

0022-3999(94)00007-R

THE ECLW COLLABORATIVE STUDY¹ II: PATIENT REGISTRATION FORM (PRF) INSTRUMENT, TRAINING AND RELIABILITY

ANTONIO LOBO,* FRITS J. HUYSE,† THOMAS HERZOG,‡ ULRİK F. MALT§
and BRENT C. OPMEER and the ECLW²

(Received 9 June 1994; accepted 10 January 1995)

Abstract—This paper describes the development and testing of the Patient Registration Form (PRF), a standardized instrument for the description of patients seen by consultation–liaison (C–L) psychiatrists and psychosomatists in general hospitals, the referral patterns, the C–L interventions and their outcomes. The PRF study is part of a large multi-centre, European investigation on the effectiveness of mental health service delivery, conducted by the European C–L Workgroup for General Hospital Psychiatry and Psychosomatics (ECLW) and performed in the framework of the of the E.C. 4th Medical and Health Research Program.

The final version of the PRF consists of 68 items. It was developed by the Program Management Group (PMG) and National Coordinators (NC) after long preparatory studies to assure both face and content validity and pilot testing. Two hundred and twenty consultants, who required 40 hours of training and came from 14 different European countries and 90 different sites, participated in the final reliability

* University of Zaragoza, University Hospital, Zaragoza, Spain.

† C–L Service Free University Hospital, Amsterdam, The Netherlands.

‡ Department of Psychotherapy and Psychosomatic Medicine, Alberts Ludwigs University, Freiburg, Germany.

§ Department of Psychosomatic and Behavioral Medicine, University of Oslo, The National Hospital, Oslo, Norway.

¶ ECLW Coordination Center, Free University Hospital, Amsterdam, The Netherlands.

¹ This study is initiated by the European Consultation/Liaison Workgroup for General Hospital Psychiatry and Psychosomatics (ECLW) grant supported by the European Community 4th Medical and Health Research Program COMAC—Health Service Research (Grant number: MR4*-340-NL) under the title: “The Effectiveness of Mental Health Service Delivery in the General Hospital”. In Germany grant support has been provided by Robert Bosch Stiftung (Grant number: 1-1.5.1030.0075.0) and in the Netherlands the National Fund for Mental Health Research (Grant number: 90.3594). Additional support has been provided by the Norwegian Research Council for Science and the Humanities (NAVF), the Spanish Fondo de Investigación Sanitaria (Grant # 92/0886E), Upjohn International Medical Sciences Liaison and Pfizer International.

² In addition to the authors mentioned, the following persons have been actively involved in the design of the study; as National Coordinators: Myriam van Moffaert Belgium, Pekka Tienari Finland, Paul Sakkas Greece, Graça Cardoso and Raul Guimaraes Lopes Portugal, Marco Rigatelli Italy, Maria Dolores Crespo Spain, and Richard Mayou and Francis Creed United Kingdom; as national consultants: Darius Razavi Belgium, Dorte Loldrup and Per Bech Denmark, Edmond Guillibert and Guy Marx France, Barbara Stein and Michael Wirsching Germany, Giovanni Fava Italy, Michiel W. Hengeveld The Netherlands, Inge Refne Norway, Bogdan Radanov and Roger Zumbrunnen Swiss, Torny Person Sweden, and Geoff Lloyd The United Kingdom. Data management and analyses have been taken care of by Andree JMM Rijssenbeek and Gerrit Koopmans The Netherlands, and Barbara Stein Germany. General Consultants of the study have been James J. Strain, Jeffrey S. Hammer and John S. Lyons United States, David Goldberg United Kingdom, Wim van der Brink and Maarten Koeter The Netherlands and Graeme Smith Australia.

study. The PRF was tested in 13 written case histories. A 'gold standard' for the correct answers in each item was decided by 'consensus ratings' of the PMG and a subsequent 80% agreement by the NCs. A high standard (average kappa (κ) ≥ 0.70 ; at least 2/3 of the PRF items, $\kappa \geq 0.70$) was required for the rater to be considered as 'reliable' (RR).

The consultants considered the PRF both 'feasible' and 'acceptable' and 93.2% of them fulfilled the RR criteria. The calculated rater-'gold standard' reliability was satisfactory: only four PRF items were identified with low agreement coefficients and no biases were observed cross-nationally in the ratings. Given the implications of misclassification for evaluating C-L effectiveness and services, these results are important and the achievement unprecedented.

Keywords: Clinical database; Consultation-liaison; European hospitals; Health services research; Reliability, Standardized assessment.

INTRODUCTION

The knowledge about the quality and quantity of psychiatric and psychological service delivery in the general hospital setting is very limited [1, 2]. In this context, in 1987 the European C-L Workgroup for General Hospital Psychiatry and Psychosomatics was founded (ECLW) and the study on 'The Effectiveness of Mental Health Service Delivery in the General Hospital' was programmed [3]. The primary goals of this investigation were (a) the description of the nature and organization of mental health C-L services in European general hospitals, and (b) the assessment of the relationships between the availability of these mental health services and the quality and quantity of service delivery.

