Clinical utility of ICD-11 diagnostic guidelines for high-burden mental disorders: results from mental health settings in 13 countries

Geoffrey M. Reed[1,2], Jared W. Keeley[3], Tahilia J. Rebello[1,4], Michael B. First[1,4], Oye Gureje[5], José Luis Ayuso-Mateos[6], Shigenobu Kanba[7], Brigitte Khoury[8], Cary S. Kogan[9], Valery N. Krasnov[10], Mario Maj[11], Jair de Jesus Mari[12], Pratap Sharan[13], Dan J. Stein[14], Min Zhao[15], Tsuyoshi Akiyama[16], Howard F. Andrews[1,4,17], Elson Asevedo[12], Majda Cheour[18], Tecelli Domínguez-Martínez[2,19], Joseph El-Khoury[8], Andrea Fiorillo[11], Jean Grenier[20], Nitin Gupta[21], Lola Kola[5], Maya Kulygina[10], Itziar Leal-Leturia[6], Mario Luciano[11], Bulumko Lusu[13], J. Nicolás I. Martínez-López[2], Chihiro Matsumoto[22], Mayokun Odunleye[23], Lucky Umukoro Onofa[24], Sabrina Paterniti[25], Shivani Purnima[13], Rebeca Robles[2], Manoj K. Sahu[26], Goodman Sibeko[13], Na Zhong[15], Wolfgang Gaebel[27], Anne M. Lovell[28], Toshimasa Maruta[29], Kathleen M. Pike[1], Michael C. Roberts[30], María Elena Medina-Mora[2]

[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 Psychology, Virginia Commonwealth University, Richmond, VA, USA; [4]New York State Psychiatric Institute, New York, NY, USA; [5]Department of Psychiatry, University of Ibadan, Ibadan, Nigeria; [6]Department of Psychiatry, Universidad Autonoma de Madrid; Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM); Instituto de Investigación Sanitaria La Princesa, Madrid, Spain; [7]Department of Neuropsychiatry, Kyushu University, Fukuoka City, Japan; [8]Department of Psychiatry, American University of Beirut Medical Center, Beirut, Lebanon; [9]School of Psychology, University of Ottawa, Ottawa, ON, Canada; [10]Moscow Research Institute of Psychiatry, National Medical Research Centre for Psychiatry and Narcology, Moscow, Russian Federation; [11]Department of Psychiatry, University of Campania "L. Vanvitelli", Naples, Italy; [12]Department of Psychiatry, Universidade Federal de São Paulo, São Paulo, Brazil; [13]Department of Psychiatry, All India Institute of Medical Sciences, New Delhi, India; [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]Department of Biostatistics, 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, ON, Canada; [21]Department of Psychiatry, Government Medical College and Hospital, Chandigarh, India; [22]Japanese Society of Psychiatry and Neurology, Tokyo, Japan; [23]Department of Psychiatry, University College Hospital, Ibadan, Nigeria; [24]Federal Neuropsychiatric Hospital Aro, Abeokuta, Nigeria; [25]Institute of Mental Health Research, Royal Ottawa Mental Health Centre, and Department of Psychiatry, University of Ottawa, Ottawa, ON, Canada; [26]Pt. Jawahar Lal Nehru Memorial Medical College, Raipur, Chhattisgarh, India; [27]Department of Psychiatry and Psychotherapy, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany; [28]Institut National de la Santé et de la Recherche Médicale U988, Paris, France; [29]Health Management Center, Seitoku University, Matsudo City, Japan; [30]Office of Graduate Studies and Clinical Child Psychology Program, University of Kansas, Lawrence, KS, USA

In this paper we report the clinical utility of the diagnostic guidelines for ICD-11 mental, behavioural and neurodevelopmental disorders as assessed by 339 clinicians in 1,806 patients in 28 mental health settings in 13 countries. Clinician raters applied the guidelines for schizophrenia and other primary psychotic disorders, mood disorders (depressive and bipolar disorders), anxiety and fear-related disorders, and disorders specifically associated with stress. Clinician ratings of the clinical utility of the proposed ICD-11 diagnostic guidelines were very positive overall. The guidelines were perceived as easy to use, corresponding accurately to patients' presentations (i.e., goodness of fit), clear and understandable, providing an appropriate level of detail, taking about the same or less time than clinicians' usual practice, and providing useful guidance about distinguishing disorder from normality and from other disorders. Clinicians evaluated the guidelines as less useful for treatment selection and assessing prognosis than for communicating with other health professionals, though the former ratings were still positive overall. Field studies that assess perceived clinical utility of the proposed ICD-11 diagnostic guidelines among their intended users have very important implications. Classification is the interface between health encounters and health information; if clinicians do not find that a new diagnostic system provides clinically useful information, they are unlikely to apply it consistently and faithfully. This would have a major impact on the validity of aggregated health encounter data used for health policy and decision making. Overall, the results of this study provide considerable reason to be optimistic about the perceived clinical utility of the ICD-11 among global clinicians.

