Rater Training on HAM-D, MADRS and YMRS – What were the difficult items to rate?

Richa Gaur[1], PhD; Martha Sajatovic[2], MD; Nathan Lee[1], MSc; Geetika Nath[1], MA; Luis Ramirez[3], MD; Hossein Kaviani[1], PhD

[1]The Cognition Group, Newark, Delaware, USA; [2]Case Western University, Cleveland, Ohio, USA; [3]Quality Outcomes Training, Cleveland, Ohio, USA.

Abstract

Introduction

Training of raters on pre-set rating conventions is critical to increase the reliability of ratings in CNS drug trials. Identification of items anticipated to be most challenging for raters to assess in order to optimize rater training programs. This study examined raters' ratings provided in an online rater training program on the Hamilton Depression Rating Scale (HAM-D), Montgomery Asberg Rating Scale (MADRS), and Young Mania Rating Scale (YMRS) for a Bipolar Disorder trial.

Methods

194 raters from 16 countries, 80 sites, speaking 20 different languages, with clinical experience in bipolar disorder ranging from 0 to 40 years participated in an online (via website) rater training and certification program. The website included learning modules on HAM-D, MADRS & YMRS with an integration of visual learning materials and 9 videos of three bipolar patients interviewed by two American clinicians on these three scales. Raters viewed and rated the videos on HAM-D, MADRS and YMRS, ratings were entered online, and analyzed for consistency and variability compared with gold consensus ratings (GCRs) achieved by three experts raters. Inter-rater agreement was assessed using Kappa statistics. Ratings between the raters and the GCRs for the individual scale items were assessed using McNemar test for binomial proportions.

Results

No significant difference for raters was found among countries, raters' past experience with bipolar disorder, or previous training on HAM-D, MADRS and YMRS. Inter-rater agreement for the three videos on the scales ranged from substantial to moderate (HAM-D, Kappa = .72, 65 & .43, p<.001), (MADRS, Kappa =.65, .47 & .44, p<.001), (YMRS, Kappa =.75, .64 & .54, p<.001). The McNemar results showed that HAM-D = 8/17, 4/17 &7/17, MADRS = 9/10, 9/10 & 5/10) & YMRS =3/11, 5/11 and 5/11 individual items were significantly different than GCRs. On the HAM-D, anxiety and retardation were most difficult, while most difficult items on the MADRS were apparent sadness, inner tension, concentration, lassitude, and inability to feel. Most difficult items for the YMRS were irritability, language/thought disorder, content, disruptive/aggressive behavior and appearance.

Conclusions

While overall moderate to substantial agreement was achieved for raters across countries in rating mood disorder rating scales, the MADRS appeared to be the more difficult scale to rate compared to HAM-D and YMRS. Rater training in mood disorder clinical trials should include focused training on the assessment of sleep/insomnia as well as specific modules unique to the HAM-D, MADRS and YMRS.

Introduction

Efficacy for CNS clinical trials in the Mood Disorders is typically determined by ratings provided by the raters on rating scales. These ratings are based entirely on an individual rater's interpretation, which may increase or decrease the variability in ratings in trials with multiple raters across a number of different clinical sites.

Training is provided for any clinical trial with an objective to maintain rating standardization by limiting potential for bias and errors, while improving rater's observational skills.

The objective of this poster is to identify items on MADRS, HAM-D and YMRS which raters found difficult to rate consistently when undergoing a comprehensive online rater training and certification program. .

Methods

The Raters

194 Raters from 16 countries, 80 sites, speaking 20 different languages, participated in the web-based training program. Clinical experience in bipolar disorder of raters ranged from 0 to 40 years.

Rater Training Program

A fully online training and certification program was provided for HAM-D, MADRS and YMRS.

Training included didactic presentations and real-patient videos;

1. Presentations mainly focused on scale overview, background and conventions, items and anchor points and potentially problematic issues.
2. Videos included 9 patient interviews, demonstrating the use of HAM-D (3), MADRS (3) and YMRS (3).

Patient Videos

Three different bipolar disorder patients (A, B & C) were used for the online training and certification sessions.

Each patient was interviewed on HAM-D, MADRS and YMRS using standardized interview schedules.

All 3 patients had current (or most recent) Major Depressive Episode, with a history of bipolar disorder:

Patient A

Late 20's, male, unemployed, single, homosexual and had diagnosed bipolar depression and anxiety disorder. Patient scored severe depression as per HAM-D & MADRS ratings, with mild mania on YMRS.

Overall interview observation; patient was attentive throughout the interview session, answered questions appropriately, and most of the time his answers were supported with examples from previous seven days.

Patient B

Mid 50's male, government employed, divorced/heterosexual, and has been hospitalized in the past for different stages of bipolar disorder. Patient scored severe depression as per HAM-D & MADRS ratings, with mild mania on YMRS.

Overall interview observation; patient was attentive but most of the time interviewer had to probe for clear responses.

Patient C

Mid 50's male, married/heterosexual, self employed and was receiving treatment for bipolar disorder for most of his life. Patient scored moderate depression as per HAM-D & MADRS ratings with mild mania on YMRS.

Overall interview observation; patient was distracted most of the time, provided irrelevant examples to support his responses and kept digressing from interview questions.

Raters were instructed to view all presentations and videos, and provide their ratings online.

The ratings submitted by the raters were compared to the Gold Consensus Ratings (GCRs) assigned after consensus among three US-based independent expert raters.