The background and methodology of the study are extensively described in other core papers [4-6]. In summary, representatives of European countries were selected on the basis of relevant publications in the international literature, or recommendation. A Program Management Group (PMG) was formed and national coordinators (NCs) for the study were appointed. The NCs updated C-L psychiatrists from their countries both from university and non-university hospitals, assessed their willingness and capacity and eventually recruited and trained them for participation in the study. The need for a comprehensive set of instruments capable of describing the relevant variables soon became apparent. These instruments are: (a) a standardized, manual-based comprehensive instrument for the description of referred patients, referral patterns, a diagnostic system, C-L interventions and outcomes. The new instrument would be called the Patient Registration Form (PRF). (b) Specific instruments for institution-oriented data (Hospital and C-L Service Description) and for provider-oriented data (Form IV-A: Consultant Data) [6].

The major objectives of this paper are: (a) to describe the development of the PRF and the studies undertaken to explore its cross-national acceptability, feasibility, face and content validity; and (b) to determine the reliability of raters across a large variety of settings in different European countries and evaluate possible reasons for discrepancies in the ratings of the PRF.

METHOD

Construction of the Patient Registration Form (PRF)

A valid and thus useful comparison of different C-L services around Europe demanded extensive discussion of the content of the PRF and a related manual in a series of consensus meetings, starting in 1989. While the construction of the PRF has been influenced by previous work in the US [7], The

Development of the E.C.L.W. PATIENT REGISTRATION FORM (PRF) FLOW CHART

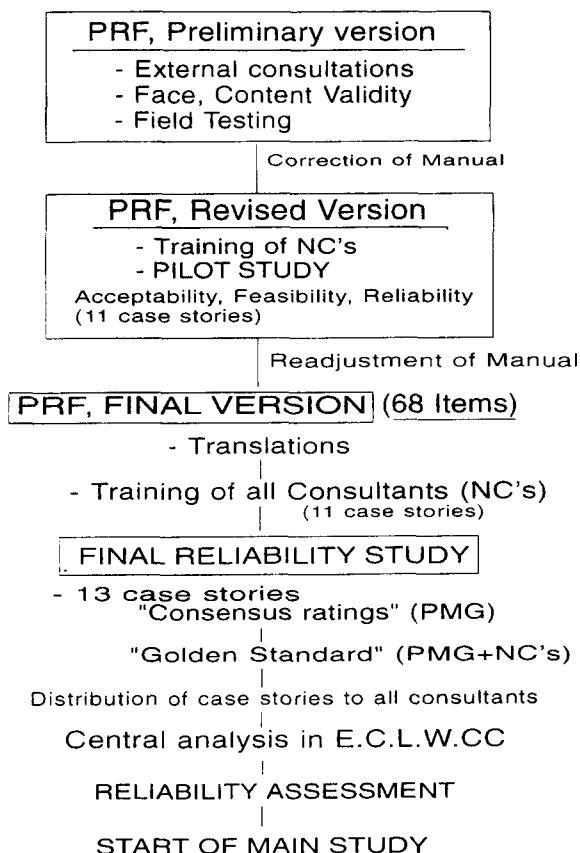


Fig. 1. The steps in the development of the Patients' Registration Form (PRF).

Netherlands [8] and Germany [9], it is essentially a completely new documentation system. Several steps were involved in its development (Fig. 1).

In the initial stage of the study, the Program Management Group (PMG) and the National Coordinators (NC) from all participating European countries consulted with external experts. Efforts were made to set minimum validity standards for the new instrument, that is, face and content validity. The PRF should: (a) take cross-national problems into account; (b) include cost-relevant aspects such as time spent on consultations, involvement of other services, previous and concurrent medical and psychiatric treatments, etc.; (c) be applicable to very different types of service provision; and (d) collect the type of information which it is possible to obtain in a reliable and valid way as part of C-L work, without the need for the use of research assistants in each hospital. As far as the diagnostic classification system is concerned, a section on somatic diagnoses was included, which incorporates the World Health Organization International Classification of Diseases [10]. For the psychiatric diagnosis, the WHO ICD-10 Chapter V 'Mental and Behavioral Disorders' has been included [11].

Pilot study of the PRF

As a second step, national coordinators (NC) and their co-workers were asked to apply the preliminary

version of the PRF in a series of clinical cases to assess acceptability and feasibility. Subsequently, a pilot study among NCs was performed to assess the inter-rater agreement as a first step towards the evaluation of the reliability of the PRF. A set of 11 case vignettes drawn from different centres and covering a large variety of service delivery patterns and of psychiatric diagnoses was compiled and distributed to NCs and scored on the forms. While the agreement was found to be essentially satisfactory ($\kappa > 0.60$) [12], some of the items proved to be unreliable. The main difficulties were encountered with regard to the description of the educational levels, the type of occupation, the employment status and the description of interventions. Consequently, some items were removed or changed and readjustments were made in the manual to improve reliability. As a result the PRF was simplified, improving its acceptability and feasibility. The final version consists of 68 items, including items involving several options, which are grouped in different sections, and appear in Table 1.

Design of the final reliability study

It was intended that the ECLW study would compile the findings of many skilled clinicians from different countries in order to identify differences and similarities between their C-L services and their ways of dealing with patients. However, a useful comparison of different European C-L services required a reliable method of collecting information. There are several ways of measuring reliability, each of which has its own strengths and weaknesses [13]. In this study, which involved a large number of raters, an important objective was to identify, relatively quickly, clinicians who used the PRF in a markedly different way from the others (outliers). More training could then be given to the less reliable raters. The method based on ratings of written case histories was judged to be the most suitable. Therefore, in the fourth and last step of the validity study, a new set of 13 test cases was compiled following the procedure described in the pilot study. The PMG discussed in detail their own ratings of these vignettes and it was unanimously decided which items could not be used in the vignette applying to a particular item, due to insufficient information. For the remaining items of the PRF the correct answers were decided upon 'consensus ratings' (CRs).

