REAL TIME QUALITY ASSURANCE OF DEPRESSION RATINGS IN PSYCHIATRIC CLINICAL TRIALS

Marc Korczykowski, MS[1], Philip T. Lavin, PhD, FASA, FRAPS[1], Courtney Kolesar[1], Ian R. Sharp, PhD[1], Michael T. Sapko, MD, PhD[1], and Jonathan C. Javitt, MD, MPH[1]

[1]Affiliation not available

August 14, 2023

Competing Interests: MK, CK, IRS, MTS, and JCJ are compensated by NRx Pharmaceuticals, Inc. Lavin Statistical Associates is paid of independent statistical analysis by NRx Pharmaceuticals, Inc.

Introduction

Clinician-administered rating scales are a universal endpoint required by regulators around the world for ascertainment of primary endpoint in psychiatric clinical trials. Signal detection in multi-site trials requires strong inter-rater reliability on these instruments; poor inter-rater reliability is associated with increased error variance, reduced study power[2] and, ultimately, failed trials. Poor inter-rater reliability, or unreliability, in psychometric rating scales has many sources, including a lack of adherence to structured and semi-structured interviews, rater scoring differences, and inconsistent interview duration.[3] Williams & Koback correctly state "The importance of reliability of assessments in a clinical trial cannot be overestimated. Without good interrater agreement the chances of detecting a difference in effect between drug and placebo are significantly reduced."[4] Commonly used methods for establishing and maintaining strong inter-rater reliability include site-rater training, external evaluation and monitoring of site-raters, and centralized rating.

Monitoring of endpoint ascertainment in clinical trials is routinely outsourced to Clinical Research Organizations (CROs) and to central laboratories. While psychometric assessments are often monitored by specialized CROs, this may not always be the best choice for a clinical trial. The unique rigor required to ensure valid and reliable clinical scale ratings means CROs must employ enough expert psychometricians who are familiar both with the rating instruments and the unique aspects of the disease and drug being studied. CRO raters must review site assessments within a day of completion to ensure rater quality and accuracy and provide remediation in a timely manner, if needed. Since personnel turnover at CROs may be as high as 20% per year[5], outsourcing the day-to-day management of highly nuanced psychometric ratings becomes impractical when there is turnover and inter-rater variation among the "master raters."

The Sponsor Rating Monitoring System (SRMS) was developed as a pre-defined, protocol-specific, data-driven method to optimize psychometric training, data validity and reliability in the context of a clinical trial of a novel antidepressant targeting bipolar depression with suicidality. In this system, the Sponsor employs expert raters with extensive experience in conducting, analyzing, and training others in the rating scales used to ascertain primary and secondary endpoints. In SRMS, these master raters help the clinical operations team select suitable clinical trial sites, document site rater qualifications, oversee rater training and qualification, and confirm that all data management conforms to the Study Protocol and GDP & GCP guidelines. Most importantly, the Sponsor "master raters" review psychometric assessments within 24 to 48

hours and provide corrective feedback, as needed. This approach further allows for referral of an aberrant rating to an adjudicating rater in real time, prior to data unblinding. The centralized SRMS model does not transfer regulatory obligations to an outside CRO or engage multiple data quality systems, which minimizes oversight and subsequent audit responsibilities.

We examined the Inter-rater Reliability (IRR), i.e., the concordance between site raters and Sponsor "master raters" on MADRS scores on patients participating in the Phase 2b/3 clinical trial "NRX101 for Suicidal Treatment Resistant Bipolar Depression" (ClinicalTrials.gov Identifier: NCT03395392) to assess the potential efficacy of the SRMS.

Methods

MADRS as a Clinical Trial Endpoint

The Montgomery-Åsberg Depression Rating Scale (MADRS) is a primary assessment instrument widely used to measure depression in clinical trials. The MADRS will often be used in conjunction with the Structured Interview Guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA). Thase et al (2021) have reported that MADRS is generally more sensitive than the HAM-D and is the scale used most often in depression trials. The MADRS has been the primary endpoint in a number of bipolar disorder clinical trials, including lumateperone[6], olanzapine plus fluoxetine[7], cariprazine[8], quetiapine plus lithium[9], and adjunctive lurasidone (NCT01284517).[10]

The MADRS is a 10-item, semi-structured assessment with scores ranging from 0-6 for each item where 0 represents absence or denial of symptom and 6 represents the highest symptom severity. The 10-item MADRS/SIGMA addresses questions related to apparent sadness, reported sadness, inner tension, reduced sleep, reduced appetite, concentration difficulties, lassitude, inability to feel, pessimistic thoughts, and suicidal thoughts. The MADRS is focused on mood symptoms whereas the HAM-D measures somatic and behavioral symptoms, which have lower reliability.[11] The MADRS also generally takes less time to administer than the HAM-D, which reduces subject burden.

