Quality Assurance of Depression Ratings in Psychiatric Clinical Trials

Michael T. Sapko, MD, PhD,[1] Cortney Kolesar, MS,[1] Ian R. Sharp, PhD,[1,2] and Jonathan C. Javitt, MD, MPH[1,3]

Abstract:

Background: Extensive experience with antidepressant clinical trials indicates that interrater reliability (IRR) must be maintained to achieve reliable clinical trial results. Contract research organizations have generally accepted 6 points of rating disparity between study site raters and central "master raters" as concordant, in part because of the personnel turnover and variability within many contract research organizations. We developed and tested an "insourced" model using a small, dedicated team of rater program managers (RPMs), to determine whether 3 points of disparity could successfully be demonstrated as a feasible standard for rating concordance.

Methods: Site raters recorded and scored all Montgomery-Åsberg Depression Rating Scale (MADRS) interviews. Audio files were independently reviewed and scored by RPMs within 24 to 48 hours. Concordance was defined as the absolute difference in MADRS total score of 3 points or less. A MADRS total score that differed by 4 or more points triggered a discussion with the site rater and additional training, as needed.

Results: In a sample of 236 ratings (58 patients), IRR between site ratings and blinded independent RPM ratings was 94.49% (223/236). The lowest concordance, 87.93%, occurred at visit 2, which was the baseline visit in the clinical trial. Concordance rates at visits 3, 4, 5, and 6 were 93.75%, 96.08%, 97.30%, and 100.00%, respectively. The absolute mean difference in MADRS rating pairs was 1.77 points (95% confidence interval: 1.58–1.95). The intraclass correlation was 0.984 and an $\eta^2 = 0.992$ ($F = 124.35$, $P < 0.0001$).

Conclusions: Rigorous rater training together with real-time monitoring of site raters by RPMs can achieve a high degree of IRR on the MADRS.

Key Words: Montgomery-Åsberg Depression Rating Scale, interrater reliability, psychometric testing, intraclass correlation, concordance, bipolar disorder, psychiatric clinical trial design

(*J Clin Psychopharmacol* 2024;00: 00–00)

Clinician-administered rating scales are the standard tools for ascertaining the primary endpoint in clinical trials of antidepressants and other psychiatry drugs. Signal detection in multisite trials requires strong interrater reliability (IRR) on these instruments. Poor IRR is associated with increased error variance, reduced study power,[1] and, ultimately, failed trials. Williams and Kobak 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." Unfortunately, a 2020 analysis of 179 randomized controlled antidepressant trials found that only 4.5% of trials reported IRR coefficients,[2] indicating a widespread methodological gap and potential source of clinical trial failure.

Poor IRR in clinician-administered rating scales has many sources including a lack of adherence to structured and semistructured interviews, rater scoring differences, inconsistent interview duration, poor interview quality, and rater bias.[3,4] Commonly used methods for establishing and maintaining strong IRR include site-rater training, external evaluation and monitoring of site-raters, and centralized rating. However, the widely used industry threshold of 10% disparity between study site ratings and centralized master ratings may be too lenient and may introduce excess variance by itself.

Outsourcing clinical assessments is advantageous for certain endpoints, for example, using a central laboratory for a bioassay. Although psychometric assessments are also routinely outsourced to contract research organizations (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 expert psychometricians. CRO raters must review site assessments soon after they are completed 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 specialized psychometric work to CROs tends to be expensive, time-consuming, and impractical in the broader clinical operations workflow.

Rather than use master raters at a CRO, the sponsor employed expert raters with extensive experience in training, conducting, and analyzing the clinician-rated scale of interest. In this patient rating system, these rater program managers (RPMs) worked closely with the clinical operations team to select suitable clinical trial sites, documented site rater qualifications and training, and provided training when needed. RPMs also reviewed psychometric assessments within 24 to 48 hours and provided corrective feedback, as needed.