Key words: International Classification of Diseases, ICD-11, diagnosis, mental disorders, clinical utility, ease of use, goodness of fit, treatment selection, assessing prognosis

(World Psychiatry 2018;17:306–315)

The World Health Organization (WHO) has released the 11th revision of the International Classification of Diseases and Related Health Problems to its member states to prepare for implementation[1]. The new classification will be presented for approval by the World Health Assembly, the WHO's governing body, in May 2019.

As we have previously described[2-5], an important focus in the development of the ICD-11 chapter on Mental, Behavioural and Neurodevelopmental Disorders by the WHO Department of Mental Health and Substance Abuse has been to improve its clinical utility.

For the purpose of developing the ICD classification of mental disorders, the WHO has defined the clinical utility of a classification construct, category, or system as depending on: a) its value in communicating (e.g., among practitioners, patients, families, administrators); b) its implementation characteristics in clinical practice, including its goodness of fit (i.e., accuracy of description), its ease of use, and the time required to use it (i.e., feasibility); and c) its usefulness in selecting interventions and in making clinical management decisions[2]. This definition is based in part on those proposed by M. First and colleagues[6,7].

Similar concepts had also been included in the ICD-10 field trails[8,9], which asked clinicians to provide ratings of goodness of fit, confidence in their selected diagnosis, ease or difficulty of making a diagnosis, and adequacy of the diagnostic guidelines for cases evaluated as a part of the study.

In a recent study[10], we expanded the operationalization of clinical utility considerably to include an assessment of utility in relation to specific components of the diagnostic guidelines as well as to specific uses of the guidelines (e.g., meeting administrative requirements, assigning a diagnosis, treatment selection, communication, teaching).

Moreover, the WHO Department of Mental Health and Substance Abuse has conducted a major programmatic field studies effort for ICD-11 focusing on clinical utility[3]. This program of research extends the concept of clinical utility to include diagnostic accuracy and diagnostic consistency, as diagnoses that are neither accurate nor reliable are unlikely to be useful.

Thus, there are both subjective and objective components to clinical utility, and these overlap to some extent with both reliability and validity[2]. Clinical utility is not simply a matter of clinician preferences. Nonetheless, the subjective components are important because clinicians who do not feel that a classification system provides them with useful and valuable information are unlikely to apply it carefully, with major implications for the quality of health encounter data related to diagnosis.

Finally, with the goal of improving clinical utility, the Department of Mental Health and Substance Abuse has made a series of substantive changes in the Clinical Descriptions and Diagnostic Guidelines (CDDG) for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders as compared to the ICD-10 CDDG[11]. The CDDG is the version that is intended to be used by mental health professionals in clinical settings. Many of these changes have involved ensuring that the ICD-11 CDDG provide consistent and relatively uniform diagnostic information across the various categories[4], something that has been identified as a shortcoming of the ICD-10 CDDG. Diagnostic guidelines have been drafted so as to allow for the appropriate exercise of clinical judgment, minimizing the use of arbitrary or pseudo-precise symptom counts and cutoffs when these are not strongly supported by evidence. The new structure of groupings and categories for ICD-11 is also intended to be more logical and more consistent with how clinicians conceptualize mental disorders[12,13].

The data presented in this paper were collected as a part of the ICD-11 developmental field study of reliability of diagnoses of high-burden mental disorders, undertaken in 13 countries around the world. The initial reliability data have been published in this journal[14], indicating that the joint-rater reliability of the ICD-11 diagnostic guidelines ranged from moderate to almost perfect (.45 to .88)[15], and was generally superior to results obtained for ICD-10[8]. The current paper focuses on clinicians' evaluations of the clinical utility of the diagnostic guidelines, using a scale that is based in part on clinical utility concepts from the ICD-10 field trial, but that more fully operationalizes the WHO's definition of clinical utility for ICD-11.

METHODS

Study design and procedures

Two study protocols were implemented to assess the clinical utility and the reliability of the proposed ICD-11 diagnostic guidelines. Protocol 1 tested the utility and 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 a participating study site were eligible to participate in Protocol 1, while adult 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 to participate in Protocol 2. Prospective participants who met these criteria were excluded only if they could not reasonably be expected to participate in the diagnostic assessment (e.g., for reasons of language or cognitive impairment).

These relatively loose criteria were in part intended to more closely approximate the natural circumstances under which the ICD-11 will be implemented in mental health settings.

Study protocols were implemented at 28 sites in 13 countries[14]. The local language was always used for the diagnostic assessments. The ICD-11 guidelines, training materials, and all material for the study were developed in English and then translated into four other languages: Chinese, Japanese, Russian and Spanish. For Tunisia, the guidelines, but not all of the other training materials, were translated into French. In other sites where English was not the local language (e.g., Brazil, Italy), the English guidelines and training materials were used even though the interviews were conducted in the local language, again replicating the circumstances under which the ICD-11 will be implemented in many settings. Details on clinician recruitment and training, study implementation processes, data collection, and ethical clearance have been provided previously[14].

