Arch Public Health
1998, 56, 333-344

The reliability of the clinical assessment of psychiatric emergency referrals

by

Spooren D. J., van Heeringen K., Jannes C.

Key-words

Emergency psychiatry, assessment, reliability, interviewer-observer, test-retest.

Abstract

The reliability of the assessment of patients referred to the psychiatric emergency department of four public hospitals was studied under two different conditions. An interviewer-observer design was used in two hospitals and a test-retest study in the other hospitals. In each hospital at least 50 patients were included.

Address correspondence to: Daniel J. Spooren (clin.psych.), Department of Psychiatry, University Hospital, De Pintelaan 185, B-9000 Gent, Belgium. Tel. +32/9/240 43 93. Fax: +32/9/240 49 89. E-mail: Daniel.Spooren@rug.ac.be

Results showed good reliability for sociodemographic variables. However, considerable differences in inter-observer agreement between the four hospitals were found for clinical characteristics. It appeared that disagreement was mainly related to the method used and to the level of previous training of the clinicians, which participated in the study. The findings indicate the need of a previous formal training of clinicians before taking part in a monitoring project.

Introduction

Although the limited reliability of the assessment of psychiatric patients at the emergency department (ED) is an often mentioned problem (1-3), relatively few studies have systematically investigated interobserver or interrater reliability. Because of the lack of objective external criteria, such as laboratory tests, the problem of reliable data collection is inherent to psychiatry. Sources of disagreement between interviewers can be related to inconsistency on the part of the patient, inconsistency on the part of the interviewer or inadequacies of the instrument used (4). Patients who use the ED frequently belong to poorer socio-economic groups, and may experience more difficulties in defining their problems or in providing accurate and relevant psychiatric and personal histories (5). Secondly, their condition may hinder accurate data gathering as the presence of intoxication, of acute medical conditions, or of agitated or threatening behaviour may render a first interview nearly impossible. On other occasions the patient and his family may be so distressed that their ability to offer valid information is limited. Other patients may refuse to offer valid information because of various reasons. Inconsistencies of the patient cannot be controlled, but information to confirm the information obtained from the patient can be retrieved from other sources. However, the availability of assessment time is often limited, and frequently previous knowledge is absent within the context of a first psychiatric assessment at the ED. Therefore, the clinician frequently has to rely only on self-report data and clinical impressions.

Given this situation it is important to attempt to reduce the other sources of unreliability as much as possible. Inconsistencies due to the interviewer can be reduced by training (6-7) and by the provision of clearly specified inclusion criteria to assign a patient to a certain category, thus reducing the degree of interpretation.

The first studies on reliability in the psychiatric ED usually concentrated on psychiatric diagnosis. Lieberman and Baker (1) found an

acceptable level for the assignment of broad major diagnostic categories, but more specific subtypes of mental illness could not be distinguished reliably. More recently, authors investigated the reliability of the assessment of other patient characteristics. Van Heeringen et al (8) showed that although the assessment of sociodemographic variables and main psychiatric diagnoses of suicide attempters at the ED could be made reliable, the assessment of other clinical variables was unreliable. Flannigan et al (9) reported good agreement for approximately half of the sections of an audit information schedule developed for a survey of acute health services, but their conclusion was based on a very small patient sample. Garbrick et al (10) demonstrated that different professional background of emergency physicians and psychiatrists results in low agreement on key variables such as dangerousness, the presence of substance abuse and the need for inpatient psychiatric treatment.

The present study aimed at the evaluation of the reliability of the assessment of psychiatric emergency referrals with the use of a standardised monitoring form in four public general hospitals in Belgium (11). Although the results of this monitoring study with respect to the sociodemographic and clinical profile were consistent with comparable epidemiological studies, for some items (the presence of life threatening circumstances, prior use of services other than hospitalisation) inconsistencies were found between the participating hospitals. In one hospital the prevalence of life threatening conditions was much higher, and the data with reference to the prior use of services showed high variability among the four hospitals. This could indicate a potential lack of reliability in data collection. Therefore the initial form was reviewed and the interrater reliability of the assessment with this revised form was investigated, by determining the degree of concordance between two different psychiatric clinicians in the four participating hospitals. A second objective was to study the role of two conditions on reliability: agreement on the basis of a single interview versus two separate assessments, and the level of training of the clinicians. Since the reliability of the assessment was studied under different conditions in each hospital, an inference could be made about the influence of these conditions.

