ABSTRACT - Aims: Given resource constraints in conducting clinical trials, it is critical that rater training focuses on scale items wherein standardization is most challenging. This analysis examined mood disorder symptom ratings submitted in an online rater training program conducted preparatory to the initiation of a multi-site, international mood disorder treatment trial. Ratings were entered online and analyzed for consistency and variability, and compared to established standards (Gold Consensus Ratings/GCRs).

Methods: Raters participated in web-based rater training on the Hamilton Depression Rating Scale (HAM-D), Montgomery Asberg Rating Scale (MADRS), and Young Mania Rating Scale (YMRS). Training included integration of didactic materials and videos of two bipolar depressed patients interviewed by two U.S. clinicians. Raters viewed the videos and rated the mood scales. Inter-rater agreement was assessed using Kappa statistics. Ratings between the raters and the GCRs for individual scale items were assessed using McNemar test for paired binomial proportions.

Results: 194 raters from 16 countries, 80 sites and speaking 20 different languages participated. Interrater agreement on videos ratings ranged from substantial to moderate (HAM-D, Kappa video A = 0.72, video B = 0.65, p < 0.001), (MADRS, Kappa = 0.65 and 0.47, p < 0.001), (YMRS, Kappa = 0.75, and 0.64, p < 0.001). There was no significant difference on agreement based upon on English proficiency, clinical experience, or by country. Scale items that differed from the GCR on the HAM-D were depressed mood, delayed insomnia, retardation, and anxiety (psychic). Items that differed on the MADRS

Key Words: rating scales, clinical trials, depression, bipolar disorder
were apparent sadness, inner tension, concentration difficulties, lassitude and inability to feel. Items that differed on the YMRS were irritability and disruptive behavior. **Conclusions:** Identification of specific rating scale items in which rater variability is greatest may facilitate training approaches that target these areas for more efficient training in international clinical trials.


**INTRODUCTION**

Pharmacotherapy clinical trials for central nervous system (CNS) disorders generally rely on ratings conducted by study investigators to document a drug effect. In depression treatment trials where depression severity outcome measures are a subjective assessment using a rating scale, rater training and the reliability of the outcome measures is particularly important. While it is known that repeated participation in rater training programs improves competency of administration of mental health rating scales, optimal training methodology has not been clearly established with identified concerns regarding variability across raters with respect to experience administering rating scales, self-reported proficiency in the language of the training, and cultural or geographic background, and factors related to how rater training should be implemented and assessed.

Clinical trial research has also become increasingly globalized in recent years, with the number of countries providing trial sites outside the U.S. more than doubling in the period between 1995–2005, while the proportion of clinical trials conducted in the U.S. and Western Europe has decreased. Cost savings for conducting clinical trials in developing countries such as India and in South America can be substantial, and it is likely that this trajectory of globalization will continue.

Given trends for globalization and multiple clinical site involvement, as well as cost and resource restraints in conducting clinical trials, it is critical that rater training focuses on domains or rating-scale items wherein standardization of applied outcome measures is most challenging. The aim of this analysis was to examine variability of mood disorder symptom ratings submitted as part of an online rater training program conducted preparatory to the initiation of a multi-site, international mood disorder treatment trial. Specifically, we evaluated inter-rater agreement on the Hamilton Depression Rating Scale (HAM-D), grid version, Montgomery Asberg Depression Rating Scale, and Young Mania Rating Scale (YMRS) across raters from different countries and in relation to established standards. Identification of specific rating scale items in which rater variability is greatest may facilitate training approaches that target these areas for more efficient training in international clinical trials.
METHODS

Raters in this multi-site international clinical trial participated in a secure web-based rater training program intended to increase knowledge and competence with administration of the 17-item HAM-D, and the MADRS and YMRS prior to the initiation of the clinical trial (Figure 1). The website included learning modules in English on HAM-D, MADRS and YMRS with an integration of didactic learning materials and six videos of bipolar depressed patients being interviewed by two U.S. clinicians on these scales. Raters viewed and rated the videos on HAM-D, MADRS and YMRS. The ratings were entered online and analyzed for consistency and variability. This analysis focuses on ratings for the first (training) video (Video A) and a second (qualification) video (Video B) which was conducted immediately subsequent to the training video. There were additional training procedures that were part of the study start-up which were implemented for raters who failed to achieve acceptable scoring for the qualification video; however, these additional training/remediation processes are not the focus of this report.

