The ICD-11 developmental field study of reliability of diagnoses of high-burden mental disorders: results among adult patients in mental health settings of 13 countries

Geoffrey M. Reed[1,2*], Pratap Sharan[3*], Tahilia J. Rebello[1,4], Jared W. Keeley[5], María Elena Medina-Mora[2], Oye Gureje[6], José Luis Ayuso-Mateos[7], Shigenobu Kanba[8], Brigitte Khoury[9], Cary S. Kogan[10], Valery N. Krasnov[11], Mario Maj[12], Jair de Jesus Mari[13], Dan J. Stein[14], Min Zhao[15], Tsuyoshi Akiyama[16], Howard F. Andrews[4,17], Elson Asevedo[13], Majda Cheour[18], Tecelli Domínguez-Martínez[2,19], Joseph El-Khoury[9], Andrea Fiorillo[12], Jean Grenier[20], Nitin Gupta[21], Lola Kola[6], Maya Kulygina[11], Itziar Leal-Leturia[7], Mario Luciano[12], Bulumko Lusu[14], J. Nicolas I. Martínez-López[2], Chihiro Matsumoto[22], Lucky Umukoro Onofa[23], Sabrina Paterniti[24], Shivani Purnima[3], Rebeca Robles[2], Manoj K. Sahu[25], Goodman Sibeko[14], Na Zhong[15], Michael B. First[1,4], Wolfgang Gaebel[26], Anne M. Lovell[27], Toshimasa Maruta[28], Michael C. Roberts[29], Kathleen M. Pike[1]

[1]Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA; [2]National Institute of Psychiatry Ramón de la Fuente Muñiz, Mexico City, Mexico; [3]Department of Psychiatry, All India Institute of Medical Sciences, New Delhi, India; [4]New York State Psychiatric Institute, New York, NY, USA; [5]Department of Psychology, Virginia Commonwealth University, Richmond, VA, USA; [6]Department of Psychiatry, University of Ibadan, Nigeria; [7]Department of Psychiatry, Universidad Autonoma de Madrid, IIS-P and Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain; [8]Department of Neuropsychiatry, Kyushu University, Fukuoka City, Japan; [9]Department of Psychiatry, American University of Beirut Medical Center, Beirut, Lebanon; [10]School of Psychology, University of Ottawa, Ottawa, Ontario, Canada; [11]Moscow Research Institute of Psychiatry, National Medical Research Centre for Psychiatry and Narcology, Moscow, Russian Federation; [12]Department of Psychiatry, University of Campania "L. Vanvitelli", Naples, Italy; [13]Department of Psychiatry, Universidade Federal de São Paulo, São Paulo, Brazil; [14]Department of Psychiatry, University of Cape Town and South African Medical Research Council Unit on Risk and Resilience in Mental Disorders, Cape Town, South Africa; [15]Shanghai Mental Health Center and Department of Psychiatry, Shanghai Jiao Tong University School of Medicine, Shanghai, People's Republic of China; [16]NTT Medical Center Tokyo, Tokyo, Japan; [17]Departments of Biostatistics and Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA; [18]Department of Psychiatry, Tunis Al Manar University and Al Razi Hospital, Tunis, Tunisia; [19]Cátedras CONACYT, National Council for Science and Technology, Mexico City, Mexico; [20]Institut du Savoir Montfort - Hôpital Montfort & Université d'Ottawa, Ottawa, Ontario, Canada; [21]Department of Psychiatry, Government Medical College and Hospital, Chandigarh, India; [22]Japanese Society of Psychiatry and Neurology, Tokyo, Japan; [23]Federal Neuropsychiatric Hospital Aro, Abeokuta, Nigeria; [24]Institute of Mental Health Research, Royal Ottawa Mental Health Centre, and Department of Psychiatry, University of Ottawa, Ontario, Canada; [25]Pt. Jawahar Lal Nehru Memorial Medical College, Raipur, Chhattisgarh, India; [26]Department of Psychiatry and Psychotherapy, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany; [27]Institut National de la Santé et de la Recherche Médicale U988, Paris, France; [28]Health Management Center, Seitoku University, Matsudo City, Japan; [29]Office of Graduate Studies and Clinical Child Psychology Program, University of Kansas, Lawrence, KS, USA

*Joint first authors

Reliable, clinically useful, and globally applicable diagnostic classification of mental disorders is an essential foundation for global mental health. The World Health Organization (WHO) is nearing completion of the 11th revision of the International Classification of Diseases and Related Health Problems (ICD-11). The present study assessed inter-diagnostician reliability of mental disorders accounting for the greatest proportion of global disease burden and the highest levels of service utilization – schizophrenia and other primary psychotic disorders, mood disorders, anxiety and fear-related disorders, and disorders specifically associated with stress – among adult patients presenting for treatment at 28 participating centers in 13 countries. A concurrent joint-rater design was used, focusing specifically on whether two clinicians, relying on the same clinical information, agreed on the diagnosis when separately applying the ICD-11 diagnostic guidelines. A total of 1,806 patients were assessed by 339 clinicians in the local language. Intraclass kappa coefficients for diagnoses weighted by site and study prevalence ranged from 0.45 (dysthymic disorder) to 0.88 (social anxiety disorder) and would be considered moderate to almost perfect for all diagnoses. Overall, the reliability of the ICD-11 diagnostic guidelines was superior to that previously reported for equivalent ICD-10 guidelines. These data provide support for the suitability of the ICD-11 diagnostic guidelines for implementation at a global level. The findings will inform further revision of the ICD-11 diagnostic guidelines prior to their publication and the development of programs to support professional training and implementation of the ICD-11 by WHO member states.

Key words: International Classification of Diseases, ICD-11, diagnosis, mental disorders, reliability, schizophrenia, mood disorders, anxiety disorders, disorders specifically associated with stress

(World Psychiatry 2018;17:174–186)

A classification system that ensures satisfactorily reliable, clinically useful, and globally applicable diagnosis of mental disorders provides an essential foundation for global mental health. Such a system facilitates efficient identification of people with the greatest mental health needs when they seek health care and supports access to appropriate and cost-effective treatment[1].

Classification systems form the interface between health encounters and health information, and are an important foundation for decisions related to health policy and resource allocation at system, national and global levels. A classification that is too cumbersome to use at the encounter level or does not provide clinically useful information to the treating health professional will not be used as intended, cannot provide valid aggregate data, and will fail to support good clinical practice, research, and policy making[2].

The World Health Organization (WHO) is nearing completion of the 11th revision of the International Classification of Diseases and Related Health Problems (ICD-11), to be released for use by WHO member states in 2018. The revision has provided a major opportunity to bring the ICD classification of mental and behavioural disorders in line with current empirical evidence and clinical practice.

To achieve these aims, the WHO Department of Mental Health and Substance Abuse appointed a series of Working Groups to focus on different disorder areas, and these groups

have conducted comprehensive reviews of available evidence, on which their recommendations are based[3-8]. In addition, the Department of Mental Health and Substance Abuse has undertaken a systematic and comprehensive program of formative and evaluative field studies focused particularly on the global applicability and clinical utility of the revised Clinical Descriptions and Diagnostic Guidelines (CDDG) for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders. These field studies are substantially different from previous efforts in their use of innovative methodologies to investigate the application of the guidelines in the context of clinical decision making[9].

The effectiveness of the ICD-11 CDDG in producing more consistent clinical judgments as compared to the ICD-10 CDDG[10] is currently being tested in a series of Internet-based, multilingual case controlled field studies using standardized case material in the form of vignettes, as these allow for experimental manipulation of the clinical information in a way that isolates the effects of the classification system on diagnostic decision making[11,12]. The use of the written vignettes offers many advantages in terms of standardization and experimental control.