The CRs were then compared with the ratings of the national coordinators (NCs). The consensus ratings were accepted as the criterion or 'gold standard' (GS) for a particular item when 80% of the NCs fully agreed on the rating. Any disagreement between the PMG and the NC's ratings on a particular item in a given case resulted in excluding the item, in order to ensure reliability.

Next, a set for reliability testing, including guidelines plus a booklet with the 13 test cases and scoring forms was forwarded to each consultant who wished to participate in the study. Where necessary, translations of the case vignettes were provided.

Reliability of raters

The NCs, together with the PMG provided the training, which was approximately 40 hr for each consultant, including testing. During the training the PRF was extensively discussed, and rated on a series of training cases.

A high reliability standard had to be met for acceptance as a participant in the study. Reliability is defined here as agreement between raters applying the same instrument (PRF) in the same cases and a 'gold standard' (CR). In order to participate in the main study, the consultants had to be considered 'reliable raters' (RR) by fulfilling the following criteria, based on the results of rating the 13 test cases:

- 1.1) The reliability of the psychiatric diagnosis for any rater should be at least $\kappa = 0.70$; this part of the study is described separately, due to the specific interest of the use of a psychiatric diagnostic system in medical settings [5].
- 1.2) The average κ (percentage agreement or PA when the calculation of κ was considered to be inappropriate) of all the items for a particular rater should be equal to or higher than 0.70.
- 1.3) In addition, for each single rater at least two-thirds of all the items in the PRF had to have a κ (PA when κ was not calculated) equal to or higher than 0.70.
- 1.4) If a rater fulfilled all the criteria 1.1–1.3, but nevertheless had single items with a reliability lower than 0.40, additional training for that particular item, or cluster of items, was necessary.

For the purpose of the present report, the raters fulfilling both criteria 1.2 and 1.3 will be considered to be reliable raters (RR).

In order to ensure reliability throughout the study, the NCs were requested to supervise the use of the PRF and Manual regularly, by discussing individual cases with each participating centre and its raters.

Statistical analysis

The items of the PRF were categorized as nominal (Nom) or ordinal (Ord). Furthermore, the distribution of each variable according to the distribution of the answers in the gold standard (GS) was categorized as symmetrical (S), asymmetrical (AS) or very asymmetrical (ASS). κ coefficients of agreement were

Table I.—Domains and variables of the Patient Registration Form

HOSPITALIZATION	DIAGNOSTIC
Administrative Date of admission Date of consultation request Date of consultation Date of last consultation Date of discharge Consultant Id Time spent on first consultation Number of follow-ups Average time spent during follow-ups Referral data Referring department Patient staying on multiple departments Type of referring service (In-patient, ICU etc.) Type of referral (normal ad hoc, contract etc.) Timing of referral Urgency Staff consultation Primary reason for referral Additional reasons for referral BACKGROUND Sociodemographic Sex Age Marital status Present living situation Educational level Type of occupation Employment status Health care organization and patient status before admission Psychiatric care last 5 yr Physical care last 5 yr Global Assessment of Functioning Scale best and worst last year (DSM-III-R) Motility status best and worst last year Mental health out-patient treatment status at admission In-patient treatment status at admission Known at own service	Health care organization and patient status at first consult Reaction level scale (RLS85) Global Assessment of Functioning Scale (GAF) Motility status Other psychosocial services involved Somatic diagnosis (ICD-9) Somatic diagnosis at first consult Additional somatic diagnoses (two) Etiology Tracts Specific treatment modalities Pregnancy Psychiatric diagnosis (ICD-10) Psychiatric disturbance leading to referral Additional psychiatric diagnoses (two) V-codes INTERVENTION Interventions/management of care as initiated by consultant Diagnostic action Obtain information from external sources Influence level of medical management Medication initiated by consultant Medication changed or stopped by consultant Psychological and behavioural approach Most important target of intervention Written information for the consultee/ward staff Non-medical consultations OUTCOMES Health care organization and patient status at discharge Reaction level scale (RLS85) Global Assessment of Functioning Scale Motility status Death of patient Influence discharge date Formulation of post-discharge treatment plan Way of communication with post-discharge health care In-patient health care status after discharge Out-patient mental health care arrangements after C-L intervention

calculated primarily for nominal variables, but also for ordinal variables, provided their distribution was not very asymmetrical. Percentage agreement (PA) was calculated both for ordinal and nominal variables, including those whose distribution was very asymmetrical. Finally, intraclass correlation coefficients (ICC) were calculated for continuous, but also for some ordinal variables, provided their distribution was not very asymmetrical. The statistical methods selected are discussed below.