Raters

Information on rater background (country, language, self-reported English proficiency), clinical experience with bipolar disorder, and

<table>
<thead>
<tr>
<th>Format for Web-Based Training on 3 Mood Symptom Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Specific Procedures</strong></td>
</tr>
<tr>
<td><strong>Didactic presentations</strong>: Overview of the HAM-D, MADRS, and YMRS via PowerPoint presentations accompanied by a voice-over in English.</td>
</tr>
<tr>
<td><strong>Training video viewing and scoring</strong>: Raters listened without discussion to a real patient vignette (Video A, severe bipolar depression with mild mania). The interview was conducted in English by a U.S.-based interviewer with local language sub-titles. Raters then scored the HAM-D, MADRS, and YMRS scales based upon the vignette.</td>
</tr>
<tr>
<td><strong>Interactive feedback</strong>: Feedback was provided to the rater on the basis of their ratings compared with gold consensus ratings (GCRs). Rater provides his/her scores for the video, and is then given intranet access to the GCRs and item scoring rationale provided by three expert raters. Feedback on individual rater scores were provided by the training site and communication with expert raters was made available.</td>
</tr>
<tr>
<td><strong>Qualification video viewing and scoring</strong>: Raters listened without discussion to a real patient vignette (Video B, severe bipolar depression with mild mania). The interview was conducted in English by a U.S.-based interviewer with local language sub-titles. Raters scored the HAM-D, MADRS, and YMRS scales based upon the vignette. Rater provides his/her scores for the video, and is then given intranet access to the GCRs and item scoring rationale provided by three expert raters. Feedback on individual rater scores were provided by the training site and communication with expert raters was made available.</td>
</tr>
</tbody>
</table>
experience with the measures, was recorded and assessed. One-hundred ninety-four raters from 16 countries across 80 sites, speaking 20 different languages, participated in the training. Rater overall clinical experience in bipolar disorder ranged from 0 to 40 years, with a mean of 4.48 years (SD = 1.22); experience with clinical trials ranged from 0 to 26 years, with a mean of 4.37 years (SD = 1.35); and experience with clinical trials in bipolar disorder ranged from 0 to 26 years, with a mean of 4.36 years (SD = 1.35).

Training Format

Figure 1 identifies the specific procedures in the training. An introductory didactic presentation for each scale focused on an overview of scale background and scoring conventions, specific scale items, and potentially problematic issues in the administration of that scale. This was followed by a total of six videos (two for each scale) that demonstrated two real patients with bipolar depression being interviewed on the HAM-D, MADRS, and YMRS. Raters were instructed to view all presentations and videos, and provide their ratings online. All video interviews were conducted in English by U.S. interviewers with local language sub-title translation. After submitting their ratings on-line, raters were given feedback in English via interactive email correspondence on their ratings in relation to those ratings made by experts.

Expert Ratings

Rater scoring on the videos was compared with standard ratings (‘gold consensus ratings/GCRs) achieved by three expert U.S. raters. The senior U.S. expert has over 15 years of experience in rating scales training and was the interviewer featured in the videos. GCRs were derived via a consensus-process involving accepting all items with the same scores across expert raters and attempts to obtain a consensus score for items with discrepancies across experts. If no consensus could be reached on a single score, there was a consensus determination of an acceptable range of scores.

Data Analysis

Trainee demographic characteristics were assessed using Pearson chi-squared analyses. Inter-rater agreement between raters and the GCR was assessed using Kappa statistics (Table 1). Ratings between the raters and the GCRs for the individual scale items were assessed using McNemar test for paired binomial proportions.16 As noted in Table 2,
significant results meant that the majority of raters scored that item more than 1 point above or below that of the GCR.

