidemiology. Wiley-Liss, New York, pp 213–227

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

WHO (1990) Composite International Diagnostic Interview. World Health Organization, Division of Mental Health, Geneva

WHO (1991) Tenth revision of the International Classification of Diseases, Chapter V (F): Mental and behavioral disorders. Clinical descriptions and diagnostic guidelines. World Health Organization, Geneva

Winefield HR, Goldney RD, Winefield AH, Tiggemann M (1989) The General Health Questionnaire: reliability and validity for Australian youth. Aust NZ J Psychiatry 23: 53–58

Wittchen HU (1994) Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review. J Psychiatr Res 28: 57–84

Zaudig M, Mittelhammer J, Hiller W, Pauls A, Thora C, Morinigo A, Mombour W (1991) SIDAM – a Structured Interview for the Diagnosis of Dementia of the Alzheimer type, Multi-infarct Dementia and Dementias of Other Aetiology according to ICD-10 and DSM-III-R. Psychol Med 21: 225–236