# DSM-5 Field Trials in the United States and Canada, Part I: Study Design, Sampling Strategy, Implementation, and Analytic Approaches

Diana E. Clarke, Ph.D., M.Sc.

William E. Narrow, M.D., M.P.H.

Darrel A. Regier, M.D., M.P.H.

S. Janet Kuramoto, Ph.D., M.H.S.

David J. Kupfer, M.D.

Emily A. Kuhl, Ph.D.

Lisa Greiner, M.S.S.A.

Helena C. Kraemer, Ph.D.

**Objective:** This article discusses the design, sampling strategy, implementation, and data analytic processes of the DSM-5 Field Trials.

**Method:** The DSM-5 Field Trials were conducted by using a test-retest reliability design with a stratified sampling approach across six adult and four pediatric sites in the United States and one adult site in Canada. A stratified random sampling approach was used to enhance precision in the estimation of the reliability coefficients. A web-based research electronic data capture system was used for simultaneous data collection from patients and clinicians across sites and for centralized data management. Weighted descriptive analyses, intraclass kappa and intraclass correlation coefficients for stratified samples, and receiver operating curves were computed. The DSM-5 Field Trials capitalized on advances since DSM-III and DSM-IV in statistical measures of reliability (i.e., intraclass kappa for stratified samples) and other recently developed measures to determine confidence intervals around kappa estimates.

**Results:** Diagnostic interviews using DSM-5 criteria were conducted by 279 clinicians of varied disciplines who received training comparable to what would be available to any clinician after publication of DSM-5. Overall, 2,246 patients with various diagnoses and levels of comorbidity were enrolled, of which over 86% were seen for two diagnostic interviews. A range of reliability coefficients were observed for the categorical diagnoses and dimensional measures.

**Conclusions:** Multisite field trials and training comparable to what would be available to any clinician after publication of DSM-5 provided "real-world" testing of DSM-5 proposed diagnoses.

*(Am J Psychiatry 2013; 170:43–58)*

For more than 10 years, and more specifically over the past 4 years, the American Psychiatric Association (APA) has been revising the diagnostic criteria in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM). The DSM-5 revision process has aimed to use evidence from clinical practice and existing epidemiological, neurobiological, clinical, and genetics literature to develop revised or new diagnostic criteria that better capture the various mental disorders to help clinicians provide more accurate diagnoses. Effective detection and treatment of mental illnesses depend strongly on the accuracy of the conceptualization and operationalization of the diagnostic criteria used in their assessments. However, evidence from the literature indicates that the current diagnostic criteria for a number of mental disorders are unclear and do not adequately capture their complexities, thereby compromising diagnosis and treatment potential (1–5). In particular, persons whose symptom presentations are mixed may exhibit pronounced declines in functioning and quality of life (1–5), but the current categorical structure of DSM diagnoses does not facilitate the assessment of symptoms across disorders. As part of the DSM revision process, the integration of cross-cutting dimensional measures has been proposed. This is seen as a way of addressing the realities of comorbid symptom presentations, allowing clinicians to better assess variations within diagnoses (e.g., accounting for mood and manic symptoms within schizophrenia) and symptoms across diagnoses, and providing longitudinal tracking of patients' symptoms over time (6).

The face and construct validity of the revised DSM-5 diagnoses were subjectively confirmed by the work groups that proposed the diagnostic changes. The diagnostic changes were supported by evidence from literature reviews and secondary data analyses conducted by the work groups. Additional reviews by the general public and by mental health professionals of varied clinical disciplines were done when the criteria were released for public commentary on the DSM-5 web site (www.dsm5.org).

The DSM-5 Field Trials were proposed to objectively evaluate the clinical utility and feasibility and to estimate the reliability and, where possible, validity of the proposed

This article is discussed in an **Editorial** by Dr. Freedman et al. (p.1)

diagnoses and dimensional measures in the environments in which they will be used (7). This entailed testing in clinical populations across multiple sites and using clinicians of various mental health disciplines. The use of multiple sites was necessary to capture the diversity of clinicians who will use the manual in clinical assessments, the diversity of patients who will seek assessments and treatments for their mental illnesses, and the diversity of clinical settings that will require the use of DSM-5. The results of the field trials were intended to inform the DSM-5 decision-making process, but in and of themselves would not determine inclusion or exclusion of diagnoses in the final version of DSM-5.

The most difficult issue to address was the estimation of the reliability coefficients of the categorical diagnoses (i.e., intraclass kappas). The goal was to estimate intraclass kappas with standard errors less than 0.1 for the diagnoses evaluated (7, 8). The design of the field trials was therefore driven primarily by the need to estimate these intraclass kappa coefficients well, which in turn meant that the reliability coefficients of the dimensional measures (i.e., intraclass correlation coefficients [9]) would be well estimated, given the need for smaller sample sizes for those goals. These sample sizes were also sufficient to allow for the examination of clinician assessments of the clinical utility and feasibility of the proposed changes to DSM-5. The aim of this article is to describe and discuss the design, sampling strategy, implementation, and data analytic processes of these field trials.

## Method

### *Study Design, Sample Size, Sampling Strategy*

The DSM-5 Field Trials were conducted over a 7- to 10-month time period in six adult and four pediatric sites in the United States and one adult site in Canada using centrally designed protocols (Table 1). The centrally designed protocols, associated measures, study information sheets, and consent or assent forms were approved by the institutional review boards at the American Psychiatric Institute for Research and Education and the 11 field trial sites. All participating clinicians, principal investigators, and research coordinators completed human subjects training before participating.

The main interest was to determine the degree to which two clinicians would agree on the same diagnosis for patients representative of the DSM clinical population; therefore a design was chosen that was comparable to that used for the DSM-III Field Trials in that the DSM-5 Field Trials were designed, conducted, and analyzed centrally to avoid any biases associated with the work groups evaluating their own work. In contrast to the DSM-III and DSM-IV Field Trials, which were split between interobserver and test-retest reliabilities, the DSM-5 Field Trials focused entirely on the test-retest design. This required that a representative sample of patients from the relevant population be independently evaluated twice using DSM-5 criteria for the diagnoses being tested, ensuring independence of errors—crucial to the estimation of reliability coefficients (8). Specifically, two independent evaluations of each patient were required, with a short (4 hours to 2 weeks) interval between the evaluations. This interval was determined to be long enough to warrant the assumption of independence of the diagnoses at the two study visits but short enough to ensure the occurrence of very few new-onset diagnoses or spontaneous recoveries.

If a simple random sample is used, with prevalences as low as those of many of the diagnoses being evaluated in the DSM-5 Field Trials, the sample size per diagnosis, per site, that is necessary to obtain a standard error less than 0.1 is very large (Figure 1). For example, for a rare diagnosis with a prevalence of 0.05, estimating kappa with a standard error of less than 0.1 requires 28 cases of individuals with the diagnosis, which would require a sample size of 560 patients (Figure 1). This is much larger than was feasible at individual sites in a limited period of time. Furthermore, there are often site differences in reliabilities, depending on the nature of the clinical population samples, clinician experience, and so on (10). As such, an adequate sample size per diagnosis had to be planned at each site so that the reliability of the diagnoses could be estimated, which would then enable comparison of reliabilities across sites and pooling the estimates where appropriate.

To increase the precision of estimation, a stratified random sampling approach was used. This enabled the estimation of kappa with a standard error of less than 0.1 using smaller total sample sizes (Figure 1). Each of the 11 field trial sites was to field test two to five target diagnoses, but some sites, when asked by the APA, chose to test four to seven target DSM-5 diagnoses. The classification into strata was based on the patient's DSM-IV diagnoses corresponding to each of the target DSM-5 diagnoses at the site. For diagnoses that were new to DSM, screening questions on existing symptoms that had a high probability of indicating the new diagnoses were used to stratify patients (Table 2). Consecutive patients at each site were classified into four to seven different strata, one corresponding to each target diagnosis. Patients having DSM-IV diagnoses other than those targeted at the site were classified into a stratum labeled "other diagnosis." Therefore, five to eight strata were assembled at each site.

Because of comorbidity, patients were often eligible for two or more strata, in which case they were assigned for sampling to the stratum that was rarest at that site. In instances where a patient had comorbid conditions that were equal in prevalences, he/she was randomly assigned to either of the strata. Within each stratum, patients were then sampled for testing. This was done to oversample for the target diagnoses and to increase the chance that representative samples of relatively rare categorical diagnoses would be obtained. With the stratified sampling approach, it was found that sampling 50 subjects per stratum would likely result in a standard error less than 0.1 regardless of the prevalence (yet unknown) or the true population kappa (yet unknown). Fifty subjects per stratum was a fail-safe sample size that would work well for all values regardless of the true prevalence and population kappa (Figure 1). However, smaller sample sizes could suffice in some cases, but a lower limit of seven was set.

### *Site Selection and Description*

The seven adult and four pediatric sites were selected from a pool of 49 institutions that submitted applications in response to the Request for Applications posted by the APA in April 2010. Criteria for site selection included overall quality of the application; past experience conducting large clinical studies; and site characteristics that included patient volume, clinician staffing (i.e., minimum of eight participating clinicians at a site), prevalence and type of mental disorders typically seen at the site, and an adequate research infrastructure to accommodate the complexities of the study design.