**RESULTS**

**Variability Across Sites and Experience**

Raters with differing levels of self-reported English language proficiency, experience with bipolar disorder patients, and from diverse countries were not significantly different from each other with respect to inter-rater agreement. No significant difference for raters was found on previous experience with the HAM-D, MADRS, and YMRS.

**Agreement Across Raters**

Raters rated patient A with more consistency when compared to patient B. Inter-rater agreement on ratings for the videos ranged from substantial to moderate (HAM-D, Kappa video A = 0.72 and video B = 0.65, \( p < 0.001 \) for both patients), (MADRS, Kappa = 0.65, and 0.47, \( p < 0.001 \) for both patients), (YMRS, Kappa = 0.75, and 0.64 \( p < 0.001 \) for both patients). These Kappa values were derived by using the modal rating among all raters to represent an item score, for each item in a respective measure. These mode values were then compared with the corresponding GCR item values, and agreement was assessed among the items. The reported p-values indicate that the null hypothesis (that the agreement is random) can be rejected.

**Variability by Item Within Each Scale**

The McNemar results for HAM-D items in the training and qualification videos are shown on Table 2. Rating scale items that differed for both videos on the HAM-D were depressed mood, delayed insomnia,
retardation, and anxiety (psychic). Rating scale items that differed for both videos on the MADRS were apparent sadness, inner tension, concentration difficulties, lassitude and inability to feel. Rating scale items that differed for both videos on the YMRS were irritability and disruptive behavior.

**DISCUSSION**

The analysis presents rater scoring results from a mood disorder rating scale training program conducted preparatory to a multi-center mood disorder treatment trial. Secure Internet access is a widely available and cost-effective approach for engaging and training clinical researchers in preparation for conducting large clinical trials in which investigators are...
spread over a large geographic area and have varying levels of clinical and research experience.\textsuperscript{17,18} Formal training of clinical investigators can reduce rating variability, and may increase the likelihood of detecting a signal/treatment effect.\textsuperscript{5,19,20} Similar to findings by other investigators, we did not find there was significant variability in rater agreement based upon clinical experience\textsuperscript{2} or rater country of origin.\textsuperscript{21}

While overall moderate to substantial agreement was achieved for raters across countries in rating mood disorder rating scales, rater agreement with expert ratings was less on the MADRS compared to HAM-D and YMRS. Specific illness symptoms that appeared difficult to rate (less rater agreement) on the HAM-D were depressed mood, anxiety and the two vegetative symptoms of sleep and motor retardation. Lo and colleagues (www.prophase.com)\textsuperscript{22} examined the results from 54 U.S. and 106 Japanese raters in an international depression treatment trial. In the report by Lo and colleagues, raters watched and rated two videos of HAM-D interviews conducted in English with Japanese subtitles. The first video depicted a severely depressed patient and the second video depicted a moderately depressed patient. There were no differences between American and Japanese raters on the more depressed patient, but for the less depressed patient video Japanese clinicians rated the patient as more severely ill on the items for Insomnia (late), Psychic Anxiety and Insight items. The authors concluded that assessment of severe depression does not seem to be impacted by cultural differences, but that rating mild-to-moderate levels of depression may pose more of a challenge.

In the analysis reported here, late insomnia and psychic anxiety were also included in the HAM-D items that seemed particularly challenging to rate. Additionally, in our analysis, ratings on the first patient video (video A) showed greater agreement across sites than ratings on the second video (video B). Illness severity in the patient videos did not appear to be related to level of agreement with GCR ratings in this analysis. Possibly rater agreement may have been greater in video A because video A presented a patient who was relatively “easy” to interview, responded readily to all questions, and did not digress in response to interviewer queries. In contrast, video B demonstrated a clinical presentation that was more disorganized and less cooperative.