As an important next step in evaluating the CDDG, studies of their implementation in clinical settings provide a fuller approximation of the subtleties of assessment, interpretation and decision making involved in making diagnoses in real patients. Accordingly, ecological implementation field studies (EIFS) are being conducted in clinical settings in a range of countries to investigate the diagnostic reliability and clinical utility of the proposed ICD-11 CDDG. The EIFS centers are located in countries that diverge widely in terms of languages, culture, and resource levels. The initial results of these studies are described in this paper.

The proposed structure and content of the ICD-11 CDDG were designed to enhance their clinical utility, validity and reliability[13]. The WHO has focused on improving clinical utility in the ICD-11 because it is critical to the WHO's public health goals related to reducing the global burden of mental disorders[1]. The WHO defines clinical utility for classificatory systems as including their value in communicating among stakeholders, their implementation characteristics in clinical practice (e.g., goodness of fit, time required to use them), and their usefulness in making clinical management decisions[14].

Thus, clinical utility, validity and reliability are distinct but overlapping constructs[15]. An example of the relationship between reliability and clinical utility of diagnoses was provided by the ICD-10 CDDG field trials[16], which showed that diagnoses with lower reliability were accompanied by lower-than-average ratings of clinical utility (e.g., diagnostic fit, confidence in diagnosis, ease of use, and adequacy of description). Similarly, aspects of the validity of diagnostic constructs also relate to their inherent clinical usefulness in the care of patients, for example in predicting treatment response or course of illness[17].

The reliability of mental disorders diagnoses has been a focus of attention in the revision processes of both the ICD and the Diagnostic and Statistical Manual of Mental Disorders (DSM) of the American Psychiatric Association since the 1970s[18]. Both classificatory systems adopted a descriptive approach to providing diagnostic guidance[19-21], in part based on studies suggesting that deficiencies in pre-DSM-III classification systems were major sources of unreliability[22-24].

In general, studies of the reliability of diagnostic classifications following the publication of DSM-III documented improved results[18]. However, the lower diagnostic reliability documented in the DSM-5 field trials[25] compared to previous field trials has highlighted the profound influence of methodology on estimates of diagnostic reliability[26]. That is, reliability is not solely a property of the classification, but also a product of the method used to estimate it. This makes comparisons across studies with different methodologies quite difficult.

The current study has used a naturalistic, joint-rater design to estimate inter-diagnostician reliability. Unlike some previous studies of the reliability of classification systems[27,28], structured interviews, which could be expected to increase reliability substantially[29-31], were not used. No instruction or training was provided regarding how clinician raters should perform the diagnostic interview, and clinician raters received relatively minimal training on the new ICD-11 guidelines. The attempt was therefore to approximate the conditions under which the guidelines will be applied in clinical settings after their publication.

The joint-rater design was used in order to minimize information variance and to focus specifically on the question of whether two clinicians, relying on the same clinical information, agree on the diagnoses to be assigned to the patient when separately applying the ICD-11 diagnostic guidelines.

Similar to the naturalistic design of the current ICD-11 EIFS, the developers of the field studies for DSM-III, ICD-10 and DSM-5 also chose not to employ structured diagnostic interviews because they are not commonly used in general clinical settings[16,25,32,33]. The DSM-III and ICD-10 CDDG field trials demonstrated good diagnostic reliability for most major classes of disorders. However, reliability estimates were likely inflated in the case of DSM-III by presentation of estimates only for disorder groupings (rather than individual diagnoses)[32] and, in the case of the ICD-10 CDDG, by the use of case conferences – in which one diagnostician interviewed the patient and then presented the case to other assessors – for establishing inter-diagnostician reliability[16].

The DSM-5 field trials also used a naturalistic design, employing two diagnosticians to assess inter-rater agreement on diagnoses and computing reliability at the level of individual diagnoses[34]. However, those field trials used a sequential, test-retest design (two diagnosticians interviewing the participant at different time points) to establish inter-clinician reliability, rather than the concurrent, joint-rater design (two clinicians interviewing the participant together) employed in the ICD-11 EIFS. The DSM-5 design, therefore, did not control for information variance and so would almost certainly yield lower reliabilities[26,35]. Therefore, reliability estimates of the recent DSM-5 field trials and the current ICD-11 EIFS are not comparable.

Arguably, the DSM-5 design is a test of the diagnostic reliability of psychiatric diagnoses more generally and not specifically of the new diagnostic manual.

The ICD-11 EIFS were designed as developmental studies with the goal of using the results in the final revision of the guidelines, rather than solely as evaluative field studies, which aim to assess what users can expect in terms of the classification's psychometric properties after the classification has been completed[36]. The concurrent joint-rater reliability design was preferred for the EIFS because it made it possible to focus on variation in the application or interpretation of the diagnostic guidelines, controlling for variance due to patient factors (e.g., giving different histories to the diagnosticians) and extraneous clinician factors (e.g., variations in the thoroughness of the interview).

The concurrent joint-rater design employed in ICD-11 EIFS focused specifically on the role of the diagnostic guidelines themselves as a source of unreliability. In a developmental field study, identification of high levels of clinician-criterion incongruity should prompt changes to the diagnostic guidelines, whereas clinician errors are likely better addressed through training on the use of the classification and clinical interviewing.

The reliability arm of EIFS described in this paper specifically targeted four groups of disorders among adult patients: schizophrenia and other primary psychotic disorders, mood disorders (including both depressive and bipolar disorders), anxiety and fear-related disorders, and disorders specifically associated with stress. These diagnoses account for the greatest proportion of global disease burden among mental disorders[37] and the highest levels of service utilization in mental health settings.

This paper describes the EIFS results concerning reliability of the proposed ICD-11 CDDG among adult patients in 13 countries.

METHODS

Study design and procedures

Two study protocols were implemented to assess the reliability of the proposed ICD-11 diagnostic guidelines. Protocol 1 tested the reliability of the guidelines for schizophrenia and other primary psychotic disorders and for mood disorders, while Protocol 2 tested the guidelines for mood disorders, anxiety and fear-related disorders, and disorders specifically associated with stress.

Adult (≥18 years of age) patients exhibiting any psychotic symptoms and presenting for care at the participating field study center were eligible for Protocol 1. Adult (≥18 years of age) patients exhibiting mood symptoms, anxiety symptoms, or stress-related symptoms but no psychotic symptoms and presenting for care at the participating field study center were eligible for Protocol 2. These requirements were intended to produce an enriched sample that was likely to have at least one of the conditions being tested, but whose diagnostic status was not determined in advance.

Exclusion criteria for both protocols were the following: communication difficulty sufficient to interfere with participation in the diagnostic interview (e.g., lack of proficiency in the language of the clinicians at the site); cognitive dysfunction to an extent that would interfere with participation in the diagnostic interview; current incapacitation due to severe physical illness or pain; current substance intoxication or withdrawal or serious medication side effects; and current imminent risk of harm to self or other. These criteria essentially functioned to allow any consenting patient exhibiting the index symptoms to be recruited, unless they could not reasonably be expected to participate in the diagnostic interview.

Protocols were implemented at 28 sites in 13 countries. Additional site information is presented in Table 1.

The local language was always used for the diagnostic interviews. The ICD-11 guidelines, training materials, and all material for the study were developed in English. Materials were then translated into four other languages – Chinese, Japanese, Russian and Spanish – with the collaboration of field study centers, using a thorough forward and back-translation process. In other sites, the English guidelines and training materials were used even though the interviews were conducted in other languages, again replicating the circumstances under which the ICD-11 will be implemented.

All sites obtained ethical clearance from their institutional review boards prior to study implementation. Research teams defined local procedures for obtaining consent and for reporting and addressing any adverse events that might be experienced by participants who were being interviewed as part of the study (e.g., inability to complete the interview due to high levels of symptoms or distress). Participants were assigned unique identification numbers, and no confidential or identifying information was reported to anyone outside the site.