Following informed consent, patients were interviewed by two clinicians with whom they had not had any prior clinical contact. 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. Based on the interview, 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 in the ICD-11. Participating clinicians could also specify a non-mental or behavioural disorder diagnosis, or no diagnosis.

After finalizing their selected diagnostic formulation, clinicians were asked twelve detailed questions about the clinical utility of the diagnostic guidelines as applied to that particular patient. These included: core clinical utility questions (ease of use, goodness of fit, clarity and understandability), questions on implementation characteristics of the guidelines (level of detail, feasibility of assessment requirements, time required), questions about the utility of specific sections of the guidelines (boundary with normality and differential diagnosis), and questions about the utility of the guidelines for specific purposes (selecting a treatment, predicting prognosis, communicating with other professionals, educating patients and family members). Specific wording of the questions and the Likert-type response options for each question are shown in Table 1.

Clinicians provided clinical utility ratings for the specific categories that were part of diagnostic groupings which were the focus of Protocols 1 and 2, i.e., schizophrenia and other primary psychotic disorders, mood disorders (including depressive disorders and bipolar disorders), anxiety and fear-related disorders, and disorders specifically associated with stress. If more than one diagnosis from these groupings was applied to a particular patient, clinical utility ratings were made for all such selected diagnoses taken together rather than for each diagnosis separately.

Participants

A total of 339 clinicians from the 28 study sites in 13 countries served as clinician raters for Protocol 1 and/or Protocol 2. The mean age of clinician raters was 37.2±8.3 years, and their 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 health care professionals (1.5%). Clinicians had an average of 7.6±7.5 years of professional clinical experience following completion of their clinical training (including post-graduate training).

As shown in Table 2, 1,806 patients participated in 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 was comparable across countries. The global patient sample had an equal gender distribution. The marital status of the majority of patients across countries was single (54.9%); 33.1% were married/cohabitating, 9.8% were separated or divorced, and 2.2% were widowed. More than half of the patients in the global sample were unemployed (55.9%) and only 22.3% had full time employment. A slight majority of patients who participated in the study were inpatients (55.0%) and the remainder were mostly 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

Clinician interviewers entered interview data using the Electronic Field Study System (EFSS), a secure web-based data collection system developed using Qualtrics™ (Provo, UT, USA) survey software, made available in all five study languages. 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 programmed 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 discovered during the review process and these were passed on to the DCC for correction.

Data analysis

A total of 3,608 sets of clinical utility ratings were made by the 339 clinicians. Because there were two raters for each patient, the N for each analysis should be double that of the number of patients (N=1,806; see Table 1), but in four cases only one set of ratings was available for a particular patient.

Clinician raters' responses to each of the 12 clinical utility variables were summarized using frequency counts for each response. To provide a metric of overall favorable responses, ratings of "Quite" and "Extremely" were combined for questions where this was appropriate (see Table 1). Responses to the clinical utility variables by country were also calculated (not all reported; available from the authors by request), as were responses to clinical utility for the five most commonly used diagnoses.

For reliability analyses, intraclass kappa coefficients were calculated with bootstrapped 95% confidence intervals, based on 1,000 resamples, for each country. Reliability coefficients were calculated for only the most common diagnoses within the study (i.e., N≥130), to maximize the chance of having a sufficient number of diagnoses within a country to estimate kappa. Per-diagnosis ratings of clinical utility were also calculated for these same diagnoses.

RESULTS

Clinical utility ratings across countries are shown in Table 1. Evaluations were overwhelmingly positive, though with some differences between items.

For the three core clinical utility questions (ease of use, goodness of fit, clarity and understandability), the overwhelming majority of participants (82.5 to 83.9%) provided ratings of "Quite" or "Extremely", indicating favourable clinical utility.

Table 1 Clinical utility questions and responses across countries (N=3,608)

Core clinical utility questions

Please rate the overall ease of use of the diagnostic guidelines with respect to this patient:

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
32 (0.9%)	556 (15.4%)	2,471 (68.5%)	549 (15.2%)	3,020 (83.7%)

Please rate the overall goodness of fit or accuracy of the diagnostic guidelines with respect to this patient:

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
28 (0.8%)	604 (16.7%)	2,497 (69.2%)	479 (13.3%)	2,976 (82.5%)

Please rate the extent to which the diagnostic guidelines were clear and understandable overall as applied to this patient:

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
14 (0.4%)	567 (15.7%)	2,473 (68.5%)	554 (15.4%)	3,027 (83.9%)

Implementation characteristics

Which of the following statements best describes your evaluation of the level of detail and specificity of the essential features for the diagnosis or diagnoses that you applied to this patient?

