How We Are Measuring Up: Comparisons of Symptom Severity Assessments Performed by Independent vs. Site-Based Raters in a Clinical Trial for Major Depressive Disorder

Nash AI[1], Bleys C[2], Xi L[3], Gause, A[4], Moyer, J[4], Van Neuten[4], L, Drevets, W[5], Opler M[6], Salvadore G[4]

[1]Janssen Scientific Affairs, LLC, Titusville, NJ, USA

[2]Janssen Research & Development, LLC, Beerse, Belgium

[3]Janssen Research & Development, LLC, Malvern, PA, USA

[4]Janssen Research & Development, LLC, Titusville, NJ, USA

[5]Janssen Research & Development, LLC, La Jolla, CA, USA

[6]MedAvante-ProPhase Inc., New York, NY, USA

Methodological Question: How do independent rater assessments correlate with site-based over time?

Introduction: Independent raters are utilized in clinical trials due to hypothesized improvements in signal detection, interrater reliability and lower risk of baseline score inflation. However, they come at a high financial cost. Here, we present statistical correlations between remote telephone-administered and site-administered symptom severity scales in a Phase 2 trial in major depressive disorder (MDD).

Methods: The CNTO136MDD2001 study investigated the efficacy of sirukumab as adjunctive treatment for MDD compared to adjunctive placebo based on Hamilton Depression Rating Scale (HDRS17) total score change from baseline to 12-week endpoint. All subjects were diagnosed with MDD and had a suboptimal response to their current standard oral antidepressant. A minimum severity criterion of HDRS17 score ≥ 18 was set at both screening and baseline as assessed by an independent rater. Approximately 1 week prior to baseline visit, subjects completed the SAFER interview, assessed by a second independent rater, which also included an HDRS17 to confirm maintenance of the severity threshold. Primary outcome was change in remote-rater assessed HDRS17 scores from baseline to week 12. Secondary outcomes included change in score for the 30-item Inventory of Depressive Symptomatology (IDS-C30) as assessed by site-based rater. Both assessments were performed at weeks 1, 4, 8 and 12 and were used to calculate number of responders (≥ 50% change) and remitters (HDRS17 ≤ 7or IDS-C30 ≤ 11) at week 12.

Results: Pearson correlations between remote-rater HDRS17 scores were r=0.545 between screening and baseline and r=0.498 between SAFER interview and baseline. Similarly, the Pearson correlation between HDRS17 and IDS-C30 at baseline was r= 0.581 but improved over the course of the trial to r=0.824 at week 8 and 0.863 at week 12. The number of responders at week 12 was significantly higher using IDS-C30 (90% CI, -0.0974, -0.0117), however, the number of remitters was not significantly different using the two scales (90% CI, -0.0357, 0.0237). Score distributions for both scales were comparable at both baseline and week 12.

Conclusions: Pearson correlations between remote raters administering identical scales were only modest and similar to correlations between different scales administered by remote versus site raters. Notably, the number of site-based raters was 84 compared to 37 remote raters (plus an additional 22 SAFER raters) and all raters, remote or site-based, underwent similar training. Similar distributions of scores for HDRS17 and IDS-C30 at week 12 suggests little to no influence or bias by site-raters on final outcomes and the similar distributions at baseline would suggest subject quality was comparable using both ratings. The improvement in correlation between site-based and remote raters over the course of this trial may be due to several factors, including better patient reporting over time. These results suggest that the use of remote raters should be purpose driven, e.g. reducing functional unblinding, and other hypothesized benefits on signal detection may be limited.

Disclosures: Authors, except for M.O., are employed by Janssen and shareholders in the company. M.O. is a full-time employee and shareholder of MedAvante-ProPhase.