Rater Training

The Sponsor required all clinical trial sites to ensure all raters were qualified with a minimum of 5 years of psychometric assessment experience. Additionally, all Raters must be certified in MADRS administration specifically for this trial prior to performing clinical trial assessments. Rater training and certification is a multi-step process consisting of reviewing professional qualifications, years of clinical experience, bipolar disorder clinical trial diagnostic experience, and protocol-specific scale administration, including the total number of MADRS administrations and administrations within the past year.

Standardized training was provided for all ratings used in the trial, including MADRS, MINI, C-SSRS, which are used as key primary or secondary trial endpoints. Protocol-specific MINI training was administered via video by Dr. David Sheehan, the author and publisher of the MINI. MADRS training consisted of reviewing and scoring one (or more) test cases and achieving a minimum inter-rater or intra-class reliability correlation coefficient (or 'IRR score') of 0.80 or greater. IRR scores quantify the degree of Rater agreement on bipolar disorder 'gold standard' training case(s). Further, all Raters were certified in the administration of the C-SSRS by completing online training via the BlueCloud® system.

'Real Time' Psychometric Rating Review

A 'Master Rater' with an average of 20 years of clinical research experience in the neuropsychiatry therapeutic area supervised and reviewed psychometric scale administrations of all site raters (MK or IRS). Each Master Rater listened to digital audio recordings of the Site Rater's interviews and provided an independent assessment without knowing the site rater's assigned score. The Rating Review Plan calls for 100% review of all MINI, MADRS, and C-SSRS data at Screening, for all new Sites and new Raters—with a 24- to

48-hour review time. The initial plan was to randomly review 50% of MADRS site ratings between midpoint and endpoints; however, NRx Rater Program Managers increased the review rate to 100% to ensure data reliability across all assessments. This process of continuous monitoring and review of concordance rates is intended to produce valid and reliable assessments, and reduce rater inflation, drift, or fatigue over time.[12]

Measurement of Concordance and Congruence

All measurements of Concordance and Congruence as defined below were ascertained in the context of a triple-blinded study in which neither participants, treating physicians, or raters were aware of treatment group assignment. Moreover, the site raters were further blinded as to the clinical chart of the study participant, previous rating scores, compliance with study drug, or any other clinical characteristics. Sponsor master raters had no participant contact or clinical information and assigned rating scores solely from audio files of site rating sessions.

Three master raters were responsible for independently evaluating MADRS interviews completed by site raters. At a clinical trial visit, a total MADRS score was obtained from the site rater and the Sponsor rater to create a pair of ratings. If the site-rater assigned MADRS score was within 3 points higher or lower than the Sponsor-rater assigned score, it was deemed concordant. If the pair of MADRS scores differed by four points or more, it was considered discordant. This 3-point measure of Concordance is a stricter standard than has recently been advocated by others (see discussion).

"Congruence" or Inter-rater reliability (IRR) was defined as the percent of sampled rating that were concordant. A Congruence or IRR standard of 90% was established by the sponsor for ongoing participation of a study site in the clinical trial. If the site rater and sponsor rater manager's review did not meet the above criteria for IRR, the reviewer contacted the site rater for a consultation on the interview and scores. This consultation provided an opportunity for the resolution of discrepancies and, potentially, site rater training or remediation. Additionally, the sponsor rater manager may contact a site rater to discuss any remediation triggers, specifically observed interviews that led to concerns over scale administration, e.g., lack of adherence to the structured interview guide, numerous leading questions, unusually brief interview duration, etc. If a lack of agreement or other issues with scale administration were identified, the SRMS worked with the rater to remediate performance.

The Congruence or Inter-Rater Reliability (IRR) was calculated as the total number of subjects in concordance divided by the number of subjects assessed multiplied by 100. Intraclass correlation was calculated (VassarStats) to assess the absolute correlation between the raters within the same patient population.[13] A within-subjects ANOVA was used to determine type 1 error. The mean of the absolute difference between site and Sponsor raters was calculated along with a 95% confidence interval.

If rating congruence could not be brought within the required 3 points, the protocol directed referral of the rating to an external adjudicating rater who was trained and standardized with the sponsor master raters on a common rating training set.

Results

Thirty-seven patients of the first 50 randomized patients received two MADRS scores at each post-randomization study visit, one by a rater at one of 12 clinical trial sites and by a Sponsor rater. A total of 113 pairs of paired MADRS assessments were conducted. The absolute difference between the site rater and the Sponsor rater is shown in Figure 1. Overall concordance between site and Sponsor raters was 93.4%. The intraclass correlation was 0.9837 and an $eta^2 = 0.9918$ ($r = 0.9839$, $F = 121.91$, $p < 0.0001$). The absolute mean difference in MADRS rating pairs was 1.78 points (95% CI: 1.50-2.06).