Eight or more volunteer clinicians of varied psychiatric/mental health disciplines, levels of training (a minimum of 2 years of postgraduate psychiatric training [i.e., PGY-2 or greater]), and years in practice were recruited. All clinicians within a study site were eligible to participate provided they had current human

**TABLE 1. Summary of the DSM-5 Field Trial Recruitment Sites**

| Site | Setting Type | Patient Population | Field Trials Period | Patient Age (years)[a] | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Percentile 25th | 50th | 75th |
| Adult sites | | | | | | | | |
| Centre for Addiction and Mental Health, Toronto, Ont., Canada (CAMH) | General and specialty psychiatry programs | Outpatients | Feb. 11, 2011–Oct. 31, 2011 | 40.04 | 13.28 | 28 | 39 | 51 |
| Dallas VA Medical Center, Dallas, Tex. (Dallas VA) | Veterans Health Administration hospital | Outpatients | Mar. 15, 2011–Oct. 31, 2011 | 49.73 | 13.93 | 39 | 52 | 61 |
| Michael E. DeBakey VA Medical Center and the Menninger Clinic, Houston, Tex. (Houston VA/Menninger) | Veterans Health Administration hospital (Houston VA) and a specialty psychiatric and behavioral hospital (Menninger) | Inpatients and outpatients | Mar. 15, 2011–Oct. 31, 2011 | 43.20 | 14.90 | 30 | 43 | 56 |
| Integrated Mood Clinic & Unit and the Behavioral Medicine Program at Mayo Clinic, Rochester, Minn. (Mayo) | Not-for-profit integrated medical practice and research group—specialty psychiatric and behavioral medicine program | Inpatients and outpatients | Jan. 10, 2011–Oct. 31, 2011 | 48.38 | 17.60 | 35 | 48 | 60 |
| University of Pennsylvania School of Medicine, Philadelphia, Pa. (Penn) | General and specialty psychiatry programs | Outpatients | Jan. 6, 2011–Oct. 31, 2011 | 42.19 | 13.68 | 30 | 42 | 53 |
| Semel Institute for Neuroscience and Human Behavior, Geffen School of Medicine, University of California Los Angeles, Los Angeles, Calif. (UCLA) | Geriatric psychiatry and neuroscience and human behavior programs | Outpatients | Dec. 15, 2010–Sept. 29, 2011 | 73.65 | 10.17 | 67 | 74 | 81 |
| University of Texas San Antonio School of Medicine (UT-SA) | General psychiatry and Veterans Health Administration settings | Inpatients and outpatients | Jan. 7, 2011–Oct. 31, 2011 | 38.23 | 12.60 | 28 | 37 | 48 |
| Pediatric sites | | | | | | | | |
| Child Behavioral Health, Baystate Medical Center, Springfield, Mass. (Baystate) | General child psychiatry | Outpatients | Feb. 17, 2011–Oct. 31, 2011 | 11.02 | 3.44 | 8 | 10 | 14 |
| The Children's Hospital, Aurora, Colo. (Colorado) | General child psychiatry | Inpatients and outpatients | Mar. 3, 2011–Oct. 31, 2011 | 11.74 | 3.48 | 9 | 12 | 15 |
| New York State Psychiatric Institute at Columbia University, New York, N.Y.; Weill Cornell Department of Psychiatry at Payne Whitney, Manhattan Division and Westchester Division, New York and White Plains, N.Y.; North Shore Child and Family Guidance Center, Roslyn Heights, N.Y. (Columbia/Cornell/North Shore) | General child psychiatry | Outpatients | Mar. 15, 2011–Oct. 31, 2011 | 11.81 | 3.43 | 9 | 12 | 15 |
| Stanford University Child & Adolescent Psychiatry Clinic and the Behavioral Medicine Clinic, Palo Alto, Calif. (Stanford) | General child and adolescent psychiatry and specialty psychiatry services | Outpatients | Feb. 24, 2011–Oct. 31, 2011 | 13.32 | 4.13 | 10 | 14 | 16 |

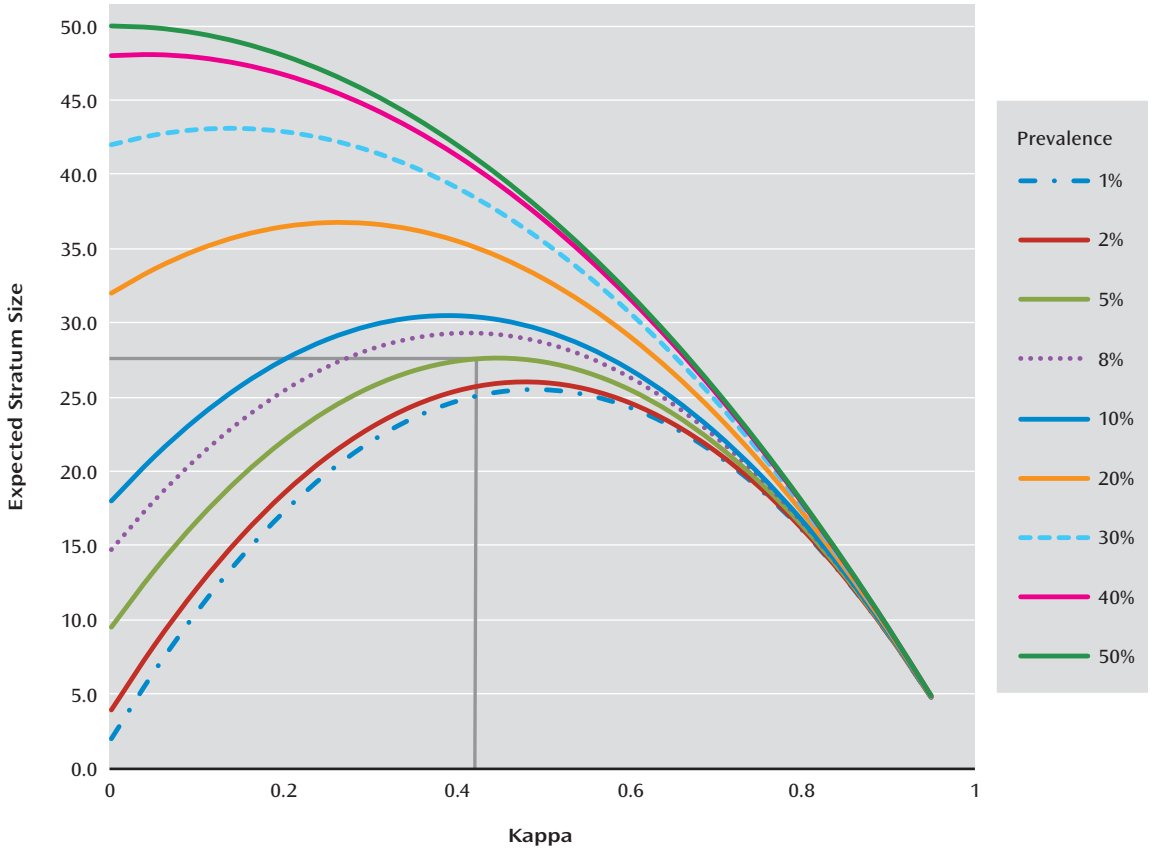
[a] Patient age data missing at some sites (CAMH: N=3; Dallas VA: N=35; UT-SA: N=1; Columbia: N=2).

subjects training and were willing to participate in the DSM-5 Field Trial clinician training sessions. The level of training provided was comparable to what would be available to any clinician after publication of DSM-5 and involved orientation to changes in diagnostic criteria across the DSM, particularly new diagnoses or those with major changes. Participating physicians were provided continuing medical education credits, and all other clinicians were provided certificates of participation that could be used toward obtaining continuing education units from the licensing body for their disciplines. All participating clinicians received remuneration for each patient assessed ($100 per adult patient interview, $150 per child/adolescent patient interview) and were informed that their participation would be acknowledged in DSM-5.

### *Patient Recruitment Process and Sampling Frame*

Patient Recruitment Screening Forms (PRSFs) were completed by intake or treating clinicians on all consecutive patients seen at the site for routine clinic visits during the study period. The PRSF inquired about the patient's age, sex, date of clinic visit, clinic status (i.e., new versus existing patient at the site), length of time in the care of the treating clinician (for existing patients), and whether the patient was currently symptomatic for any DSM-IV diagnoses or had high-probability symptoms associated with the

**FIGURE 1. Expected Number of "Cases" in a Representative Sample From the Population Necessary to Achieve a Standard Error Less Than 0.1 for Various Values of Prevalence and Kappa[a]**




[a] Gray lines represent the example of a diagnosis with a prevalence of 5%, which to estimate kappa with a standard error less than 0.1 would require 28 individuals with the diagnosis and a sample size of 560 patients.

DSM-5 proposed diagnoses being tested at the site (Table 2). "Currently symptomatic" was defined as having enough symptoms to meet criteria for the diagnoses at the time the PRSF was being completed.

In order to maintain blindness to the patient's stratum assignment, clinicians who completed the PRSF were not eligible to complete diagnostic interviews for the patients they screened. Completion of the PRSF on consecutive patients was necessary to obtain the information needed to define the totality of patients seen in the clinic (i.e., the sampling frame) and to obtain the prevalence estimates for each DSM-IV diagnosis that defined a stratum associated with the DSM-5 diagnosis targeted at the site. This information was later used in the calculation of sampling weights.

### *Eligibility Criteria, Stratum Assignments, DSM-IV Prevalence Estimates, and Sampling Weights*

All interested patients, identified on the PRSF, were referred to the research coordinator to determine eligibility and stratum assignment. Eligible patients were those who were currently symptomatic for any DSM-IV diagnoses or high-probability symptoms associated with the DSM-5 diagnoses being tested at the site (i.e., target diagnosis) irrespective of the number of diagnoses and the type and status of treatment. Adult patients without cognitive impairment or other impaired capacity were also required to be able to read and communicate in English. Patients with cognitive impairment or other impaired capacity had to have a caregiver who could read and communicate in English. In the pediatric version of the field trials, patients had to be at least 6 years old and were required to have a parent or guardian who could read and communicate in English, would accompany the patient to the study appointments, and would complete the parent/guardian version of the study measures. At the Colorado site, the lower age limit was 5 years, given the testing of the diagnostic criteria for PTSD in children and adolescents.

Patients who were currently symptomatic with one or more of the target diagnoses being tested at a site were eligible for potential assignment to one of the target diagnosis strata. Patients who were currently symptomatic for any other DSM-IV diagnoses (not including the target diagnoses) were eligible for potential assignment to the "other diagnosis" stratum.

Each enrolled patient was assigned to two randomly selected participating clinicians, who were new to the patient and blinded to the patient's stratum membership for the test (visit 1) and retest (visit 2) diagnostic assessments. Clinicians were blinded to each other's ratings. Each adult patient was offered a remuneration of $40 per study visit. The participating parent or guardian of each pediatric patient was also offered remuneration ($40 per study visit) as was the participating child/adolescent (a $25 gift card).

The estimated DSM-IV prevalence of a diagnosis in each clinic population was the proportion of all "currently symptomatic" patients with that diagnosis as indicated by the patient's intake or treating clinician. Individuals with more than one of the diagnoses being field tested at a site qualified for more than one DSM-IV stratum and contributed to the prevalence estimate for each condition.

For purposes of sampling weights, each patient who qualified for more than one stratum was assigned to the rarest stratum at that site. The sampling weight for each target diagnosis stratum was the proportion of those in the sampling frame assigned for sampling to that stratum. Patients with comorbid conditions contributed only to the sampling weight for the stratum to which they were assigned. The number of patients included in the sampling frame, the DSM-IV prevalence of the targeted diagnoses, and sampling weights for each site are outlined in Table 3.