Insufficient:	About the right amount:	Too much:
148 (4.1%)	3,275 (90.8%)	185 (5.1%)

Please rate the extent to which the guidelines imposed assessment requirements that were difficult to apply to this patient (e.g., requirements that rely too much on the patient's memory of remote events or the patient's ability to report temporal relationships between symptoms):

Very difficult:	Somewhat difficult:	Quite easy:	Extremely easy:	Quite + extremely easy:
35 (1.0%)	518 (14.4%)	2,752 (76.3%)	303 (8.4%)	3,055 (84.7%)

How would you describe the amount of time that it took you to apply all of the Essential Features to this patient for the diagnosis or diagnoses that you selected, in comparison to your usual clinical practice?

Much longer:	Somewhat longer:	About the same:	Shorter:
30 (0.8%)	472 (13.1%)	2,669 (74.0%)	437 (12.1%)

Specific sections

Please rate the extent to which the description of the boundary between disorder and normality contained in the guidelines was useful as applied to this patient:

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
78 (2.2%)	770 (21.3%)	2,304 (63.9%)	456 (12.6%)	2,760 (76.5%)

Please rate the extent to which the description of the boundary between this patient's disorder and other disorders (section on differential diagnosis) was useful as applied to this patient:

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
49 (1.4%)	762 (21.1%)	2,322 (64.4%)	475 (13.2%)	2,797 (77.5%)

Specific uses

How useful would the diagnostic guidelines be in helping you to select a treatment for this patient?

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
70 (1.9%)	887 (24.6%)	2,223 (61.6%)	428 (11.9%)	2,651 (73.5%)

How useful would the diagnostic guidelines be in helping you to assess this patient's prognosis?

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
83 (2.3%)	1,055 (29.2%)	2,104 (58.3%)	366 (10.1%)	2,470 (68.5%)

How useful would the diagnostic guidelines be in helping you to communicate about this patient with a colleague or other health care professional?

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
49 (1.4%)	746 (20.7%)	2,216 (61.4%)	597 (16.5%)	2,813 (78.0%)

How useful would the diagnostic guidelines be in helping you to educate this patient and/or family about his or her condition?

Not at all:	Somewhat:	Quite:	Extremely:	Quite + extremely:
52 (1.4%)	884 (24.5%)	2,236 (62.0%)	436 (12.1%)	2,672 (74.1%)

For implementation characteristics, a large majority indicated that the guidelines did not impose assessment requirements that were difficult to apply (84.7%), provided about the right level of detail (90.4%), and took about the same amount of time or less time than their usual practice (86.1%).

Regarding specific sections, the manner in which the guidelines provided guidance about differentiating disorders from

Table 2 Patient demographics 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

normality and from other disorders was also rated very positively, with 76.5% and 77.5% of participants, respectively, indicating that these sections were quite or extremely useful.

Regarding the clinical utility of the guidelines for specific purposes, 78.0% of participants indicated that they would be quite or extremely useful for communicating with colleagues or other professionals. The lowest, though still positive overall, ratings were provided for potential usefulness in selecting a treatment (73.5%) and assessing prognosis (68.5%).

We also examined variations in clinical utility ratings across countries. Table 3 shows ratings by country for the three core clinical utility questions. Ratings by country for other clinical utility variables (see Table 1) are not reported here, but are available upon request. The most apparent variation across these three questions is that the ratings shown are substantially lower for Japan (47.9 to 49.7% answering "Quite" or "Extremely") and somewhat lower for Tunisia (69.0 to 70.4%) as compared to the proportion of participants answering "Quite" or "Extremely" for other countries (81.5 to 97.9%).

If variability in perceived clinical utility were directly related to the adequacy of the guidelines, it might be expected that perceived clinical utility and inter-rater reliability would vary together. Table 4 shows concurrent reliability or joint rater agreement, represented by interclass kappa with bootstrapped 95% confidence intervals, for the five most common diagnoses among the sample: schizophrenia, schizoaffective disorder, bipolar type I disorder, single episode depressive disorder and recurrent depressive disorder. While there is clearly variability in reliability by country, there is not a discernible relationship between lower ratings of clinical utility by Japanese and Tunisian participants and the reliability of their diagnostic assignments. Conversely, lower reliability coefficients (e.g., for the Russian Federation) did not correspond with low perceptions of clinical utility.

Clinical utility ratings by diagnosis are shown for these same five diagnoses in Table 5. Across the three core overall clinical utility questions, depressive disorders had slightly lower ratings than schizophrenia, schizoaffective disorder, and bipolar disorder. Slightly lower reliability estimates for single episode depressive disorder and recurrent depressive disorder appear to correspond to slightly lower clinical utility ratings for these categories, but schizoaffective disorder had very high clinical utility ratings in spite of having similarly lower reliability.