Seven of the 113 assessment pairs were discordant, i.e., the MADRS scores differed by 4 points or more between the site rate and the Sponsor rater. Of the 106 concordant pairs, 17 pairs had identical scores, 36 pairs differed by one point, 36 pairs differed by two points, and 17 pairs differed by three points (range 0 to

9). The seven discordant pairs occurred across five clinical trial sites. In six of seven discordant pairs, the site rater assigned a higher MADRS score than the Sponsor rater. In each case of discordant pairs, the Sponsor rater contacted the site rater to provide feedback about the MADRS assessment.

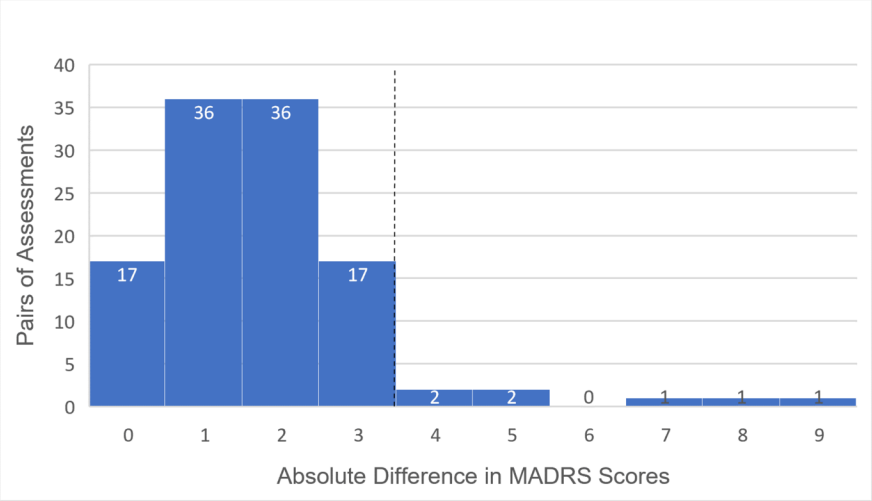

Figure 1. Absolute Difference Between Site and Sponsor Raters. The dashed line indicates the cutoff between concordant and discordant pairs of MADRS scores.

Discussion

The high IRR observed in this trial, specifically 93.4%, supports the utility of "in-sourced" psychometric review. This result, to our knowledge, provides first evidence that this method is practical and implementable with complex psychiatric patients with bipolar depression and subacute suicidal ideation or behavior. This result also replicates and extends the findings of Targum and Catania (2019), who examined concordance between site and site-independent raters using digital audio recording of 3,736 MADRS interviews. They report concordance rates between 89.5% and 95.8% with lower concordance occurring during earlier visits and higher concordance occurring at later visits. The average concordance across all visits was 93.3%. In a separate paper, Targum et al (2014) report a concordance rate of 93.8% between site and site-independent raters, however the discordance cutoff score was 6 or more points on the total MADRS score.[15] However, Targum and Catania defined discordance as a deviation of greater than 6 points on the MADRS, which was equal to one standard deviation of the mean total MADRS score.[14]

In contrast to Targum, the SRMS method used a more rigorous definition of 3 points to achieve a similar concordance rate of 93.4%. If the SRMS criteria were relaxed to define 6 points as the discordant cutoff, only 3 discordant pairs would occur out of 133 assessments yielding a 97.7% IRR rate. Importantly, we did not include Screening visits in this analysis; screening visit MADRS data are used to confirm participant inclusion by Study Protocol, not IRR scores. However, Targum and Catania report the highest discordance rates in Screening Visits (11.5%). The ICC was also very high, consistent with or exceeding results published in similar studies.[16]

Targum and Catania reported that for MADRS scores equal to or greater than 30, site raters tend to assign higher (more severe scores) than site-independent raters.[14] The converse is true for MADRS scores less than 20. Our results confirm this finding. In most (6 out of 7) of our discordant pairs, the site rater score was higher than the Sponsor rater, and in all but one of the pairs, the Sponsor rater-assigned MADRS score was greater than 30. Additionally, when examining interview length, Targum and Catania noted that

MADRS interviews less than or equal to 12 minutes were associated with significantly higher rates of scoring discordance.