### Assessment Method and Familiarization

An important decision in the planning process was to have central protocol development, implementation, data collection, as well as ongoing monitoring of the DSM-5 Field Trials, all of which required the use of an electronic data capture system. The National Institutes of Health-funded Research Electronic Data Capture (REDCap) system at Vanderbilt University (11) was modified to meet the needs of the DSM-5 Field Trials. The DSM-5 Field Trial REDCap system included a patient component with all patient-rated measures, programmed for easy access by multiple simultaneous users, and scoring of measures with real-time transmission of results. The clinician component included all clinician-rated dimensional measures and diagnostic checklists and was accessible by multiple clinicians at the same time across different time zones. A research coordinator component enabled careful coordination and monitoring of the workflow within sites while enabling central monitoring of the workflow across sites by the DSM-5 Field Trial Project Manager. Patients could only access their own information, and clinicians could only access information on patients assigned to them. The functionality and ease of use of the patient and clinician components of the DSM-5 REDCap system were pilot tested and the systems modified accordingly before implementation in the DSM-5 Field Trials.

**Familiarization: clinician training.** Clinician training occurred in two parts. Part 1 involved a 1-hour web-based training session introducing the batteries of patient- and clinician-rated DSM-5 cross-cutting dimensional measures, including information on their development and function and how the results should be interpreted and potentially used as diagnostic interviews. Clinicians also had a brief orientation to some of the changes in DSM-5, such as new diagnoses and those with major reconceptualization. The training also included orientation to the DSM-5 Field Trial's REDCap system, including how to log on and access the various DSM-5 diagnostic checklists, clinician-rated dimensional measures, and the results of the patient-rated measures. Clinicians were given unique usernames, passwords, and access to a practice version of the system and encouraged to practice with the system prior to part 2 of the training session. They were also encouraged to familiarize themselves with the proposed changes to the diagnostic criteria across DSM.

Part 2 of the clinician training was an in-person, 3-hour session conducted by the DSM-5 Research Team at APA (D.E.C., W.E.N., D.A.R.). Clinicians were provided with a training manual that outlined the study protocol and study visit workflow (Figure 2), the DSM-5 diagnostic checklists, and clinician- and patient-rated measures. The session included more detailed information on DSM-5 criteria and the DSM-5 dimensional measures being incorporated into the diagnostic schema. A mock clinical interview was conducted to demonstrate the diagnostic interview process and how to incorporate the clinician component of the DSM-5 Field Trials REDCap system. Ongoing interactive Web-based training sessions were provided on an as-needed basis or as new clinicians joined the study.

**Familiarization: research coordinator training.** Given the multisite nature of the DSM-5 Field Trials, it was important to have centralized training of the research coordinators across field trial sites. Each site's lead research coordinator or primary back-up attended a full-day in-person training session conducted by the DSM-5 Research Team at the APA (D.E.C., W.E.N., D.A.R., L.G.). The goal of the session was to familiarize the lead research coordinators with the study protocol (Figure 2), including their roles and responsibilities throughout the study. All research coordinators had to attend a 2-hour interactive web-based session during which they were oriented to the functionality of the research coordinator component of the DSM-5 REDCap system and its connectedness to the patient and clinician components of the system. Ongoing interactive web-based training was available to the sites as needed or as new research coordinators joined the study. Biweekly meetings were held throughout the course of the field trials to immediately address any concerns. Real-time troubleshooting assistance was provided by the APA research team.

### Data Analysis Plan

All analyses were based on the sampling weights associated with the strata at each site and conducted by using SAS statistical software and SUDAAN, where necessary. Descriptive statistics (mean, standard deviation, quartiles, correlation coefficients, and frequency distributions) were estimated for the study population at each site (i.e., patients and clinicians) and for each dimensional measure.

**Reliability of the categorical diagnoses/variables.** Test-retest reliability for the categorical (binary) diagnoses was based on the intraclass kappa (estimated for a stratified sample) and presented with a two-tailed 95% confidence interval (CI) using bootstrap methods (12, 13). Intraclass kappa is the difference between the probabilities of getting a second positive diagnosis between those with a first positive and those with a first negative diagnosis (14), thus reflecting the predictive value of a first test to a second. Given the stratified sampling approach for the study, sampling weights for each site were used to obtain unbiased site-specific estimates of intraclass kappa for each categorical diagnosis tested. Equations 1 and 2 below were used to calculate intraclass kappa coefficients for a stratified sample, for each diagnosis tested.

$$K_x = \frac{\sum_i w_i^*(Q_{i2} + Q_{i0}) - \left(P_x^2 + (1 - P_x)^2\right)}{2P_x(1 - P_x)} \qquad \text{(Eq. 1)}$$

$$P_x = \sum w_i^* * Q_{i2} + 0.5 \sum w_i^* * Q_{i1} \qquad \text{(Eq. 2)}$$

Where:

$w_i^*$ = the sample weight = proportion assigned for sampling in a particular stratum. If a patient was eligible for two or more strata, he/she was assigned to the rarest stratum.

$Q_{i2}$ = proportion of those in stratum $i$ where both clinicians diagnosed the particular diagnosis $X$.

$Q_{i1}$ = proportion of those in stratum $i$ where only one of the two clinicians diagnosed the particular diagnosis $X$.

$Q_{i0}$ = proportion of those in stratum $i$ where both clinicians did NOT diagnose the particular diagnosis $X$.

Note: $Q_{i2} + Q_{i1} + Q_{i0} = 1$

$P_x$ = the overall prevalence of the target diagnosis in the population.

To obtain the 95% CIs on the kappa, a bootstrap method was utilized (12). The simple meta-analytic approach, which involved the weighted average of the reliability coefficients, was used to calculate pooled intraclass kappa estimates and their 95% CIs for diagnoses tested at two or more sites. In instances where the 95% CIs for intraclass kappas for the same diagnosis did not overlap for all sites at which it was tested, a cautionary note was associated with the pooled estimate. The results of the test-retest reliability of the DSM-5 categorical diagnoses tested in these field

**TABLE 2. Criteria Defining Each Stratum in the DSM-5 Field Trials[a]**

| Site Type and Stratum | Criteria Defining Stratum | Site Assigning to Stratum |
|---|---|---|
| Adult sites | | |
| Alcohol use disorder | DSM-IV diagnosis of any alcohol use disorder (i.e., any substance abuse or dependence) | Houston VA/Menninger |
| Antisocial personality disorder | DSM-IV diagnosis of antisocial personality disorder OR personality disorder with antisocial features | Dallas VA |
| Attenuated psychosis syndrome | Evidence of delusions, hallucinations, and/or disorganized speech in attenuated form with intact reality testing, but of sufficient severity and/or frequency so as to be beyond normal variation | CAMH; UT-SA |
| Binge eating disorder | Evidence of uncontrollable binge eating (i.e., discrete episodes in which the individual uncontrollably eats a larger amount than most people would in a similar time and under similar circumstances) | Penn |
| Bipolar I disorder | DSM-IV diagnosis of bipolar I disorder | Mayo; UT-SA |
| Bipolar II disorder | DSM-IV diagnosis of bipolar II disorder | Mayo |
| Borderline personality disorder | DSM-IV diagnosis of borderline personality disorder OR personality disorder with borderline features | CAMH; Houston VA/Menninger |
| Complex somatic symptom disorder | DSM-IV diagnosis of somatization disorder<br>DSM-IV diagnosis of hypochondriasis<br>DSM-IV diagnosis of pain disorder associated with psychological factors<br>DSM-IV diagnosis of pain disorder associated with psychological factors and a general medical condition<br>AND/OR<br>DSM-IV diagnosis of undifferentiated somatoform disorder | Mayo |
| Generalized anxiety disorder | DSM-IV diagnosis of generalized anxiety disorder | Penn |
| Hoarding disorder | Evidence of persistent difficulties discarding or parting with possessions regardless of their value | Penn |
| Major depressive disorder | DSM-IV diagnosis of major depressive disorder | Houston VA/Menninger; UT-SA; Dallas VA; UCLA |
| Major neurocognitive disorder | Evidence or a history of ANY dementia disorder (e.g., dementia of Alzheimer's disease type, vascular dementia, dementia due to general medical condition, dementia due to multiple etiologies, or dementia not otherwise specified) or amnestic disorder | Mayo; UCLA |
| Mild neurocognitive disorder | Evidence of mild cognitive impairment, cognitive complaints, or memory complaints (but not a diagnosis of dementia such as dementia, dementia not otherwise specified, Alzheimer's disease, vascular dementia, Lewy body dementia, Parkinson's dementia, etc.) | Dallas VA; Mayo; UCLA |
| Mild traumatic brain injury (TBI) | History of head injury, mild TBI, or postconcussional syndrome | Houston VA/Menninger; Dallas VA |
| Mixed anxiety-depressive disorder | Current co-occurring subsyndromal depression and generalized anxiety (note: regardless of having a past diagnosis of major depressive episode, major depressive disorder, or generalized anxiety disorder) | Penn; UCLA |
| Narcissistic personality disorder | DSM-IV diagnosis of narcissistic personality disorder or personality disorder with narcissistic features | Houston VA/Menninger |
| Obsessive-compulsive personality disorder | DSM-IV diagnosis of obsessive-compulsive personality disorder OR personality disorder with obsessive-compulsive features | Houston VA/Menninger; Penn |
| Posttraumatic stress disorder | DSM-IV diagnosis of posttraumatic stress disorder | Houston VA/Menninger; Dallas VA |
| Schizoaffective disorder | Current DSM-IV diagnosis of schizoaffective disorder | CAMH |
| Schizophrenia | Current DSM-IV diagnosis of schizophrenia | CAMH; UT-SA |
| Schizotypal personality disorder | DSM-IV diagnosis of schizotypal personality disorder OR personality disorder with schizotypal features | CAMH |
| Other diagnosis | Currently symptomatic for any DSM-IV disorder(s) not including any of the disorders/conditions listed above | All sites |
| Pediatric sites | | |
| Attention deficit hyperactivity disorder (ADHD) | DSM-IV diagnosis of ADHD | Baystate; Columbia/Cornell/North Shore |