Method

Patient selection

During approximately three months a non-probability sample of patients was included in each hospital. Sample size was determined at

50 patients in each hospital. Inclusion of patients depended on practical circumstances, e.g. the availability of two clinicians at the time of the first interview. Therefore, patients were only included from Monday till Friday during office hours. Reasons for exclusion were extreme agitation or violent outbursts, or severe non-responsiveness due to various conditions (mutism, confusion, ...). The final decision about inclusion depended on the clinicians' personal judgement about the patient's state at the time of the referral.

Study Design

Two different designs were used to assess the inter-rater reliability of the monitoring. In the interviewer-observer design (applied in hospital A and C) monitoring was done independently by two observers on the basis of one interview. The observer was present during the course of the interview. Clinicians changed roles as interviewer in order to prevent systematic bias in data gathering.

TABLE 1
Conditions of reliability assessment

	Single interview	Two interviews within < 2 hours	Two interviews within 8hrs-24hrs
With previous training	Hospital C		Hospital D (training on DSM-IV diagnosis)
Without previous training	Hospital A	Hospital B	

In the test-retest design the two clinicians interviewed the same patient independently. Preferably the interviews are scheduled with a short time interval. In hospital B, both interviews for all patients were performed within two hours. In hospital D, however, due to practical circumstances, a second interview was possible only after the patient was admitted to the psychiatric department. Therefore the interviews were performed in two different locations (emergency department and psychiatric department), and the time interval between the two interviews ranged between 8 and 24 hours. It is clear that under this condition inter-rater agreement may be effected by changes in the patient's state.

A second difference between hospitals was related to the level of training of the clinicians participating in the study. In hospital A two residents (third and fourth year of training) participated, without the presence of a senior resident. In hospital B higher-grade residents performed interviews, but a senior psychiatrist was present for ready supervision. In hospital C interviews were performed by a senior psychiatrist or by a higher grade resident who was trained by the former.

The study in hospital D was started six months after termination of the study in the other hospitals. Prior to the study in this hospital, sources of unreliability due to the checklist were adapted and the reviewed standardised monitoring form was tested. Furthermore, both clinicians agreed to review and study the diagnostic criteria of the DSM-IV in order to increase their diagnostic effectiveness. Both interviewers were residents in their fourth year of training.

Analysis

Reliability was assessed by calculating kappa-coefficients (12) for nominal variables. In judging the kappa values, the following cut-off points were used (13): 0.75 = excellent; 0.65-0.74 = good; 0.50-0.64 = fair; 0.40-0.49 = moderate; < 0.40 = poor. Kappa values of zero or negative values indicate that the observed percentage of agreement is equal to or less than the agreement expected on a random basis.

Monitoring form

During the interview clinicians used a revised standardised monitoring form including sociodemographic variables, variables related to the circumstances of the referral (moment, intoxication, modality, and main reason for the referral), and clinical variables (previous use of services, psychiatric diagnosis, precipitating life events, and decision). In addition the form contained a number of ordinal scales to estimate the degree of emergency of the referral and the availability of internal and external resources to the patient (Spooren, van Heeringen, Jannes, submitted). The assignment of psychiatric diagnoses was made according to the criteria of the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and was limited to the major categories of axis I.

Results

Characteristics of the samples

In each hospital 50 patients were included (except for hospital B where 59 patients were included). Compared to an earlier descriptive study of the population of patients referred to the ED (11), no differences were found between the sample and the population on sociodemographic characteristics. Differences were observed for the proportion of certain diagnoses and for referral characteristics. Compared to the population study mood and adjustment disorders in two hospitals, and voluntary referrals of patients in three hospitals were more common. These differences were probably related to the conditions of sampling (inclusion was limited to office hours, exclusion of certain problem groups, only patients referred for further hospitalisation in hospital D).

