Rating the Raters

Assessing the Quality of Hamilton Rating Scale for Depression Clinical Interviews in Two Industry-sponsored Clinical Drug Trials

Nina Engelhardt, PhD, Alan D. Feiger, MD,† Kenneth O. Cogger, PhD,‡ Dawn Sikich, BA,†*
*David J. DeBrota, MD,§ Joshua D. Lipsitz, PhD,|| Kenneth A. Kobak, PhD,**
Kenneth R. Evans, PhD,¶ and William Z. Potter, MD, PhD#

Objective: The quality of clinical interviews conducted in industry-sponsored clinical drug trials is an important but frequently overlooked variable that may influence the outcome of a study. We evaluated the quality of Hamilton Rating Scale for Depression (HAM-D) clinical interviews performed at baseline in 2 similar multicenter, randomized, placebo-controlled depression trials sponsored by 2 pharmaceutical companies.

Methods: A total of 104 audiotaped HAM-D clinical interviews were evaluated by a blinded expert reviewer for interview quality using the Rater Applied Performance Scale (RAPS). The RAPS assesses adherence to a structured interview guide, clarification of and follow-up to patient responses, neutrality, rapport, and adequacy of information obtained.

Results: HAM-D interviews were brief and cursory and the quality of interviews was below what would be expected in a clinical drug trial. Thirty-nine percent of the interviews were conducted in 10 minutes or less, and most interviews were rated fair or unsatisfactory on most RAPS dimensions.

Conclusions: Results from our small sample illustrate that the clinical interview skills of raters who administered the HAM-D were below what many would consider acceptable. Evaluation and training of clinical interview skills should be considered as part of a rater training program.

(*J Clin Psychopharmacol* 2006;26:71–74)

Results of clinical drug trials sponsored by the pharmaceutical industry have a direct and often rapid impact on public health. If the severity of disease in clinical trial patients is not accurately measured, the clinical trial's result, whether favoring or disfavoring a given treatment, may be misleading.

*MedAvante, Inc., MedAvante, Ewing, NJ; †Research Training Associates of Colorado, Lakewood, CO; ‡Peak Consulting, Conifer, CO; §Eli Lilly and Company, Indiana University, Indianapolis, IN; ||New York State Psychiatric Institute, College of Physicians and Surgeons, Columbia University, New York, NY; ¶Ontario Cancer Biomarker Network, Toronto, Ontario, Canada and #Merck Research Laboratories, West Point, PA.

Received March 1, 2005; accepted after revision September 28, 2005.

Address correspondence and reprint requests to Nina Engelhardt, PhD, 7162 North Pennsylvania Street, Indianapolis, IN 46240. E-mail: ne@medavante.net.

ISSN: 0271-0749/06/2601-0071
DOI: 10.1097/01.jcp.0000194621.61868.7c

Outpatient clinical trials of antidepressants involve multiple research sites. At each site, it is common for several different raters to assess the severity of depression in study patients. This measurement of severity is typically required on multiple occasions for each patient. Detection of a change in a patient's condition may require the comparison of assessments from 2 or more different raters, and pooled analysis of data from multiple sites requires that ratings of patients by many different raters be combined. Differences between raters in how they assess patients may introduce measurement error and, consequently, increase the risk for Type II error, resulting in misleading conclusions as to how treatments compare in effectiveness.[1,2]

It is surprising that the competency of raters, particularly with respect to clinical interview skill, is relatively neglected in industry-sponsored trials. Current approaches to training raters assume a high degree of clinical interview skill and focus instead on achieving consistency in the application of scale conventions and interrater reliability. Available evidence, however, indicates that such training fails to address the variability in proficiency of raters used in industry-sponsored trials.[3–6] Variability might be predicted given the great diversity in the background and experience of raters participating in clinical drug trials conducted in the United States, ranging from psychiatrists to study coordinators with degrees in fields unrelated to psychiatry and little, if any, clinical experience.[7]

The Hamilton Rating Scale for Depression (HAM-D)[8] is the most commonly used measure in clinical trails for the evaluation of treatment of depression. Conventions for scoring each item and discriminating among levels of severity within an item are brief and oriented to the clinician familiar with assessing and treating patients with depression. However, many raters who participate in industry-sponsored clinical trials today do not have comparable levels of experience and/or interview skills. For example, in a recent survey of 29 raters at 12 sites in an industry-sponsored, multisite depression trial, 72% of the raters had learned to administer the HAM-D at investigator meetings and only 38% reported having ever been observed actually conducting a HAM-D as part of their HAM-D training.[3]

Structured interview guides have been created to assist raters in conducting interviews in a uniform fashion.[9,10] However, if the guides are being used inconsistently, there will obviously be more variance.