Specific illness symptoms that appeared difficult to rate (less rater agreement) on the MADRS were the observation of sadness, inner tension, concentration problems, lassitude and a perceived inability to feel emotions. Williams and Kobak\textsuperscript{23} have developed a semi-structured version of the MADRS that may improve the quality of clinical trial ratings. In our analysis, specific illness symptoms that appeared difficult to rate on the YMRS were irritability and disruptive behavior. Mackin and
colleagues \(^{24}\) assessed YMRS ratings administered as part of a rater qualification program for global clinical trials. In the study by Mackin \(^{24}\) raters included 126 trained English-speaking clinician raters from three different countries. Two video interviews, each of an American person with moderate/severe mania, were shown to psychiatrists from three countries (U.S., U.K., and India). Total YMRS scores differed significantly between the U.S. and U.K. groups \((p < 0.001)\) and between India and UK \((p < 0.001)\) groups. Differences between India and U.S. rater groups were less marked \((p = 0.28)\). Overlapping somewhat with our own analysis, the report by Mackin \(^{24}\) found that greatest differences were found for YMRS items on mood elevation \((p < 0.001)\), irritability \((p < 0.001)\), thought content \((p < 0.001)\), and disruptive behavior \((p < 0.001)\). In the Mackin \(^{24}\) report, Indian raters identified American patients with mania as significantly more ill and inappropriate than did American raters. In contrast, British raters generally rated these same American patients with mania significantly lower compared with the American raters. \(^{24}\) While the analysis presented here found that international raters differed from U.S. expert raters on the YMRS only on the two items of irritability and disruptive behavior, it is possible that this could have been related to the relatively low levels of mania in both videos.

Limitations of this study include the fact that most raters were experienced clinicians, that GCR ratings were established from a relatively small (albeit highly experienced) group of expert raters from a single country, and that raters scored only two videos for each rating scale. Additionally, it has been suggested that the use of videotapes may artificially inflate estimates of reliability by reducing the information variance that would result if each rater interviewed the patient independently. \(^{17,25}\) Also, it is possible that assessment of English proficiency based upon self-report may not be accurate, and some of the difficulty with rating specific items could have been due to raters not understanding all elements of the interview. It is expected that the local language sub-titles provided for this training would have maximized comprehension, but it is possible that this was not completely successful.

On the basis of our findings, we advocate that rater training programs for use in international clinical trials should employ a selection of patients representing a spectrum of cooperativeness, illness severity and symptoms. Vignettes/videos that demonstrate varying symptoms may aid in the identification of potential weakness in rater competency or in cultural differences that should be addressed. Additionally vignettes should include, and emphasize, items identified as most challenging for a given mood disorder scale. A focus on sleep problems and psycholog-
ical manifestations of anxiety may be particularly helpful in training on the HAM-D, and a focus on tension, concentration, anhedonia and lat-
titude may be particularly helpful for the MADRS. A focus on irri-
tability and on disruptive behaviour may be particularly helpful for
YMRS training. Use of semi-structured/standardized interview formats
may improve inter-rater agreement as well.

It should also be noted that the rater training program described here
only addressed scale scoring conventions. An equally important aspect
of rater training is an emphasis on clinical interviewing skills required
for administration of mood disorders scales. Engelhardt and col-
leagues26 reported an analysis of rater performance in administration of
the HAM-D from two large, multi-site clinical trials, and noted that
HAM-D interviews were brief and cursory in 39% of interviews, and
rated as fair or unsatisfactory on a standardized measure of rater per-
formance (the Rater Applied Performance Scale/RAPS).27 It is possi-
ble that a greater understanding of which items may be particularly
difficult to rate on mood disorder rating scales could also be used to
train raters in interview skills for a specific outcome measure. Focused
attention to challenging items may prepare raters to assess these items
in ways that can better elicit information from patients and yield a more
accurate rating.

ACKNOWLEDGEMENTS

Portions of this manuscript have been presented at the New Clinical
Drug Evaluation Unit (NCDEU) meeting, Boca Raton, Florida on

FUNDING SUPPORT

Support for this analysis was provided by a grant to the first author
from The Cognition Group.

REFERENCES

1. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater Reliability in
4. Kobak KA, Lipsitz JD, Williams JB, Engelhardt N, Bellew KM. A new approach to rater training and
5. Kobak KA, Lipsitz JD, Feiger AD. Development of a standardized training program for the Hamilton