We examined the IRR concordance between site raters and RPMs on Montgomery-Åsberg Depression Rating Scale (MADRS) scores assessed from patients participating in the phase 2b/3 clinical trial "NRX101 for Suicidal Treatment Resistant Bipolar Depression" (ClinicalTrials.gov identifier: NCT03395392) to assess the efficacy of this novel patient rating system.

METHODS

MADRS as a Clinical Trial Endpoint

The MADRS and the Hamilton Rating Scale for Depression (HAM-D) are the 2 primary assessments used to measure depression in clinical trials. Both scales are administered by trained clinicians to detect depressive symptom change. The MADRS will often be used in conjunction with the Structured Interview Guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA).[6] The MADRS has been the primary endpoint in a number clinical trials of new drugs for treating bipolar disorder, including lumateperone,[7] olanzapine plus fluoxetine,[8] cariprazine,[9] quetiapine

From the [1]NRx Pharmaceuticals, Inc, Wilmington, DE; [2]Department of Psychology, Chestnut Hill College, Philadelphia, PA; and [3]Department of Ophthalmology, Johns Hopkins University, Baltimore, MD.
Michael T. Sapko, ORCID: 0000-0002-8381-2367

Received March 20, 2024; accepted after revision September 27, 2024.

Address correspondence to: Michael T. Sapko, MD, PhD, NRx Pharmaceuticals, Inc, Wilmington, DE (e-mail: msapko@nrxpharma.com).

Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.psychopharmacology.com).

ISSN: 0271-0749
DOI: 10.1097/JCP.0000000000001936

plus lithium,[10] and adjunctive lurasidone (ClinicalTrials.gov identifier: NCT01284517).[11]

The MADRS is a 10-item, semistructured 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.[12] The MADRS also generally takes less time to administer than the HAM-D, which reduces patient burden.

Rater Training

Raters must be certified in MADRS administration prior to performing clinical trial assessments. Rater training and certification is a multistep 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. The Sponsor required all clinical trial sites to ensure all raters were qualified with a minimum of 5 years of psychometric assessment experience in a clinical trial setting. All approved raters also demonstrated prior clinical trial experience with bipolar disorder patients.

Standardized training was provided for all ratings used in the trial, including MADRS, MINI, and 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 consists of reviewing and scoring 1 (or more) test cases and achieving a minimum interrater or intraclass 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

Three RPMs with an average of 20 years of neuropsychiatry clinical research experience and MADRS administration (6 years minimum) supervised and reviewed site rate assessments. An RPM listened to digital audio recordings of the site rater's interviews and provided an independent assessment without knowing the site rater's assigned score, that is, a blinded rating. 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; however, The sponsor decided that the veracity of the MADRS scoring required a 100% review rate. The sponsor hired another RPM and increased the RPM review rate to 100%. 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.[13]

Remediation

If the site rater and RPM'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, otherwise known as "adjudication," provided an opportunity for the resolution of scoring discrepancies and, potentially, site rater training or remediation. Additionally, the RPM may contact a site rater to discuss any remediation triggers, specifically observed interviews that led to concerns over scale administration, for example, lack of adherence to the structured interview guide, numerous leading questions, unusually brief interview duration, and so on. If a lack of agreement or other issues with scale administration were identified, the RPMs worked with the rater to remediate performance by identifying specific scoring issues and reviewing compliance to training documents and the study protocol.

Data Analysis

The clinical trial protocol and its associated informed consent agreement were reviewed and approved by the central institutional review board of this study, Advarra, Inc. 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, that is, absolute difference, than the sponsor-rater assigned score, it was deemed concordant. If the pair of MADRS scores differed by 4 points or more, it was considered discordant. The concordance rate was calculated as the total number of subjects in concordance by the number of subjects assessed multiplied by 100. To test for systematic differences between site raters and RPMs, both the magnitude and the direction of difference were calculated, that is, the site rater score higher or lower than the RPM. Skewness was tested using the MS Excel SKEW command. Intraclass correlation (correlation for unordered pairs; VassarStats) was calculated to assess the absolute correlation between the raters within the same patient population.[14] A 1-way analysis of variance for independent samples was used to determine a *P* value.

