

S022-3999(96)00213-9

## THE ECLW COLLABORATIVE STUDY: III. TRAINING AND RELIABILITY OF ICD-10 PSYCHIATRIC DIAGNOSES IN THE GENERAL HOSPITAL SETTING—AN INVESTIGATION OF 220 CONSULTANTS FROM 14 EUROPEAN COUNTRIES

ULRIK F. MALT,\* F. J. HUYSE,† T. HERZOG,‡ A. LOBO,§  
A. J. M. M. RIJSSENBECK† and THE ECLW

(Received 4 October 1994; accepted 26 June 1996)

**Abstract**—A comprehensive training program for reliable use of the ICD/10 in Consultation–Liaison (C-L) psychiatry was conducted with 220 psychiatrists and psychologists from 14 European countries. The training included rating of written test cases and development of a coding manual to avoid diagnostic pitfalls not addressed in the ICD-10 manual. Following this training, all consultants rated 13 written case histories. One hundred sixty-seven consultants (76%) had a kappa ( $\kappa$ ) of at least 0.70. Only 13 (6%) had a  $\kappa$  0.40. The percentage of high reliability raters was evenly distributed among the different countries. Consultants had some problems in the differentiation between adjustment disorders and depressive disorders, and in the classification of disorders where ICD-10 differs from the DSM-III-R system. National biases in diagnostic practice were found with regard to the “case” concept and the role of alcohol in confusional states. Finnish consultants coded “no psychiatric disorder” significantly more often, whereas German and Italian consultants attributed delirious state more often to alcohol than consultants from other European countries. The study demonstrates that it is possible to achieve acceptable interrater reliability in applying the ICD-10 guidelines, through training programs designed for C-L psychiatrists and psychologists. Nevertheless, this first cross-national study shows the importance of addressing differences in national diagnostic practice. Copyright © 1996 Elsevier Science Inc.

**Keywords:** Diagnosis; Classification; ICD-10; Consultation–liaison psychiatry; Reliability.

### INTRODUCTION

In 1987, the European Consultation Liaison Workgroup (ECLW) was organized as an informal workgroup by European consultation–liaison (C-L) psychiatrists in the field of psychosomatics and psychiatry. A major task of the ECLW was to describe and compare the patient, service, and treatment characteristics of C-L services in

---

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

† Free University Hospital, Free University, Amsterdam, The Netherlands.

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

§ Psychiatric Service, University of Zaragoza, University Hospital, Zaragoza, Spain.

Address correspondence to: Frits J. Huyse, Department of C-L Psychiatry, Free University Hospital Amsterdam, P.O. Box 7057, 1071 MB Amsterdam, The Netherlands. Tel: (011) 31-20-444-0196; Fax: (011) 31-20-444-0197.

different European general hospitals [2]. Within this study, reliable clinical descriptions were mandatory.

A fundamental problem was the lack of agreement that arose in Europe about the most valid way of classifying mental disorders in the different diagnostic traditions and systems used in the different European countries. Diagnostic practice is not only sensitive to political differences but also to the prevailing theoretical orientation of the countries' mental health workers and each country's history of philosophy of science [3]. These differences in theoretical and philosophical tradition still influence classification. Accordingly, the attitudes toward the DSM-III system of classification among professionals varied within Europe [4-6] and the ICD-9 was applied in different ways in each of the European countries [7-8].

The introduction of the ICD-10 [9] with the provision of diagnostic guidelines [10] and research criteria [11] offers new opportunities for reliable diagnostic communication between European countries. However, neither the ICD-10 nor the DSM-III-R systems of classification of mental disorders are particularly tailored for use in consultation/liaison psychiatry and psychology, and several problems of classification commonly seen in C-L settings are not explicitly addressed [1]. We conducted a computer-assisted literature review (*Medline; Psychological Abstract*) in August 1994 and did not find any study addressing the validity or the reliability of the ICD-10 in C-L psychiatry.

One of the aims of the ECLW Collaborative Study (CS) was to overcome these problems by identifying unclear areas and performing practical training in the classification of the disorders frequently seen by C-L psychiatric services.

The aims of the present study are to: (1) describe the methodology used to secure acceptable interrater reliability; (2) report the actual reliability obtained; (3) identify areas of concern which should require special attention in the future teaching of ICD-10 of C-L practitioners; and (4) identify possible national biases with regard to diagnostic practice when using the ICD-10 in a general hospital setting.

## METHOD

### *Participants*

Participants included professionals working within the general hospital setting. Most were psychiatrists, some were psychologists or psychologically trained medical doctors, and a few were C-L nurses. All had reported an interest in participating in the ECLW CS [2, 12]. To qualify for participation in the study, participants had to be trained in the use of the study's patient registration form (PRF) [13] which included items of ICD-10 diagnoses.