DISCUSSION

In the current analyses, clinician ratings of clinical utility of the proposed ICD-11 diagnostic guidelines proved to be very positive overall. This was likely in part related to the attention to clinical utility in the construction of the guidelines[4], as well as the fact that they had already been tested in Internet-based studies in global, multilingual studies via the Global Clinical Practice Network (https://gcp.network) and refined on that basis[16,17].

The guidelines were perceived as easy to use, corresponding accurately to patients' presentations (i.e., goodness of fit), clear and understandable, providing an appropriate level of detail, taking about the same or less time than clinicians' usual practice, and providing useful guidance about distinguishing disorder from normality and from other disorders. Clinicians evaluated the guidelines as relatively less useful for treatment selection and assessing prognosis than for communicating with other health professionals, though the former ratings were still positive overall.

As described, two of the core clinical utility questions used in this study were based on questions used in the ICD-10 field study[8,9]. In that study, 82.5% of participating global clinicians rated the goodness of fit of ICD-10 guidelines as good or very good, and 85.0% said that they were moderately or very easy to use[18]. These percentages are nearly identical to the ones obtained in this study for the ICD-11 guidelines, but differences in the scaling (see Table 1) suggest that the current results could be viewed as more positive.

It should be noted that participating clinicians would likely have been disposed to view the guidelines positively, given that they were participating in a WHO field study about the new global classification system in which their institutions were specifically involved. There may have been both a positive cognitive bias and a social desirability element to their responses. It is possible that clinicians not participating in this type of study will greet the ICD-11 guidelines with less enthusiasm when asked to implement them within their clinical settings. However, this would be true of any parallel assessment of clinical utility such as those for ICD-10[8,9,18] and DSM-5[19], and does not change the overall interpretation of the results.

The pattern of results related to the usefulness of guidelines for specific functions (e.g., treatment selection, prognosis, communicating with other professionals) is entirely consistent with the pattern of results from a separate survey regarding clinicians' current use of the ICD-10, DSM-IV, and DSM-5[10]. It is expected that ratings of the utility of treatment selection and prognosis might not be as high as other uses of the ICD-11, as many treatments are not specific to a single diagnostic label[20], nor is the ICD-11 intended to be a treatment guide.

It is nonetheless reassuring that, although following the same pattern, clinicians' ratings of the usefulness of the ICD-11 diagnostic guidelines they had just used for treatment selection, assessing prognosis, and educating patients and families were substantially higher than the ratings clinicians participating in the other study made about the ICD-10 or the DSM-IV or the DSM-5[10]. Even so, this may be an inherent limitation of current categorical classification systems (i.e., ICD-11, ICD-10, and DSM-5), which are not organized around the most meaningful typologies for selecting treatment or establishing prognosis[20,21]. Future efforts at creating a closer link between

Table 3 Clinical utility ratings by country for three core questions

Ease of use

	Not at all	Somewhat	Quite	Extremely	Quite + extremely
Brazil (N=200)	4 (2.0%)	30 (15.0%)	125 (62.5%)	41 (20.5%)	166 (83.0%)
Canada (N=106)	0	19 (17.9%)	71 (67.0%)	16 (15.1%)	87 (82.1%)
PR China (N=405)	3 (0.7%)	62 (15.3%)	306 (75.6%)	34 (8.4%)	340 (84.0%)
India (N=418)	3 (0.7%)	46 (11.0%)	291 (69.6%)	78 (18.7%)	369 (88.3%)
Italy (N=200)	0	13 (6.5%)	125 (62.5%)	62 (31.0%)	187 (93.5%)
Japan (N=336)	13 (3.9%)	161 (47.9%)	147 (43.8%)	15 (4.5%)	162 (48.2%)
Lebanon (N=206)	1 (0.5%)	15 (7.3%)	147 (71.4%)	43 (20.9%)	190 (92.2%)
Mexico (N=306)	1 (0.3%)	25 (8.2%)	213 (69.6%)	67 (21.9%)	280 (91.5%)
Nigeria (N=264)	0	13 (4.9%)	185 (70.1%)	66 (25.0%)	251 (95.1%)
Russian Fed. (N=208)	0	25 (12.0%)	166 (79.8%)	17 (8.2%)	183 (88.0%)
Spain (N=140)	0	3 (2.1%)	133 (95.0%)	4 (28.6%)	137 (97.9%)
South Africa (N=413)	3 (0.7%)	25 (6.1%)	303 (73.4%)	82 (19.9%)	385 (93.2%)
Tunisia (N=406)	4 (1.0%)	119 (29.3%)	259 (63.8%)	24 (5.9%)	283 (69.7%)