*continued*

**TABLE 2. Criteria Defining Each Stratum in the DSM-5 Field Trials[a] (*continued*)**

| Site Type and Stratum | Criteria Defining Stratum | Site Assigning to Stratum |
|---|---|---|
| Autism spectrum disorder | DSM-IV diagnosis of:<br>a. Autistic disorder (autism)<br>b. Asperger's disorder<br>c. Childhood disintegrative disorder<br>AND/OR<br>d. Pervasive developmental disorder not otherwise specified | Baystate; Stanford |
| Avoidant/restrictive food intake disorder | Exhibiting an eating or feeding disturbance manifested by persistent failure to meet appropriate nutritional and/or energy needs<br>AND a positive response to one or more of the following:<br>a. eating or feeding disturbance associated with significant weight loss (or if the individual is a child, failure to achieve expected weight gain or faltering growth)<br>b. eating or feeding disturbance associated with significant nutritional deficiency<br>c. eating or feeding disturbance associated with dependence on enteral feeding<br>d. eating or feeding disturbance associated with marked interference with psychosocial functioning | Stanford |
| Bipolar disorder | DSM-IV diagnosis of:<br>a. Bipolar I disorder<br>OR<br>b. Bipolar disorder not otherwise specified | Baystate |
| Conduct disorder | DSM-IV diagnosis of conduct disorder | Colorado; Columbia/Cornell/North Shore |
| Disruptive mood dysregulation disorder | Evidence of explosive aggressive behaviors<br>AND/OR<br>If the patient is a child/adolescent, a history of severe temper tantrums<br>AND answered NO to,<br>DSM-IV diagnosis of mental retardation or pervasive developmental disorder | Baystate; Colorado; Columbia/Cornell/North Shore |
| Major depressive disorder | DSM-IV diagnosis of major depressive disorder | Colorado; Stanford |
| Mixed anxiety-depressive disorder | Current co-occurring subsyndromal depression and generalized anxiety (note: regardless of having a past diagnosis of major depressive episode, major depressive disorder, or generalized anxiety disorder) | Colorado; Stanford |
| Nonsuicidal self-injury | Current or history of self-injurious ideation or behaviors | Baystate; Colorado; Columbia/Cornell/North Shore |
| Oppositional defiant disorder | DSM-IV diagnosis of oppositional defiant disorder | Baystate; Colorado; Columbia/Cornell/North Shore |
| Posttraumatic stress disorder in children/adolescents | DSM-IV diagnosis of posttraumatic stress disorder | Baystate; Colorado |
| Substance use disorder | DSM-IV diagnosis of any substance use disorder (i.e., any substance abuse or dependence)<br>If Yes and currently symptomatic, specify the substance use disorder or dependence (e.g., alcohol abuse, alcohol dependence, etc.) | Columbia/Cornell/North Shore |
| Other diagnosis | Currently symptomatic for any DSM-IV disorder(s) not including any of the disorders/conditions listed above | All sites |

[a] Information is based on what is known about the patient before DSM-5 criteria are applied.

trials are presented in an accompanying article by Regier and colleagues (15).

The following standards were set for the reliability coefficients for DSM-5 categorical diagnoses: intraclass kappas of 0.8 and above were "excellent"; from 0.60 to 0.79 were "very good"; from 0.40 to 0.59 were "good"; from 0.20 to 0.39 were "questionable"; and values below 0.20 were "unacceptable" (8). The goal of the DSM-5 Field Trials was to attain intraclass kappas at least in the "good" reliability range (8).

**Reliability of the dimensional measures.** Test-retest reliability estimates for dimensional measures (continuous and ordinal) were estimated using parametric intraclass correlation coefficients (ICCs) and presented with their two-tailed 95% confidence intervals, using sampling weights and bootstrap methods. The parametric ICC is a measurement of agreement or consensus between two or more raters on the same set of subjects where the measures are assumed to be ordinal or continuous and to be normally distributed (9). The ICC is a "relative measure of reliability" in that it reflects a ratio of the variability between subjects to the total variability in the population sampled (16, 17). The parametric ICC was used because of its reported robustness (9) and because it reflects the predictive value of a first measure to a second.

**TABLE 3. Summary of Patients Screened, Stratified for Sampling, and Seen at Visits 1 and 2 and Sample Weight per Stratum by Field Trial Site**

| Site | Target Diagnosis | Screened Into Stratum[a] N | Screened Into Stratum[a] % | Number Assigned to Stratum for Sampling | Number Reassigned to Stratum for Sampling | Sample Weight | Completed Visit 1 | Completed Visits 1 and 2 |
|---|---|---|---|---|---|---|---|---|
| Adult sites | | | | | | | | |
| CAMH (N=878) | Schizophrenia | 469 | 0.534 | 458 | | 0.522 | 56 | 53 |
| | Schizoaffective disorder | 125 | 0.142 | 120 | | 0.137 | 50 | 49 |
| | Attenuated psychosis syndrome | 28 | 0.032 | 25 | | 0.029 | 18 | 17 |
| | Schizotypal personality disorder | 23 | 0.026 | 23 | | 0.026 | 12 | 9 |
| | Borderline personality disorder | 52 | 0.059 | 52 | | 0.059 | 43 | 39 |
| | Other | 199 | 0.227 | 199 | 200 | 0.228 | 63 | 62 |
| | Antisocial personality disorder[b] | 2 | 0.002 | 1 | | | | |
| Dallas VA (N=1,119) | PTSD | 555 | 0.496 | 197 | | 0.176 | 53 | 51 |
| | Mild neurocognitive disorder | 233 | 0.208 | 162 | | 0.145 | 48 | 48 |
| | Major depressive disorder | 546 | 0.488 | 442 | | 0.395 | 33 | 31 |
| | Mild TBI | 85 | 0.076 | 79 | | 0.071 | 26 | 25 |
| | Antisocial personality disorder | 53 | 0.047 | 53 | | 0.047 | 30 | 26 |
| | Other | 186 | 0.166 | 186 | | 0.166 | 53 | 51 |
| Houston VA/ Menninger (N=868) | PTSD | 405 | 0.467 | 159 | | 0.183 | 44 | 43 |
| | Alcohol use disorder | 229 | 0.264 | 168 | | 0.194 | 47 | 46 |
| | Major depressive disorder | 293 | 0.338 | 161 | | 0.186 | 60 | 57 |
| | Mild TBI | 132 | 0.152 | 128 | | 0.148 | 40 | 39 |
| | Borderline personality disorder | 115 | 0.132 | 115 | | 0.133 | 45 | 44 |
| | Other | 132 | 0.152 | 132 | 137 | 0.158 | 36 | 35 |
| | Obsessive-compulsive personality disorder[b] | 2 | 0.002 | 0 | | | | |
| | Narcissistic personality disorder[b] | 8 | 0.009 | 5 | | | | |
| Mayo (N=458) | Bipolar I disorder | 114 | 0.249 | 108 | | 0.236 | 26 | 25 |
| | Mild neurocognitive disorder | 60 | 0.131 | 58 | | 0.127 | 45 | 40 |
| | Complex somatic symptom disorder | 47 | 0.103 | 46 | | 0.100 | 42 | 37 |
| | Major neurocognitive disorder | 30 | 0.066 | 30 | | 0.066 | 24 | 23 |
| | Bipolar II disorder | 82 | 0.179 | 80 | | 0.175 | 30 | 25 |
| | Other | 136 | 0.297 | 136 | | 0.297 | 42 | 34 |
| Penn (N=582) | Mixed anxiety-depressive disorder | 190 | 0.326 | 161 | | 0.277 | 47 | 45 |
| | Generalized anxiety disorder | 200 | 0.344 | 99 | | 0.170 | 48 | 47 |
| | Binge eating disorder | 59 | 0.101 | 52 | | 0.089 | 29 | 28 |
| | Hoarding | 32 | 0.055 | 32 | | 0.055 | 17 | 16 |
| | Obsessive-compulsive personality disorder | 41 | 0.070 | 36 | | 0.062 | 8 | 8 |
| | Other | 202 | 0.347 | 202 | | 0.347 | 54 | 53 |
| UCLA (N=488) | Mixed anxiety-depressive disorder | 73 | 0.150 | 73 | | 0.150 | 49 | 48 |
| | Mild neurocognitive disorder | 88 | 0.180 | 84 | | 0.172 | 46 | 43 |
| | Major depressive disorder | 128 | 0.262 | 107 | | 0.219 | 50 | 50 |
| | Major neurocognitive disorder | 105 | 0.215 | 98 | | 0.201 | 29 | 26 |
| | Other | 126 | 0.258 | 126 | | 0.258 | 46 | 42 |
| UT-SA (N=735) | Bipolar I disorder | 206 | 0.280 | 204 | | 0.278 | 46 | 38 |
| | Schizophrenia | 114 | 0.155 | 114 | | 0.155 | 31 | 30 |
| | Major depressive disorder | 157 | 0.214 | 156 | | 0.212 | 43 | 35 |
| | Other | 251 | 0.341 | 251 | 261 | 0.355 | 56 | 48 |
| | Attenuated psychosis syndrome[c] | 10 | 0.014 | 10 | | | | |