A total of 220 consultants from the following 14 countries participated in the reliability training and examination: Belgium ( $n=10$ ); Finland ( $n=31$ ); France ( $n=2$ ); Germany ( $n=43$ ); Greece ( $n=3$ ); Italy ( $n=11$ ); Lithuania ( $n=1$ ); The Netherlands ( $n=34$ ); Norway ( $n=12$ ); Portugal ( $n=48$ ); Sweden ( $n=1$ ); Switzerland ( $n=3$ ); Spain ( $n=7$ ); and the United Kingdom ( $n=14$ ).

### *Developing the training program*

The training of the potential consultants for the study is outlined in Table I. The training of the consultants had to address several problems. At the time this study was started (1990), most consultants had little or no knowledge of the ICD-10. The ICD-10 was not printed, but a preliminary version of the diagnostic guidelines published in 1990 was available, which was later accepted with only minor adjustments. Neither the 1990 version [9] nor the final 1992 version [10], however, addressed diagnostic problems common in general hospital settings. Furthermore, it was impossible to train more than 200 consultants from 14 different countries at the same time.

Table I.—An overview of the methodology of the reliability training and assessment

Content	Target group
1. Introductory symposium on ICD-10	NC
2. Training course ICD-10	NC
Training cases ( $n=11$ )	NC
Test cases ( $n=13$ ) + reliability assessment	NC
Establishing "golden standard" for test cases	NC
3. Booklet with cases	PC
ICD-10 coding guidelines for C-L	PC
Suggestions for specific training courses	PC
4. Assessment of reliability and identification of outliers	PC

NC = national coordinators; PC = participating consultants.

### 1. Introductory symposium

To resolve these problems, we organized a seminar on the use of ICD-10 in C-L psychiatry settings including representatives from each of the countries having expressed an interest in the ECLW CS (national coordinators). The seminar was led by the first author who was involved in the provision of clinical guidelines and research criteria for ICD-10 [10, 11], and had extensive experience with issues of classification and diagnosis [3, 8, 14, 15].

The aims of the seminar were to get to know the ICD-10 and perform diagnostic exercises on written case studies obtained from C-L psychiatry and psychosomatic services in Europe. These cases dealt with the following disorders: Dissociative disorder (F44); Alcohol-induced delirium (F10); Somatoform autonomic dysfunction (F45.3); Delirium not induced by alcohol of psychoactive substance (F05); Anorexia nervosa (F50.0); Adjustment disorder (F43.2); Dissociative disorder (F44); Other mental disorders due to brain damage and dysfunction and to physical disease (F06); Delirium (F05); Bipolar disorder (F31); and Alcohol-induced mental disorder (F10).

The emphasis was placed on identifying, discussing, and clarifying areas of confusion in the ICD-10 guidelines with regard to classification decisions. The differences between the ICD-10 and national traditions and the DSM-III-R system, were stressed.

### 2. Training course for national coordinators

Based on these theoretical and practical experiences, a specific training course on ICD-10 was organized by the first author at the University of Freiburg 1990. During this course, all national coordinators discussed and agreed upon the ICD-10 diagnosis of 11 written case studies (Table II). These 11 case stories were later distributed in a booklet as training cases to all consultants who expressed interest in participating in the ECLW CS. The booklet was accompanied by a separate list of correct diagnoses with comments as decided upon by the national coordinators.

During this course, another 13 test cases were diagnosed by the national coordinators without prior discussion. The 13 cases were chosen from clinical cases presented from different countries. They were selected to give broad and sufficient coverage of the main diagnostic problems consultants face in consultation/liaison work. After the ratings, the national coordinators discussed the cases extensively and agreed upon the correct ICD-10 diagnoses, the "gold standard." Following this procedure, the overall kappa ( $\kappa$ ) of each of the NC was calculated based on the 13 cases. These 13 cases were later used as cases for the ICD-10 interrater reliability test of the study (Table III).

### 3. A training booklet with cases

During the discussions with the national coordinators, unclear aspects of the ICD-10 guidelines for use in C-L psychiatry and psychosomatics were recorded and the solutions agreed upon. These comments and guidelines were used to develop coding guidelines for ICD-10 use in C-L psychiatry and psychosomatics [1]. The coding guidelines correspond to the 1992 ICD-10 clinical guidelines in content, but with additional guidelines for problems not addressed in the ICD-10 manual:

1. The clinicians could record up to three diagnoses.
2. Clinicians were to give priority to diagnosis of the disorder leading to referral or contact with the C-L service. When reviewing the patient's whole case history the most important diagnosis may well be the "lifetime" diagnosis, which may be different from the one most relevant to the immedi-

Table II.—Training cases used in a study of the reliability of ICD-10 diagnosis in consultation liaison work