Reliability of assessments

Good to excellent agreement was found for the assessment of sociodemographic characteristics in all hospitals. Assessment of the referral circumstances showed heterogeneous results. Good to excellent reliability was found for most variables in the hospitals where

TABLE 2

Proportion of agreement and kappa-values (κ) of sociodemographic characteristics, variables related to the referral circumstances, and previous treatment

Hospital	A		C		B		D	
Characteristic	% Agr.	κ	% Agr.	κ	% Agr.	κ	% Agr.	κ
Situation of living	94%	.90	94%	.87	96%	.92	76%	.62
Source of income	86%	.82	100%	1.00	93%	.92	83%	.78
Modality*	86%	.40	98%	.93	72%	.46	94%	.54
Intoxication	93%	.73	94%	.76	91%	.77	91%	.82
Source of referral*	78%	.66	98%	.96	79%	.67	63%	.43
Main reason for referral	78%	.74	83%	.80	58%	.52	74%	.67
History of complaint	92%	.78	98%	.95	84%	.62	84%	.34
Previous hospitalizations	76%	.67	95%	.93	93%	.90	95%	.57
Time since last discharge	78%	.67	96%	.94	90%	.85	96%	.61

* Categories: Modality: voluntary, involuntary admissions, with active opposition, civil commitment. Source of referral: self-referred, referred by family or environment, professional source of referral (health care professional, institution, legal). Reason for referral: parasuicide, depressive complaints, alcohol, medication or drug use, acute psychotic confusion, other confusional states, conduct problem, anxiety or panic attack, psychosomatic complaints, situational problems (family, social, professional).

TABLE 3
Proportion of agreement and kappa-values (κ) of major psychiatric diagnoses, psychosocial stressors and disposition

Hospital	A		C		B		D	
Characteristic	% Agr.	κ	% Agr.	κ	% Agr.	κ	% Agr.	κ
Axis I diagnosis								
Substance use disorder	56%	.08	90%	.80	90%	.77	96%	.92
Psychotic disorder	86%	.29	98%	.85	91%	.71	96%	.83
Affective disorder	72%	-.03	96%	.87	74%	.41	88%	.74
Adjustment disorder	90%	.49	94%	.76	80%	.51	86%	.19
Acute stressor	81%	.58	91%	.79	96%	.46	71%	.38
Domain of the stressor	58%	.42	97%	.95	73%	.61	81%	.57
Judgement of reliability of the information	46%	.14	68%	.53	47%	.05	40%	-.05
Crisis unit*	100%	1.00	100%	1.00	98%	.96		
Decision**	82%	.76	88%	.85	93%	.91		

* Further management of the patient by a multidisciplinary psychiatric team for a maximum of 72 hours. The treatment alternative "crisis unit" was not available in hospital D.
** Decision for further inpatient psychiatric treatment. Since all patients included in hospital D were referred for inpatient treatment no analysis was performed.

judgement was done during one interview (A and C). Kappa values were moderate to good in the hospitals where the assessment was based on two separate interviews. The judgement of the previous care was good to excellent in three hospitals and fair in one (hospital D).

Agreement on the assignment of the three major Axis I diagnoses was good to excellent in three hospitals, except for the diagnosis of mood disorder in hospital B. In hospital A, however, kappa values for all diagnoses were poor. The assignment of adjustment disorder showed only acceptable agreement in hospital C. Clinicians were further asked to judge if there was a precipitating psychosocial stressor at the time of the referral, and if so to what domain this stressor pertained. Both variables showed only moderate agreement in three out of four hospitals. The variable "reliability of information" assessed the clinician's judgement about the information obtained from the patient. All kappa values for this variable were low, indicating the subjective nature of this judgement. Finally, clinicians showed excellent agreement about the clinical decisions regarding the need of further crisis intervention and disposition.

The results offer an indication of the influence of both study conditions on reliability. First, good to excellent agreement could only be

obtained in hospital C where both conditions were satisfied, e.g. in case both assessments were performed in a single interview, and when the psychiatric resident participating in the study was trained previously by the senior psychiatrist in the use of the form. If both assessments were performed during a single interview by residents without specific training (hospital A), good agreement was obtained for most non-clinical variables, but their clinical assessment proved to be unreliable. In both hospitals where the assessments was based on two separate interviews (hospital B and D) less agreement was found for variables related to the referral circumstances and to previous treatment, but in both hospitals the assessment of clinical variables was more reliable than in hospital A.

Discussion

Sociodemographic characteristics of patients were monitored reliably. Reliability of the other variables was less consistent. Overall we noticed that an assessment of the patient during one interview with well-trained clinicians (situation of hospital C) resulted in good to excellent agreement in practically all the observed characteristics. However, if the training aspect is neglected (hospital A) the interrater agreement for most variables relying strongly on the judgement of the rater (did the patient present himself voluntarily or not, presence of a psychiatric diagnosis) becomes low. In hospital A good to excellent agreement was found in only 56% (10/18) of the observed variables. If the assessment was performed on different occasions (hospital B and D) agreement between clinicians further decreased as the time interval between the two assessments became longer. While results in hospital B were comparable with hospital A (56% or 10/18 variables with good to excellent agreement), the proportion of variables with good to excellent agreement is reduced to 38% (6/16) if the time interval between the first and second interview is more than eight hours.

