

Impact of a Comprehensive HAMD Inter-Rater Reliability Training in a Multi-Site Trial

Alan D. Feiger^{1,2}, M.D., Joshua D. Lipsitz^{1,3}, Ph.D., Kenneth A. Kobak^{1,4}, Ph.D., Kenneth R. Evans⁵, Ph.D., Terrence Sills⁵, Ph.D.

¹Research Training Associates, 3555 Lutheran Pkwy, Suite 320, Wheat Ridge, CO 80033

²Feiger Health Research Center, ³New York State Psychiatric Institute, ⁴Healthcare Technology Systems, ⁵Boehringer-Ingelheim

Background: Increasing inter-rater reliability is a crucial step in decreasing error variance in clinical trials. In spite of its importance, the question as to whether pre-study training improves reliability during the trial has not been examined empirically. In addition, several sets of HAMD scoring conventions exist, with little empirical data supporting specific sets of conventions and training conventions with improved outcomes. The current study examined the impact of a comprehensive training protocol on inter-rater reliability and study outcome.

Method: Twenty-one raters from 6 sites met for 2 ½ days of intensive HAMD training. The training involved a) a thorough review and discussion of conventions; b) group role-play exercises; c) individualized small group hands-on training using professional medical actors as patients, with participants conducting interviews with feedback from trainers; d) ongoing monitoring and feedback during the trial using audiotapes; e) a specific set of scoring conventions and modification of the SIGH-D developed for the trial by a panel of experts. All HAMD interviews were audio taped and were later rated, independently, by three HAMD experts (except the visually rated items).

Results: Correlations between HAMD scores conducted by raters during the trial and three HAMD experts who reviewed the audiotapes were .93, .94 and .89 respectively (Cohen's Weighted Kappa, 95% CI .91-.96, .90-.97, and .85-.93). Reliability correlations on individual HAMD items ranged from .87 for Item 5 (weight loss) to -.02 for item 17 (insight) (psychomotor items were not rated). Several items were identified as having lower reliabilities and skewed distributions:

depressed mood, work and activities, guilt, psychic anxiety, hypochondriasis, and insight.

Table 1. Total HAMD Score: Study Raters vs Training Consultants

Trainer	Cohen's Kappa (95% C.I.)	Spearman Rank Correlation (95% C.I.)
KK (N=114)	0.93 (0.91, 0.96)	0.93 (0.90, 0.96)
AF (N=108)	0.94 (0.90, 0.97)	0.93 (0.90, 0.96)
JL (N=129)	0.89 (0.85, 0.93)	0.88 (0.83, 0.93)

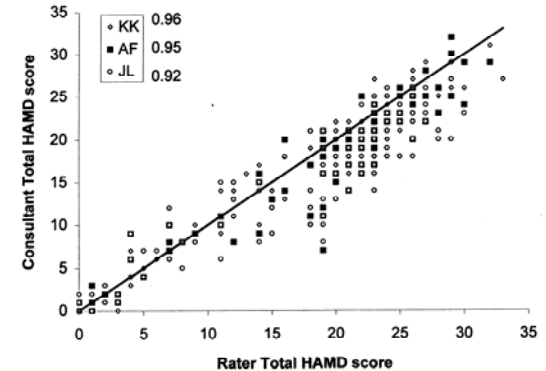
¹Weighted Kappa (quadratic weights) which treats smaller disagreements less seriously than larger disagreements

Table 2 Total HAMD Score: Comparison Between Consultants

Comparison	Cohen's Kappa ¹ (95% C.I.)	Spearman Rank Correlation (95% C.I.)
KK vs AF (N=83)	0.95 (0.92, 0.97)	0.90 (0.84, 0.97)
KK vs JL (N=78)	0.93 (0.90, 0.96)	0.89 (0.83, 0.96)
JL vs AF (N=91)	0.95 (0.93, 0.97)	0.90 (0.84, 0.96)

¹Weighted Kappa (quadratic weights) which treats smaller disagreements less seriously than larger disagreements

Figure 1. Correlation Between Study Raters and Three HAMD Expert Reviewers



Conclusions: In a multi-site antidepressant trial, raters demonstrated very high inter-rater reliability and higher than usual drug placebo separation. Audibility of tapes may have caused some variance, as did lack of observational data for Item 1. Results suggest that intensive pre-study training and audiotape monitoring may increase both reliability and validity (as measured by active drug vs placebo separation). We propose further study of this training methodology in a controlled trial in which half the raters will receive training as usual and half receive comprehensive training.

This study was supported by a grant from Boehringer-Ingelheim