Case #	Background information	Reason for referral of patient to C-L service	ICD-10 diagnosis
1	Male, 29 years; Neurosurgery; injury with blindness	Behavioral disturbance	F05 Delirium; F10 Alcohol (dependence)
2	Female, 19 years; Neurology; injury, apallic	Parents' behavior	No diagnosis
3	Male, 28 years; Neurology; neuropathy	Hallucinations, delusions	F18 Volatile solvent (psychotic)
4	Female, 49 years; ear nose throat pain	Claims fishbone as cause; no physical findings	F22 Persistent delusional disorder
5	Female, 50 years; Dermatology; psoriasis	Dysphoric	F10 Alcohol (harmful or dependent), F34 Persistent mood disorder
6	Male, 38 years; Neurology; volatile solvent encephalopathy(?)	Rule out of psychiatric etiology	F33 Recurrent depressive disorder
7	Male, 20 years; Dermatology; erythema	Excessive emotional responses to his skin problems	F40.1 Social phobia
8	Female, 28 years; Burn Unit; burn injury	Emotional numbness and withdrawal	F43.1 Posttraumatic stress disorder
9	Male, 61 years; Surgery; Wilson's disease	Liver transplant (routine assessment)	F43.2 Adjustment disorder (prolonged depressive reaction)
10	Male, 48 years; Neurology; epilepsy	Symptoms partly unrelated to EEG findings	F54 Psychological factors associated with physical disorder
11	Male, 35 years; Urology; spinal cord injury, uropathy	Behavioral dysfunction	First: F60 Personality disorder (dissocial and emotionally unstable); later: F06 Mental disorder due to brain damage and dysfunction and physical disease

Table III.—Consultation liaison cases used to assess the interrater reliability and national bias in the ECLW CS

Case #	Background information	Reason for referral of patient to C-L service	ICD-10 diagnosis (primary)
1	Male, 29 years; Neurology; paraparesis	Neurological findings inconclusive	F44 Dissociative motor disorder
2	Female, 79 years; Pulmonology; dyspnea	Agitation	F32 Depressive episode
3	Female, 70 years; Neurology; low back pain	Case on routine screening with GHQ-28	F32 Depressive episode
4	Female, 35 years; Neurology; migraine	Emotional problems?	F54 Psychological factors associated with physical disorder
5	Male, 28 years; Neurology; headache	Fear of cancer	F45.2 Hypochondriacal disorder
6	Male, 22 years; Neurology; asthenia	Negative findings	F45.0 Somatization disorder of F48 Neurasthenia
7	Female, 21 years; Medicine; abdominal pain	Physical findings do not explain symptom	F43.2 Adjustment disorder (brief depressive reaction)
8	Female, 19 years; Medicine; systemic lupus erythematosus	Agitated	F06 Mental disorder due to brain damage, dysfunction, or physical disease
9	Female, 35 years; Neurology; multiple sclerosis	Wants to talk about her emotional response	No psychiatric diagnosis
10	Female, 64 years; Dermatology; infested with parasites	Not verified	F22: Persistent delusional disorder
11	Male, 86 years; Ear Nose Throat	History of drinking	F10 Disorder due to alcohol
12	Male, 30 years; Thoracic Surgery; stenosis arteria pulmonalis	Anxiety	F40.01 Agoraphobia with panic disorder
13	Female, 71 years; Traumatology; hip fracture	Confused and anxious	F05 Delirium (Sec. diagnosis: F32 Depressive episode)

ate consultation. In such instances, the lifetime diagnosis should be recorded as the second or third diagnosis.

3. If there was any doubt about the order in which to record several diagnoses, the clinicians were asked to record the diagnoses in the numerical order in which they appear in the classification. For instance, the classification of "Organic mental disorders" (F0) was given priority over "Neurotic, stress-related, and somatoform disorders" (F4).
4. The clinicians were urged to come to a diagnostic conclusion. Therefore, "diagnosis deferred" should be avoided whenever possible. In the few cases that diagnoses could not be made, however, "diagnosis deferred" was a diagnostic possibility (item 42 in the manual).
5. The participants were asked to be cautious in making diagnoses. If there was presence of a psychiatric disorder, the patient should be rated as having "no diagnosis."
6. In C-L practice, we very often see patients with physical conditions to which psychological factors are judged to be contributory; for example, tension headaches and peptic ulcers. The reliability of this category is doubtful, because discussions suggested clearly that this diagnosis was influenced by the theoretical training of the consultants. Psychodynamically trained consultants would more easily use this category as compared to more biologically or behaviorally oriented consultants. Therefore, we advocated that this category only be used when the patient did not qualify for another mental disorder.

In addition to these general coding guidelines, we made specific coding guidelines with regard to the diagnosis and differential diagnosis of mood disorders versus stress-related and somatoform disorders, the delineation of mood disorders from adjustment disorders and grief, and the classification of somatoform and dissociative disorders [1].