RESULTS

Fifty-eight patients received at least 1 pair of independently assessed MADRS scores, 1 by the rater at a clinical trial site and 1 by the RPM. A total of 236 pairs of MADRS assessments were conducted. The absolute difference between the site rater and the RPM MADRS scores is shown in Figure 1. Overall concordance between site raters (n = 23) and RPMs was 4.49%. The percentages in the abstract and discussion are correct.%. Thirteen of the 236 assessment pairs were discordant, that is, the MADRS scores differed by 4 points or more between the site rate and the RPM. Of the 223 concordant pairs, 39 pairs had identical scores, 72 pairs differed by 1 point, 68 pairs differed by 2 points, and 46 pairs differed by 3 points (range 0 to 9). The 13 discordant pairs occurred across 9 clinical trial sites. The lowest concordance, 87.93%,

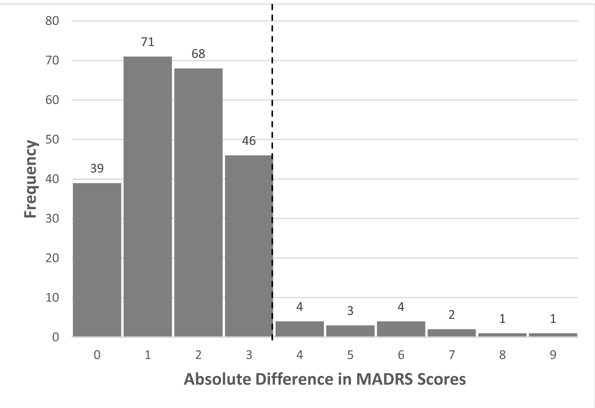

FIGURE 1. Absolute difference between site and sponsor raters. The dashed line indicates the cutoff between concordant and discordant pairs of MADRS scores.

TABLE 1. Relation of Site Rater to Rater Program Manager MADRS Scores

	Site Rater < RPM	Site Rater = RPM	Site Rater > RPM	Skewness
All visits	64	39	133	0.378
Visit 2 (baseline)	13	6	39	0.888
Visit 3	15	9	24	0.073
Visit 4	15	10	26	−0.285
Visit 5	10	5	22	0.348
Visit 6 (end of treatment)	11	9	22	−0.342
MADRS ≥30	31	10	60	0.559
MADRS <30	33	29	73	0.047
MADRS <20	20	23	45	0.358

occurred at visit 2, which was the baseline visit in the clinical trial. Concordance rates at visits 3, 4, 5, and 6 were 93.75%, 96.08%, 97.30%, and 100.00%, respectively. The absolute mean difference in MADRS rating pairs was 1.77 points (95% confidence interval: 1.58–1.95). The intraclass correlation was 0.984 and an $\eta^2 = 0.992$ ($F = 124.35$, $P < 0.0001$).

To determine whether the site raters were consistently higher or lower than sponsor raters, we examined the relative difference between the ratings and assessed skewness. Skewness across all visits was 0.378 (Table 1). A score between −0.5 and 0.5 is considered symmetric. Skewness at visits 2, 3, 4, 5, and 6 were 0.888, 0.073, −0.285, 0.348, and −0.342, respectively. Skewness of site rater scores when RPM-scored MADRS was ≥30, <30, and <20 was 0.559, 0.047, and 0.358, respectively.

Concordance rates by MARDS item, by site, and by site rater are provided in Tables S1, S2, and S3 (Supplemental Digital Content, SDC 1: http://links.lww.com/JCP/A933), respectively.