The clinical trial from which these concordance data were collected comprised 12 sites with a target enrollment of 72 participants, which is the typical size of a phase 2 trial. Thus, the SRMS system is likely applicable to most Phase 2 and smaller Phase 3 psychiatric trials. The SRMS method with its focus on real time review approach worked well with a limited number (n=12) of experienced clinical trial sites, where participants were interviewed in their primary language. Experience in one site where patients who were not primarily English-speaking were interviewed in English resulted in ratings that were technically uninterpretable and lower than allowable IRR. The clinical trial sites were selected, among other things, for rater experience, particularly with respect to MADRS administration. High concordance rates are likely due, at least in part, to working with experienced site raters (minimum 5 years of experience) who were willing to engage in initial and ongoing training during the trial, if needed. It remains to be seen whether this approach can scale to larger trials with is unclear how well this approach would scale for larger trials. A study with more sites and a larger number of participants would likely require more than the three Master Raters to ensure 100% assessment review at all sites. At the same time, if the SRMS approach results in reduced variance associated with the primary and secondary study endpoints, study sample sizes can be reduced without sacrificing study power.

References

1. Dome P, Rihmer Z, Gonda X. Suicide Risk in Bipolar Disorder: A Brief Review. *Medicina (Kaunas).* 2019;55(8). 10.3390/medicina55080403

2. Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharmacol.* 2002;22(3):318-325. 10.1097/00004714-200206000-00013

3. Kobak KA, Brown B, Sharp I, et al. Sources of unreliability in depression ratings. *J Clin Psychopharmacol.* 2009;29(1):82-85. 10.1097/JCP.0b013e318192e4d7

4. Williams JB, Kobak KA. Development and reliability of a structured interview guide for the Montgomery Asberg Depression Rating Scale (SIGMA). *Br J Psychiatry.* 2008;192(1):52-58. 10.1192/bjp.bp.106.032532

5. Fassbender M. CRO Industry Still Plagued by Turnover: Report. https://www.outsourcing-pharma.com/Article/2019/01/03/CRO-industry-still-plagued-by-CRA-turnover-Report.

6. Calabrese JR, Durgam S, Satlin A, et al. Efficacy and Safety of Lumateperone for Major Depressive Episodes Associated With Bipolar I or Bipolar II Disorder: A Phase 3 Randomized Placebo-Controlled Trial.*Am J Psychiatry.* 2021;178(12):1098-1106. 10.1176/appi.ajp.2021.20091339

7. Tohen M, Vieta E, Calabrese J, et al. Efficacy of olanzapine and olanzapine-fluoxetine combination in the treatment of bipolar I depression. *Arch Gen Psychiatry.* 2003;60(11):1079-1088. 10.1001/archpsyc.60.11.1079

8. Earley WR, Burgess MV, Khan B, et al. Efficacy and safety of cariprazine in bipolar I depression: A double-blind, placebo-controlled phase 3 study. *Bipolar Disord.* 2020;22(4):372-384. 10.1111/bdi.12852

9. Young AH, McElroy SL, Bauer M, et al. A double-blind, placebo-controlled study of quetiapine and lithium monotherapy in adults in the acute phase of bipolar depression (EMBOLDEN I). *J Clin Psychiatry.* 2010;71(2):150-162. 10.4088/JCP.08m04995gre

10. Suppes T, Kroger H, Pikalov A, Loebel A. Lurasidone adjunctive with lithium or valproate for bipolar depression: A placebo-controlled trial utilizing prospective and retrospective enrolment cohorts. *J Psychiatr Res.* 2016;78:86-93. 10.1016/j.jpsychires.2016.03.012

11. Iannuzzo RW, Jaeger J, Goldberg JF, Kafantaris V, Sublette ME. Development and reliability of the HAM-D/MADRS interview: an integrated depression symptom rating scale. *Psychiatry Res.*2006;145(1):21-37. 10.1016/j.psychres.2005.10.009

12. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater reliability in clinical trials of depressive disorders.*Am J Psychiatry.* 2002;159(9):1598-1600. 10.1176/appi.ajp.159.9.1598

13. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation - A discussion and demonstration of basic features. *PLoS One.*2019;14(7):e0219854. 10.1371/journal.pone.0219854

14. Targum SD, Catania CJ. Audio-digital recordings for surveillance in clinical trials of major depressive disorder. *Contemp Clin Trials Commun.* 2019;14:100317. 10.1016/j.conctc.2019.100317

15. Targum SD, Pendergrass JC, Toner C, Asgharnejad M, Burch DJ. Audio-digital recordings used for independent confirmation of site-based MADRS interview scores. *Eur Neuropsychopharmacol.*2014;24(11):1760-1766. 10.1016/j.euroneuro.2014.08.016

16. Targum SD, Daly E, Fedgchin M, Cooper K, Singh JB. Comparability of blinded remote and site-based assessments of response to adjunctive esketamine or placebo nasal spray in patients with treatment resistant depression. *J Psychiatr Res.* 2019;111:68-73. 10.1016/j.jpsychires.2019.01.017