#### 4. Guidelines for training courses for participants

Based on the national coordinators' experience with ICD-10 diagnostic skills, and the first author's previous experience with the teaching of diagnostic classification systems to psychiatrists and psychologists [14], a standardized training program for teaching ICD-10 and its application to consultation-liaison work was designed [1]. All national coordinators found able to apply ICD-10 reliably were asked to organize national training courses along these lines with necessary adjustments. The participants in these national courses were advised to spend most of their time on those areas of the ICD-10 that differed from ICD-9 and DSM-III-R. The necessity of emphasizing the communication aspects of the training (i.e., reliability), and not the validity of the classification system, was restressed.

#### 5. Assessment of reliability and national biases

After the national training, all consultants who expressed their interest in participating in the ECLW CS were given a book with the 13 cases for the assessment of reliability (Table III). The diagnosis of these 13 were to be mailed directly to the ECLW coordination center in Amsterdam, which computed the results and calculated the interrater reliability coefficient of each consultant.

Although the distinction between reliability and validity seems clear in theory, in practice it is often subtle. In this study, we use the diagnosis of the criterion raters as the "gold standard" to which we compare consultant diagnosis. This procedure may be considered an assessment of validity [16]. If one accepts this assumption, the calculation of sensitivity and specificity of consultant diagnosis would be the most appropriate way to present the results of this study.

On the other hand, the present study design emphasizes the reliable use of the ICD-10 system. There was no intention to assess the validity of consultant diagnosis by means of structured interviews with patients whose case reports were considered for this study. From this point of view, one may argue that the results should be presented by means of coefficients of reliability [8] comparing participants to national coordinators.

The best way of assessing reliability has been a source of disagreement among experts. In a study like this, one may argue that a phi-coefficient is most appropriate for calculating reliability [17]. The reliability coefficients obtained by reviewing case records are known to be lower than those obtained by structured psychiatric interviews, because the case records often provide ambiguous information [16]. It is possible to accept the Maxwell model: When in doubt, the consultant randomly assigns questionable cases to different diagnostic categories.

Other investigators strongly argue against this approach and favor the use of kappa statistics because it is appropriate for change agreement. A  $\kappa$  of 0.70 is commonly considered to reflect good reliability and a  $\kappa$  of 0.40 is often considered to reflect the lower limit of reliable interrater reliability. However, an even lower  $\kappa$  value may reflect reliable ratings because of its dependence on the base rate [18]. To avoid this problem we also calculated the overall agreement for concordance. This is analogous to the recommendation of using positive (P pos) and negative (P neg) agreement as two separate indices of proportionate agreement in raters' positive and negative decisions [19].

## RESULTS

In Table IV, percent agreement data for each of the 13 cases are shown together with associated diagnoses. It can be seen that, for most diagnoses, the overall agreement is very good. Less than optimal agreement occurs when the cases address the differentiation of depressive disorder from adjustment disorder (cases 2 and 3); psychological factors affecting physical disorder versus no psychiatric diagnosis (case 4); classification of psychotic disorder with possibly organic etiology (case 10); classification of agoraphobia with panic (case 12); and when deciding about the role of alcohol in a delirious state (case 13). It can also be seen that subclassification of depressive episodes causes some problems. Despite these problems, the majority of cases are classified with acceptable overall agreement using the "gold standard" when F45 and F48 are accepted as correct answers in case 6, and F22 and F06 in case 10 (see Discussion).

On the consultant level, 167 of 220 (76%) consultants were reliable ( $\kappa > 0.70$ ). Another 11% had a  $\kappa = 0.66$ . Only 13 (6%) had  $\kappa < 0.40$ . In countries with more than one consultant, the percentage of reliable consultants varied from 53.5% to 85.3%. A closer look at diagnoses, given by country, did suggest some national biases, however. Of the 403 classifications made by the Finnish consultants, 13.2% were "no mental disorder" compared to 7.7% among the German consultants' classifications and 8.2% of the Portuguese classifications. Delirium was more often attributed to alcohol by the German participants than those from other countries (Table V). The diagnosis of F10 Alcohol was indicated more frequently by German consultants than by those from Finland, Norway, and the United Kingdom.

## DISCUSSION

### *Overall reliability*

This is the first cross-national study addressing interrater characteristics of ICD-10 (Chapter F, mental disorders) diagnoses made by professionals working in consultation-liaison psychiatry and psychosomatics. The main finding is that there is very good overall diagnostic reliability, considering all the diagnostic alternatives available and the limited information available in the test cases. The percent agreement and  $\kappa$  values are better than the reliability coefficients for somatoform disorders found in the DSM-III field trials [20] and better than the reliability coefficients reported from the German ICD-10 interrater reliability studies [21, 22].