Goodness of fit

	Not at all	Somewhat	Quite	Extremely	Quite + extremely
Brazil (N=200)	6 (3.0%)	31 (15.5%)	120 (60.0%)	43 (21.5%)	163 (81.5%)
Canada (N=106)	1 (0.9%)	28 (26.4%)	63 (59.4%)	14 (13.2%)	77 (72.6%)
PR China (N=405)	4 (1.0%)	58 (14.3%)	293 (72.3%)	50 (12.3%)	343 (84.6%)
India (N=418)	3 (0.7%)	49 (11.7%)	293 (70.1%)	73 (17.5%)	366 (87.6%)
Italy (N=200)	0	11 (5.5%)	123 (61.5%)	66 (33.0%)	189 (94.5%)
Japan (N=336)	7 (2.1%)	168 (50.0%)	149 (44.3%)	12 (3.6%)	161 (47.9%)
Lebanon (N=206)	1 (0.5%)	20 (9.7%)	139 (67.5%)	46 (22.3%)	185 (89.8%)
Mexico (N=306)	2 (0.7%)	37 (12.1%)	209 (68.3%)	58 (19.0%)	267 (87.3%)
Nigeria (N=264)	0	22 (8.3%)	195 (73.9%)	47 (17.8%)	242 (91.7%)
Russian Fed. (N=208)	0	28 (13.5%)	162 (77.9%)	18 (8.7%)	180 (86.5%)
Spain (N=140)	0	7 (5.0%)	127 (90.7%)	6 (4.3%)	133 (95.0%)
South Africa (N=413)	2 (0.5%)	27 (6.5%)	360 (87.2%)	24 (5.8%)	384 (93.0%)
Tunisia (N=406)	2 (0.5%)	118 (29.1%)	264 (65.0%)	22 (5.4%)	286 (70.4%)

Clarity and understandability

	Not at all	Somewhat	Quite	Extremely	Quite + extremely
Brazil (N=200)	1 (0.5%)	20 (10.0%)	141 (70.5%)	38 (19.0%)	179 (89.5%)
Canada (N=106)	0	18 (17.0%)	65 (61.3%)	23 (21.7%)	88 (83.0%)
PR China (N=405)	2 (0.5%)	55 (13.6%)	296 (73.1%)	52 (12.8%)	348 (85.9%)
India (N=418)	2 (0.5%)	51 (12.2%)	281 (67.2%)	84 (20.1%)	365 (87.3%)
Italy (N=200)	0	7 (3.5%)	115 (57.5%)	78 (39.0%)	193 (96.5%)
Japan (N=336)	5 (1.5%)	164 (48.8%)	154 (45.8%)	13 (3.9%)	167 (49.7%)
Lebanon (N=206)	0	22 (10.7%)	147 (71.4%)	37 (18.0%)	184 (89.3%)
Mexico (N=306)	1 (0.3%)	25 (8.2%)	214 (69.9%)	66 (21.6%)	280 (91.5%)
Nigeria (N=264)	0	17 (6.4%)	191 (72.3%)	56 (21.2%)	247 (93.6%)
Russian Fed. (N=208)	0	26 (12.5%)	159 (76.4%)	23 (11.1%)	182 (87.5%)
Spain (N=140)	0	6 (4.3%)	127 (90.7%)	7 (5.0%)	134 (95.7%)
South Africa (N=413)	1 (0.2%)	32 (7.7%)	328 (79.4%)	52 (12.6%)	380 (92.1%)
Tunisia (N=406)	2 (0.5%)	124 (30.5%)	255 (62.8%)	25 (6.2%)	280 (69.0%)

Table 4 Concurrent reliability (joint rater agreement, represented by interclass kappa) and bootstrapped 95% confidence interval (CI) for five most common diagnoses by country

	Kappa (95% CI)				
Country	**Schizophrenia**	**Schizoaffective disorder**	**Bipolar type I disorder**	**Single episode depressive disorder**	**Recurrent depressive disorder**
Brazil (N=100)	.61 (.39 to .79)	.45 (.14 to .73)	.85 (.56 to 1.00)	.43 (−.03 to .78)	-
Canada (N=53)	-	-	-	.65 (.30 to .90)	.85 (.68 to .96)
PR China (N=203)	.96 (.92 to .99)	-	.87 (.78 to .95)	.32 (−.02 to .66)	.71 (.55 to .84)
India (N=209)	.90 (.82 to .96)	.59 (−.01 to .91)	.88 (.78 to .96)	.76 (.61 to .87)	.85 (.70 to .97)
Italy (N=100)	.85 (.74 to .96)	.79 (.59 to .93)	.95 (.84 to 1.00)	-	-
Japan (N=168)	.90 (.82 to .97)	-	.77 (.53 to .94)	.77 (.61 to .90)	.75 (.61 to .87)
Lebanon (N=103)	.95 (.86 to 1.00)	.82 (.64 to .95)	.82 (.67 to .93)	-	.64 (.29 to .88)
Mexico (N=153)	.87 (.76 to .96)	.38 (−.02 to .74)	-	.46 (.27 to .62)	.64 (.52 to .76)
Nigeria (N=132)	.93 (.86 to .98)	.71 (.45 to .89)	.83 (.68 to .94)	.93 (.72 to 1.00)	-
Russian Fed. (N=104)	.54 (.33 to .73)	.45 (.20 to .66)	.52 (−.02 to .88)	-	-
South Africa (N=208)	.71 (.60 to .81)	.68 (.55 to .80)	.80 (.71 to .88)	-	.76 (.40 to 1.00)
Spain (N=70)	.84 (.51 to 1.00)	-	.86 (.70 to .97)	.58 (.24 to .84)	.83 (.58 to 1.00)
Tunisia (N=203)	.84 (.75 to .92)	.59 (.30 to .80)	.69 (.52 to .84)	.63 (.41 to .80)	.50 (.24 to .71)
Overall	.87 (.84 to .89)	.66 (.58 to .72)	.84 (.81 to .87)	.64 (.57 to .77)	.74 (.69 to .79)