We evaluated the quality of HAM-D clinical interviews in 2 similar depression trials sponsored by 2 different pharmaceutical companies. The studies were conducted between 1999 and 2001 in the United States and Canada. The sponsors requested that all HAM-D interviews performed at the baseline visit in each study be audiotaped. A subset of tapes was evaluated for interview quality and adherence to study guidelines regarding administration of the HAM-D. Findings are discussed in terms of the association between rater behaviors and quality of clinical interview. Current approaches to rater training are critiqued in light of our findings.

MATERIALS AND METHODS

Study Design

A total of 790 outpatients in 2 double-blind, randomized, placebo-controlled and active comparator-controlled multicenter studies of antidepressants were asked if their baseline HAM-D interviews could be audiotaped. A separate informed consent document was used for this purpose. Refusal by the study patient to grant permission to audiotape the interview was not considered grounds for exclusion from participation in either study. The sponsors provided sites with tape recorders. Each rater was instructed to label the tape recording with the patient number, site number, date, and rater initials.

The maximum treatment duration in both studies was 8 weeks. The 2 study designs shared the following characteristics: double-blind, randomized, multisite, placebo-controlled, and active comparator-controlled. Both studies included outpatients who were at least 18 years and met criteria from the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) for Major Depressive Disorder. The primary efficacy outcome measure was the 17-item HAM-D total score. The larger study (Study A) was conducted at 19 United States sites. Study B comprised 3 Canadian sites.

Study B required the use of the *Structured Interview Guide for the Hamilton Rating Scale for Depression* (SIGH-D).[9] In this study, the SIGH-D was printed on the forms used by raters to record HAM-D scores. Study A requested, but did not require, the use of the SIGH-D, and provided copies of the SIGH-D in addition to a HAM-D scoring form.

In our study, the first 104 tapes that were audible and contained sufficient information for a blinded expert reviewer to generate a total score with a reasonable degree of confidence were reviewed. Study A yielded 74 taped interviews and Study B, 30 taped interviews. The sample of 104 audiotaped interviews represented a minimum of 13 unique sites and 28 unique raters. Due to inconsistent or illegible writing on the audiotape labels, it is unclear how many interviews were conducted by the same rater in Study A or Study B, or how many interviews represented additional unique sites and raters.

The tapes were reviewed by an expert reviewer who has 15 years experience administering the HAM-D and has developed training materials and served as a rater training consultant to the pharmaceutical industry for the past 10 years.

Interview quality was evaluated using the Rater Applied Performance Scale (RAPS).[11] The RAPS assesses 6 dimensions of rater performance: adherence, follow-up, clarification, neutrality, rapport, and accuracy. In this study, adequacy of information obtained was substituted for the RAPS dimension of accuracy. The substitution was made because in many cases, the accurate rating was difficult to determine because the interviewer did not follow up or clarify responses sufficiently. Each dimension was rated on a 4-point scale: excellent, good, fair, or unsatisfactory. Scores were assigned based on quality and consistency of performance throughout the interview. A description of the RAPS dimensions and rating anchors is published elsewhere.[11]

Adherence pertains to whether the rater followed specified guidelines and probes provided on the SIGH-D. Adhering to the interview guide is important, because it increases the likelihood that questions will be asked in a uniform manner within and across research sites. Prior research has shown that use of the SIGH-D increases the reliability of individual items in the scale compared to when the HAM-D is administered in an unstructured fashion.[9,12]

Follow-up evaluates the use of follow-up questions to obtain more information. This usually means going beyond ''yes'' or ''no'' responses.

Clarification involves the use of questions to clarify ambiguous information presented by the patient, and rephrasing and repeating back to the patient in the form of a question what s/he has said.

Neutrality refers to the degree to which the rater's interview style and demeanor do not unduly influence or bias a patient's response. For example, neutrality is often achieved through the use of open-ended questions.

Rapport assesses the ability of the rater to maintain an appropriate relationship during the interview. Although it is essential to establish good rapport with the patient, the rater must be careful not to cross the boundary into therapy, which could confound the experimental intervention.

Adequacy of information obtained was defined as whether the expert reviewer felt there were sufficient data gathered for the expert to rate each item with confidence. The rating for adequacy was based on how many of the 15 items evaluated could be rated with certainty. This ranged from excellent (where 15 items could be rated with confidence) to unsatisfactory (where fewer than 50% of the items could be rated with confidence).

Interview length was calculated on only 102 of 104 interviews in the original sample. Interview length was measured in minutes from beginning to end of recorded interview by the expert reviewer using a stopwatch.