Our results are also much better than those obtained by the case vignette method in primary care [23], however. As expected, the  $\kappa$  values found in our study are somewhat lower than what has been reported in studies applying checklists or structured psychiatric interviews [24, 25]. The reliability coefficients obtained are even more impressive considering that all participants (except Germans) had to rate case vignettes written in English. Although most consultants had a working knowledge of English, the consultants do not use English in their practice. This language problem may impair reliability.

The most important causes of variability that influence agreement are variations between diagnosticians in interpretation of criteria and interpretation of vignettes. Other possible factors such as patients' accounts, interviewing style, and different interpretations of patients' nonverbal behavior cannot be varied in case vignette de-

Table IV.—Overall agreement of diagnoses based on 220 consultants from 14 European countries.

Case #	Background information	Percent correct diagnosis (most frequent alternative)	ICD-10 diagnosis (primary)
1	Male, 29 years; paraparesis; neurology neg.	90% (none 3.2%)	F44 Dissociative motor disorder
2	Female, 79 years; depressed, agitated, dyspnea	53.6% (F43: 37.7%)	F32 Depressive episode
3	Female, 70 years; depressed with low back pain	61.8% (F33: 17.7%) (F43: 6.8%) (none 4.5%)	F32 Depressive episode
4	Female, 35 years; migraine + current emotional problems	74.5% (none 13.2%) (F45: 4.1%)	F54 Psychological factors associated with physical disorder
5	Male, 28 years; headache, fear of cancer, no finding	92.3% (F40: 3.2%) (none 2.3%)	F45.2 Hypochondriacal disorder
6	Male, 22 years; asthenia, multiple somatic symptoms, no findings	77.7%, 19.1% (none 1.4%)	F45.0 Somatization disorder of F48 Neurasthenia
7	Female, 21 years; distress, depressed, abdominal pain	80.9% (F32: 6.8%) (F45: 6.4%)	F43.2 Adjustment disorder (brief depressive reaction)
8	Female, 19 years; manic-agitated, 70 mg/day steroids, SLE	82.3% (F30: 8.2%) (F05: 3.2%)	F06 Mental disorder due to brain dysfunction (steroids)
9	Female, 35 years; MS, wants to talk about emotions/thoughts	91.4% (F43: 3.6%) (F54: 1.4%)	No psychiatric diagnosis
10	Female, 64 years; believes infested with parasites, no findings	65.0%, 25.0%	F22: Persistent delusional disorder F06: organic mental disorder (see text)
11	Male, 86 years; drinking problem, uses sleeping pills	92.7% (F51: 2.7%) (none 2.7%)	F10 Disorder due to alcohol
12	Male, 30 years; Thoracic Surgery; anxiety	53.9% (F41: 42.9%)	F40.01 Agoraphobia with panic disorder
13	Female, 71 years; Operated, confused and anxious, alcohol in the past but not now	49.3% (F10: 44.2%) (F06: 3.7%)	F05 Delirium (sec. diagnosis: F32 Depressive episode)



Table V.—National biases in psychiatric diagnoses (including only countries with 10 or more consultants)

Country (diagnoses)	No diagnosis	F10 Alcohol	F05 Delirium
Finland ( <i>n</i> = 403)	13.2*	8.2†	4.2%‡
The Netherlands ( <i>n</i> = 441)	11.1	9.3†	7.0%‡§
Norway ( <i>n</i> = 156)	10.9	8.3	6.4%‡
United Kingdom ( <i>n</i> = 179)	10.6	7.8†	3.9%‡
Belgium ( <i>n</i> = 130)	9.0	11.5	3.8%‡
Portugal ( <i>n</i> = 624)	8.2*	10.9	4.2%‡
Italy ( <i>n</i> = 140)	7.9	13.6	1.4%
Germany ( <i>n</i> = 559)	7.7*	13.4†	0.7%‡

For each of the three diagnostic categories, the expected percentage based on the "gold standard" is 7.7%.

\*Finnish consultants coded "no diagnosis" significantly more often than German ( $\chi^2 = 7.8$ ;  $p = 0.005$ ) and Portuguese ( $\chi^2 = 6.67$ ;  $df = 1$ ;  $p = 0.009$ ) consultants.

†German consultants coded F10 ("Alcohol-related disorder") significantly more often than the consultants from Finland ( $\chi^2 = 6.42$ ;  $df = 1$ ;  $p = 0.01$ ), The Netherlands ( $\chi^2 = 4.08$ ;  $df = 1$ ;  $p = 0.043$ ) and the United Kingdom ( $\chi^2 = 4.0$ ;  $df = 1$ ;  $p = 0.045$ ).

‡German consultants coded F05 ("Delirium due to organic mental dysfunction not caused by alcohol") significantly less frequently than consultants from all other countries.