A site director was responsible at each site for recruiting clinician raters. According to the practice standards of their countries, all clinician raters were qualified to make mental disorders diagnoses independently as a part of their scope of practice. Advanced residents in psychiatry (following completion of first two years of residency) could function as interviewers but were always paired with a fully qualified individual. Training was organized either at the level of the site or for multiple sites within a given country.

Clinician raters were provided with the ICD-11 diagnostic guidelines being tested and were asked to review them prior to the training session. The training session reviewed central features of the ICD-11 diagnostic guidelines in those areas covered by the protocols and their differences with ICD-10. The sessions used a standard set of slides developed by the WHO. Interactive exercises provided an opportunity for practice in applying the guidelines to case vignettes. The only difference between Protocol 1 and Protocol 2 was that, for the former,

Table 1 Participating country and study site information

Country	Protocol(s) implemented	N. sites	Site names	N. raters
Brazil	1	1	Universidade Federal de São Paulo	21
Canada	2	1	Royal Ottawa Mental Health Centre/University of Ottawa Institute of Mental Health Research	7
China	1 and 2	1	Shanghai Mental Health Center	25
India	1 and 2	3	All India Institute of Medical Sciences, New Delhi	44
			Government Medical College Hospital, Chandigarh	
			Pandit Jawaharlal Nehru Memorial Medical College, Raipur	
Italy	1	1	University of Campania "L. Vanvitelli", Naples	14
Japan	1 and 2	11	Kyushu University	90
			Hokkaido University	
			University of Occupational & Environmental Health, Kitakyushu	
			Tokyo Medical Dental University	
			Tokyo Metropolitan Matsuzawa Hospital	
			Nihon University School of Medicine, Tokyo	
			Nagoya University	
			Hizen National Psychiatric Center, Yoshinogari	
			NTT Medical Center Tokyo	
			Tokyo University	
			Tokushima University	
Lebanon	1 and 2	2	American University of Beirut	14
			Hôpital Psychiatrique De La Croix, Jal El Dib	
Mexico	1 and 2	1	National Institute of Psychiatry Ramón de la Fuente Muñiz, Mexico City	23
Nigeria	1	2	University College Hospital, Ibadan	32
			Federal Neuropsychiatric Hospital, Aro, Abeokuta	
Russia	1	2	Moscow Research Institute of Psychiatry	41
			First Saint Petersburg City Mental Hospital	
South Africa	1 and 2	1	Valkenberg Psychiatric Hospital, Cape Town	10
Spain	1 and 2	1	Hospital Universitario La Princesa, Madrid	6
Tunisia	1 and 2	1	Razi Hospital, Tunis	12

clinician raters were informed that they were required to assess for schizophrenia and other primary psychotic disorders and for mood disorders, as well as for any other area they deemed relevant in arriving at a diagnostic formulation, while for the latter they were required to assess for mood disorders, anxiety and fear-related disorders, and disorders specifically associated with stress. No other instruction was given about how to approach the interview, and it was left to the judgment of the clinician raters to determine how best to perform the assessment, according to their professional training and usual practice, as will be the case when the ICD-11 is implemented.

Training sessions lasted for approximately two hours per protocol (i.e., approximately four hours for sites that were doing both Protocol 1 and Protocol 2). Training sessions were therefore not dissimilar to those that clinicians might realistically receive when the ICD-11 is implemented in their countries. The sessions also covered the study flow and data collection procedures. Post-training and prior to start of data collection, clinician raters registered to participate using an online registration system, providing demographic information as well as details regarding their clinical experience (see Table 2).

A broader group of clinicians at each study site were given information on the study inclusion and exclusion criteria and referral procedures, and asked to refer qualifying patients to either Protocol 1 or Protocol 2. At most sites, clinician raters who conducted joint-rater interviews were also part of the pool of referring clinicians, in which case they were not permitted to

interview any patient they had referred. Referring clinicians were invited to participate in the training sessions for interviewers, though this was not mandatory.

Upon referral, a research coordinator explained the study to referred patients and obtained their informed consent. Following informed consent, patients were interviewed by two clinicians who had no prior clinical contact with the patient. One clinician rater served as the primary interviewer and the second as an observer. The observer was allowed to ask additional follow-up questions at the end of the interview. Clinician pairings were varied as much as possible given constraints of availability and scheduling, and participating clinicians alternated as primary interviewer and observer.

The clinician raters were instructed to set aside 60-90 min for the joint-rater interview. They were asked to approach assessments as they would in routine practice. The range and length of the diagnostic interviews were therefore substantially consistent with usual practice in participating mental health centers.

Based on the interview, and in some cases additional supplementary material provided to both clinicians (e.g., patient file excluding current or prior psychiatric diagnoses and psychotropic medication prescriptions, interviews with family members), clinician raters independently arrived at a diagnostic formulation consisting of up to three diagnoses. Diagnoses were non-hierarchical (i.e., not specified as primary, secondary or tertiary) and could fall within any mental, behavioural or neurodevelopmental disorder diagnostic grouping. Participating clinicians could also specify a non-mental or behavioural disorder diagnosis, or no diagnosis. For diagnoses included in Protocol 1 and Protocol 2, additional detailed questions were asked about symptom presentation and clinical utility of the guidelines.

Following the interview, both clinician raters independently provided data based on the interview using a secure web-based data collection system. Participating clinicians were instructed to record their data within 24 hours. Information provided included each clinician rater's diagnostic formulation, and ratings of the presence or absence of each element of any disorder from the diagnostic groupings that were the focus of Protocol 1 or Protocol 2. Data provided by each clinician also included responses to detailed questions about the clinical utility of the diagnostic guidelines as applied to that particular patient.

Participants

A total of 339 clinicians from the 28 study sites in 13 countries (see Table 2) served as clinician raters for Protocol 1 and/or Protocol 2. The mean age of clinician raters was 37.2 ± 8.3 years, and the ages were comparable across countries. There was a slight majority of male clinician raters in the global sample (56.6%). The overwhelming majority of clinician raters in the study were psychiatrists (93.2%), with a small representation of psychologists (3.8%), nurses (1.5%) and other medical professionals (1.5%). The average clinical experience of the clinician raters was 7.6 ± 7.5 years.

As shown in Table 3, 1,806 patients were recruited into the study for Protocol 1 (N=1,041) or Protocol 2 (N=765). The average age of participating patients was 39.9 ± 13.7 years, and ages were comparable across countries. The global sample had an equal gender distribution. The marital status of the majority of patients in the global sample was single (54.9%); 33.1% were married/cohabitating, 9.8% were separated/divorced and 2.2% were widowed. More than half of the patients in the global sample were unemployed (55.9%) and only 22.3% of the patients had full time employment. A slight majority of recruited patients in the global sample were inpatients (55.0%) and the remainder were nearly all outpatients (44.4%). The small remaining proportion (0.6%) were enrolled in other types of programs such as partial day hospitalization.

Data collection, management and processing

Data reported by clinician interviewers were securely collected using the Electronic Field Study System (EFSS), a web-based data collection system developed using Qualtrics™ (Provo, UT, USA) and made available in five study languages. Clinicians logged onto the EFSS using a unique password to report all study data.

Data from the sites were stored and managed centrally by the Data Coordinating Center (DCC) at Columbia University. Data quality was established through continuous monitoring of the data collection procedures by local research staff at each site and through use of programming functions within Qualtrics™, such as forced response and content validation options. This provided a mechanism for collecting data in a standardized, uniform format from all sites. Site-based research teams kept records of any errors in data entry that were passed on to the DCC for correction.