*continued*

**TABLE 3. Summary of Patients Screened, Stratified for Sampling, and Seen at Visits 1 and 2 and Sample Weight per Stratum by Field Trial Site (*continued*)**

| Site | Target Diagnosis | Screened Into Stratum[a] N | Screened Into Stratum[a] % | Number Assigned to Stratum for Sampling | Number Reassigned to Stratum for Sampling | Sample Weight | Completed Visit 1 | Completed Visits 1 and 2 |
|---|---|---|---|---|---|---|---|---|
| Pediatric sites | | | | | | | | |
| Baystate (N=569) | Disruptive mood dysregulation disorder | 134 | 0.236 | 111 | | 0.195 | 30 | 25 |
| | ADHD | 336 | 0.590 | 157 | | 0.276 | 49 | 41 |
| | Nonsuicidal self-injury | 45 | 0.079 | 41 | | 0.072 | 10 | 7 |
| | Autism spectrum disorder | 132 | 0.232 | 112 | | 0.197 | 37 | 34 |
| | Bipolar disorder | 36 | 0.063 | 36 | | 0.063 | 14 | 12 |
| | Other | 95 | 0.167 | 95 | 112 | 0.197 | 28 | 27 |
| | PTSD[b] | 30 | 0.053 | 11 | | | | |
| | Oppositional defiant disorder[b] | 37 | 0.065 | 6 | | | | |
| Colorado (N=1,047) | Disruptive mood dysregulation disorder | 425 | 0.406 | 171 | | 0.163 | 54 | 51 |
| | Major depressive disorder | 221 | 0.211 | 155 | | 0.148 | 37 | 32 |
| | Mixed anxiety-depressive disorder | 293 | 0.280 | 207 | | 0.198 | 43 | 40 |
| | PTSD | 152 | 0.145 | 122 | | 0.116 | 14 | 13 |
| | Conduct disorder | 82 | 0.078 | 82 | | 0.078 | 19 | 16 |
| | Other | 294 | 0.281 | 294 | 310 | 0.296 | 53 | 47 |
| | Oppositional defiant disorder[b] | 68 | 0.065 | 13 | | | | |
| | Nonsuicidal self-injury[b] | 24 | 0.023 | 3 | | | | |
| Columbia/Cornell/ North Shore (N=582) | Disruptive mood dysregulation disorder | 125 | 0.215 | 103 | | 0.177 | 20 | 19 |
| | ADHD | 320 | 0.550 | 216 | | 0.371 | 55 | 53 |
| | Nonsuicidal self-injury | 55 | 0.094 | 55 | | 0.094 | 8 | 7 |
| | Oppositional defiant disorder | 129 | 0.222 | 66 | | 0.113 | 16 | 14 |
| | Other | 128 | 0.220 | 128 | 142 | 0.244 | 32 | 29 |
| | Substance use disorder[c] | 24 | 0.041 | 10 | | | | |
| | Conduct disorder[c] | 17 | 0.029 | 4 | | | | |
| Stanford (N=463) | Avoidant/restrictive food intake disorder | 70 | 0.151 | 70 | | 0.151 | 32 | 27 |
| | Major depressive disorder | 96 | 0.207 | 68 | | 0.147 | 31 | 28 |
| | Mixed anxiety-depressive disorder | 71 | 0.153 | 57 | | 0.123 | 23 | 21 |
| | Autism spectrum disorder | 119 | 0.257 | 98 | | 0.212 | 31 | 30 |
| | Other | 170 | 0.367 | 170 | | 0.367 | 45 | 43 |

[a] Ns exceed total N seen at site and percentages exceed 1 because of comorbidity.
[b] Stratum added to the site on July 25, 2011, but was folded into "other" stratum for sample weights and visit completion counts because of sample size of 6 or less who completed at least one visit.
[c] Stratum assigned from the start of the study but was folded into "other" stratum for sample weights and visit completion counts because of sample size of 6 or less who completed at least one visit.

Two ICC models were used in this study: Type- (1, 1), a one-way random model of absolute agreement and Type- (2, 1), a two-way random model of absolute agreement. A one-way random model of absolute agreement was used when determining the reliability estimates for each clinician-rated dimensional measure given that each patient was rated by a different and randomly selected clinician from a pool of participating clinicians within each site (9). The two-way random model of absolute agreement was used when determining the reliability estimate for each patient-rated dimensional measure since each patient was rated by the same raters (i.e., self or proxy [9]). Site-specific reliability coefficients were calculated for each dimensional measure. Pooled estimates, based on a meta-analytic approach, were also calculated since the same measures were used across sites. In instances where the 95% CI for the ICC estimates for the same dimensional measure did not overlap for all sites, a cautionary note was associated with the pooled estimate.

The robustness of the parametric ICC was checked by using a nonparametric ICC for comparison. In the nonparametric approach, patient scores on each dimensional measure were ranked and the one-way and two-way random ICC models of absolute agreement were used to estimate the reliability coefficient. In general, the parametric ICC method was more conservative and therefore reported for the field trials of the DSM-5 cross-cutting dimensional measures in the accompanying paper by Narrow et al. (6).

The standards proposed for the DSM-5 dimensional measures were as follows: ICCs over 0.80 were "very good"; from 0.60 to 0.79 were "good"; from 0.40 to 0.59 were "questionable"; and values below 0.40 were "unacceptable" (8). These standards correspond to those for IQ testing, for example (16, 17). These standards, like any standards, are suggestions. In this case, they are based on existing reliability estimates of psychometric tests that yield dimensional outcomes (16–19).

**FIGURE 2. Steps Involved in the Baseline Study Visit (TEST) and First Follow-Up Study Visit (RETEST: 4 hours to 2 weeks later) in the DSM-5 Field Trials**

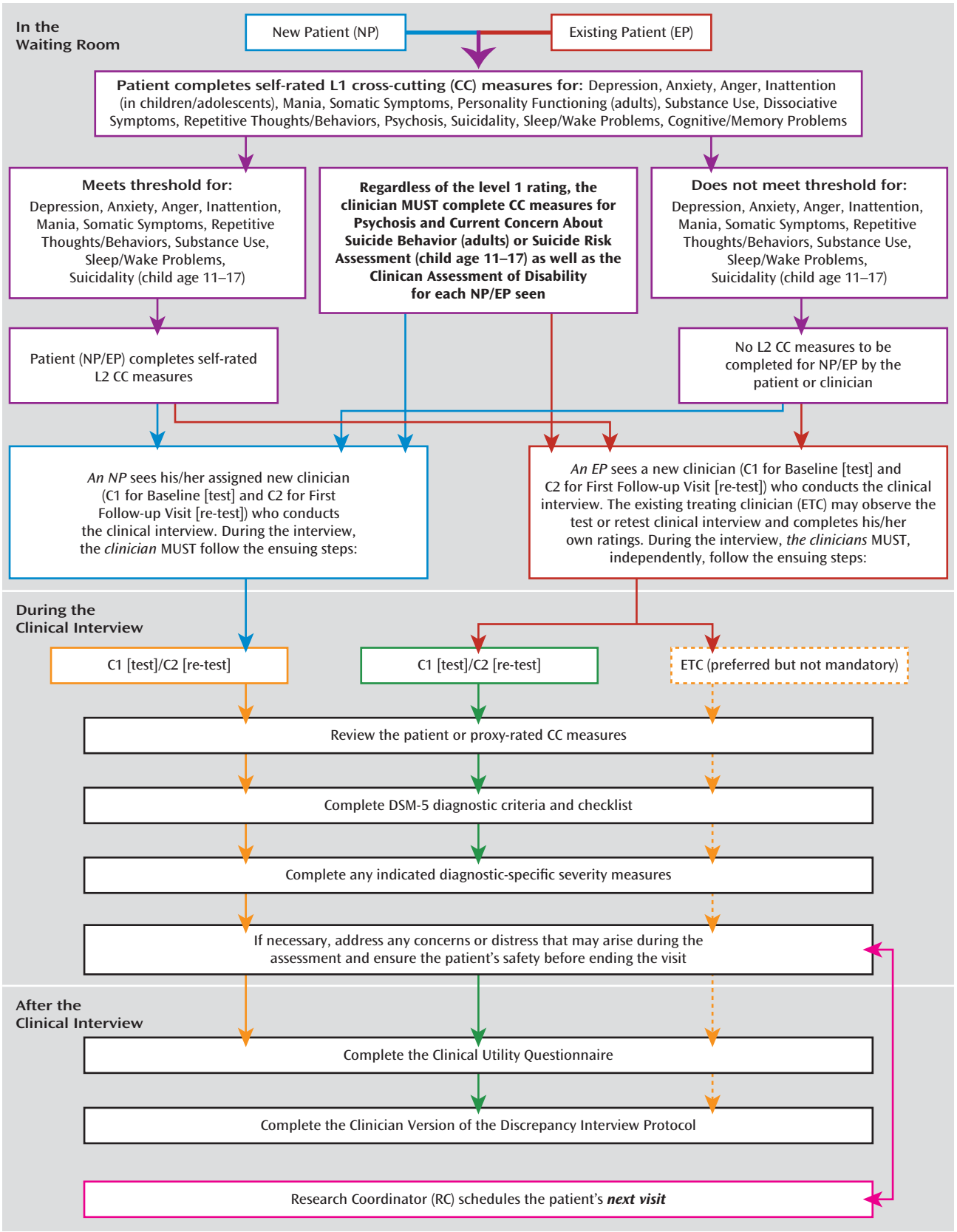


**TABLE 4. Patient Demographic Characteristics Across the DSM-5 Field Trial Adult Study Sites**

| Characteristic | CAMH (N=242) | | Dallas VA[a] (N=243) | | Houston VA/ Menninger (N=272) | | Mayo (N=209) | | Penn (N=203) | | UCLA (N=220) | | UT-SA (N=176) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean[b] | SD | Mean[b] | SD | Mean[b] | SD | Mean[b] | SD | Mean[b] | SD | Mean[b] | SD | Mean[b] | SD |
| Age (years) | 40.2 | 6.3 | 52.2 | 4.5 | 38.0 | 5.6 | 51.9 | 7.0 | 43.6 | 6.4 | 70.5 | 4.6 | 37.2 | 6.1 |
| 25th percentile | 28 | | 46 | | 26 | | 40 | | 31 | | 65 | | 26 | |
| 50th percentile | 39 | | 54 | | 35 | | 51 | | 45 | | 70 | | 36 | |
| 75th percentile | 51 | | 61 | | 51 | | 63 | | 54 | | 78 | | 47 | |
| | N | %[b] | N | %[b] | N | %[b] | N | %[b] | N | %[b] | N | %[b] | N | %[b] |
| Male | 120 | 60.4 | 208 | 85.0 | 172 | 64.3 | 72 | 32.9 | 62 | 32.7 | 73 | 32.8 | 101 | 58.0 |
| Hispanic or Latino origin | 4 | 2.8 | 18 | 7.7 | 29 | 9.1 | 3 | 2.0 | 12 | 6.3 | 17 | 7.7 | 95 | 51.8 |
| Mexican | 0 | | 12 | | 11 | | 1 | | 3 | | 11 | | 82 | |
| Puerto Rican | 0 | | 1 | | 7 | | 0 | | 5 | | 2 | | 4 | |
| Cuban | 0 | | 1 | | 3 | | 0 | | 1 | | 0 | | 0 | |
| Other | 4 | | 4 | | 8 | | 2 | | 3 | | 4 | | 9 | |
| Married/cohabiting | 34 | 12.0 | 101 | 47.9 | 65 | 23.8 | 142 | 63.9 | 47 | 22.6 | 74 | 32.3 | 34 | 20.2 |
| Race/ethnicity | | | | | | | | | | | | | | |
| White/Caucasian | 175 | 68.1 | 129 | 55.8 | 182 | 66.2 | 200 | 95.2 | 112 | 52.3 | 177 | 80.1 | 139 | 77.2 |
| Black, African descent | 20 | 9.2 | 104 | 40.3 | 70 | 26.5 | 2 | 0.9 | 73 | 38.0 | 21 | 9.5 | 21 | 13.2 |
| Other | 47 | 22.8 | 9 | 3.9 | 20 | 7.3 | 7 | 4.0 | 18 | 9.8 | 22 | 10.4 | 16 | 9.6 |
| Level of education | | | | | | | | | | | | | | |
| Less than high school | 46 | 22.7 | 11 | 4.6 | 6 | 2.2 | 6 | 2.1 | 19 | 10.4 | 7 | 3.3 | 38 | 19.7 |
| Completed high school | 43 | 21.6 | 78 | 29.6 | 54 | 20.2 | 31 | 14.8 | 39 | 20.4 | 22 | 9.8 | 48 | 27.0 |
| Greater than high school | 149 | 53.4 | 149 | 64.0 | 211 | 77.3 | 171 | 82.4 | 144 | 68.8 | 190 | 86.5 | 88 | 52.0 |
| Other | 4 | 2.2 | 4 | 1.7 | 1 | 0.3 | 1 | 0.7 | 1 | 0.4 | 1 | 0.5 | 2 | 1.2 |

[a] Data missing for one subject.
[b] Weighted.