§Italian consultants coded F05 significantly less frequently than consultants from The Netherlands ( $\chi^2 = 5.22$  [using the Yates correction];  $df = 1$ ;  $p = 0.022$ ).

sign. There is some evidence that the high reliability obtained by rating case vignettes correlates to reliability rating video interviews [23]. Nevertheless, strictly speaking, the present method does not tell us if the consultants are reliable in their day-to-day work. However, case vignettes have many useful features, which are of crucial importance for this type of cross-national study. The method quickly establishes whether raters know the criteria and interpret case records in the same way. By this method, outliers can be identified [16]. These outliers can be given more training while more reliable raters can receive positive feedback. This was exactly the procedure followed in the present study.

Thus, the results obtained should provide sufficient indications to assume, with reasonable confidence, that the overall results of this large study of C-L practice should be reliable enough to produce meaningful results on the group level. Indirectly, our results also support clinical studies demonstrating a good interrater reliability of the ICD-10 system used in clinical settings [26].

### Coding problems

Despite our effort to provide the clinicians with additional guidelines to differentiate depression from adjustment disorder, it was difficult to reach acceptable agreement on this issue (Table IV, case 2). We found no national bias. This suggests that it is the clinical guidelines that are insufficient. In fact, this problem has also been identified in the United States using the DSM-IV system [27]. The problems related to the subclassification of depressive disorder may be further increased by the differences in contents between ICD-10, DSM-III-R, DSM-IV, and research diagnostic criteria (RDC) [28], although the present method does not allow any conclusive statement on this issue.

Subclassification of a depressive episode and correct classification of agoraphobia with panic disorder (Table IV, cases 3 and 12) showed only 61.8% and 53.9% agreement. In both cases, the ICD-10 clinical guidelines contained accurate enough information to make a correct classification, and the case vignettes also contained the information needed for accurate classification. These observations, and the lack of any national bias in the diagnoses, point to problems related to training. In the DSM-III, DSM-III-R, and DSM-IV systems of classification, first and recurrent episodes of depression are differentiated by a fourth digit. In ICD-10, however, these types of depressive episodes are classified as F32 (single episode) and F33 (recurrent episode). In DSM-III, panic disorder was separated from agoraphobia with or without panic disorder as is the case in ICD-10. DSM-III-R changed this practice, however, using panic disorder as the main principle for classification. DSM-IV has again adopted the ICD-10 and DSM-III approach. These formal changes in classification represent the most likely explanation for the lower reliability found in cases 3 and 12. These findings indicate that, when presenting results, F32 and F33 should be combined. Furthermore, agoraphobia with and without panic disorder should be considered as a group in future presentations of results.

### *Problems related to the interpretation of the ICD-10 manual*

In case 6, the patient could be classified as both a somatization disorder and neurasthenia. We encouraged the use of neurasthenia when appropriate. This seems to have led to a tendency by some consultants (19.1%) to consider neurasthenia as the most appropriate diagnosis because fatigue was one of the main symptoms. Others considered somatization to be primary since this number is lower (F45 versus F48). This problem emphasizes the difficult issue of how to classify and diagnose multiple somatic symptoms. The solution to this problem, given the current ICD-10 guidelines and research criteria, is probably to encourage double diagnoses in cases where both disorders are present.

In case 10, the patient was delusional. From a descriptive point of view, F22 should therefore be correct. However, at the time of this reliability study, a complete text of the ICD-10 guidelines became available and, in that text, the diagnosis of "Delusional syndrome" (F06). This exception to the general rule of descriptive coding caused problems and led 25% of the consultants to code F06. Most of these consultants were German who had had access to an updated ICD-10 version [10] in German. Those who relied on the ECLW version [1] rated the case correctly as F22. For this reason, we accepted both F06 and F22 as acceptable diagnoses in case 10. This situation does bring to light the problem of when clinical guidelines deviate from an overall descriptive approach on a specific issue.

### *National diagnostic biases*

Another problem highlighted by this study is the national variation in the classification of alcohol-related disorders and delirium. For many years, the concept of delirium was closely related to alcohol (e.g., delirium tremens). This historical tradition, and the clinical experience in several countries that confusion and agitation are often related to alcohol abuse, may explain why alcohol was considered more often to be the cause of delirium in Germany (Table V).

The higher rating of alcohol problems may also be explained by clinical experience. When exposed to a case vignette in which alcohol is mentioned, the German

consultants may have emphasized alcohol more than those consultants living in countries with less per-capita alcohol consumption. Thus, national bias may be an artifact of the test procedure and may not necessarily mirror true differences in real-life practice.

Whatever the explanation, neither the DSM nor the ICD system has specific laboratory requirements for the diagnosis of alcohol-related delirium. The best judgment of the clinician is the key to diagnosis. Our experiences therefore suggest that national biases may occur in the etiological categorization of confusional states in C-L practice. Other studies have identified major differences in the way ICD-10 and DSM-III-R/DSM-IV classify harmful use of alcohol [29–31]. Therefore, these results and ours do point to a potential problem of reliable use of the F10 category in ICD-10, particularly in relation to alcohol.