Statistical Methods

Standard descriptive statistics were calculated for all variables, and measures of association and correlation were computed for all relevant combinations of variables. For all items comprising the RAPS, and for all other variables measured on an ordinal scale, Kendall's Tau was used to

TABLE 1. Percentage of Interviews Judged Excellent, Good, Fair, or Unsatisfactory on 6 Dimensions of the RAPS

	Excellent	Good	Fair	Unsatisfactory
Adherence	11%	17%	29%	43%
Clarification	1%	35%	37%	27%
Adequacy	2%	14%	24%	60%
Follow-up	5%	14%	36%	45%
Neutrality	22%	32%	29%	17%
Rapport	2%	49%	38%	11%

measure association. For variables measured on a continuous scale, such as interview length, Pearson correlation was used. Results are presented in terms of percentage of interviews, not raters.

RESULTS

There were no statistically significant differences between Study A and Study B with respect to any of the RAPS dimensions or interview length ($P < 0.05$ for all comparisons). Descriptive statistics on adherence, clarification, follow-up, rapport, and neutrality were calculated on a pooled sample of 104 interviews. A sample of 102 interviews was used to calculate adequacy of information obtained and interview length. Two interviews from the original sample of 104 were excluded from this analysis due to missing values on interview length and adequacy.

Rater Performance

Table 1 displays the percent of interviews judged on the RAPS to be excellent, good, fair, or unsatisfactory on the dimensions of adherence, clarification, adequacy of information obtained, follow-up, neutrality, and rapport.

A majority of interviews (72%) were judged either fair or unsatisfactory in adherence to the SIGH-D. Only 28% were rated good or excellent.

A total of 64% of interviews were rated fair or unsatisfactory for use of clarification of patient responses. Thirty-five percent were rated good, and only 1 interview was rated excellent.

Eighty-four percent of interviews were rated either fair or unsatisfactory in terms of the adequacy of information obtained by the interviewer for the expert reviewer to make a confident rating. More than half (54%) of the interviews did not provide adequate information for the expert reviewer to rate Item 1, Depressed Mood, with confidence.

Only 19% of the interviews were judged to be good or excellent on use of follow-up questions to obtain more information.

Approximately half (54%; 51%) of the interviews demonstrated good or excellent neutral interview style and rapport, respectively.

Interview Length

Baseline interviews ranged in duration from 2 to 35 minutes. Mean interview length was 13 minutes. Over one-third of the interviews, or 39%, were conducted in 10 minutes or less. Severity of illness, defined as baseline HAM-D total score, was not significantly correlated with length of interview ($P = 0.4$).

Table 2 illustrates the correlations among the RAPS dimensions and between the RAPS dimensions and interview length. With the exception of interview length and rapport and interview length and adequacy of information obtained, all correlations were significant at the $P = 0.001$ level (2-sided, 0.223 critical Tau value).

DISCUSSION

Blind evaluation of our sample of audiotapes by an expert reviewer revealed that most baseline HAM-D interviews were brief and cursory, with 39% of the interviews lasting 10 minutes or less. Hamilton wrote ''An adequate interview will surely not be less than half an hour, for that gives an average time of about two minutes per item, which is not really sufficient.''[13] This recommendation seems reasonable especially when conducting the first, or baseline, interview. We noted that simply reading the SIGH-D without pausing for patient responses takes an average of 3.5 minutes. In the 8% of interviews lasting 5 minutes or less, patients would have had to respond to each of 14 HAM-D items (the number of items that requires a question rather than observation only) in a total of 90 seconds, or 6.4 seconds per item. The finding that only 28% of interviews demonstrated adequate (excellent or good) adherence to the SIGH-D may partially account for the brevity of recorded interviews.

TABLE 2. Kendall (Tau) Correlation Matrix of Items of the RAPS

	Int Length	Adherence	Follow-up	Clarification	Neutrality	Rapport	Adequacy
Int Length	1.000						
Adherence	0.318	1.000					
Follow-up	0.283	0.497	1.000				
Clarification	0.256	0.323	0.428	1.000			
Neutrality	0.291	0.608	0.326	0.266	1.000		
Rapport	0.212	0.320	0.359	0.350	0.305	1.000	
Adequacy	0.206	0.505	0.616	0.570	0.283	0.365	1.000

Eighty-one percent of interviews failed to demonstrate adequate use of follow-up questions to obtain sufficient information from the patient to make a confident rating, and 64% of interviews failed to adequately clarify ambiguous patient responses. It is not surprising, therefore, that the expert reviewer judged 83% of the interviews to contain insufficient information to arrive at item scores with confidence (a rating of either unsatisfactory or fair). Sixty percent of interviews received a rating of unsatisfactory, meaning there were insufficient data to confidently rate at least half of the 15 HAM-D items. The correlation between adherence and adequacy of information obtained (0.505; Table 2) suggests that the more likely a rater was to adhere to the SIGH-D, the more likely that rater was to elicit sufficient information to accurately rate the presence and severity of depressive symptoms.