Data analysis

The main analysis of the study addressed the reliability of diagnoses included in Protocols 1 and 2. Data from both protocols were combined in the current analyses. Diagnostic reliability was estimated based on agreements between clinician raters irrespective of whether the diagnosis was listed first, second or third. For example, if for a particular patient one clinician rater diagnosed single episode depressive disorder, panic disorder, and agoraphobia, and the other clinician rater diagnosed agoraphobia and single episode depressive disorder, both clinician raters would have agreement on the diagnosis of single episode depressive disorder and agoraphobia, but disagreement for panic disorder.

Only diagnoses that occurred at least 30 times across the study were included in these analyses, as diagnoses assigned less frequently were not considered to have sufficient stability for the present evaluation. To estimate diagnostic reliability,

Table 2 Demographics of clinician raters by country

	Total (N=339)	Brazil (N=21)	Canada (N=7)	China (N=25)	India (N=44)	Italy (N=14)	Japan (N=90)	Lebanon (N=14)	Mexico (N=23)	Nigeria (N=32)	Russia (N=41)	South Africa (N=10)	Spain (N=6)	Tunisia (N=12)
Age, years (mean±SD)	37.2 ± 8.3	35.5 ± 8.6	44.4 ± 13.8	32.6 ± 4.8	34.1 ± 7.4	39.8 ± 6.2	38.9 ± 7.7	36.1 ± 8.6	37.6 ± 7.9	37.8 ± 6.0	39.2 ± 11.1	35.5 ± 7.0	32.0 ± 5.9	38.3± 9.3
Gender, N (%)														
Male	192 (56.6)	10 (47.6)	1 (14.3)	5 (20.0)	29 (65.9)	9 (64.3)	72 (80.0)	7 (50.0)	12 (52.2)	25 (78.1)	14 (34.1)	5 (50.0)	2 (33.3)	1 (8.3)
Female	145 (42.8)	11 (52.4)	6 (85.7)	20 (80.0)	15 (34.1)	5 (35.7)	16 (17.8)	7 (50.0)	11 (47.8)	7 (21.9)	27 (65.9)	5 (50.0)	4 (66.7)	11 (91.7)
Clinical profession, N (%)														
Psychiatry	316 (93.2)	21 (100)	2 (28.6)	25 (100)	44 (100)	14 (100)	88 (97.8)	11 (78.6)	22 (95.7)	32 (100)	39 (95.1)	3 (30.0)	5 (83.3)	10 (83.3)
Psychology	13 (3.8)	0	5 (71.4)	0	0	0	0	3 (21.4)	1 (4.3)	0	1 (2.4)	1 (10.0)	1 (16.7)	1 (8.3)
Nursing	5 (1.5)	0	0	0	0	0	0	0	0	0	0	5 (50.0)	0	0
Other medical	5 (1.5)	0	0	0	0	0	2 (2.2)	0	0	0	1 (2.4)	1 (10.0)	0	1 (8.3)
Years of experience (mean±SD)	7.6 ± 7.5	6.6 ± 7.4	13.3 ± 11.9	4.2 ± 3.9	5.4 ± 6.4	7.7 ± 7.2	7.1 ± 6.7	7.8 ± 7.4	9.2 ± 8.2	5.8 ± 4.1	13.6 ± 10.3	6.4 ± 3.9	6.0 ± 4.8	6.5 ± 6.6

Table 3 Demographics of patients by country

	Total (N=1,806)	Brazil (N=100)	Canada (N=53)	China (N=203)	India (N=209)	Italy (N=100)	Japan (N=168)	Lebanon (N=103)	Mexico (N=153)	Nigeria (N=132)	Russia (N=104)	South Africa (N=208)	Spain (N=70)	Tunisia (N=203)
Age, years (mean±SD)	39.9 ± 13.7	32.9 ± 9.6	39.8 ± 14.2	43.9 ± 15.6	36.5 ± 11.4	41.4 ± 11.2	47.0 ± 15.1	36.4 ± 12.5	38.1 ± 13.0	37.5 ± 12.2	36.3 ± 11.7	35.1 ± 11.0	52.0 ± 16.2	43.2 ± 12.6
Gender, N (%)														
Male	908 (50.3)	62 (62.0)	19 (35.8)	123 (60.6)	120 (57.4)	50 (50.0)	72 (42.9)	38 (36.9)	48 (31.4)	65 (49.2)	44 (42.3)	133 (63.9)	26 (37.1)	108 (53.2)
Female	897 (49.7)	38 (38.0)	33 (62.3)	80 (39.4)	89 (42.6)	50 (50.0)	96 (57.1)	65 (63.1)	105 (68.6)	67 (50.8)	60 (57.7)	75 (36.1)	44 (62.9)	95 (46.8)
Relationship status, N (%)														
Single	992 (54.9)	81 (81.0)	22 (41.5)	110 (54.2)	66 (31.6)	71 (71.0)	77 (45.8)	68 (66.0)	91 (59.5)	68 (51.5)	65 (62.5)	167 (80.3)	28 (40.0)	78 (38.4)
Married/ cohabitating	597 (33.1)	12 (12.0)	17 (32.1)	75 (36.9)	133 (63.6)	19 (19.0)	64 (38.1)	20 (19.4)	42 (27.5)	41 (31.1)	22 (21.2)	25 (12.0)	28 (40.0)	99 (48.8)
Separated/ divorced	177 (9.8)	6 (6.0)	13 (24.5)	15 (7.4)	4 (1.9)	7 (7.0)	21 (12.5)	15 (14.6)	20 (13.1)	18 (13.6)	13 (12.5)	13 (6.3)	9 (12.9)	23 (11.3)
Widowed	40 (2.2)	1 (1.0)	1 (1.9)	3 (1.5)	6 (2.9)	3 (3.0)	6 (3.6)	0	0	5 (3.8)	4 (3.8)	3 (1.4)	5 (7.1)	3 (1.5)
Employment, N (%)														
Full time	403 (22.3)	4 (4.0)	14 (26.4)	47 (23.2)	69 (33.0)	11 (11.0)	26 (15.5)	16 (15.5)	17 (11.1)	41 (31.1)	22 (21.2)	22 (10.6)	26 (37.1)	88 (43.3)
Part time	142 (7.9)	5 (5.0)	6 (11.3)	3 (1.5)	12 (5.7)	9 (9.0)	14 (8.3)	11 (10.7)	31 (20.3)	11 (8.3)	6 (5.8)	8 (3.8)	3 (4.3)	23 (11.3)
Unemployed	1009 (55.9)	76 (76.0)	30 (56.6)	80 (39.4)	110 (52.6)	74 (74.0)	109 (64.9)	66 (64.1)	79 (51.6)	64 (48.5)	53 (51.0)	167 (80.3)	20 (28.6)	81 (39.9)
Student	136 (7.5)	6 (6.0)	4 (7.5)	15 (7.4)	15 (7.2)	4 (4.0)	10 (6.0)	15 (14.6)	30 (19.6)	10 (7.6)	7 (6.7)	12 (5.8)	2 (2.9)	6 (3.0)
Retired	152 (8.4)	10 (10.0)	1 (1.9)	62 (30.5)	3 (1.4)	2 (2.0)	15 (8.9)	0	5 (3.3)	8 (6.1)	18 (17.3)	0	22 (31.4)	6 (3.0)
Treatment setting, N (%)														
Outpatient	801 (44.4)	82 (82.0)	53 (100)	0	122 (58.4)	67 (67.0)	48 (28.6)	14 (13.6)	135 (88.2)	84 (63.6)	4 (3.8)	0	49 (70.0)	143 (70.4)
Inpatient	994 (55.0)	18 (18.0)	0	203 (100)	87 (41.6)	33 (33.0)	120 (71.4)	89 (86.4)	17 (11.1)	48 (36.4)	91 (87.5)	207 (99.5)	21 (30.0)	60 (29.6)
Other	11 (0.6)	0	0	0	0	0	0	0	1 (0.7)	0	9 (8.7)	1 (0.5)	0	0

