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[2], Gause, A[2], Moyer, J[2], Opler M[3], Van Nueten L[2], Drevets, WC[2], Salvadore G[2]

[1]Janssen Scientific Affairs, LLC; [2]Janssen Research and Development, LLC; [3]MedAvante-ProPhase Inc.

Background

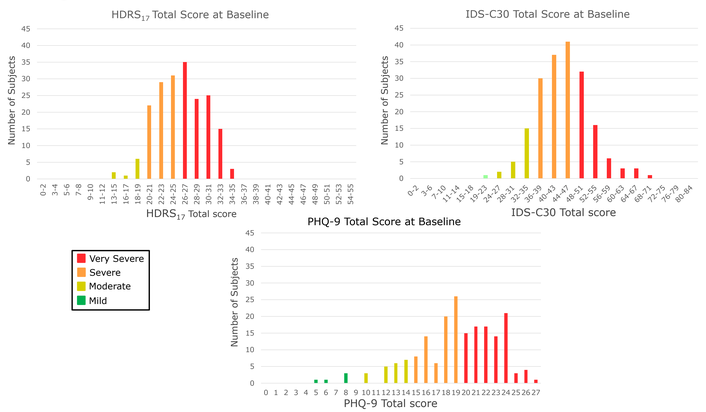
- A high degree of placebo response can impair our ability to detect superiority of active compound vs placebo even if the active compound is effective.
- Use of independent remote raters in clinical trials for MDD, rather than site-based raters, has been reported to reduce placebo effect, improve internal consistency of rating scales and lower baseline score inflation for subjects who are randomized in clinical trials[1], however, the performance of independent raters vs site-based raters has not been systematically investigated. Further, scheduling independent rater assessments outside of clinic visits adds an additional level of complexity for a study and acceptance of phone-based raters by patients may vary across study populations.
- In addition to clinician-based rating scales, patient-rated symptom severity scales also report on changes in disease over the course of a trial, but rely on the direct report of the patient without utilizing a separate rater. To date, only clinician-rated assessments are accepted as primary endpoints for regulatory agency approval in MDD.
- Through rigorously testing the characteristics of these different rating systems, and their relative reliabilities, we may improve trial design and allow better signal detection in clinical trials of MDD.
- Here, we utilize data from a Phase 2 clinical trial of a novel adjunctive treatment for MDD that included site-based, remote rater, and patient-rated symptom severity scales to search for potential scale-dependent differences in patient outcomes and examine the relationships between scales over the course of a 12-week trial.

OBJECTIVE

- To determine if use of different types of raters for outcomes assessments in a trial for adjunctive MDD leads to differences in the characteristics of enrolled subjects and outcome measures and how well these scales are correlated.

RESULTS

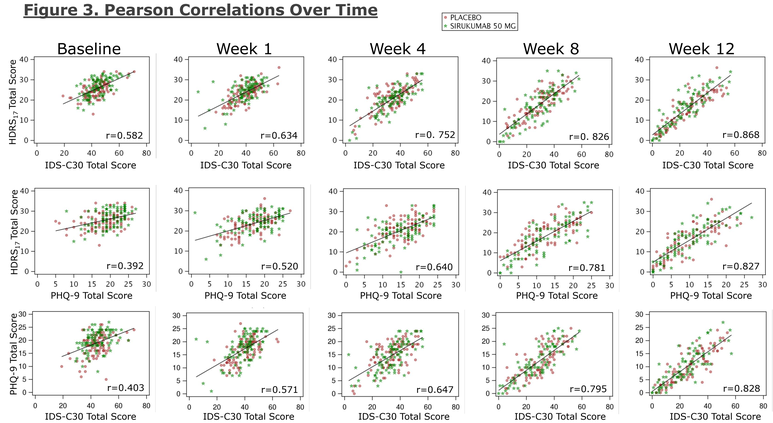
Figure 1. Score Distributions at Baseline



Both remote-rater $HDRS_{17}$ and site-based rater IDS-C30 show bell-curved distribution of scores at baseline, whereas the patient-rated PHQ-9 skews towards higher severity at baseline. $HDRS_{17}$ identified significantly more patients rated as "very severe" (p=0.0003) and significantly less as "moderate" (p=0.0002) than IDS-C30 while PHQ-9 identified more patients rated "very severe" (p=0.0241) than IDS-C30. These findings are in contrast to those published previously[1]. $HDRS_{17}$ and PHQ-9 did not identify significantly different proportions of patients within any particular group.