**Convergent validity.** To examine convergent validity, receiver operating characteristic (ROC) curves (20, 21) were used. ROC curves were used to examine the association between the clinician- and patient-rated dimensional measures and their associated categorical diagnoses. To maintain the assumption of independence of the ratings, clinician-rated dimensional measures of one clinician were compared with the categorical ratings by the second, independent, and "blinded" clinician. Similarly, since clinicians were privy to the patient-rated results simultaneous to completion of the categorical diagnoses, patient-rated dimensional measures at visit 1 were compared with the categorical diagnoses at visit 2 and vice versa. These results will be presented in an upcoming article.

## Results

Overall, 7,789 patients were seen across the 11 field trial sites and screened during the study period (5,128 in adult sites combined and 2,661 in pediatric sites combined [Table 3]). Of these, 4,110 were interested, eligible, and assigned for sampling (N=2,791 and 1,319 across the adult and pediatric sites, respectively). Written informed consent was obtained for 1,755 of 2,791 adult patients and 689 of 1,319 child/adolescent patients. The majority of the patients who provided written consent for field trial participation completed visit 1 (N=2,246 of 2,444 patients overall; 78%–98% in the adult field trials and more than 98% in the pediatric field trials). The demographic characteristics of these patients are presented in Table 4 (for adult sites) and Table 5 (for pediatric sites). Overall, more than 86% of the patients who completed visit 1 also completed visit 2.

The patient population across adult (Table 4) and pediatric (Table 5) field trial sites varied. For instance, compared with the other six adult field trial sites, UT-SA had a larger proportion of Hispanic patients (51.2% relative to less than 15% in the other sites). Indeed, the high proportion of Hispanic/Latino patients was one factor in the site being selected for the field trials. The proportion of patients of black/African-American descent varied from 0.9% at the Mayo site to 40% at the Dallas VA site. Similarly, the proportion of male patients varied from 32.7% at the Penn site to 85% at the Dallas VA site. Among the pediatric sites, the patient populations at Baystate and Columbia/Cornell consisted of greater than 40% Hispanics compared with 15% and 12% at Colorado and Stanford. The proportion of patients of Black/African-American descent was about 10% at Baystate, Colorado, and Columbia compared with <1% at Stanford. At Stanford, a majority of the patients (73.1%) lived in two-parent households compared with 48.5%, 52.4%, and 57.0% at Baystate, Columbia, and Colorado respectively. Differences in the patient population across sites were expected given the variability in the sites selected for the DSM-5 Field Trials (e.g., general psychiatry, Veterans Health Administration, and geriatric psychiatry settings).

Two hundred eighty-six clinicians from various clinical disciplines participated in the DSM-5 Field Trials.

**TABLE 5. Patient Demographic Characteristics Across the DSM-5 Field Trial Pediatric Study Sites**

| Characteristic | Baystate[a] (N=168) | | Colorado[a] (N=220) | | Columbia/Cornell/North Shore (N=131) | | Stanford (N=162) | |
|---|---|---|---|---|---|---|---|---|
| | Mean[b] | SD | Mean[b] | SD | Mean[b] | SD | Mean[b] | SD |
| Age | 11.0 | 1.5 | 10.9 | 1.4 | 11.5 | 1.7 | 13.8 | 2.0 |
| 25th percentile | 8 | | 8 | | 9 | | 10 | |
| 50th percentile | 10 | | 11 | | 11 | | 15 | |
| 75th percentile | 14 | | 14 | | 14 | | 17 | |
| | N | %[b] | N | %[b] | N | %[b] | N | %[b] |
| Age group | | | | | | | | |
| Under 11[c] | 81 | 51.4 | 98 | 46.4 | 58 | 48.0 | 37 | 26.5 |
| 11-17[d] | 86 | 48.9 | 122 | 53.6 | 68 | 49.1 | 89 | 54.0 |
| ≥18[e] | | | | | 5 | 2.9 | 36 | 19.5 |
| Male | 114 | 69.7 | 133 | 60.7 | 79 | 63.8 | 82 | 54.6 |
| Hispanic or Latino origin | 64 | 41.2 | 35 | 15.5 | 57 | 44.1 | 20 | 12.1 |
| Mexican | 0 | | 20 | | 2 | | 13 | |
| Puerto Rican | 57 | | 5 | | 20 | | 0 | |
| Cuban | 1 | | 0 | | 3 | | 0 | |
| Other | 6 | | 10 | | 32 | | 7 | |
| Race/ethnicity | | | | | | | | |
| White/Caucasian | 89 | 49.6 | 168 | 77.3 | 69 | 52.7 | 126 | 77.0 |
| Black, African descent | 17 | 11.0 | 21 | 9.7 | 10 | 9.6 | 1 | 0.6 |
| Other/mixed | 61 | 39.4 | 31 | 13.0 | 51 | 37.7 | 35 | 22.4 |
| Current living situation | | | | | | | | |
| Lives with both parents | 83 | 48.5 | 120 | 57.0 | 70 | 52.4 | 116 | 73.1 |
| Lives with only one parent | 67 | 41.4 | 86 | 36.8 | 54 | 44.0 | 28 | 16.1 |
| Lives with other relative/s (not including parent/s) | 8 | 5.0 | 6 | 2.4 | 2 | 1.8 | 0 | 0.0 |
| Lives with foster parents | 2 | 1.2 | 0 | 0.0 | 0 | | 1 | 0.4 |
| Other | 7 | 3.9 | 8 | 3.8 | 4 | 1.8 | 17 | 10.4 |
| Level of education | | | | | | | | |
| None | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 10 | 6.7 |
| Kindergarten | 10 | 5.9 | 16 | 7.4 | 1 | 0.7 | 3 | 1.6 |
| Elementary/grade school | 75 | 46.3 | 100 | 47.7 | 57 | 45.9 | 38 | 27.2 |
| Junior high/middle school | 35 | 22.7 | 49 | 21.9 | 26 | 19.5 | 24 | 13.7 |
| High school | 41 | 22.0 | 53 | 22.3 | 42 | 31.1 | 63 | 38.4 |
| College | 6 | 3.1 | 2 | 0.8 | 4 | 2.7 | 24 | 12.4 |

[a] Data missing for one subject.
[b] Weighted.
[c] Patients who were younger than age 11 were not required to complete the patient-completed measures and therefore have no associated patient-completed measures. For these patients, only parent-completed measures are available.
[d] Some of the patients aged 11-17 who completed visit 1 were unable to read and understand what they read or what was read to them and therefore have no patient-completed measures (Baystate: N=6; Colorado: N=4; Columbia: N=7; Stanford: N=8).
[e] Although the patients were 18 and older, Columbia's IRB mandate stated that these patients had to be accompanied by a parent/guardian who completed the parent-completed measures. At Stanford, patients 18-25 were not required to have a parent/guardian present, and so the majority have no associated parent-completed measures, but two were accompanied to the interview by a parent/guardian who completed the associated parent-completed measures. Of these two patients, one was unable to read or understand what was read and therefore has no patient-completed measures.

Participating clinicians included board-certified psychiatrists and trainees (PGY 2+), licensed clinical and counseling psychologists and neuropsychologists (i.e., doctorate-level training) and those in supervised practice, master's-level counselors, licensed clinical social workers, and advanced practice licensed mental health nurses. Of the 286 clinicians, seven functioned purely as intake or referring clinicians and did not complete any diagnostic interviews. The remaining 279 clinicians completed, on average, seven or more diagnostic interviews. The characteristics of the clinicians who completed the diagnostic interviews in the DSM-5 Field Trials are presented in Table 6 (for the adult sites) and Table 7 (for the pediatric sites). Clinicians who participated in diagnostic interviews in the DSM-5 Field Trials varied by clinical discipline, years in practice, and other clinician characteristics. Having the diagnostic changes to the DSM tested by clinicians of varied disciplines and other characteristics was a goal of the DSM-5 Field Trials. The variations in the clinician discipline and experience may, however, limit the ability to compare reliability estimates across field trial sites and should be taken into consideration when considering the results presented in subsequent articles (6, 15). The results of the quantitative and qualitative analyses of the