A total of 19 PRF items were not included in the calculations of agreement coefficients. Nine of them relate to dates, time spent in consultation, etc. The others are: secondary or additional items; those whose options are too obvious, such as sex or age; items with multiple answers; and, finally, the items considered

Table II.—Number of sites and participating consultants in ECLW study and proportion of “reliable raters” (RR), by European country

Country	Sites <i>N</i>	Consultants <i>N</i>	RR <i>N</i> (%)
Belgium	7	10	10 (100%)
Finland	6	31	28 (90.3%)
France	2	2	2*
UK	11	14	13 (92.9%)
Italy	5	11	9 (81.8%)
Netherlands	9	34	32 (94.1%)
Norway	5	12	12 (100%)
Portugal	17	48	47 (97.9%)
Spain	3	7	6 (85.7%)
Sweden	1	1	1*
Switzerland	2	3	2*
Germany	18	43	39 (90.7%)
Greece	3	3	3*
Lithuania	1	1	1*
Totals	90	220	205 (93.2%)

* Per cent not calculated, because of low number of raters.

unsuitable on the basis of disagreement between the PMG and NCs, namely: ‘Type of occupation’, ‘Written information for the consultee’ and ‘Out-patient mental health care arrangements after C–L intervention’.

RESULTS

Two hundred and twenty consultants from 14 different European countries and 90 different sites participated in the reliability study (Table II). Portugal and Germany were the countries with the highest number of consultants and centres involved. The consultants reported that the PRF is both, ‘feasible’ and ‘acceptable’. Once they had rated 5–10 PRFs, the time taken to score the form ranged between 10 and 15 min after completing the consultation process. When the ratings of the 13 case vignettes were completed, 205 consultants (93.2%) fulfilled the criteria required for consideration as ‘reliable raters’ (RR) (Table II).

Only eight of the PRF items selected for the calculation of agreement coefficients between rater and ‘gold-standard’ (nine items if the psychiatric diagnosis is included [6]) fulfilled the criteria for use of κ . These mean agreement coefficients range from 0.70 to 0.88 (Table III). Percentage agreement (PA) in these eight items, which does not appear in the table, was also calculated, the coefficients being only slightly above the κ ones. The agreement with the ‘gold standard’ tends to be best in highly structured questions, such as the variable ‘Referring Department’ or questions with only two choices, such as ‘Medication initiated’. The mean agreement in ‘Somatic Diagnosis’ is $\kappa = 0.72$. The proportion of raters having satisfactory κ coefficients (≥ 0.7) is quite high in most items; it was 56.5% in the item ‘somatic diagnosis’ (Table III).

Table III also reports the mean percentage agreement (PA) coefficients between rater and ‘gold-standard’ calculated for the selected items which did not fulfil the

Table III.—“Rater-Gold standard” reliability coefficients ($N = 220$) of Patients’ Registration Form (PRF) and number and proportion of raters scoring $\kappa \geq 0.7$ (percentage agreement ≥ 0.7 where κ was not calculated) for each individual item on the PRF

Variable	Mean	SD	Kappa (κ)	
			Raters scoring	PA ≥ 0.7
			N	(%)
Referring department	0.88	0.05	214	97.3
Urgency	0.76	0.19	187	85.0
Marital status	0.73	0.07	176	80.0
Present living situation	0.70	0.13	145	66.0
Somatic diagnosis	0.72	0.14	124	56.5
Medication initiated	0.88	0.13	185	84.1
Psychological & behavioral approach family general	0.83	0.18	168	76.4
Way of communication	0.72	0.15	128	58.3

Variable	Mean	SD	Percentage Agreement (PA)	
			Raters scoring	PA ≥ 0.7
			N	(%)
Multiple departments	0.96	0.08	218	99.1
Type of service	0.98	0.04	219	99.5
Type of referral	0.81	0.10	194	88.3
Staff consultation	0.90	0.13	196	89.1
Timing of referral	0.86	0.30	182	82.7
Primary reason referral	0.75	0.08	134	61.1
Employment status	0.60	0.13	38	17.4
Educational level	0.62	0.18	62	28.2
Psychiatric care before admission	0.88	0.08	209	95.0
Physical care before admission	0.62	0.17	57	26.0
Mental health out-patient treatment status at admission	0.87	0.10	201	91.4
Known at own service	0.95	0.11	214	97.3
Inpatient treatment status at admission	0.93	0.12	205	93.2
Reaction level scale 85 (at first consultation)	0.91	0.07	218	99.1
Additional diagnostic procedures	0.91	0.03	219	99.5
Medical/Surgical consultation	0.89	0.07	217	98.6
Psychometric assessment	0.93	0.09	213	96.8
Information from medical sources	0.83	0.16	170	77.3
Information from social service sources	0.98	0.04	219	99.5
Information from mental health sources	1.00	0.03	219	99.5
Information from family sources	0.93	0.07	218	99.1
Information from other sources	0.99	0.04	219	99.5
Influence medical treatment	0.72	0.21	118	53.6
Psychological & behavioural treatment, Patient general	0.92	0.10	212	96.4
Psychological & behavioural treatment, Staff general	0.70	0.16	108	49.1
Psychological & behavioural treatment, None	0.98	0.05	219	99.5
Most important target of intervention	0.79	0.12	168	76.4
Reaction level scale 85 (at last consultation)	0.84	0.13	199	90.5
Death of patient	0.99	0.04	219	99.5
Influence discharge date	0.55	0.14	23	10.5
Formulation post-discharge treatment plan	0.83	0.12	186	84.5
In-patient health care status after discharge	0.95	0.09	214	97.3