Table 4 Concurrent reliability of ICD-11 diagnoses

	Number of diagnoses (N)	Joint rater agreement (intraclass kappa)	Standard error	Bootstrapped 95% CI
Schizophrenia	725	0.87	0.012	0.84-0.89
Schizoaffective disorder	189	0.66	0.035	0.58-0.72
Acute and transient psychotic disorder	40	0.45	0.087	0.27-0.60
Delusional disorder	30	0.69	0.084	0.51-0.84
Bipolar I disorder	351	0.84	0.017	0.81-0.87
Bipolar II disorder	95	0.62	0.048	0.52-0.70
Single episode depressive disorder	191	0.64	0.035	0.57-0.77
Recurrent depressive disorder	267	0.74	0.025	0.69-0.79
Dysthymic disorder	57	0.45	0.073	0.28-0.58
Generalized anxiety disorder	129	0.62	0.044	0.53-0.70
Panic disorder	59	0.57	0.069	0.42-0.69
Agoraphobia	46	0.62	0.072	0.47-0.75
Social anxiety disorder	38	0.88	0.045	0.78-0.95
Post-traumatic stress disorder	51	0.49	0.076	0.33-0.64
Complex post-traumatic stress disorder	45	0.56	0.077	0.40-0.71
Adjustment disorder	82	0.73	0.046	0.63-0.81

intraclass kappa coefficients for diagnoses weighted by site and study prevalence were calculated. Bootstrapped 95% confidence intervals for kappa, based upon 1,000 resamples, were then calculated. All analyses were conducted using SPSS.

Landis and Koch[38] adjectives were used to describe ranges of reliability values for kappa: slight (from 0 to 0.20), fair (from 0.21 to 0.40), moderate (from 0.41 to 0.60), substantial (from 0.61 to 0.80), and almost perfect (from 0.81 to 1.0).

RESULTS

Estimates of joint-rater agreement are shown in Table 4, along with bootstrapped 95% confidence intervals. The point estimate of kappa ranged from 0.45 (dysthymic disorder) to 0.88 (social anxiety disorder) and would be considered moderate to almost perfect according to Landis and Koch adjectives for all diagnoses for which it was calculated.

The kappa estimates were almost perfect for schizophrenia (0.87) and bipolar I disorder (0.84); substantial for schizoaffective disorder (0.66), delusional disorder (0.69), bipolar II disorder (0.62), single episode depressive disorder (0.64), recurrent depressive disorder (0.74), generalized anxiety disorder (0.62), agoraphobia (0.62), and adjustment disorder (0.73); and moderate for acute and transient psychotic disorder (0.45), dysthymic disorder (0.45), panic disorder (0.57), post-traumatic stress disorder (0.49), and the newly introduced diagnosis of complex post-traumatic stress disorder (0.56).

In general, point estimates of kappa were lower for disorders for which smaller samples were obtained. The higher number of diagnoses of primary psychotic and mood disorders reflects the type of settings (55% inpatient) and the nature of the centers (tertiary and secondary care) involved in the reliability arm of EIFS.

The estimates of kappa were precise for all diagnoses for which it was calculated (confidence interval <0.5; standard error <0.1). The lower bound estimates of the confidence interval for kappa were higher than 0.4 (fair reliability) for 13 of the 16 disorders. However, the lower bound estimates were only in the fair range (from 0.2 to 0.4) for acute and transient psychotic disorder (0.27), dysthymic disorder (0.28), and post-traumatic stress disorder (0.33). All diagnoses with lower bound confidence interval estimates of kappa (<0.4) were made less often, suggesting that higher reliability for these disorders might accrue in samples of larger sizes.

Table 5 provides a comparison of the results of joint-rater agreement in the current study of the ICD-11 CDDG with the results of the ICD-10 CDDG field trial[16]. This comparison is intended to be illustrative rather than exact because of major differences in study methodologies. Unlike the ICD-11 EIFS, which used two raters for face-to-face joint rater interviews, the ICD-10 field study used case conferences, in which one rater conducted a face-to-face interview and then presented the case to other raters as a basis for establishing inter-rater reliability. The case conference methodology is likely to produce more consensus-based results, in which reliability would be correspondingly higher. Further, though most ICD-10 diagnoses correspond closely to proposed ICD-11 diagnoses, they are not identical.

While statistical comparisons of the two studies are not justified, in 10 of the 14 possible comparisons between the ICD-

Table 5 Comparison of reliability estimates in ICD-11 CDDG EIFS and ICD-10 CDDG field trials

ICD-11 EIFS		ICD-10 CDDG field trail	
	Kappa (N)		**Kappa (N)**
Schizophrenia	0.87 (725)	F20 Schizophrenia	0.81 (490)
Schizoaffective disorder	0.66 (189)	F36 Schizoaffective disorder	0.48 (148)
Acute and transient psychotic disorder	0.45 (40)	F23 Acute/transient psychotic disorders	0.65 (146)
Delusional disorder	0.69 (30)	F22.0 Delusional disorder	0.62 (83)
Bipolar I disorder	0.84 (351)	F30 Manic episode	0.69 (53)
		F31 Bipolar affective disorders	0.81 (259)
Single episode depressive disorder	0.64 (191)	F32 Depressive episode	0.66 (353)
Recurrent depressive disorder	0.74 (267)	F33 Recurrent depressive disorders	0.69 (302)
Dysthymic disorder	0.45 (57)	F34.1 Dysthymia	0.36 (101)
Generalized anxiety disorder	0.62 (129)	F41.1 Generalized anxiety disorder	0.48 (67)
Panic disorder	0.57 (59)	F41.0 Panic disorder	0.74 (31)
Agoraphobia	0.62 (46)	F40.0 Agoraphobia	0.51 (22)
Social anxiety disorder	0.88 (38)	F40.1 Social phobias	0.41 (22)
Post-traumatic stress disorder	0.49 (51)	F43.1 Post-traumatic stress disorder	0.62 (23)
Adjustment disorder	0.73 (82)	F43.2 Adjustment disorder	0.54 (107)

CDDG – Clinical Descriptions and Diagnostic Guidelines, EIFS – Ecological Implementation Field Studies

11 CDDG EIFS and the ICD-10 CDDG field study, the kappa values were higher for ICD-11. These differences tended to be modest.

DISCUSSION

The 11th revision of the Mental, Behavioural and Neurodevelopmental Disorders chapter of the ICD has proposed substantive changes to the conceptualization of many disorders, which may impact their reliability, validity and clinical utility. Field studies that assess how well the proposed changes perform in the hands of the intended users are crucial to this revision process. Accordingly, the EIFS for proposed ICD-11 CDDG were conducted in a broad spectrum of secondary and tertiary mental health care settings across countries with varied languages, cultures, and resource levels.

The results of the ICD-11 EIFS show that all common and high-burden disorders in the adult population covered in the current study were diagnosed with at least satisfactory – and in most cases excellent – reliability by a sample of clinician raters that included advanced trainees in psychiatry as well as more experienced clinicians. This suggests that the proposed ICD-11 CDDG are suitable for use at a global level and that their satisfactory implementation extends beyond application to written vignettes to application to real patients in clinical settings.

Reliability of diagnosis impacts clinical communication, generalizability of the guidelines across patient populations, and tailoring of treatments according to diagnosis, in addition to the selection of samples for research. The DSM-III had introduced fully operationalized diagnostic criteria in the classification of mental disorders as a way of improving diagnostic reliability[30,31]. The ICD-11 CDDG were designed to align with the overarching principles of categorization emerging from earlier studies analyzing how clinicians naturally organize clinical conditions[2]. ICD-11 disorders are presented in terms of the essential features that clinicians could reasonably expect to find in all cases, in an effort to communicate the essence of the disorder, with greater flexibility for clinical and cultural variation[13]. The ICD-11 CDDG avoid fully operationalized criteria characterized by precise cutoffs and symptom counts, unless these are specifically empirically supported. The present results challenge the assumption that the more clinician-friendly, less concretely algorithmic, and less precisely specified approach adopted for the ICD-11 CDDG is inherently less reliable.