An example of an audiotaped HAM-D clinical interview from our sample is posted on the *Journal of Clinical Psychopharmacology* Web site (www.psychopharmacology.com). This interview lasted 5 minutes, and was judged unsatisfactory on all 6 RAPS dimensions.

Our study had a number of significant limitations that prevented us from adequately describing our sample and answering important questions raised by our findings. The small sample size prohibits any generalization regarding the quality of clinical interview skill of the raters who participated in the 2 industry-sponsored clinical drug trials, and certainly, to raters in general. Indeed, we were unable to adequately characterize the sample of interviews in terms of number of unique raters and sites. Furthermore, there may have been differences between patients who agreed to participate and those who did not. In addition, findings cannot be generalized to other clinician-administered scales, such as the Montgomery-Asberg Depression Rating Scale.[14]

We were also unable to evaluate the consistency of rater behaviors from one interview, or patient, to the next, despite the fact that our sample included more than 1 interview per rater for some raters. Nor could we cite the total number of audiotapes that were not included in the analysis due to insufficient information obtained in the interview for the expert reviewer to make a rating with a reasonable degree of confidence. Such information would, of course, be highly relevant in an evaluation of clinical interview skill.

Future studies could shed light on some very important questions that remain unanswered regarding clinical trial methodology. For example, is there a relationship between quality of clinical interview and trial outcome? A recent study[5] attempted to answer this question. A total of 216 baseline HAM-D interviews in a multicenter depression trial were recorded and evaluated for interview quality using the RAPS. Overall, the study was a failed trial (ie, the active comparator failed to separate from placebo). However, post hoc analyses found that those interviews rated ''good'' or ''excellent'' showed a large and significant placebo separation (6.8 points, $P = 0.017$), whereas those interviews rated ''fair'' or ''poor'' on interview quality failed to separate (−2.8 points, $P = 0.266$) (negative number reflects greater change with placebo than with drug).

Current rater training methods may not address the needs of many raters today. The common practice of showing, days before the start of a clinical drug trial, videotaped clinical interviews to a large group of raters who score the interviews and later discuss discrepancies in scoring at best provides an estimate of a rater's skill at passive rating and knowledge of scale conventions, but reveals nothing about a rater's skill at actually administering the scale. A comprehensive rater training program should include, at a minimum, evaluation of raters' ability to administer the scale in question.

Our findings raise serious questions about the quality of HAM-D assessments conducted in clinical drug trials. Considerably more attention needs to be paid to evaluating the quality of clinical assessments in industry-sponsored drug trials and to investigating the relationship between quality of clinical interview and trial outcome.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of Eli Lilly and Company and Boehringer-Ingelheim.

REFERENCES

1. Porter R, Frampton C, Joyce PR, et al. Randomized controlled trials in psychiatry, Part I: Methodology and critical evaluation. *Aust N Z J Psychiatry*. 2003;37(3):257–264.
2. Robinson DS, Rickels K. Concerns about clinical trials. *J Clin Psychopharmacol*. 2000;20(6):593–596.
3. Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet-based technologies: a comparison to traditional rater training. *J Psychiatr Res*. In press.
4. Demitrack MA, Faries D, Herrera JM, et al. The problem of measurement error in multi-site clinical trials. *Psychopharmacol Bull*. 1998;34:19–24.
5. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection. *Am J Psychiatry*. 2005;162:3.
6. Kobak KA, Lipsitz JD, Williams JBS, et al. A new approach to rater training and certification in a multicenter clinical trial. *J Clin Psychopharmacol*. 2005;25(5):1–6.
7. Targum SD. Rater competency for mood disorders scales. Poster presented at the 45th Annual Meeting of the National Institute of Mental Health, New Clinical Drug Evaluation Unit (NCDEU); June 2005; Boca Raton, FL.
8. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56–62.
9. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry*. 1988;45:742–747.
10. Shear MK, Vander Bilt J, Rucci P, et al. Reliability and validity of a structured interview guide for the Hamilton Rating Scale for Anxiety (SIGH-A). *Depress Anxiety*. 2001;13(4):166–178.
11. Lipsitz JD, Kobak KA, Feiger A, et al. The Rater Applied Performance Scale (RAPS): development and reliability. *Psychiatry Res*. 2004;127:147–155.
12. Moberg PJ, Lazarus LW, Mesholam RI, et al. Comparison of the standard and structured interview guide for the Hamilton Depression Rating Scale in depressed geriatric outpatients. *Am J Geriatr Psychiatry*. 2001;9:35–40.
13. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychiatry*. 1967;6:278–296.
14. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979;132:382–389.