Another important diagnostic problem is the definition of a case, an issue of major importance in C-L practice. What is a normal emotional response to having a physical disorder? Our findings point to national biases with regard to this issue. Consultants from “northern” Europe (e.g., Finland) more often use “no diagnosis” compared to Germany and Portugal (Table V). This may be related to the theoretical training of the consultants. Countries like Finland, The Netherlands, and Norway have been heavily influenced by British psychiatry with its emphasis on “if in doubt, don’t” attitude to psychiatric diagnosis. DSM-III has also been most influential in Finland and Norway [5, 14]. This may lead consultants to code “no diagnosis” when all criteria are not fulfilled. Considering the fact that continental and southern countries tended to be more accurate on this issue, our findings may indicate that consultants from countries with a tradition of being “psychologically minded” may run the risk of underestimating psychopathology. Such observations are important when issues like sampling and diagnoses in clinical studies are considered. Interestingly, a Dutch study found the Present State Examination (PSE), a UK-originated instrument, to have low sensitivity for nonpsychotic disorders like OCD and several anxiety disorders, as compared to DSM-III-R–based diagnoses [24].

### *Methodological limitations*

The strength, but also the main flaw, of the ECLW CS is its diversity of raters. The unbalanced sample of participating nations and physicians within the participating countries resulted in a convenience sample that is not representative. The bias introduced by this sampling is unknown. Hence, the reliabilities found in our study are not generalizable. Variation between physicians within individual countries may be greater than intercountry variation. Furthermore, the small sample size in some countries may indicate that the current study does not have sufficient statistical power to detect significant reliability variations.

The training manual did not include comorbidity despite the fact that some of the training and examination cases did include comorbidity cases. Increased agreement may be expected in cases with single medical and psychiatric disease. Multiple medical and psychiatric comorbidity may decrease agreement. This important problem should be included in future studies.

### *Conclusions*

In the largest cross-national diagnostic interrater reliability study ever done in C-L we have been able to train consultants with different diagnostic traditions to

achieve acceptable high interrater reliability. We have also been able to identify outliers, and offer these clinicians more training. On the national level, feedback about risks of biases have been given. It is therefore reasonable to assume that the training programs, the manuals, and the exercises have increased the "diagnostic communication skills" among C-L practitioners in Europe. To develop closer clinical and research collaboration between countries, such communication skills are crucial. We believe this is a major achievement, and very encouraging for the future of C-L psychiatry in Europe.

*Acknowledgments*—In addition to the authors, the following persons were actively involved in the design and conduct of the study as national coordinators: Myriam van Moffaert (Belgium), Pekka Tienari (Finland), Paul Sakkas (Greece), Graca Cardoso and Raul Guimares Lopes (Portugal), Marco Rigatelli (Italy), Maria Dolores Crespo (Spain), Richard Mayou, and Francis Creed (United Kingdom). Data Management and analyses were carried out by Andree J. M. M. Rijssenbeek, Brent C. Opmeer, and Gerrit Koopmans (The Netherlands), and Barabara Stein (Germany). This study was initiated by the European Consultaion/Liaison Workgroup for General Hospital Psychiatry and Psychosomatics (ECLW) grant supported by the European Community's Fourth Medical and Health Research Program COMAC-Health Service Research (Grant No. 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 No. 1-1.5.1030.0075.0) and in The Netherlands by the National Fund for Mental Health Research (Grant No. 90.3594). Additional support has been provided by the Norwegian Research Council for Science and the Humanities (NAVF), the Spanish Fondo de Investigacion Santaria, the Upjohn International Medical Sciences Liaison, and Pfizer International. James J. Strain, Jeffrey S. Hammer, and John S. Lyons (United States); Wim van der Brink and Maarten Koeter (The Netherlands); Graeme Smith (Australia); and David Goldberg (United Kingdom) provided advice. The authors thank Allan House, MD, The General Infirmary, Leeds, for valuable help in preparing the final version of this article.