One variable was unreliable in all hospitals: the practice of asking the clinician to judge the reliability of the obtained information turned out to be completely subjective. This variable probably tells us more about the nature of the relation between the clinician and his patient, than about the patient. Therefore using this practice as an alternative to a separate reliability study should be abandoned.

A number of limitations should be considered when interpreting the results of this study. A first limitation is related to sampling. Because

inclusion of patients was not performed on a random basis, the generalizability of the results is limited. Moreover, because of the under-representation of difficult patient groups, reliability of the assessment on a random sample would probably have been lower. Secondly, because the different study conditions were not systematically distributed over the different settings, it was impossible to determine the exact influence of these conditions on reliability. Thirdly the use of kappa as a measure for interrater agreement has been criticised because, when the number of positive ratings made by either clinician is much smaller than the number of negative ratings, kappa tends to be low, even when good agreement was obtained. The reason is that kappa does not reflect degree of agreement when the characteristic is absent. Several authors suggested the use of the indices of positive and negative agreement (6) or the random error coefficient (REC) (14) as better indices of reliability in these cases. The rationale behind this criticism is that clinicians do not operate on a random basis. As could be observed in table 2 and 3, for a number of variables a high overall interrater agreement resulted in low kappa values (modality of the referral, a diagnosis of adjustment disorder). In these cases we may have underestimated the degree of reliability. However, in case of a first assessment at the ED, clinicians often are in a state of complete ignorance about whether or not a characteristic is present for a particular patient. This implies that in this particular situation allocation of a patient to a certain category might well be a random allocation, arguing for the more stringent application of kappa. Finally, we should born in mind that even in the hospital where good to excellent agreement was found, nothing can be said about the validity of the assessment, due to the lack of a golden standard.

The results of this study were consistent with a previous study on the assessment of patients referred for attempted suicide (8), where it was found that monitoring of sociodemographic characteristics can be performed reliably. However, as has been recently demonstrated (15) reliability of demographic patient characteristics becomes lower in case of change. Therefore it is better to include patient items that remain stable over time, such as the level of education in standard monitoring forms.

In performing the study under different conditions in the four settings we could investigate various hypotheses about the sources of disagreement. A common explanation for the unreliability of diagnosis at the ED is related to the limited experience of the clinicians working in these settings (2). In this study, with the exception of the senior psychiatrist of

hospital C, all participants were junior psychiatrists in their third or fourth year of training. It therefore seems unlikely that differences in the level of experience of the clinicians explain the variation between hospitals. Probably more important were differences in specific training in the use of the instrument before the study commenced. The importance of previous training has been repeatedly demonstrated in studies about the reliability of diagnostic interviews (6-7). It appears from this study that, in order to guarantee a reliable use of the monitoring checklist, a specific instrument-related training is required. Several findings in favour of this argument were found. In hospital C the senior could train his junior colleague before the start of the study. Furthermore, it appears that prior to the study in hospitals B and D, special attention was given to the correct application of the assignment of diagnoses according to the criteria of DSM-IV. This resulted in reasonable kappa values for the main diagnoses in these hospitals, despite the fact that two interviews were used. Moreover, the lack of specific training and the absence of close supervision of a senior psychiatrist in hospital A may partially explain the poor agreement among clinicians for the assignment of diagnoses, despite the condition of a single interview.

The unreliability of some characteristics related to the referral circumstances can be attributed to poor documentation as well as to problems of judgement. For example the poor result for the item about the modality of the referral in hospitals A and B was related to a problem clinicians experienced with the distinction between the categories "involuntary referral" and "involuntary referral with active opposition". Therefore we tried to reduce error by suppressing this distinction in the final revision of our checklist. However, even with this adapted version clinicians in hospital D still obtained a low reliability score for this variable. This time it was due to a lack of knowledge of the referral circumstances of the clinician who performed the second interview at the psychiatric ward.

Conclusion

It appears from this and previous studies that monitoring of patients by clinicians in the ED can only be performed reliably under limited conditions. The use of a standardised monitoring with clear and specific criteria, is a first but insufficient measure to increase reliability and comparability between hospitals. However, in order to obtain good to excellent reliability, adequate training of clinicians should precede the use of the monitoring form.