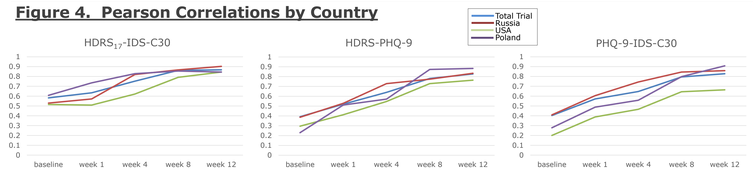
RESULTS

Figure 3. Pearson Correlations Over Time



Positive correlation between scales was observed. Correlations between scales improved over time from baseline to week 12 regardless of rater. This phenomenon has been reported previously[1,2] and may reflect changes in how patients respond to assessments over time, either in improved awareness of symptom severity or improved ability to communicate symptoms. Pearson correlation between $HDRS_{17}$ and IDS-C30 conducted by the same rater has previously been reported as r=0.95 in IDS-C30 validation studies[3].

Figure 4. Pearson Correlations by Country



Pearson correlations between scales analyzed by country for those countries that contributed the majority of subjects to the trial. Similar trend of improved correlations over the duration of the trial. US correlations appear to be consistently less than those for the total trial population and Russia or Poland individually.

Table 1. Cronbach Coefficient Alpha

Cronbach coefficient Alpha measures internal consistency for each scale. Given several symptoms of depression are highly correlated, we would expect the severity scores for those items to be consistent within a scale as well. Lower Cronbach coefficient Alpha score indicates less internal consistency and less reliability of the scale.

		$HDRS_{17}$	IDS-C30	PHQ-9
Cronbach Alpha (α)	Baseline	0.552	0.735	0.840
	Week 12	0.855	0.923	0.919

Notably, all scales have acceptable internal consistency (α>0.7) with the exception of the remote-rater $HDRS_{17}$ at baseline. This suggests remote rater $HDRS_{17}$ assessments are less reliable than either the site-based IDS-C30 or patient-rated PHQ-9 at the same timepoint.

Table 2. Association and Agreement in $HDRS_{17}$ Total Scores Between Different Remote Raters

Population	$HDRS_{17}$ Screening 1 vs. Screening 2		$HDRS_{17}$ Screening 2 vs. Baseline	
	Pearson (r)	ICC (ρ)	Pearson (r)	ICC (ρ)
Total Trial	0.545	0.51	0.605	0.45
Russia	0.678	0.54	0.523	0.42
USA	0.648	0.63	0.652	0.67
Poland	0.405	0.33	0.686	0.24

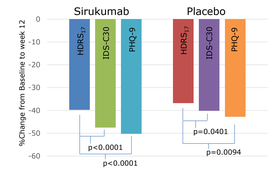
- Screening 1 and Screening 2 $HDRS_{17}$ conducted by different remote raters (MedAvante-Prophase (screening 1 and baseline) and CTNI (Screening 2)
- Screening 1 and Screening 2 were conducted ~3 weeks apart
- Screening 2 and Baseline were conducted ~1 week apart

Correlations between different remote raters administering the same scale have positive moderate association for total trial population and each individual country for which correlations can be calculated. Pearson correlations for the same scale by two different remotes raters is very similar to that between baseline $HDRS_{17}$ and IDS-C30 in Fig. 3 (r= 0.582, $HDRS_{17}$ performed by MedAvante-Prophase, IDS-C30 by site rater). ICCs range from poor to moderate and are more variable by country.

Table 3. Response and Remission Rates **Figure 5. Percent Change from Baseline by Scale**

	Response Rate		Remission Rate	
	Sirukumab	Placebo	Sirukumab	Placebo
$HDRS_{17}$	38.3%*†	35.7%#	21.0%	20.0%
IDS-C30	47.5%*	37.5%‡	22.2%	19.1%
PHQ-9	49.4%†	48.8%#‡	22.2%	24.7%

- Response defined as ≥ 50% improvement in total score
- Remission defined as total score $HDRS_{17}$ ≤ 7, IDS-C30 ≤ 11, PHQ-9 ≤ 4
- Significant differences, *$HDRS_{17}$-IDS-C30 p=0.0196, †$HDRS_{17}$-PHQ-9 p=0.0201, #$HDRS_{17}$-PHQ-9 p=0.0023 ‡IDS-C30-PHQ-9 p=0.0126



Response rate, not remission rate, is significantly different between scales. Though no scale demonstrated significant differences between sirukumab and placebo arms at week 12, $HDRS_{17}$ demonstrates significantly reduced total percent change from baseline to week 12 within both treatment arms vs. IDS-C30 or PHQ-9.