## REFERENCES

1. The European Consultation-Liaison Work Group. Manual and training guidelines for clinical description and diagnosis of disorders seen by the C-L service within the general hospital. Fourth revision. Amsterdam: ECLW, Free University Hospital, 1990:72.
2. Huyse FJ, Herzog T, Malt UF, Lobo A and the ECLW. The European Consultation-Liaison Workgroup (ECLW) collaborative study. I: General outline. *Gen Hosp Psychiatry* 1996; 18:44-55.
3. Malt UF. Philosophy of science and DSM-III. *Acta Psychiatr Scand* 1986;73(suppl. 328):35-44.
4. Spitzer RL, Williams JBW, Skodol AE, eds. International perspectives on DSM-III. Washington, DC: American Psychiatric Press 1983.
5. Maser JD, Kaelber C, Weise RE. International use and attitudes toward DSM-III and DSM-III-R: growing consensus in psychiatric classification. *J Abnorm Psychol* 1991;100:271-279.
6. Mezzich JE, von Cranach M. International classification in psychiatry. New York: Cambridge University Press 1988.
7. Zigmond AS, Sims ACP. The effect of the use of the ICD 9th revision upon hospital in-patient diagnoses. *Br J Psychiatry* 1983;142:409-413.
8. Torgersen T, Rosseland LA, Malt UF. Coding guidelines for ICD-9 section on mental disorders and reliability of chart clinical diagnosis. *Acta Psychiatr Scand* 1990;81:62-67.
9. Sartorius N, Jablensky A, Cooper JE, Bruke JD. Psychiatric classification in an international perspective with special reference to Chapter V (F) of the 10th revision of the International Classification of Diseases "Mental, behavioural and developmental disorders." *Br J Psychiatry* 1988; 152(suppl 1).
10. Diagnostic description and clinical guidelines for the ICD-10 Chapter F (Mental disorders and behavioural dysfunction). Geneva: WHO 1992.
11. Research criteria for the ICD-10 Chapter F (Mental disorders and behavioural dysfunction). Geneva: WHO 1993.
12. Herzog T, Huyse FJ, Malt UF, 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* (in press).
13. Lobo A, Huyse FJ, Herzog T, Malt UF, Opmeer BC and the ECLW. The ECLW collaborative

- study. II: Patient registration form (PRF) instrument, training and reliability. *J Psychosom Res* 1996;40:143–156.
14. Malt UF. Teaching DSM-III to clinicians. *Acta Psychiatr Scand* 1986;73(suppl 328):68–75.
  15. Bech P, Malt UF, Dencker SJ, Ahlfors UG, Elgen K, Lewander T. Rating scales for psychiatric disorders. *Acta Psychiatr Scand* 1993;87(suppl 372).
  16. Grove WM, Aadreasen NC, McDonald-Scott P, Keller MB, Shapiro RW. Reliability studies of psychiatric diagnosis. Theory and practice. *Arch Gen Psychiatry* 1981;38:408–413.
  17. Maxwell AE. Coefficients of agreement between observers and their interpretation. *Br J Psychiatry* 1977;130:79–83.
  18. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–549.
  19. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;43:551–558.
  20. American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 3rd ed., rev. Washington, DC: American Psychiatric Press, 1987.
  21. Freyberger HJ, Dittmann V, Stieglitz R-D, Dilling H. ICD-10 in der Erprobung: Ergebnisse einer multizentrischen Feldstudie in den deutschsprachigen Ländern. *Nervenarzt* 1990;61:271–275.
  22. Hiller W, Dichtl G, Hecht H, Hundt W. An empirical comparison of diagnoses and reliabilities in ICD-10 and DSM-III-R. *Eur Arch Psychiat Clin Neurosci* 1993;242:209–217.
  23. Jenkins R, Smeeton N, Shepherd M. Classification of mental disorder in primary care. *Psychol Med* 1988 (suppl 12).
  24. van den Brink W, Koeter MWJ, Ormel J, Dijkstra W, Giel R, Slooff CJ, *et al.* Psychiatric diagnosis in an outpatient population. *Arch Gen Psychiatry* 1989;46:369–372.
  25. Janca A, Ustün TB, Early TS, Sartorius N. The ICD-10 symptom checklist: a comparison to the ICD-10 classification of mental and behavioural disorders. *Soc Psychiatry Psychiatr Epidemiol* 1993;28:239–242.
  26. Okasha A, Sadek A, Al-Haddad MK, Abdel-Mawgoud M. Diagnostic agreement in psychiatry. A comparative study between ICD-9, ICD-10 and DSM-III-R. *Br J Psychiatry* 1993; 162:621–626.
  27. American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 4th ed. (DSM-IV). Washington, DC: American Psychiatric Association Press 1994.
  28. Philipp M, Maier W, Delmo CD. The concept of major depression. II. Agreement between six competing operational definitions in 600 psychiatric inpatients. *Eur Arch Psychiat Clin Neurosci* 1991;240:266–271.
  29. Rapaport MH, Tipp JE, Schuckit MA. A comparison of ICD-10 and DSM-III-R criteria for substance abuse and dependence. *Am J Drug Alc Abuse* 1993;19:143–151.
  30. Grant BF. ICD-10 harmful use of alcohol and the alcohol dependence syndrome: prevalence and implications. *Addiction* 1993;88:413–420.
  31. Grant BF. ICD-10 and proposed DSM-IV harmful use of alcohol/alcohol abuse and dependence, United States 1988. A nosological comparison. *Alc Clin Exper Res* 1993;17:1093–1101.