Cells without values are those with an insufficient number of observations to calculate kappa

Table 5 Clinical utility ratings for three core questions for five most common diagnoses

Ease of use					
	Not at all	**Somewhat**	**Quite**	**Extremely**	**Quite + extremely**
Schizophrenia	4 (0.3%)	127 (10.0%)	896 (70.9%)	237 (18.8%)	1133 (89.6%)
Schizoaffective disorder	0	24 (11.1%)	166 (76.5%)	27 (12.4%)	193 (88.9%)
Bipolar type I disorder	1 (0.2%)	64 (10.8%)	412 (69.8%)	113 (19.2%)	525 (89.0%)
Single episode depressive disorder	1 (0.4%)	56 (21.5%)	165 (63.5%)	38 (14.6%)	203 (78.1%)
Recurrent depressive disorder	4 (0.9%)	78 (18.4%)	290 (68.6%)	51 (12.1%)	341 (80.6%)
Goodness of fit					
	Not at all	**Somewhat**	**Quite**	**Extremely**	**Quite + extremely**
Schizophrenia	3 (0.2%)	141 (11.2%)	897 (71.0%)	223 (17.6%)	1120 (88.6%)
Schizoaffective disorder	0	33 (15.2%)	163 (75.1%)	21 (9.7%)	184 (84.8%)
Bipolar type I disorder	1 (0.2%)	65 (11.0%)	446 (75.6%)	78 (13.2%)	524 (88.8%)
Single episode depressive disorder	1 (0.4%)	58 (22.3%)	173 (66.5%)	29 (11.2%)	202 (77.7%)
Recurrent depressive disorder	3 (0.7%)	81 (19.1%)	284 (67.1%)	55 (13.0%)	339 (80.1%)
Clarity and understandability					
	Not at all	**Somewhat**	**Quite**	**Extremely**	**Quite + extremely**
Schizophrenia	1 (0.1%)	134 (10.6%)	890 (70.4%)	239 (18.9%)	1129 (89.3%)
Schizoaffective disorder	0	26 (12.0%)	161 (74.2%)	30 (13.8%)	191 (88.0%)
Bipolar type I disorder	0	61 (10.3%)	434 (73.6%)	95 (16.1%)	529 (89.7%)
Single episode depressive disorder	0	48 (18.5%)	174 (66.9%)	39 (15.0%)	213 (81.9%)
Recurrent depressive disorder	0	82 (19.4%)	283 (66.9%)	58 (13.7%)	341 (80.6%)

This analysis excluded diagnostic formulations in which more than one of the five index diagnoses included in the table had been assigned (N=853)

mental health diagnosis and treatment planning would be a worthwhile endeavor from the perspective of enhancing public health, but would need to take a variety of other factors into account (e.g., functional status, treatment availability and acceptability).

Looking at country-level clinical utility ratings, it is clear that clinicians' perceptions of the utility of the diagnostic guidelines were similarly positive across a very diverse set of countries: Brazil, Canada, China, India, Italy, Lebanon, Mexico, Nigeria, Russia, Spain, and South Africa. This may reflect the substantial international participation in the development of the guidelines, with all WHO regions represented and a substantial number of experts from low- and middle-income countries included in all ICD-11 Working Groups, as well as prior international multilingual testing via the Global Clinical Practice Network.

It is encouraging that conducting the clinical assessment in a wide range of local languages did not seem to impact the perceived utility of the diagnostic guidelines. The main deviation from this was the substantially lower ratings of clinical utility made by Japanese participants and the somewhat lower (though still positive) ratings made by Tunisian participants. For Japan, it is possible that these differences are partly related to a cultural tendency not to make extreme ratings, either positive or negative[22], and for both countries this may have been affected by the particular characteristics of the clinician raters involved. For Tunisia, not having all of the training materials available in French may have affected the outcome. However, it is also possible that the proposed ICD-11 diagnostic guidelines specifically correspond less well to presentations of mental disorders more characteristic of Japanese and Tunisian patients as compared to patients from other countries. Further research will be necessary to understand more about global variation in the perceived clinical utility of diagnostic guidelines.

