Two for the Price of One: HAM-D Rating Variability Does Not Necessarily Increase with the Number of Raters

Submitter Zachary Cole

Affiliation Adams Clinical

SUBMISSION DETAILS

I agree to provide poster pdf for attendee download. Yes

What is the Methodological Issue Being Addressed? Increased variability between raters reduces inter-rater reliability (IRR) and inflates Type II Error rate[1,2]. To our knowledge, there is no direct evidence in the current literature demonstrating the impact of these factors in clinical trials. To address this, the current study assessed how the use of multiple clinician raters per participant evaluating antidepressant treatment (ADT) response affected outcomes and variability in Major Depressive Disorder (MDD) clinical trials.

Introduction Understanding and minimizing placebo response is a primary concern for MDD clinical trials, as roughly half of trials fail to demonstrate efficacy. IRR is considered a contributing factor due to potential variability across ratings[1,2]. As such, sponsors and sites often emphasize rater consistency, particularly across baseline and primary endpoints. However, this can present practical challenges to sites (e.g., staffing, scheduling), which can impede study enrollment. The current study explored whether participants assessed by different raters at baseline and week six/end of treatment (EOT) showed different outcomes and increased variability in ADT response compared to those assessed by the same rater.

Methods A sample of 122 MDD trial participants were prescribed an FDA approved ADT. Depressive symptoms were assessed at baseline and EOT using the HAM-D, either by the same rater (n = 25) or different rater (n = 97).

Results ADT response (≥50% reduction in HAM-D total score) was exhibited by 24% of participants in the Same Rater condition and 39% of the Different Rater condition. A 2 (Rater: Same, Different) x 2 (Time: Baseline, EOT) mixed-effects factorial ANOVA indicated that EOT ratings (M = 14.54, SD = 9.09) were significantly lower than Baseline ratings (M = 21.37, SD = 6.14; $F(1, 120) = 57.383$, $p < .001$). There was no difference between Same and Different raters ($F(1, 120) = 0.436$, $p = .436$, $BF10 = 0.209$), with Bayes Factors indicating strong support for H0. The interaction was not statistically significant ($F(1, 120) = 0.054$, $p = .817$).

Rating variability was assessed using Levene's Test for Homogeneity of Variances. These analyses revealed no statistically significant differences in variance between the Same and Different raters at Baseline ($F(1, 120) = 0.007$, $p = .934$, $BF10 = 0.314$) and EOT ($F(1, 120) = 0.2616$, $p = .610$, $BF10 = 0.243$), with the Bayes Factors indicating strong support for H0. The difference between the Baseline and EOT ratings for the Same (M = 7.24, SD = 9.79) and Different (M = 6.72, SD = 9.99) raters also revealed no statistically significant differences: $F(1, 120) = 0.201$, $p = .655$, $BF10 =$

0.238.

Conclusion The Same and Different Rater groups had equivalent ratings and variance at Baseline and EOT (indicated by BF10) despite a higher response rate in the Different Rater group. These results suggest that assigning multiple raters to clinical trial participants does not necessarily reduce the ADT response rate or increase rating variability. This striking pattern of results is potentially attributable to an emphasis on IRR training and continual refinement/optimization. It is important to note that these findings did not result from blinded trial data, and thus cannot directly inform conclusions regarding the placebo response.

Co-Authors

Zachary Cole[1], Miriam Evans[1], Sarah Starling[1]

[1] Adams Clinical

Keywords

Keywords
Inter-rater reliability
Inter-rater variability
MDD clinical trials
HAM-D
Clinician-reported outcomes

Guidelines I have read and understand the Poster Guidelines

Disclosures The authors report no conflicts of interest for this work.

Related Tables and Supporting Materials isctm abstract - revised -- References.pdf