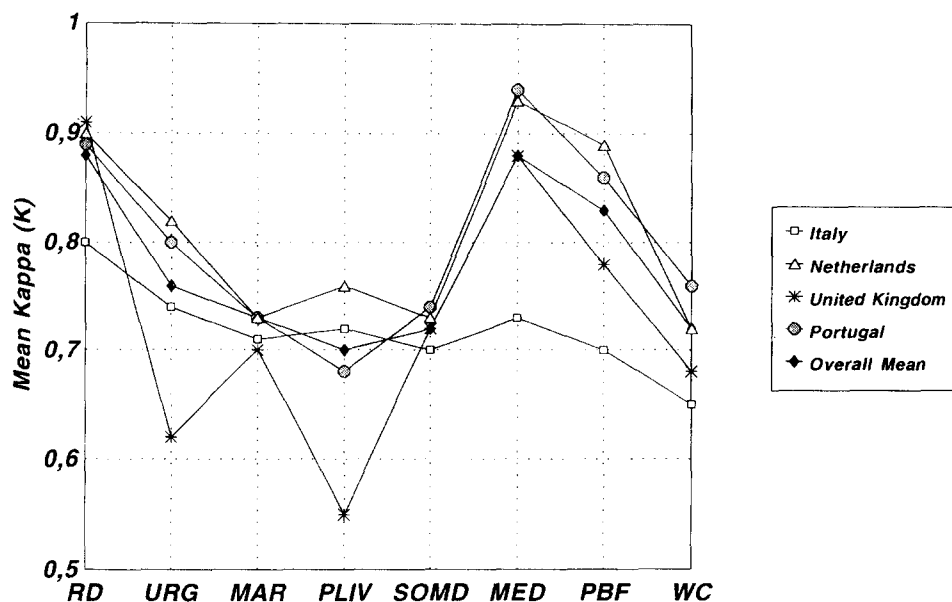


Fig. 2. Overall mean kappa (κ) coefficients per item and mean κ coefficients per country, only in countries where extreme deviations from the mean were observed. RD = Referring Department; URG = Urgency; MAR = Marital Status; PLIV = Present Living Situation; SOMD = Somatic Diagnosis; MED = Medication Initiated; PBF = Psychological and Behavioral Approach, Family General; WC = Way of Communication.

mentioned criteria for calculation of κ . The coefficients were quite acceptable in most items, but were low (<0.7) in the following: 'employment status', 'educational level', 'physical care before admission' and 'influence discharge date'. The standard deviations tend to be rather high in such cases, yet further analysis seemed to be appropriate owing to the 80% agreement by the NCs: the proportion of raters having satisfactory agreement coefficients (≥ 0.7) was low in all four of these items (Table III). The proportion of consultants with satisfactory coefficients, however, was quite high in most other items. Only the items related to motility status and the DSM-III-R's Global Assessment of Functioning scale (GAF) [14] fulfilled our criteria for calculation of intraclass correlation coefficients (ICC). This part of the study will merit a separate report in the future.

The possibility of national biases influencing the reliability coefficients on the PRF has been assessed in different ways. Previously, nations with less than seven participating consultants were excluded from this part of the analysis. Firstly, the proportion of 'reliable raters' (RR) per country was calculated. Table II shows that the proportion of RR ranges from 81.8 to 100% and in no participating country is far from the mean.

Secondly, the mean kappa coefficients per item were calculated for each of the participating countries. No important differences were observed between nations. Figure 2 shows only the results for the countries which deviate most from the overall mean. On the one hand, the coefficients for both Portugal and The Netherlands were only slightly higher than the mean; on the other hand, both those of Italy and

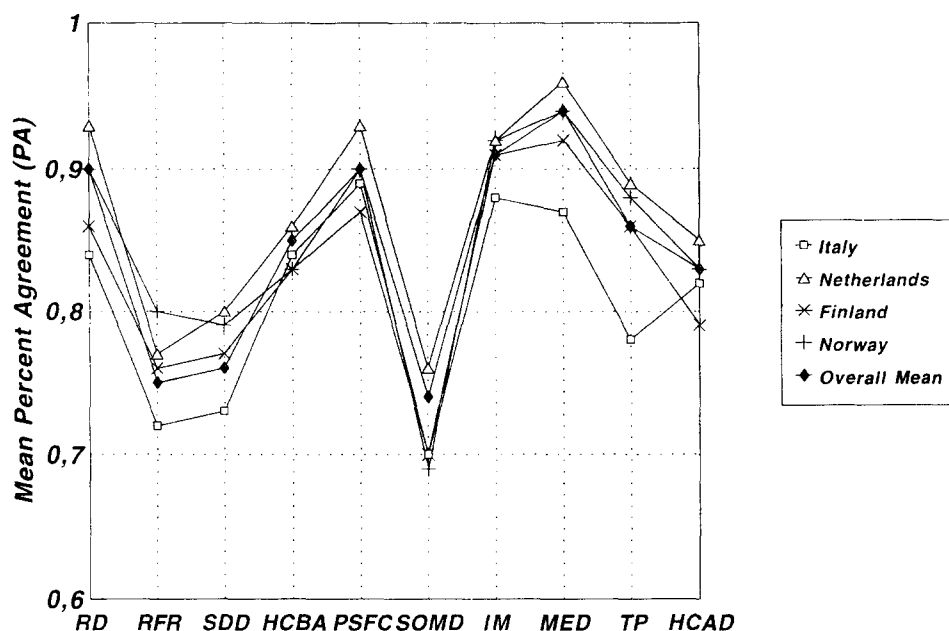


Fig. 3. Overall mean percentage agreement (PA) coefficients per section and mean PA coefficients per country, only in countries where extreme deviations from the mean were observed. RD = Referral Data; RFR = Reason for Referral; SDD = Sociodemographic Data; HCBA = Health Care Organization and Patient Status Before Admission; PSFC = Health Care Organization and Patient Status at First Consultation; SOMD = Somatic Diagnosis; IM = Intervention/Management of Care as Initiated by Consultant; MED = Treatment Medication; TP = Treatment Psychological and Behavioral Approach; HCAD = Health Care Organization and Patient Status After Discharge.

the UK were only slightly lower than the mean. However, the deviations are unequally distributed: they are minimal in items such as 'marital status' and are maximum in items such as 'urgency', 'present living situation' or 'medication initiated'. The deviations occur in the same items in which high standard deviations had been observed in the overall calculations (Table III).