**TABLE 6. Clinician Characteristics Across the DSM-5 Field Trial Adult Study Sites**

| Characteristic | CAMH (N=21) | | Dallas VA (N=21) | | Houston VA/ Menninger (N=33) | | Mayo (N=21) | | Penn[a] (N=23) | | UCLA (N=10) | | UT-SA (N=19) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % | N | % | N | % | N | % |
| Discipline | | | | | | | | | | | | | | |
| Board-certified psychiatrist | 11 | 52.4 | 8 | 38.1 | 12 | 36.4 | 11 | 52.4 | 4 | 17.4 | 3 | 30.0 | 5 | 26.3 |
| Psychiatrists in training (PGY2-5) | 0 | 0.0 | 0 | 0.0 | 1 | 3.0 | 4 | 19.0 | 3 | 13.0 | 0 | 0.0 | 4 | 21.0 |
| Licensed doctorate-level psychologist | 8 | 38.1 | 13 | 61.9 | 12 | 36.4 | 6 | 28.6 | 13 | 56.5 | 6 | 60.0 | 6 | 31.6 |
| Supervised practice | 2 | 9.5 | 0 | 0.0 | 3 | 9.1 | 0 | 0.0 | 0 | 0.0 | 1 | 10.0 | 0 | 0.0 |
| Licensed counselor (master's-level) | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 1 | 5.3 |
| Licensed clinical social worker | 0 | 0.0 | 0 | 0.0 | 4 | 12.1 | 0 | 0.0 | 3 | 13.0 | 0 | 0.0 | 0 | 0.0 |
| Licensed advanced mental health nurse | 0 | 0.0 | 0 | 0.0 | 1 | 3.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 1 | 5.3 |
| Other (e.g., Pharm.D.; diagnostician) | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 2 | 10.5 |
| Male | 14 | 66.7 | 7 | 33.3 | 11 | 33.3 | 14 | 66.7 | 13 | 56.5 | 5 | 50.0 | 9 | 47.4 |
| Race/ethnicity | | | | | | | | | | | | | | |
| White/Caucasian | 16 | 76.2 | 12 | 57.1 | 25 | 75.8 | 17 | 81.0 | 15 | 78.9 | 8 | 80.0 | 15 | 78.9 |
| Black, African descent | 1 | 4.8 | 4 | 19.1 | 5 | 15.1 | 1 | 4.8 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 |
| Other/mixed | 4 | 19.0 | 5 | 23.8 | 3 | 9.1 | 3 | 14.3 | 4 | 21.0 | 2 | 20.0 | 4 | 21.1 |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Years in practice | 10.3 | 11.4 | 9.4 | 8.8 | 10.3 | 10.1 | 13.7 | 10.0 | 11.9 | 12.2 | 8.6 | 6.5 | 11.2 | 10.9 |
| 25th percentile | 3 | | 3 | | 3 | | 5 | | 4 | | 4 | | 3 | |
| 50th percentile | 5 | | 6 | | 7 | | 10 | | 8 | | 6 | | 6 | |
| 75th percentile | 18 | | 12 | | 14 | | 22 | | 12 | | 14 | | 20 | |
| Age (years) | 40.9 | 11.2 | 41.9 | 9.6 | 43.3 | 11.6 | 46.2 | 9.7 | 43.2 | 15.1 | 41.1 | 6.2 | 42.7 | 13.0 |
| 25th percentile | 33 | | 34 | | 34 | | 42 | | 31 | | 36 | | 32 | |
| 50th percentile | 37 | | 38 | | 41 | | 48 | | 36 | | 40 | | 37 | |
| 75th percentile | 46 | | 50 | | 49 | | 52 | | 54 | | 46 | | 52 | |
| Patient interviews completed | 22.4 | 28.9 | 22.6 | 9.2 | 16.4 | 15.2 | 19.0 | 14.8 | 17.5 | 14.1 | 43.4 | 33.1 | 17.9 | 16.5 |
| 25th percentile | 5 | | 18 | | 7 | | 8 | | 4 | | 16 | | 4 | |
| 50th percentile | 11 | | 25 | | 12 | | 16 | | 18 | | 38 | | 16 | |
| 75th percentile | 27 | | 27 | | 20 | | 25 | | 27 | | 70 | | 24 | |
| Time between visit 1 and visit 2 (days) | 5.6 | 3.8 | 4.8 | 3.9 | 6.2 | 3.9 | 2.9 | 3.9 | 7.4 | 3.3 | 8.8 | 3.7 | 5.4 | 4.7 |
| 25th percentile | 2 | | 1 | | 4 | | 0 | | 6 | | 6 | | 1 | |
| 50th percentile | 6 | | 5 | | 6 | | 1 | | 7 | | 8 | | 4 | |
| 75th percentile | 7 | | 7 | | 8 | | 4 | | 9 | | 12 | | 8 | |

[a] Data on some variables missing for four participants.

clinicians' evaluation of the clinical utility and feasibility of the diagnostic changes to the DSM will be presented in an upcoming article.

As can be seen in both adult and pediatric sites (Table 3), very few of the diagnostic strata achieved the fail-safe sample size goal of 50 patients. As noted earlier, a lower sample size may be adequate in some cases, but a lower limit of seven was set for the estimation of reliability for the field trials. At a sample size of six or less, the field trial for a target diagnosis at a site was declared "unsuccessful," in which case intraclass kappa was not estimated and the stratum was folded into the "other diagnosis" group. Of the 60 strata across the 11 field trial sites, 10 were unsuccessful by this definition (four across adult and six across pediatric sites). For example, at the UT-SA site, only six patients completed visits 1 and 2 in the attenuated psychosis syndrome stratum.

The DSM-5 Field Trials aimed to obtain precise estimates of the reliability of the categorical diagnoses and the dimensional measures (i.e., a standard error ≤0.1 as indicated by 95% CI sizes no greater than 0.5 [i.e., used to define a "successful" field trial]) (3). Of the remaining 50 categorical diagnostic strata with stratum sample size greater than six across the 11 field trial sites, 11 were not "successful" (seven across adult and four across pediatric sites). Some of these 11 field trials had high kappa coefficients, but even so, the wide confidence intervals indicated that the true kappas could not be estimated with precision (see Regier et al. [15]). Results of field trials declared "unsuccessful" were excluded in any pooled estimate for a DSM-5 diagnosis.

The field trials for dimensional measures that were completely missing for more than 25% of the sample or had missing data for more than 25% of the items were

**TABLE 7. Clinician Characteristics Across the DSM-5 Field Trial Pediatric Study Sites**

| Characteristic | Baystate (N=12) | | Colorado (N=54) | | Columbia/ Cornell/North Shore (N=32) | | Stanford (N=33) | |
|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % |
| Discipline | | | | | | | | |
| Board-certified psychiatrist | 4 | 33.3 | 13 | 24.1 | 9 | 28.1 | 11 | 33.3 |
| Psychiatrists in training (PGY2-5) | 0 | 0.0 | 7 | 13.0 | 1 | 3.1 | 4 | 12.1 |
| Licensed doctorate-level psychologist | 4 | 33.3 | 13 | 24.1 | 8 | 25.0 | 6 | 18.2 |
| Supervised practice | 0 | 0.0 | 12 | 22.2 | 1 | 3.1 | 12 | 36.4 |
| Licensed counselor (master's-level) | 0 | 0.0 | 0 | 0.0 | 1 | 3.1 | 0 | 0.0 |
| Licensed clinical social worker | 3 | 25.0 | 7 | 13.0 | 12 | 37.5 | 0 | 0.0 |
| Licensed advanced mental health nurse | 1 | 8.3 | 2 | 3.7 | 0 | 0.0 | 0 | 0.0 |
| Other (e.g., Pharm.D.; diagnostician) | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 |
| Male | 6 | 50 | 16 | 29.6 | 6 | 18.8 | 7 | 21.2 |
| Race/ethnicity[a] | | | | | | | | |
| White/Caucasian | 11 | 91.7 | 44 | 81.5 | 24 | 75.0 | 19 | 65.5 |
| Black, African descent | 0 | 0.0 | 2 | 3.7 | 2 | 6.3 | 3 | 10.3 |
| Other/mixed | 1 | 8.3 | 6 | 11.1 | 6 | 18.8 | 7 | 24.1 |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Years in practice | 12.2 | 6.4 | 6.8 | 7.5 | 11.8 | 9.5 | 4.2 | 5.7 |
| 25th percentile | 6 | | 1 | | 5 | | 0 | |
| 50th percentile | 12 | | 5 | | 8 | | 2 | |
| 75th percentile | 18 | | 9 | | 15 | | 7 | |
| Age (years) | 43.8 | 9.1 | 37.3 | 9.0 | 42.3 | 10.6 | 35.5 | 6.7 |
| 25th percentile | 38 | | 31 | | 33 | | 31 | |
| 50th percentile | 44 | | 34 | | 41 | | 34 | |
| 75th percentile | 50 | | 42 | | 48 | | 38 | |
| Patient interviews completed | 26.2 | 24.3 | 7.8 | 8.0 | 7.9 | 7.8 | 9.4 | 9.2 |
| 25th percentile | 10 | | 2 | | 3 | | 3 | |
| 50th percentile | 17 | | 5 | | 5 | | 7 | |
| 75th percentile | 39 | | 10 | | 11 | | 13 | |
| Time between visit 1 and visit 2 (days) | 7.2 | 3.6 | 8.7 | 4.5 | 7.4 | 3.9 | 5.2 | 3.7 |
| 25th percentile | 5 | | 5 | | 5 | | 2 | |
| 50th percentile | 7 | | 9 | | 7 | | 5 | |
| 75th percentile | 10 | | 14 | | 10 | | 7 | |

[a] Data missing for two participants in Colorado and four participants in Stanford.

declared unsuccessful and the reliability estimates were not calculated. As with the categorical diagnoses, a field trial for a particular dimensional measure was unsuccessful in estimating the reliability coefficient with precision if the size of the 95% CI was greater than 0.5, even if the reliability coefficient was high (see Narrow et al. [6]). Results of a field trial for a dimensional measure that was declared unsuccessful were not included in the pooled estimate for that measure.

## Discussion

The DSM-5 Field Trials were crucial for testing the feasibility, clinical utility, test-retest reliability, and (where possible) the validity of DSM-5 diagnoses that were new to DSM, represented major changes from their previous versions, or had minor changes but were of significant clinical and public health importance. These field trials were a multisite study that utilized a rigorous test-retest reliability design with stratified sampling, thereby improving upon previous DSM field trials' sampling methods and generalizability. The DSM-5 Field Trials can be most closely compared with the DSM-III Field Trials in that both attempted to generate representative samples of patients and clinicians. The DSM-5 Field Trials' stratified sampling approach is in contrast to the approximation of simple random sampling used in the DSM-III Field Trials. The sampling used in the DSM-III Field Trials resulted in small sample sizes, below the standards set for the DSM-5 Field Trials, for the low-prevalence diagnoses tested.