DISCUSSION

The goal of the current study was to evaluate a novel "insourced" patient rating system in a phase 2 clinical trial of bipolar disorder patients with subacute suicidal ideation and behavior. The high IRR observed in this trial, specifically 94.49%, suggests that an "insourced" psychometric review is an effective option in CNS clinical trials. 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,[15] who examined concordance between site and site-independent raters using digital audio recording of 3736 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 93.7%. However, Targum and Catania[15] defined discordance as a deviation of greater than 6 points on the MADRS, which was equal to 1 standard deviation of the mean total MADRS score. Our method used a more rigorous cutoff of 3 points to achieve a similar concordance rate of 93.7%. If we were to apply 6 points as the discordant cutoff in our dataset, 6 discordant pairs would have occurred out of 236 assessments, yielding a 97.46% concordance 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%). In a separate article, Targum et al[16] 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.[16] The intraclass correlation for our dataset was also very high, consistent with or exceeding results published in similar studies.[17]

Targum and Catania[15] reported that, for MADRS scores equal to or greater than 30, site raters tend to assign higher (more severe) scores than site-independent raters. The converse is true for MADRS scores less than 20 according to their analysis. We found that the site scores were higher than RPM scores whether the RPM-assigned MADRS score was greater than 30, 30 or less, or less than 20. Like Targum and Catania, we also found the largest magnitude of "score inflation" for MADRS scores greater than 30 (skewness = 0.577).

When examining interview length, Targum and Catania also noted in previous research that MADRS interviews less than or equal to 12 minutes were associated with significantly higher rates of scoring discordance. Anecdotally, our reviewers noticed an inverse correlation between interview length and the magnitude of score discrepancy with considerable decreases in quality when interviews were less than approximately 10 minutes. It has also been suggested that site raters may consciously or unconsciously ascertain worse scores at the initial visit to assure trial entry of subjects.[18,19] This tendency introduces a near-certainty of a high placebo effect as patients will quickly regress to the mean on postrandomization visits. We found some evidence to support this assertion in our dataset. The distribution across all visits and at each visit except for visit 2 had a normal distribution. The MADRS ratings at visit 2 were slightly skewed (skewness = 0.888) such that the site raters' MADRS scores were higher than those of the RPMs.

The lowest concordance between rater pairs occurred at visit 2 and steadily improved across visits with perfect concordance occurring at visit 6. This is likely because the ratings were reviewed by the RPM within 24 to 48 hours after completion at the site. If a large discrepancy was noted, the RPM contacted the site to discuss and adjudicate the discrepancy and perform additional training, if needed. This adjudication and retraining likely led to better concordance at later visits. Furthermore, the concordance rates we report are likely better than they would have been if no adjudication took place.

"Insourcing" clinical endpoint ratings is an innovative method that may centralize, de-risk, and optimize clinical trial endpoint validity and reliability across multiple raters and sites. This approach to psychometric rating differs from most psychiatric clinical trials, which outsource assessment monitoring to specialized CROs. Our insourcing model is predicated on continuous review by sponsor RPMs who have extensive neuropsychiatry clinical trial experience and who provided frequent and direct contact with site raters and study coordinators. RPMs reviewed all screening, baseline, and subsequent visits to monitor and resolve any source of potential measurement error, specifically, poor

IRR, poor interview quality, or rater bias, as defined by Kobak et al.[4] "Real-time" data review by RPMs typically occurs with 24–48 hours of the site assessment, so that site raters can receive timely feedback on the quality of their interviews and assessments, if needed.

In our experience, outsourced rating services require at least a week to evaluate and report IRR data, by which time the trial participant may be well along the treatment protocol. Further, the insourced method streamlines and optimizes operational infrastructure with respect to having a smaller, more agile, and cost-effective team.

This novel "real-time" review approach works best with a limited number of high-performing clinical trial sites. The clinical trial sites were selected, among other things, for rater experience, particularly regarding MADRS administration. Concordance rates were high because, at least in part, the sponsor selected sites with experienced site raters (minimum 5 years of experience) who were willing to engage in initial and ongoing training during the trial, if needed. Future studies will determine how well this approach can scale for larger trials. A study with more sites and a larger number of participants would likely require more than the 3 site rater managers to ensure 100% assessment review at all sites. The clinical trial from which these concordance data were collected comprised 12 sites with a target enrollment of 74 participants, which is the typical size of a phase 2 trial. Thus, the patient rating system is likely applicable and generalizable to most phase 2 and smaller phase 3 psychiatric trials.