Finally, mean percentage agreement (PA) coefficients were calculated for each one of the PRF sections and for each of the participating countries. Once again, important differences were not observed between nations. Figure 3 only shows the results for the countries which deviate most from the overall mean. On the one hand, the coefficients for both The Netherlands and Norway were only slightly higher than the mean; on the other, the coefficients for both Italy and Finland were only slightly lower than the mean. The deviations were rather similar across the different sections.

DISCUSSION

The Patient Registration Form or PRF was intended to include data relevant to C-L work, allowing the comparison of findings in different European countries. It was also intended to allow the identification of differences between services, in their way of dealing with C-L work. Such a new instrument should fulfil at least minimal

validity standards, that is, face and content validity. The standards for these types of validity are debatable [15, 16], but we feel confident that the PRF fulfils both criteria, on the basis of the following considerations.

Face validity refers to the judgment that the new instrument makes sense to an investigator and therefore increases acceptability by clinicians. In support of this, consultants from university and non-university sites throughout Europe reported that the PRF is both 'feasible' and 'acceptable'. While these characteristics were not formally tested and some criticism emerged initially in relation to the complexity of the instrument, experienced C-L psychiatrists could successfully fill in the final version of the form in an average of 10–15 min after a regular consultation process.

Content validity refers to the systematic examination of the new instrument by experts in the discipline, to ensure that its items cover the type of information that would be needed for subsequent interpretation [15]. The question might be whether or not this registration form covers a complete description of the patient and the service delivery in consultation. This aspect has been assured in the process of development of the PRF. It was examined by different C-L investigators and experts, from different European and non-European countries. It was only after a long and careful developmental period that the final version of the instrument was decided upon.

The main purpose of the study was to assess the reliability of consultants across a large variety of European settings in rating the PRF. The overall results are satisfactory: the proportion of reliable raters was high (93.2%); where kappa were calculated, the mean agreement coefficients for each item were quite acceptable; and only four items had 'low' mean percentage agreement coefficients (<0.7). Furthermore, it is remarkable that these findings were consistent cross-nationally and, certainly, no North–South or East–West regional differences could be detected. While the agreement coefficients tended to be lower in some countries, the deviations from the overall mean were minimal. They tended to be consistently higher in The Netherlands. However, this might be explained by the fact that in this country since 1983 a group of C-L psychiatrists, the NCCP, which included the Dutch participating centres, had been involved in this process of rating their activities [17]. Cross-national issues may be important variables lost in the attempt to achieve reliability. However, we do not believe important losses have occurred in this investigation, and only three PRF items were not included in the calculations of reliability for reasons of disagreement between the PMG and NCs.

It is uncommon to find studies reporting the proportion of unreliable raters identified in the international literature [18]. This study illustrates the importance of doing so: only 15 consultants (6.8%) were not considered to be 'reliable raters' and needed further training, including supervision, in order to participate in the field study. Our investigation also illustrates that mean average coefficients may obscure some of the less positive findings: a proportion of raters had low agreement coefficients (<0.7) in some items, although the mean overall results seemed to be acceptable. However, this statement must be qualified: the proportion of raters with low coefficients may improve considerably by lowering the cut-off point only slightly. A dramatic example of this is shown in the item 'somatic diagnosis': while 43.5% of the raters had a 'low' coefficient of agreement ($\kappa < 0.7$) on this item, only 18.6% had a coefficient < 0.67 .

The following PRF items were identified in this investigation because the mean agreement coefficients were considered to be low ($PA < 0.7$): 'employment status', 'educational level', 'physical care before admission' and 'influence the discharge date'. Contrary to a priori expectations, some of the main problems were related to demographics, and not to variables such as diagnosis. We do not believe this reflects a flaw in the design of the case histories. They had been selected in such a way that the degree of difficulty varied among them. Rather, the discussion of disagreements in the research group revealed that cultural differences were responsible for the difficulties encountered with the items on employment status and educational level: the unreliability seemed to be related to the procedure of written test cases. As a result, the NCs made special national instruction adjustments for the field study. For the item 'physical care before admission', the cause for unreliability was considered to be in the area of the case vignettes; consequently, no readjustments were made. The item 'influence the discharge date', where a considerable degree of inference was necessary in the original instructions, required a more specific operational definition in the manual.

Several reasons support the decision to select the 'case-record' method for assessing reliability in this investigation. Firstly, cost-constraints and logistics, since it was a large, multi-center, cross-cultural study. Secondly, the possibility of pre-selecting the cases so that they are representative of the defined population. While 13 cases might not identify the potentially extensive number of presentations that are less reliable, both 'easy' and 'difficult' cases were included in this study. Thirdly, information variance is eliminated in the vignette design and the rules for categorization are clearly assessed. The main sources of variability that influence agreement here are variations between diagnosticians in their interpretation of criteria and in interpretation of vignettes [13]. This method permits the quick identification of outliers in need of further training.