Even with the use of the stratified sampling approach, the field trials for some DSM-5 diagnoses were unsuccessful in meeting the standards set for DSM-5 and, as such, trustworthy reliability coefficients could not be obtained. This situation resulted from unrealistic assessments of the patient flow and total staff effort needed to recruit 50 patients per stratum at the field trial sites, particularly for rare disorders. The results of the DSM-5 Field Trials were intended to help to inform the DSM-5 decision-making process (along with many other factors unrelated to field

trials), which would not be available for DSM-5 diagnoses with unsuccessful field trials. However, since reliability information from the field trials was only one of many factors to be used in the DSM-5 decision-making process, field trials were not done for every DSM-5 diagnosis, and the few that were not "successful" were added to that list.

The DSM-5 Field Trials were conducted across a variety of clinical settings and hence captured a heterogeneous overall patient population. However, since the sites were primarily large academic clinical settings with research infrastructure that enhanced the feasibility of the implementation of the complex study protocol, the results might not be generalizable to patients seen in solo or small group practices or other community-based settings. Patients who present to academic settings may be different from those in other settings in their symptom presentations. For instance, if patients present to academic/large clinical settings when they have more severe symptoms and to solo or small group practices when they have less severe or subthreshold symptom presentations, reliability of the categorical diagnoses and dimensional measures might be different.

The intended primary purpose of DSM-5 is to support clinical use. Thus, the assessment of DSM-5 diagnoses in adult and pediatric sites in the United States and Canada, with the participation of mental health professionals of varied disciplines, enhances the generalizability of findings. Clinicians of varied clinical disciplines, years in practice, race/ethnicity, and other characteristics completed the diagnostic interviews used in the estimation of the reliability coefficient for the various diagnoses. This is a major strength of the field trials in that the reliabilities of the DSM-5 diagnoses were assessed by the clinicians who would use the manual in clinical care. A weakness, however, is that the clinician population in academic/large clinical settings may differ from those in solo or small group practices or nonacademic settings. Clinicians in solo or small group practice might have less time or resources to complete the diagnostic interview in a fashion similar to that of the study clinicians. We attempted to mitigate this difference by integrating the study's diagnostic interviews into busy clinical settings and practitioner schedules and enrolling clinicians whose time was not spent solely in research endeavors.

In assessing the reliability estimates for the categorical diagnoses (i.e., intraclass kappas) in comparison to those obtained from the DSM-IV Field Trials, one needs to keep in mind the different methods used in the two field trials. The DSM-IV Field Trials enrolled carefully selected patients likely to have the target disorder, excluded patients with high levels of comorbidity and other confusing presentations, and used diagnosticians highly trained on a specific diagnostic instrument. All of these factors will tend to produce higher kappa estimates compared to the more naturalistic field trial methods employed in the DSM-5 Field Trials, for which patient exclusion criteria were minimal and diagnostic instruments requiring training were not used (8).

Further publications (6, 15) detail the outcomes of the DSM-5 Field Trial methodology as applied to specific diagnoses and dimensional assessments. The methodological approaches described herein demonstrate efforts to use an empirically sound approach to assessing diagnostic quality. Since DSM-5 is intended to be a living document, these field trials also were important in providing a stepping stone to conduct future field trials in routine clinical settings.

Presented in part at the 165th annual meeting of the American Psychiatric Association, Philadelphia, May 5–9, 2012. Received Jan. 30, 2013; revision received Aug. 31, 2012; accepted Sept. 4, 2012 (doi: 10.1176/appi.ajp.2012.12070998). From the American Psychiatric Association, Division of Research and American Psychiatric Institute for Research and Education, Arlington, Va.; the Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Md.; the Stanford University School of Medicine, Palo Alto, Calif.; and the University of Pittsburgh Medical Center, Pittsburgh, Pa. Address correspondence to Dr. Clarke (dclarke@psych.org).

All authors report no financial relationships with commercial interests.

This study was funded by the American Psychiatric Association.

The authors acknowledge the efforts of Paul Harris, Ph.D., and his research team at Vanderbilt University, including Brenda Minor, Jon Scherdin, and Rob Taylor, for assistance and support provided during the development of the DSM-5 Field Trial REDCap System and throughout the DSM-5 Field Trials. The authors would also like to acknowledge the efforts of the temporary research staff (Alison Newcomer, June Kim, and Mellisha McKitty) and research interns in the APA Division of Research and graduate students in the Department of Mental Health at Johns Hopkins Bloomberg School of Public Health (Flora J. Or and Grace P. Lee) for their research support. Last, the authors would like to acknowledge the research coordinators across the DSM-5 Field Trial sites, without whose concerted efforts in learning, implementing, and adhering to the procedures involved in this complex multisite study the field trials would not have been possible.

Research coordinators at the adult field trial sites: Natalie St. Cyr, M.A., Nora Nazarian, and Colin Shinn (The Semel Institute for Neuroscience and Human Behavior, Geffen School of Medicine, University of California Los Angeles, Los Angeles, Calif.); Gloria I. Leo, M.A., Sarah A. McGee Ng, Eleanor J. Liu, Ph.D., Bahar Haji-Khamneh, M.A., Anissa D. Bachan, and Olga Likhodi, M.Sc. (Centre for Addiction and Mental Health, Toronto, Ont., Canada); Jeannie B. Whitman, Ph.D., Sharjeel Farooqui, M.D., Dana Downs, M.S., M.S.W., Julia Smith, Psy.D., Robert Devereaux, Elizabeth Anderson, Carissa Barney, Kun-Ying H. Sung, Solaleh Azimipour, Sunday Adewuyi, and Kristie Cavazos (Dallas VA Medical Center, Dallas, Tex.); Melissa Hernandez, Fermin Alejandro Carrizales, Patrick M. Smith, Nicole B. Watson, M.A., and Martha Dahl (University of Texas San Antonio School of Medicine, San Antonio, Tex.); Kathleen Grout, M.A., Sarah Neely, Lea Kiefer, M.P.H., Jana Tran, M.A., Steve Herrera, and Allison Kalpakci (Michael E. DeBakey VA Medical Center and the Menninger Clinic, Houston, Tex.); Lisa Seymour, Sherrie Hanna, Cynthia Stoppel, Kelly Harper, Scott Feeder, and Katie Mingo (Integrated Mood Clinic & Unit and the Behavioral Medicine Program at Mayo Clinic, Rochester, Minn.); Jordan Coello and Eric Wang (University of Pennsylvania School of Medicine, Philadelphia, Pa.).

Research coordinators at the pediatric field trial sites: Kate Arnow, Stephanie Manasse, and Nandini Datta (Stanford University Child & Adolescent Psychiatry Clinic and the Behavioral Medicine Clinic, Palo Alto, Calif.); Laurie Burnside, M.S.M., C.C.R.C., Darci Anderson, Heather Kennedy, M.P.H., Elizabeth Wallace, Vanessa Waruinge, and Amanda Millar (The Children's Hospital, Aurora, Colo.); Julie Kingsbury, C.C.R.P., and Brenda Martin (Child Behavioral Health, Baystate

Medical Center, Springfield, Mass.); Zvi Shapiro, Julia Carmody, Alex Eve Keller, Sarah Pearlstein Levy, Stephanie Hundt, and Tess Dougherty (New York State Psychiatric Institute at Columbia University, New York, N.Y.; Weill Cornell Department of Psychiatry at Payne Whitney Manhattan Division, New York, N.Y.; North Shore Child and Family Guidance Center, Roslyn Heights, N.Y.; and Weill Cornell Department of Psychiatry at Payne Whitney Westchester Division, White Plains, N.Y.).

## References

1. Goldberg D, Kendler KS, Sirovatka PJ, Regier DA: Diagnostic Issues in Depression and Generalized Anxiety Disorder: Refining the Research Agenda for DSM-5. Arlington, VA, American Psychiatric Association, 2010
2. Ghaemi SN, Sachs GS, Chiou AM, Pandurangi AK, Goodwin K: Is bipolar disorder still underdiagnosed? Are antidepressants overutilized? J Affect Disord 1999; 52:135–144
3. Pope HG Jr, Lipinski JF, Cohen BM, Axelrod DT: "Schizoaffective disorder": an invalid diagnosis? A comparison of schizoaffective disorder, schizophrenia, and affective disorder. Am J Psychiatry 1980; 137:921–927
4. Grant BF, Stinson FS, Hasin DS, Dawson DA, Chou SP, Ruan WJ, Huang B: Prevalence, correlates, and comorbidity of bipolar I disorder and axis I and II disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. J Clin Psychiatry 2005; 66:1205–1215
5. Kendell R, Jablensky A: Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 2003; 160:4–12
6. Narrow WE, Clarke DE, Kuramoto SJ, Kraemer HC, Kupfer DJ, Greiner L, Regier DA: DSM-5 Field Trials in the United States and Canada, part III: development and reliability testing of a cross-cutting symptom assessment for DSM-5. Am J Psychiatry 2013; 170:71–82
7. Kraemer HC, Kupfer DJ, Narrow WE, Clarke DE, Regier DA: Moving toward DSM-5: the field trials. Am J Psychiatry 2010; 167:1158–1160
8. Kraemer HC, Kupfer DJ, Clarke DE, Narrow WE, Regier DA: DSM-5: how reliable is reliable enough? (commentary). Am J Psychiatry 2012; 169:13–15
9. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420–428
10. Streiner DL, Norman GR: Measurement Scales: A Practical Guide to Their Development and Use, 2nd ed. Oxford, Oxford University Press, 1995, pp 104–127
11. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG: Research Electronic Data Capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009; 42:377–381
12. Chmura Kraemer H, Periyakoil VS, Noda A: Kappa coefficients in medical research. Stat Med 2002; 21:2109–2129
13. Efron B, Tibshirani R: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1986; 1:54–75
14. Kraemer HC: Measurement of reliability for categorical data in medical research. Stat Methods Med Res 1992; 1:183–199
15. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ: DSM-5 Field Trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry 2013; 170:59–70
16. McGraw KO, Wong SP: Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996; 1:30-46 (correction 1996; 1:390)
17. Weir JP: Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005; 19:231–240
18. Canivez GL, Watkins MW: Long-term stability of the Wechsler Intelligence Scale for Children-III. Psychol Assess 1988; 10:285–291
19. Brown SJ, Rourke BP, Cicchetti DV: Reliability of tests and measures used in the neuropsychological assessment of children. Clin Neuropsychol 1989; 3:353–368
20. Egan JP: Signal Detection Theory and ROC Analysis. New York, Academic Press, 1975
21. Swets JA, Dawes RM, Monahan J: Better decisions through science. Sci Am 2000; 283:82–87