The reliability coefficients observed in this study were based on routine clinical assessments (lasting about one hour) using open form interviews by clinicians with diverse training and experience. The results were similar to those achieved by diagnostic assessments using more complex and time consuming structured instruments[26,39,40]. These results suggest that the use of more uniform procedures by clinicians based on a brief training may yield adequate reliability for commonly diagnosed mental disorders in clinical settings. A hypothesis that would be well worth testing – given the resources that are devoted to the refinement of diagnostic criteria – is that further gains could be obtained by focusing greater attention on

appropriate training in diagnostic skills and interviewing techniques[41], rather than on introducing greater precision in the strict operationalization of diagnostic guidelines.

In general, the reliability of diagnoses in ICD-11 CDDG was superior to that of diagnoses in ICD-10 CDDG[16], though strict comparisons are not appropriate due to differences in methodology of these field studies. Similar comparisons with the studies of ICD-10 Diagnostic Criteria for Research[33] and the DSM-III[32] were not performed because of even greater methodological differences. The ICD-10 Diagnostic Criteria for Research field trial involved the use of a structured diagnostic instrument that covered the diagnostic criteria for assessment[33]. The published results of DSM-III field trial provided kappa values for disorder groupings rather than for specific disorders[32], which would tend to maximize reliability results because disagreements within a grouping are substantially more likely than disagreements concerning disorders from different groupings.

Changes to the ICD-11 CDDG relative to the ICD-10 CDDG were proposed by expert working groups based on the available scientific evidence and with explicit attention to additional sources of information related to clinical utility and global applicability. In no case were changes proposed solely to improve reliability, though the more consistent presentation of information in the ICD-11 CDDG as compared to the ICD-10 CDDG[13] likely helped in this regard. However, had the outcome of these changes been an overall reduction in reliability of the ICD-11 CDDG relative to the ICD-10 CDDG, this would have been cause for concern.

The reliability of ICD-11 CDDG generalized anxiety disorder, agoraphobia, social anxiety disorder, and adjustment disorder seems to have improved relative to the ICD-10 CDDG. This is reassuring, because the reliability of milder disorders compared to more severe disorders (e.g., schizophrenia and bipolar disorder) was lower in ICD-10 field trials[16,33]. Data from DSM-5 field trials suggest that disorders that are more broadly defined have higher reliability[25]. A number of hierarchical exclusion rules have been eliminated for anxiety and fear-related disorders in ICD-11 CDDG because they lacked specific empirical support[7]. Similarly, the subtypes of adjustment disorder have been eliminated from ICD-11 CDDG because they lacked evidence for validity or clinical utility[5].

The conceptualization of generalized anxiety disorder has been broadened in ICD-11 CDDG to include worry as an alternative essential feature to generalized apprehension and accompanying physiological symptoms[7], based in part on studies that show that worry is a central characteristic of the disorder[42]. Agoraphobia is reconceptualized to include a broader array of feared stimuli (fear of situations, fear of specific negative outcomes) and behaviours manifested in response to these stimuli (avoidance or entering the situations under specific conditions or enduring the situation with intense fear/anxiety), partly to allow for situations that may be more representative of those reported in low- and middle-income countries[43]. The ICD-11 conceptualization of social anxiety disorder has broadened the concept of ways in which the person could fear being negatively evaluated by others to include cultural variants of the disorder (i.e., fears of humiliation, embarrassment, rejection, or being offensive) as well as the range of behaviours in response to social stimuli[44,45]. It is possible that the greater attention to the cognitive and behavioural components of anxiety disorders and their contextual and cultural features in the ICD-11 CDDG as compared to the ICD-10 CDDG[7] helped to improve the reliability of these diagnoses.

Changes made in the diagnostic guidelines for adjustment disorder based on an earlier case-controlled study of disorders specifically associated with stress[9], particularly in providing additional guidance on differentiation from normal stress reactions, likely improved its diagnostic reliability in the current study.

Schizoaffective disorder is not a rare diagnosis in clinical populations, and its reliability is subject to ongoing discussion[46]. Jager et al[47] reviewed six studies and reported kappa scores between 0.08 and 0.63, concluding that only one study showed good agreement. In a meta-analysis of studies on sequential reliability (test-retest) of schizoaffective disorders, Santelmann et al[46] documented a mean difference of approximately 0.2 for kappa between schizoaffective disorder and other diagnoses such as schizophrenia, bipolar disorder and unipolar depression. The improved reliability of ICD-11 schizoaffective disorder in comparison to ICD-10 CDDG diagnosis may be related to the decision, in the proposed ICD-11 CDDG, to even more clearly apply the diagnostic requirements to the current episode rather than to the longitudinal presentation of the illness[3]. This is different from the longitudinal approach historically and currently taken by the DSM, on which most previous studies have been based[46,47]. Again, the purpose of the changes made for ICD-11 was to increase the clinical utility of the categories, and to the extent possible their validity, but it is reassuring that improved reliability appears to have been an outcome of these changes.

Some areas of the classification merit further consideration based on these results. The ICD-11 CDDG diagnoses of acute and transient psychotic disorder, panic disorder, and posttraumatic stress disorder seemed to have lower reliability than the equivalent categories in the ICD-10 CDDG, though it was not considered appropriate to analyze these differences statistically. However, these differences are modest in size (in all cases <0.2), and the reliability estimates for the ICD-11 CDDG in these categories are still in the moderate range.

However, unlike the categories discussed previously that were broadened in the ICD-11 CDDG, the description of each of these disorders has been narrowed in terms of their essential features. ICD-11 acute and transient psychotic disorder now exclusively comprises acute psychoses with "polymorphic" presentation[3], which is not strictly comparable to the broader concept tested in the ICD-10 field trial[16]. The reliability of acute and transient psychotic disorder with polymorphic presentation in ICD-10 Diagnostic Criteria for Research field trial[33] was similar to that in the present study. Nevertheless, based on these

results, the description of acute and transient psychotic disorder has been revised for the final version of the guidelines to define this aspect of the disorder more explicitly and to provide additional guidance on how to differentiate it from other conditions.

The proposed ICD-11 CDDG for panic disorder now require a clear discrimination between panic attacks of unexpected nature and panic attacks occurring in relation to symptoms of specific mental disorders (i.e., phobic disorders, some obsessive-compulsive disorders, and disorders specifically associated with stress). If panic attacks can be explained as due to symptoms of other specific mental disorders, a "with panic attacks" qualifier should be used rather than an additional co-occurring diagnosis of panic disorder. If some panic attacks over the course of the disorder have been unexpected and not exclusively in response to stimuli associated with the focus of apprehension related to the relevant disorder, a separate diagnosis of panic disorder should be assigned. In such cases, it is not necessary to apply the "with panic attacks" qualifier[7]. The lower kappa value for the ICD-11 CDDG as compared to the ICD-10 CDDG for panic disorder suggests that clinicians may have found it difficult to differentiate between expected and unexpected panic attacks or have been unclear about when to use the "with panic attacks" qualifier and when instead apply an additional diagnosis of panic disorder. This provides an example of an apparent trade-off between validity and reliability. Based on the results of this study, the final version of the ICD-11 CDDG will contain more detailed guidance on how to differentiate between unexpected and expected panic attacks and on how to decide whether applying the "with panic attacks" qualifier or a co-occurring panic disorder diagnosis. Increased emphasis on this issue in training programs as a part of ICD-11 implementation may also be helpful.