In conclusion, the current results support the use of an insourced model and the importance of increased transparency with respect to reporting IRR reliability data, particularly primary efficacy endpoints,[2] in neuropsychiatry trials. Defining minimally acceptable concordance for clinical trial endpoints a priori and publishing the IRR results with clinical trial data would considerably increase the transparency and generalizability of findings across neuropsychiatry trials. Insourcing increases operational efficiency in screening complex psychiatric patients for clinical trial inclusion and standardizes the data management and review for subsequent trial visits. This method was particularly well suited for endpoint adjudication in the broader context of a phase 2 clinical trial running across 10–14 sites. Future studies will demonstrate the utility of this model in larger, pivotal phase 3 clinical trials with a greater number of sites and participants. Finally, we call upon clinical trial sponsors and CROs to track and publish IRR along with psychiatric clinical trial results.

AUTHOR DISCLOSURE INFORMATION

J.J. owns equity in NRx Pharmaceuticals, Inc. M.S., I.S., and C.K. are compensated by NRx Pharmaceuticals, Inc.

This work was funded by NRx Pharmaceuticals, Inc.

Data availability statement: The datasets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

REFERENCES

1. Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharmacol*. 2002;22:318–325.
2. Berendsen S, Verdegaal LMA, van Tricht MJ, et al. An old but still burning problem: inter-rater reliability in clinical trials with antidepressant medication. *J Affect Disord*. 2020;276:748–751.
3. Kobak KA, Brown B, Sharp I, et al. Sources of unreliability in depression ratings. *J Clin Psychopharmacol*. 2009;29:82–85.
4. Kobak KA, Kane JM, Thase ME, et al. Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms? *J Clin Psychopharmacol*. 2007;27:1–5.
5. Fassbender M. CRO industry still plagued by turnover: report. Available at: https://www.outsourcing-pharma.com/Article/2019/01/03/CRO-industry-still-plagued-by-CRA-turnover-Report.
6. Williams JB, Kobak KA. Development and reliability of a structured interview guide for the Montgomery-Asberg Depression Rating Scale (SIGMA). *Br J Psychiatry*. 2008;192:52–58.
7. 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;178:1098–1106.
8. 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;60:1079–1088.
9. 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;22:372–384.
10. 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;71:150–162.
11. Suppes T, Kroger H, Pikalov A, et al. 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.
12. Iannuzzo RW, Jaeger J, Goldberg JF, et al. Development and reliability of the HAM-D/MADRS interview: an integrated depression symptom rating scale. *Psychiatry Res*. 2006;145:21–37.
13. Mulsant BH, Kastango KB, Rosen J, et al. Interrater reliability in clinical trials of depressive disorders. *Am J Psychiatry*. 2002;159:1598–1600.
14. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation—a discussion and demonstration of basic features. *PloS One*. 2019;14:e0219854.
15. Targum SD, Catania CJ. Audio-digital recordings for surveillance in clinical trials of major depressive disorder. *Contemp Clin Trials Commun*. 2019;14:100317.
16. Targum SD, Pendergrass JC, Toner C, et al. Audio-digital recordings used for independent confirmation of site-based MADRS interview scores. *Eur Neuropsychopharmacol*. 2014;24:1760–1766.
17. Targum SD, Daly E, Fedgchin M, et al. 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.
18. Mundt JC, Greist JH, Jefferson JW, et al. Is it easier to find what you are looking for if you think you know what it looks like? *J Clin Psychopharmacol*. 2007;27:121–125.
19. Greenberg RP, Bornstein RF, Greenberg MD, et al. A meta-analysis of antidepressant outcome under "blinder" conditions. *J Consult Clin Psychol*. 1992;60:664–669; discussion 670-667.