Secondly, the results from this study appear to indicate that certain characteristics of the psychiatric emergency referral assessment are transient. If the time interval between two assessments exceeds 12 hours, interrater agreement tends to be poor, except for more stable characteristics such as sociodemographic variables, and the assignment of major psychiatric diagnoses.

Because of their low threshold emergency departments of general hospitals offer unique research opportunities: studies of service users and needs assessments can be performed on a regular basis. The monitoring of changes in service use can be important from a public health perspective, because they may indicate an alteration of health care needs within a region. However, before health policy decisions are based on these studies, they should be critically examined. A minimal requirement is a detailed description of the method of data collection and its limitations. A better approach is an independent study of the validity of the obtained data. Whenever possible clinician based ratings should be supplemented by other methods of assessment, such as the use of short screening instruments or tests.

Acknowledgements

This study is part of an evaluation research ordered by the federal Secretary of Social Affairs. The authors acknowledge the co-operation of the staffs and supervisors of the following psychiatric emergency departments: General hospital "Stuivenberg", Antwerp (R. Desnyder, M.D.); "Brugmann" University Hospital, Brussels (I. Pelc, PhD, M.D.; P.Minner, M.D.); Public Hospital of Marchienne-au-Pont (Wilmotte, PhD., M.D.; R. Guillaume, M.D.). We further want to express our special thanks to the clinicians who performed the assessment in the participating hospitals.

References

1. LIEBERMAN P B, BAKER F M. The reliability of psychiatric diagnosis in the emergency room, Hospital and Community Psychiatry, 1985; 36/3:291-293.
2. MARSON D C, MC GOVERN, M P & POMP H C. Psychiatric Decision Making in the Emergency Room: A Research Overview. American Journal of Psychiatry, 1988; 145/11, 918-925.
3. CLAASSEN C A, GILFILLAN S, ORSULAK P, CARMODY T J, BATTAGLIA J, RUSH A J. SUBSTANCE use among patients with a psychotic disorder in a psychiatric emergency room. Psychiatric Services, 1997; 48/3:353-358.

4. WARD C, BECK A, MENDELSON M, MOCK J, ERBAUGH J. The psychiatric nomenclature: reasons for diagnostic disagreement. Archives of General Psychiatry, 1962; 7:198-205.
5. BIRK A W, BASSUK E L. The concept of emergency care in Bassuk & Birk, (eds.), Emergency Psychiatry. Concepts, Methods, and Practices, 1984, Plenum Press, New York and London.
6. WING J, NIXON J, LEFF J. Reliability of the PSE (ninth edition) used in a population study. Psychological Medicine, 1977; 7/505-516.
7. COOPER J, COPELAND J, GROWN G, HARRIS T, GOURLAY A. Further studies on interviewer training and inter-rater reliability of the Present State Examination (PSE). Psychological Medicine, 1977; 7:517-523.
8. VAN HEERINGEN C, RIJCKEBUSCH W, DE SCHINKEL C, JANNES C. The reliability of the assessment of suicide attempters. Archives of Public Health, 1993; 51:443-456.
9. FLANNIGAN C B, GLOVER G R, FEENEY S T, WING J K, BEBBINGTON P E, LEWIS S W. Inner London collaborative audit of admissions in two health districts. I: introduction, methods and preliminary findings, British Journal of Psychiatry, 1994; 165: 734-742.
10. GARBRICK L, LEVITT A, BARRETT M, GRAHAM M. Agreement between emergency physicians and psychiatrists regarding admission decisions. Academic Emergency Medicine, 1996; 3/11:1027-1030.
11. SPOOREN D, JANNES C, HENDERICK H, VAN HEERINGEN C. Epidemiology of psychiatric emergency referrals in four regions of Belgium, Crisis, 1996; 17/1:15-21.
12. COHEN J A. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960; 20:37-46.
13. FLEISS J L. Statistical methods for rates and proportions, 1981, New York, Wiley.
14. MAXWELL A E. Coefficients of agreement between observers and their interpretation. British Journal of Psychiatry, 1977; 130:79-83.
15. VAN DEN AKKER M, FRANSSEN G H L M, BUNTINX F, METSEMAKERS J, KNOTTNERUS J. The reliability of register-based patient characteristics. Archives of Public Health, 1997; 55:231-238.