Though post-traumatic stress disorder is a well-recognized clinical entity, it has been criticized for the broad composition of its symptom clusters and high levels of co-occurrence with other disorders. Studies have also suggested that the threshold for an ICD-10 diagnosis of the disorder is relatively low[48,49]. The ICD-11 CDDG for post-traumatic stress disorder diagnosis are conceptually narrower than the ICD-10 ones, and now require the presence of re-experiencing of intrusive symptoms in the "here and now", as opposed to only experiencing intrusive memories of the traumatic event, as well as the presence of functional impairment[5]. This model has garnered increasing empirical support[50]. However, an earlier Internet-based study on disorders specifically associated with stress[9] showed that clinicians did not consistently apply the proposed ICD-11 guidelines regarding the required element of re-experiencing of the traumatic event(s). The subsequent version of the ICD-11 CDDG used in the present study provided additional clarification regarding re-experiencing in PTSD. However, the application of some of the changes introduced in the ICD-11 for post-traumatic stress disorder still appears to be difficult for practicing clinicians. Further exploration of the discrepancies between clinician raters at the level of specific symptoms may cast additional light on this issue. A specific focus on the new conceptualization of post-traumatic stress disorder as a part of ICD-11 training programs will also likely to be needed.

Some of the limitations of the ICD-11 EIFS need to be acknowledged. First, it bears repeating that the joint-rater (concurrent) method of testing reliability, which constrains the information provided to the two diagnosticians to be identical, usually generates higher kappa values compared to those obtained when separate interviews are conducted[26,51]. Second, the present study was conducted in multiple centers in diverse countries, including a very high proportion of low- and middle-income countries, but participating clinicians cannot be considered to be a globally representative sample of diagnosing mental health professionals. Participating institutions were typically high-status secondary or tertiary care centers, where the training of clinicians in diagnostic classification and interviewing is likely to meet the highest national standards. Clinician interviewers in the study would also have had some specific interest in diagnostic classification and in learning about the ICD-11. It can therefore be assumed that the reliabilities obtained in the study are higher than those that will be obtained in usual practice across all settings where the ICD-11 CDDG will be implemented. However, these problems are inherent in any field study, unless they can be overcome by a level of resources substantially in excess of those available for the EIFS.

Moreover, because the study sites were large academic settings that would tend to serve patients with moderate to severe mental health problems, the results may not be generalizable to patients with milder disorders seen in community settings. Mitigating this concern somewhat is the fact that ICD-11 CDDG include specific guidance on delineation of disorders from normal variation and have raised diagnostic thresholds for some of the conditions tested in EIFS (e.g., disorders specifically associated with stress).

Finally, the current study assessed only a relatively small proportion of the wide range of mental disorder diagnoses that may be applied to adult patients, focusing on those that are responsible for the highest level of disease burden and account for the greatest proportions of mental health services in participating centers. A much broader range of diagnostic categories is being addressed via Internet-based studies[9,12] and the overall consistency between the results of the two types of studies is reassuring in this regard.

CONCLUSIONS

As a developmental field study[36], the ICD-11 CDDG EIFS has been designed to provide information regarding the source of diagnostic disagreements through assessment of each element of the diagnostic guidelines for those disorders included in the protocols. This study has provided additional data for the WHO to use in improving the diagnostic guidelines prior to their publication. The WHO will also use the data in the development of training manuals and training courses for

clinicians in order to support member states in their implementation of ICD-11, with specific attention to the low- and middle-income countries in which the overwhelming majority of the world's population live.

The primary conclusion of this multi-country study is that the proposed ICD-11 CDDG can be interpreted in a consistent manner by diagnosing mental health professionals in a wide range of countries. The global applicability of the ICD-11 CDDG conceptualization of commonly diagnosed mental disorders is supported by the assessment of reliability of these guidelines in diverse settings (across 28 sites in 13 countries and in five languages) using a naturalistic field study design and a training approach that can easily be replicated for ICD-11 implementation. In the limited number of conditions that fell short, the findings will inform further revision prior to publication of the ICD-11.

The magnitude of this collaboration, the inclusion of clinicians in practice around the globe, the administration of the study in multiple languages, and the completion of this research in time to have the findings inform the final guidelines are major strengths of the ICD-11 research program. In addition to the specific value of this study in shaping the ICD-11, the EIFS and the WHO's Global Clinical Practice Network[52] for Internet-based ICD-11 field studies (http://gcp.network) have galvanized interest among clinicians around the world to participate in ongoing research that will continue to improve many dimensions of clinical understanding of mental illness and mental health service delivery.

ACKNOWLEDGEMENTS

The opinions contained in the paper are those of its authors and, except as specifically stated, are not intended to represent the official policies or positions of the World Health Organization. Funding was received for national activities related to this project in the following countries: Brazil - Conselho Nacional de Desenvolvimento Científico e Tecnológico; Canada - University Medical Research Fund, Royal's University of Ottawa Institute of Mental Health Research; Japan - Japanese Society of Psychiatry and Neurology, and Japan Agency for Medical Research and Development; Mexico - Consejo Nacional de Ciencia y Tecnología. Additional support for data collection in Brazil, Lebanon, Nigeria, South Africa and Tunisia was provided by the Columbia University Global Mental Health Program. Otherwise, this project was funded by in-kind contributions of the participating institutions. The authors express their gratitude to the following individuals who contributed substantially to the conduct of this research: Gustavo M. Barros, Ary Gadelha, Michel Haddad, Nuno H.P. Santos (Brazil); Huajian Ma, Zhen Wang, Jingjing Huang (China); Huma Kamal, Nidhi Malhotra (India); Gaia Sampogna, Lucia Del Gaudio, Giuseppe Piegari, Francesco Perris, Luca Steardo Jr (Italy); Tomofumi Miura, Itta Namamura, Kiyokazu Atake, Ayako Endo, Yuki Kako, Shinichi Kishi, Michihiko Koeda, Shinsuke Kondo, Akeo Kurumaji, Shusuke Numata, Naoya Oribe, Futoshi Suzuki, Masashi Yagi (Japan); Sariah Daouk, Chadia Haddad, François Kazour, Nicole Khauli (Lebanon); Francisco Juárez, Alejandra González, Omar Hernández, Carolina Muñoz (Mexico); Mayokun Odunleye (Nigeria); Tatiana Kiska, Oleg Limankin, Pavel Ponizovsky (Russian Federation); Roxanne James, Christine Lochner, Adele Pretorius (South Africa); Carolina Ávila, Cora Fernández, Julián Gómez, Ana Izquierdo, Beatriz Vicario, Rubén Vicente (Spain); Rahma Damak (Tunisia).