Some authors have considered the case record method as the least stringent one [13] or as tending to yield a false high reliability [19]. However, Kendell [20] found that the inter-rater agreement of psychiatric diagnosis was as good when the transcript was used as with the full videotape. Spitzer and Williams [21] have argued that, contrary to what one might expect, the reliability obtained by using case records is generally lower than the reliability of assessments based on live interviews. The explanation might be that live interviews provide important diagnostic information that is left out of the case records or, conversely, that case records frequently provide ambiguous information that interferes with making a differential diagnosis.

Videotape interviews are very useful for intercentre reliability work in collaborative studies. The main advantage over the previous method is that sources of variability here are not reduced to a minimum and, in this respect, the study is more stringent. However, the PRF includes a considerable amount of information which is not collected in an interview. Interrater designs are considered to be the standard training experience for some studies [13, 19]. They are most helpful when the purpose of the study is to focus on the adequacy of criteria, since they measure the same sources of variation as the videotape studies and may be more cost-effective, since no equipment is needed. In theory, a joint assessment of the patient's chart and consultation process to collect all the information included in an instrument such as the PRF could be made. However, difficulties for intercentre studies, particularly

cross-national studies would be almost insuperable. Similar difficulties would arise in the test-retest design, which introduces substantial sources of variance and is considered to be the most stringent design [13].

Most studies reporting agreement in psychiatric interviews refer to 'reliability' [18]. Authors such as Tinsley and Weiss [22] have argued that reliability represents "the degree to which the ratings of different judges are proportional when expressed as deviations from their means" and have emphasized the importance of describing the manner in which agreement is calculated. In the strict sense, our method of comparing the raters' judgment with a gold standard, resulting from 'consensus ratings' (CR) for each PRF item, would be closer to Tinsley and Weiss' concept of 'interrater agreement'. However, the same authors have indicated that the distinction between interrater reliability and interrater agreement blurs when one deals with nominal scales [22].

The measurement of agreement is difficult, particularly when the variables of interest have more than two possible values. It might be argued that the results of this study are overly optimistic, since percentage agreement (PA), which may be inflated by random and chance agreement [12], was the most frequently calculated coefficient. There is widespread consensus regarding the advantages of the more stringent kappa (κ) coefficient [16–25], but even recent studies report both, PA and κ coefficients [18]. Dewey [26] has argued that κ is also the measure of choice when compared to new measures such as Maxwell's 'random error' coefficient. However, the limitations of κ have also been discussed. The principal weakness is that κ measures the frequency of exact agreement rather than the degree of approximate agreement and might be particularly arbitrary when continuous data are grouped into ordinal data. In such situations, which only happened in our study on the PRF items related to motility status and GAF, intraclass correlation coefficients (ICC) might be the measures of choice [27]. The ICCs of such PRF items will be reported separately.

Further difficulties concerning κ arise in situations of marginal asymmetry [25, 28]. Kappa coefficients are considered to be too restrictive when the distribution of the categories is very asymmetrical. Even high percentage agreement can result in low or negative κ coefficients [29, 30]. Furthermore, for items where the characteristic to be observed is very clear, the restrictive κ would not be an appropriate measure of agreement [31]. Therefore, we have used PA rather than κ in PRF items with very asymmetrical distributions, since PA is still considered to be an acceptable agreement index if used with caution or minor corrections [22]. In fact, in the PRF items where both, PA and κ were calculated, the differences found were minimal.

The most important form of validity or construct validity [15] is difficult to document. Concurrent criterion validity should at least be attempted when developing new instruments such as the PRF. However, a criterion instrument with its own already established validity is lacking in C–L psychiatry. This may lead to instrument-centred information bias [32] and is a recurring problem in rather new fields [15]. In the meantime, it is common practice to rely on minimum validity standards and on reliability, as we did in this study. Given the implications of misclassification for evaluating C–L effectiveness and services, we feel this is an important and unprecedented achievement. The standardized PRF will permit the reliable description

of C–L service delivery, including cost-relevant aspects throughout European countries and eventually also in other countries.