It is important to note, however, that the observed variations in perceived clinical utility, either by country or by diagnosis, had no discernible relationship to variations in reliability. In particular, the lower ratings by Japanese participants of clinical utility did not seem to impact their ability to apply the guidelines consistently. Similarly, instances of lower reliability did not result in correspondingly poorer ratings of clinical utility. This finding highlights the importance of taking into account multiple characteristics of the classification system when evaluating its performance. Neither clinical utility ratings nor reliability estimates provide the whole story.

This paper adds to our previous finding that inter-diagnostician reliability using the proposed ICD-11 diagnostic guidelines was moderate to almost perfect (.45 to .88)[15] for mental disorders accounting for the greatest proportion of global disease burden and the highest levels of service utilization among adult patients presenting for treatment at 28 participating centers in 13 countries[14]. Reliability was superior overall to that previously reported for equivalent ICD-10 guidelines.

WHO's model for ICD-11[2] does not consider clinical utility as defined solely by preference ratings. Instead, it is a dynamic construct that is directly integrated with the actual use of the manual as intended. As such, adequate reliability or consistency of application across the globe is also evidence of the clinical utility of the new ICD-11 guidelines.

CONCLUSIONS

The 11th revision of the Mental, Behavioural and Neurodevelopmental Disorders chapter of the ICD has made substantive changes to the conceptualization of many disorders, which may impact their clinical utility, in addition to their reliability and validity. This study is part of a program of field studies focused on clinical utility adopted by WHO in revising the Mental and Behavioural Disorders chapter of ICD-10[3].

In clinical settings, the ICD functions partly as an interface between health encounters and health information[13], and diagnostic guidelines that are experienced by their intended users as lacking in clinical utility have little chance of being implemented faithfully and consistently. In this event, the validity of the diagnostic components of health encounter data would be seriously compromised, with downstream implications for the quality of decision-making regarding health policy and programmes and resource allocation based on those data.

Therefore, field studies that assess perceived clinical utility of the proposed ICD-11 diagnostic guidelines among its intended users have very important implications. For this reason, the study was conducted in a broad spectrum of secondary and tertiary mental health care settings across countries with varied languages, cultures, and resource levels.

Overall, the results provide considerable reason to be optimistic about the perceived clinical utility of the ICD-11 among global clinicians.

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 WHO. 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 – National Council of Science and Technology (project no. 234473). 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 (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); 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. World Health Organization. WHO releases new International Classification of Diseases (ICD 11). http://www.who.int/news-room/detail/18-06-2018-who-releases-new-international-classification-of-diseases-(icd-11).
2. 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.
3. 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.
4. 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.
5. 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.
6. First MB, Pincus HA, Levine JB et al. Clinical utility as a criterion for revising psychiatric diagnoses. Am J Psychiatry 2004;161:946-54.
7. First MB. Clinical utility in the revision of the Diagnostic and Statistical Manual of Mental Disorders (DSM). Prof Psychol Res Pr 2010;41: 465-73.
8. 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.
9. 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.
10. First MB, Rebello TJ, Keeley JW et al. Do mental health professionals use diagnostic classifications the way we think they do? A global survey. World Psychiatry 2018;17:187-95.
11. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organization, 1992.
12. 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.
13. Roberts MC, Reed GM, Medina-Mora ME et al. A global clinicians' map of mental disorders to improve ICD-11: analysing meta-structure to enhance clinical utility. Int Rev Psychiatry 2012;24:578-90.
14. Reed GM, Sharan P, Rebello TJ et al. 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. World Psychiatry 2018;17:174-86.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
16. 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.
17. Keeley JW, Gaebel W, First MB et al. Psychotic disorder symptom rating scales: are dichotomous or multi-point scales more clinically useful? An ICD-11 field study. J Schizophr Res (in press).
18. Regier DA, Kaelber CT, Roper MT et al. The ICD-10 clinical field trial for mental and behavioral disorders: results in Canada and the United States. Am J Psychiatry 1994;151:1340-50.
19. Mościcki E, Clarke DE, Kuramoto SJ et al. Testing DSM-5 in routine clinical practice settings: feasibility and clinical utility. Psychiatr Serv 2013;64: 952-60.
20. Maj M. Why the clinical utility of diagnostic categories in psychiatry is intrinsically limited and how we can use new approaches to complement them. World Psychiatry 2018;17:121-2.
21. Clark LA, Cuthbert B, Lewis-Fernandez R et al. ICD-11, DSM-5, and RDoC: three approaches to understanding and classifying mental disorder. Psychol Sci Public Interest 2017;18:72-145.
22. Dolnicar S, Grun B. Cross-cultural differences in survey response patterns. Int Market Rev 2007;24:127-43.

DOI:10.1002/wps.20581