REFERENCES

1. International Advisory Group for the Revision of ICD-10 Mental and Behavioural Disorders. A conceptual framework for the revision of the ICD-10 classification of mental and behavioural disorders. World Psychiatry 2011;10:86-92.
2. Reed GM, Roberts MC, Keeley J et al. Mental health professionals' natural taxonomies of mental disorders: implications for the clinical utility of the ICD-11 and the DSM-5. J Clin Psychol 2013;69:1191-212.
3. Gaebel W. Status of psychotic disorders in ICD-11. Schizophr Bull 2012;38: 895-8.
4. Maj M, Reed GM. The ICD-11 classification of mood and anxiety disorders: background and options. World Psychiatry 2012;11(Suppl. 1).
5. Maercker A, Brewin CR, Bryant RA et al. Diagnosis and classification of disorders specifically associated with stress: proposals for ICD-11. World Psychiatry 2013;12:198-206.
6. Stein DJ, Kogan CS, Atmaca M et al. The classification of obsessive-compulsive and related disorders in the ICD-11. J Affect Disord 2016;190: 663-74.
7. Kogan CS, Stein DJ, Maj M et al. The classification of anxiety and fear-related disorders in the ICD-11. Depress Anxiety 2016;33:1141-54.
8. Tyrer P, Reed GM, Crawford MJ. Classification, assessment, prevalence and effect of personality disorder. Lancet 2015;385:717-26.
9. Keeley JW, Reed GM, Roberts MC et al. Disorders specifically associated with stress: a case-controlled field study for ICD-11 Mental and Behavioural Disorders. Int J Clin Health Psychol 2016;16:109-27.
10. World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva: World Health Organization, 1992.
11. Evans SC, Roberts MC, Keeley JW et al. Using vignette methodologies for study clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol 2015;15:160-70.
12. Keeley JW, Reed GM, Roberts MC et al. Developing a science of clinical utility in diagnostic classification systems: field study strategies for ICD-11 mental and behavioural disorders. Am Psychol 2016;71:3-16.
13. First MB, Reed GM, Hyman SE et al. The development of the ICD-11 Clinical Descriptions and Diagnostic Guidelines for Mental and Behavioural Disorders. World Psychiatry 2015;14:82-90.
14. Reed GM. Toward ICD-11: improving the clinical utility of WHO's international classification of mental disorders. Prof Psychol Res Pr 2010;41:457-64.
15. Reed GM, First MB, Medina-Mora ME et al. Draft diagnostic guidelines for ICD-11 mental and behavioural disorders available for review and comment. World Psychiatry 2016;15:112-3.
16. Sartorius N, Kaelber CT, Cooper JE et al. Progress toward achieving a common language in psychiatry. Results from the field trial of the clinical guidelines accompanying the WHO classification of mental and behavioral disorders in ICD-10. Arch Gen Psychiatry 1993;50:115-24.
17. Mullins-Sweatt SN, Widiger TA. Clinical utility and DSM-5. Psychol Assess 2009;21:302-12.
18. Blashfield RK, Keeley JW, Flanagan EH et al. The cycle of classification: DSM-I through DSM-5. Annu Rev Clin Psychol 2014;10:25-51.
19. Stengel E. Classification of mental disorders. Bull World Health Organ 1959;21:601-63
20. Spitzer R, Sheehy M, Endicott J. DSM-III: Guiding principles. In: Rakoff V, Stancer H, Kedward H (eds). Psychiatric diagnosis. New York: Brunner/Mazel, 1977:1-24.
21. World Health Organization. Glossary of mental disorders and guide to their classification. Geneva: World Health Organization, 1974.
22. Beck AT, Ward CH, Mendelson M et al. Reliability of psychiatric diagnosis 2: a study of consistency of clinical judgments and ratings. Am J Psychiatry 1962;119:351-7.
23. Ward CH, Beck AT, Mendelson M et al. The psychiatric nomenclature: reasons for diagnostic disagreement. Arch Gen Psychiatry 1962;7:198-205.
24. Harvey PD, Heaton RK, Carpenter WT Jr et al. Diagnosis of schizophrenia: consistency across information sources and stability of the condition. Schizophr Res 2012;140:9-14.
25. Regier DA, Narrow WE, Clarke DE et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnosis. Am J Psychiatry 2013;170:59-70.
26. Chmielewski M, Clark LA, Bagby RM et al. Method matters: understanding diagnostic reliability in DSM-IV and DSM-5. J Abnorm Psychol 2015;124: 764-9.
27. Lahey BB, Applegate B, Barkley RA et al. DSM-IV field trials for oppositional defiant disorder and conduct disorder in children and adolescents. Am J Psychiatry 1994;151:1163-71.

28. Rounsaville BJ, Kosten TR, Williams JB et al. A field trial of DSM-III-R psychoactive substance dependence disorders. Am J Psychiatry 1987;144:351-5.
29. Brittain PJ, Stahl D, Rucker J et al. A review of the reliability and validity of OPCRIT in relation to its use for the routine clinical assessment of mental health patients. Int J Methods Psychiatr Res 2013;22:110-37.
30. Aboraya A, Rankin E, France C et al. The reliability of psychiatric diagnosis revisited: the clinician's guide to improve the reliability of psychiatric diagnosis. Psychiatry 2006;3:41-50.
31. First MB. Counterpoint – there isn't enough evidence available to speculate on the reliability of diagnoses in clinical settings. Psychiatry 2007;4: 24-5.
32. Spitzer R, Forman J, Nee J. DSM-III field trials, I: initial interrater diagnostic reliability. Am J Psychiatry 1979;136:815-7.
33. Sartorius N, Ustün TB, Korten A et al. Progress toward achieving a common language in psychiatry, II: Results from the international field trials of the ICD-10 diagnostic criteria for research for mental and behavioral disorders. Am J Psychiatry 1995;152:1427-37.
34. Clarke DE, Narrow WE, Regier DA et al. DSM-5 field trials in the United States and Canada, Part I: study design, sampling strategy, implementation, and analytic approaches. Am J Psychiatry 2013;170:43-58.
35. Feinn R, Gelernter J, Cubells JF et al. Sources of unreliability in the diagnosis of substance dependence. J Stud Alcohol Drugs 2009;70:475-81.
36. First MB. The importance of developmental field trials in the revision of psychiatric classifications. Lancet Psychiatry 2016;3:579-84.
37. Whiteford HA, Degenhardt L, Rehm J et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 2013;382:1575-86.
38. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
39. Andrews G, Peters L, Guzman A-M et al. A comparison of two structured diagnostic interviews: CIDI and SCAN. Aust N Z J Psychiatry 1995;29:124-32.
40. Lobbestael J, Leurgans M, Arntz A. Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). Clin Psychol Psychother 2011;18:75-9.
41. Bruehl S, Ohrbach R, Sharma S et al. Approaches to demonstrating the reliability and validity of core diagnostic criteria for chronic pain. J Pain 2016;17(Suppl. 9):T118-31.
42. Olatunji BO, Broman-Fulks JJ, Bergman SM et al. A taxometric investigation of the latent structure of worry: dimensionality and associations with depression, anxiety, and stress. Behav Ther 2010;41:212-28.
43. Stein DJ. Agoraphobia and panic disorder: options for ICD-11. World Psychiatry 2012;11(Suppl. 1):89-93.
44. Emmelkamp PMG. Specific and social phobias in ICD-11. World Psychiatry 2012;11(Suppl. 1):94-9.
45. Kinoshita Y, Chen J, Rapee RM et al. Cross-cultural study of conviction subtype Taijin Kyofu: proposal and reliability of Nagoya-Osaka diagnostic criteria for social anxiety disorder. J Nerv Ment Dis 2008;196:307-13.
46. Santelmann H, Franklin J, Bußhoff J, et al. Test-retest reliability of schizoaffective disorder compared with schizophrenia, bipolar disorder, and unipolar depression – a systematic review and metaanalysis. Bipolar Disord 2015;17:753-68.
47. Jager M, Haack S, Becker T et al. Schizoaffective disorder – an ongoing challenge for psychiatric nosology. Eur Psychiatry 2011;26:159-65.
48. Brewin CR, Fuchkan N, Huntley Z et al. Outreach and screening following the 2005 London bombings: usage and outcomes. Psychol Med 2010;40: 2049-57.
49. Peters L, Slade T, Andrews G. A comparison of ICD-10 and DSM-IV criteria for posttraumatic stress disorder. J Trauma Stress 1999;12:335-43.
50. Brewin CR, Cloitre M, Hyland P et al. A review of current evidence regarding the ICD-11 proposals for diagnosing PTSD and complex PTSD. Clin Psychol Rev 2017;58:1-15.
51. Kraemer HC. The reliability of clinical diagnoses: state of the art. Annu Rev Clin Psychol 2014;10:111-30.
52. Reed GM, Rebello TJ, Pike KM et al. WHO's Global Clinical Practice Network for mental health. Lancet Psychiatry 2015;2:379-80.

DOI:10.1002/wps.20524