CONCLUSIONS AND DISCUSSION

- This data failed to identify any reductions in baseline severity score inflation, improvement in internal consistency of symptom severity assessments (α), or a differentiation between placebo vs. sirukumab arms for remotes raters vs. site-based raters.
- Pearson correlations between scales improved over time and were similar by week 12. $HDRS_{17}$ performed by remote raters from different agencies failed to demonstrate improved correlations (r and ρ) over those between remote and site-based raters at similar time-points.
- None of the scales presented here demonstrated significant differences between sirukumab and placebo treatment arms. Though $HDRS_{17}$ did demonstrate significantly less change from baseline in the placebo arm (reduced placebo effect), the difference in change from baseline in the sirukumab arm was also significantly less than either IDS-C30 or PHQ-9.
- While the rationale for using remote ratings to address some methodologic problems (e.g. addressing functional unblinding) may be valid, these results call into question the supposed benefits to signal detection in global clinical trials.

LIMITATIONS

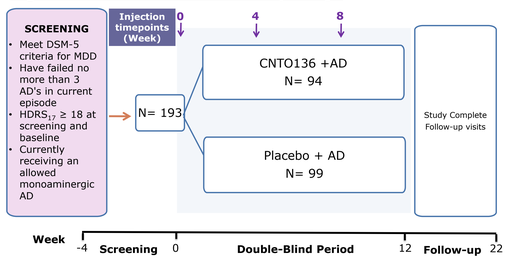
- This study failed to show a significant difference in efficacy between the adjunctive sirukumab and adjunctive placebo treatment arms in the primary analysis sample.
- $HDRS_{17}$ assessments conducted by separate remote raters were not conducted on the same day
- $HDRS_{17}$ and IDS-C30 were assessed by different raters and used a combination of severity and frequency of symptoms while PHQ9 is rated solely by frequency, limiting correlation assessments.
- These data result from post-hoc analyses. Study sample size was not powered to address these specific questions

REFERENCES

1. Kobak, K. et. al *J Clin Psychopharm* 2010; 130:193–197. **2.** Bukumiric, Z. et. al *J Affect Disord* 2016 Jan 15;190:733-743. **3.** Rush et. al *Psychol Med.* 1996; 26(3):477-486. **4.** Li Lu and Nawar Shara, "Reliability analysis: Calculate and Compare Intra-class Correlation Coefficients (ICC) in SAS", NESUG 2007; Statistics and Data Analysis. **5.** Base SAS(R) 9.4 Procedures Guide: Statistical Procedures, Third Edition **6.** Cronbach, L.J. (1951). "Coefficient alpha and the internal structure of the tests. Psychometrika. 16, 297-334.

METHODS

Figure 2. CNTO136MDD2001 Study Design



- Analyses of a Janssen R&D Phase 2a clinical trial (CNTO136MDD2001). Subjects aged 21-64 with MDD without psychotic features (Diagnostics and Statistics Manual 5 [DSM 5]) meeting severity criteria of $HDRS_{17}$ score ≥ 18 as assessed by remote, independent rater at screening visit 1 (MedAvante-Prophase) and screening visit 2 (MGH-Clinical Trials Network and Institute) and a history of inadequate response to ≥1 but ≤ 3 antidepressants in their current episode were randomized [1:1] to 12 weeks of once-monthly treatment with placebo (n=99) or sirukumab injectable (n=94) in addition to their standard oral antidepressant therapy.
- All participants continued the oral antidepressant they were receiving at study entry throughout the duration of the double-blind period.
- Changes in depression severity were measured using the Hamilton Depression Rating Scale ($HDRS_{17}$), Inventory of Depression Symptomatology – Clinician (IDS-C30) and Patient Health Questionnaire (PHQ-9). $HDRS_{17}$ was assessed by MedAvante-Prophase remote rater at all double-blind period time points and IDS-C30 was conducted by site-based clinician at all double-blind period time points. A single rater was not requested for each individual patient. PHQ-9 was performed by subject at each double-blind period time point.
- Screening visit 1 occurred 4 weeks prior to baseline visit, screening visit 2 occurred approximately 1 week prior to baseline visit. $HDRS_{17}$ and IDS-C30 were performed at screening visit 1 and baseline, only $HDRS_{17}$ was performed at screening visit 2. PHQ-9 was not administered at screening, only at baseline and subsequent visits during the double-blind period.
- Statistical Methods:
- Primary and secondary efficacy analyses were based on the Intent-to-Treat using the MMRM method.
- The associations between scales were assessed using Pearson correlation coefficient. The agreement (reliability) between scales were assessed using intra-class correlation coefficient[4].
- The internal consistency of each scale was accessed using Cronbach coefficient alpha (α)[5,6].
- Comparisons of response/remission rates between scales were carried out using McNemar test.
- Comparisons of percentage changes from baseline between scales was carried out using paired t-test.

PHARMACEUTICAL COMPANIES OF Johnson & Johnson

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.