Table 1: Organization of Online Training and Certification Sessions

Activities and objectives	Approximate time
Introduction to HAM-D, MADRS & YMRS via a flash (Power Point) presentation accompanied by a voice-over in English.	**45 minutes**
HAM-D, MADRS & YMRS videos ("training session" **Patient A**,) which entailed listening without discussion. The participant noted the scores on a paper version of the HAM-D, MADRS & YMRS scales.	HAM-D = 19 minutes MADRS = 16 minutes YMRS = 16 minutes **Total = 51 minutes**
HAM-D, MADRS & YMRS ratings from the participant's score sheets were entered into a database and and interactive feedback was provided to the rater on the basis of their ratings compared with gold consensus ratings.	25 minutes
HAM-D, MADRS & YMRS videos ("certification session" **Patient B**): required viewing without discussion and feedback.	HAM-D = 17 minutes MADRS = 21 minutes YMRS = 14 minutes **Total = 52 minutes**
HAM-D, MADRS & YMRS videos ("certification session" **Patient C**) required viewing without discussion and feedback.	HAM-D = 17 minutes MADRS = 21 minutes YMRS = 20 minutes **Total = 58 minutes**
Intranet posting of their HAM-D, MADRS & YMRS ratings using the original scales containing explanatory anchor points.	**30 minutes**
Total	**261 minutes**

Results

Inter-rater reliability was calculated using Kappa statistics

Table 2: Inter Rater Reliability

Scale	Training Session Patient A	Certification Session Patient B	Re-Certification Session Patient C
HAM-D	k=.72 substantial	k=.65 substantial	k=.43 moderate
MADRS	k=.65 substantial	k=.47 moderate	k=.44 moderate
YMRS	k=.75 substantial	k=.64 substantial	k=.54 moderate

McNemar test for paired binomial proportions

The McNemar test for paired binomial proportions was used to identify items which were rated significantly different from GCRs. Significant results meant that the majority of raters deviated from that item by more than +/- 1 of the GCR.

Table 3: McNemar Results

Items highlighted were found to be significantly different than the GCRs in all three sessions

HAM-D

Training Session Patient A		Certification Session Patient B		Re-certification Session Patient C	
Item No.	*Item Label*	*Item No.*	*Item Label*	*Item No.*	*Item Label*
01	Depressed mood	01	Depressed mood	01	Depressed mood
02	Guilt feelings	04	Initial insomnia	03	Suicide
06	Delayed insomnia	06	Delayed insomnia	06	Delayed insomnia
08	Retardation	08	Retardation	08	Retardation
10	Anxiety (psychological)	10	Anxiety (psychological)	09	Agitation
12	Loss of appetite			10	Anxiety (psychological)
14	Loss of libido			15	Hypochondriasis
15	Hypochondriasis				

MADRS

Training Session Patient A		Certification Session Patient B		Re-certification Session Patient C	
Item No.	*Item Label*	*Item No.*	*Item Label*	*Item No.*	*Item Label*
01	Apparent sadness	01	Apparent sadness	01	Apparent sadness
02	Reported sadness	02	Reported sadness	03	Inner tension
03	Inner tension	03	Inner tension	06	Concentration difficulties
04	Reduced sleep	04	Reduced sleep	07	Lassitude
05	Reduced appetite	05	Reduced appetite	08	Inability to feel
06	Concentration difficulties	06	Concentration difficulties		
07	Lassitude	07	Lassitude		
08	Inability to feel	08	Inability to feel		
09	Pessimistic thoughts	09	Pessimistic thoughts		
		10	Suicidal thoughts		

YMRS

Training Session Patient A		Certification Session Patient B		Re-certification Session Patient C	
Item No.	*Item Label*	*Item No.*	*Item Label*	*Item No.*	*Item Label*
04	Sleep	01	Elevated mood	05	Irritability
05	Irritability	02	Increased motor activity/ energy	07	Language/thought disorder
09	Disruptive/aggressive behavior	05	Irritability	08	Content
		09	Disruptive/aggressive behavior	09	Disruptive/aggressive behavior
		11	Insight	10	Appearance

Raters' background, experience and performance on HAM-D, MADRS and YMRS

Rater with different level of English language proficiency (p=ns), experience with bipolar disorder patients (p=ns) and regions (p=ns) were not significantly different from each other with respect to the kappa values.

Conclusions

Kappa values varied for all three sessions, from moderate to substantial.

Raters rated patient A with most consistency when compared to patient B and C. Possibly because patient A was easy to interview and responded appropriately to all questions. Ratings on patient C received moderate Kappa values, as this patient was difficult to interview and rate. Patient kept digressing from the interview questions.

Some items on HAM-D, MADRS and YMRS were found to be repeatedly challenging on all patient vignettes. Raters found MADRS most difficult to rate compared to HAM-D and YMRS.

Further research is needed to confirm if items identified as difficult were vignette-specific or could be generalized to other rater training programs.

Recommendations

On the basis of the results we advocate that rater training raining programs should include different patient vignettes based on the items identified as most challenging on each scale.

Using a set of vignettes that include both cooperative and somewhat problematic patients may aid in the identification of potential weaknesses in rater competency.

It should also be noted that the rater training program described only addresses scoring conventions. An equally important aspect of rater training is an emphasis on examining clinical interviewing skills required for administration of mood disorders scales.

References

1. Montgomery, S.A., Asberg, M. (1979), A New Depression Scale Designed to be Sensitive to Change. Br J Psychiatry; 134:382-9.
2. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater reliability in clinical trials of depressive disorders. Am J Psychiatry 2002, 159: 1598-1600.
3. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978;133:429-35.

TCG Advancing CNS Research