REFERENCES

1. HUYSE FJ and the ECLW: consultation–liaison psychiatry: does it help to get organized? *Gen Hosp Psychiatry* 1991; **13**: 183–187.
2. HUYSE FJ, HERZOG T, MALT UF, LOBO A: The effectiveness of mental health service delivery in the general hospital. In *Health Services Research* (Edited by Fracchia GN, Theofilatou M) pp. 227–242. Amsterdam: IOS Press, 1993.
3. HUYSE FJ and Members of the ECLW. European Consultation/ Liaison Workgroup (ECLW) for General Hospital Psychiatry and Psychosomatics. *Gen Hosp Psychiatry* 1991; **13**: 383.
4. HUYSE FJ, HERZOG T, MALT UM, LOBO A and the ECLW: The European Consultation–Liaison Workgroup (ECLW) Collaborative Study. I. General outline. *Gen Hosp Psychiatry* (in press).
5. MALT UM, HUYSE FJ, HERZOG T, LOBO A, RIJSSENBEK APMM, and the ECLW. The European Consultation–Liaison Workgroup (ECLW) Collaborative Study. III. Training of reliability and assessing of ICD-10 psychiatric diagnoses in the general hospital setting. *J Psychosom Res* (in press).
6. HERZOG T, HUYSE FJ, MALT UM, LOBO A, STEIN B, and the ECLW. The European Consultation–Liaison Workgroup (ECLW) Collaborative Study. IV. Assessment of institutional and provider factors. *Gen Hosp Psychiatry* 1994; (submitted).
7. HAMMER JS, STRAIN JJ, LYONS JS. Consortium-based Consultation/Liaison research. *Int J Psych Med* 1987; **17**: 237–248.
8. HUYSE FJ, HENGVELD MW, STRAIN JJ, HAMMER JS, ZWAAN T. Interventions in consultation–liaison psychiatry: The development of a schema and checklist for operationalized interventions. *Gen Hosp Psychiatry* 1988; **10**: 88–101.
9. HERZOG T, HARTMANN A. Psychiatrische, psychosomatische und medizinpsychologische Konsiliar- und Liaisonstätigkeit in der Bundesrepublik Deutschland. Ergebnisse einer Umfrage. *Nervenarzt* 1990; **61**: 281–293.
10. WORLD HEALTH ORGANIZATION. *International Classification of Diseases, 1975 Revision (ICD-9)*. Geneva: World Health Organization, 1977.
11. WORLD HEALTH ORGANIZATION. *The ICD-10 Classification of Mental and Behavioral Disorders. Clinical Descriptions and Guidelines*. Geneva: World Health Organization, 1992.
12. COHEN J. A coefficient of agreement for nominal scales. *Educ Psychol Measur* 1960; **20**: 37–46.
13. GROVE MW, ANDREASEN MC, MCDONALD-SCOTT P *et al*. Reliability studies of psychiatric diagnosis. *Arch Gen Psychiat* 1981; **38**: 403–413.
14. AMERICAN PSYCHIATRIC ASSOCIATION. *Diagnostic and Statistical Manual of Mental Disorders*, Third Edn, Revised. Washington DC: American Psychiatric Association, 1987.
15. REGIER DA, BURKE JD. Quantitative and experimental methods in psychiatry. In *Comprehensive Textbook of Psychiatry* Fourth Edn (Edited by Kaplan HI, Sadock BJ), pp. 308–326. Baltimore: Williams and Wilkins, 1989.
16. LOBO A, CAMPOS R, PÉREZ-ECHEVERRÍA MJ, IZUZQUIZA J, GARCÍA-CAMPAYO J, SAZ P, MARCOS G. A new interview for the multi-axial assessment of psychiatric morbidity in medical settings. *Psychol Med* 1993; **23**: 505–510.
17. HENGVELD MW, ROOIJMANS HGM, VECHT-VAN DEN BERGH R. Psychiatric consultations in a Dutch University hospital: A report on 1814 referrals, compared with a literature review. *Gen Hosp Psychiatry* 1984; **6**: 271–279.
18. WITTCHEN HU, ROBINS LN, COTTLER LB, SARTORIUS N, BURKE JD, REGIER D. Cross-cultural feasibility, reliability and sources of variance of the Composite International Diagnostic Interview (CIDI). *Br J Psychiatry* 1991; **159**: 645–653.
19. BECH P, MALT UF, DENCKER SJ, AHLFORS UG, ELGEN K, LEWANDER T, LUNDELL A, SIMPSON GM, LINGJAERDE O. Scales for assessment of diagnosis and severity of mental disorders. *Acta Psychiat Scand* 1993; **87**: Suppl. 372.
20. KENDELL RE. Psychiatric diagnosis: A study of how they are made. *Br J Psychiatry* 1973; **122**: 437–445.
21. SPITZER RL, WILLIAMS JBW: Classification of mental disorders. In *Comprehensive Textbook of Psychiatry* Fifth Edn (Edited by Kaplan HI Sadock BJ), pp. 591–613. Baltimore: Williams and Wilkins, 1985.
22. TINSLEY HEA, WEISS DJ. Interrater reliability and agreement of subjective judgements. *J Counsel Psychol* 1975; **22**: 358–375.

23. SHROUT PE, SPITZER RL, FLEISS JL. Quantification of agreement in psychiatric diagnosis revisited. *Arch Gen Psychiatry* 1987; **44**: 172–177.
24. BERRY KJ, MIELKE PW. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educ Psychol Measur* 1988; **48**: 921–933.
25. CICHETTI DV, FEINSTEIN AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990; **43**: 551–558.
26. DEWEY ME. Coefficients of agreement. *Br J Psychiatry* 1983; **143**: 487–489.
27. MACLURE M, WILLET WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; **126**: 161–169.
28. COLLIS GM. Kappa, measures of marginal symmetry and intraclass correlations. *Educ Psychol Measur* 1985; **45**: 55–62.
29. BRINK W VAN DER. *Meeting van DSM-III Persoonlijkheidspathologie. Betrouwbaarheid en validiteit van de SIDP-R en de DSM-III*. Groningen: Van Dederen B.V., 1989.
30. FEINSTEIN AR, CICHETTI DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; **43**: 543–549.
31. MAXWELL AE. Coefficients of Agreement between observers and their interpretation. *Br J Psychiatry* 1977; **130**: 79–83.
32. LEVENSON JL, COLENVA C, LARSON DB, BARETA JC. Methodology in consultation–liaison research: a classification of biases. *Psychosomatics* 1990; **31**: 367–376.