Schizophrenia Research 232 (2021) 54–60

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Audio-digital recordings to assess ratings reliability in clinical trials of schizophrenia

Steven D. Targum *, J. Cara Pendergrass, Christopher Murphy

Signant Health, Boston, MA, United States of America

ARTICLE INFO

Article history:
Received 10 November 2020
Received in revised form 29 March 2021
Accepted 2 May 2021Available online xxxx

Keywords:
Ratings reliability
Paired ratings
Surveillance
Schizophrenia
BPRS
PANSS

ABSTRACT

We examined ratings reliability in 5 clinical trials of subjects with schizophrenia experiencing an acute exacerbation of psychosis. Audio-digital recordings of site-based interviews of the Positive and Negative Syndrome Scale (PANSS) or Brief Psychiatric Rating Scale (BPRS) were used to obtain blinded, site-independent scores to evaluate paired scoring concordance.

High intraclass correlations were noted between 1810 paired site-based and site-independent PANSS scores ($r = 0.801$) and 1837 paired BPRS scores ($r = 0.897$) with high limits of agreement such that 93.9% of paired scores were within the calculated 95% confidence intervals. In 2 studies where sufficient PANSS data was available at baseline and endpoint, blinded site-independent ratings yielded a predictive value of 84.2% for replicating site-based response/nonresponse treatment outcomes.

There was a significant positive correlation between site-based scores and paired scoring deviations (PANSS: $r = 0.246$; $p < 0.0001$; BPRS: $r = 0.176$; $p < 0.0001$). The magnitude (symptom severity) of PANSS or BPRS scores affected the directionality of paired scoring deviations in each study. Site-based raters scored the most symptomatic subjects higher and less symptomatic subjects lower than the paired site-independent raters on either instrument.

This analysis affirms the utility of paired audio-digital scoring of site-based interviews as a surveillance strategy for schizophrenia studies. We noted a high predictive value of blinded site-independent raters to replicate site-based treatment outcomes. The bi-directionality of paired scoring deviations observed for both the PANSS and BPRS is consistent with findings found for depression rating instruments.

1. Introduction

Clinical trials rely on ratings reliability to achieve meaningful trial outcomes, and it is presumed that clinician raters will conduct competent, reliable interviews at each study visit. However, large clinical trials use multiple trial sites and necessarily require multiple raters whose interviewing techniques and application of scoring criteria may vary and affect the reliability of the outcome measures (Muller and Szegedi, 2002; Kobak et al., 2004, 2005, 2007; Targum, 2006). A lack of inter-rater reliability (IRR) can affect the power of the clinical trial to achieve signal detection (Muller and Szegedi, 2002). Inter-rater scoring differences may be due to a different clinical perspective, a difficult subject, a lack of precision in the application of ratings conventions, or simply a lack of time given to adequately administer the rating instrument. Ratings inaccuracy (e.g. score inflation) due to misplaced site-based rater and/or subject motivations can also reflect a deceptive practice to enroll subjects into trials (Lee et al., 2018). In an analysis of 63 published papers, Mulsant et al. (2002) noted that there were few published reports that described the reliability of the ratings that were conducted during the study, and that only 13% of these studies even mentioned IRR (Mulsant et al., 2002). Elaborate rater training and certification programs have been employed to instruct, standardize, and demonstrate inter-rater reliability for each of the commonly used rating instruments (Muller and Wetzel, 1998; Kobak et al., 2004, 2005; Targum, 2006; West et al., 2014). In an effort to assure ratings reliability, site-independent review and scoring of site-based ratings has been used as a quality assurance, surveillance strategy to monitor site-based raters during a clinical trial (Zarate Jr et al., 1997; Targum and Pendergrass, 2014; Targum et al., 2014, 2015, 2020). This surveillance method employs audio-digital recording and blinded scoring of site-based interviews to obtain "paired" scores based upon the same interview.

In a previously published analysis, we reported that audio-digital recording and paired scoring of site-based clinician interviews of the Montgomery-Asberg depression rating scale (MADRS) yielded high intra-class correlations between the site-based and paired site-independent scores from 5 different studies of major depressive

* Corresponding author at: 505 Tremont Street, #907, Boston, MA 02116, United States of America.
E-mail address: sdtargum@yahoo.com (S.D. Targum).

https://doi.org/10.1016/j.schres.2021.05.004
0920-9964/

disorder (Targum and Catania, 2019). Site-independent review and scoring of recorded site-based clinician interviews has also been employed in clinical trials of schizophrenia (Targum et al., 2014, 2020). The Positive and Negative Syndrome Scale (PANSS) and Brief Psychiatric Rating Scale (BPRS) are symptomatic rating instruments that are commonly used in clinical trials of schizophrenia (Kay et al., 1987; Overall and Gorham, 1988). Successful administration and scoring of these instruments in subjects with psychosis requires good interviewing skills, clinical judgment, some subjective reasoning, and corroborative information for nearly half of the 30 PANSS items (Kay et al., 1987). In this current report, we examined the utility of the audio-digital recording surveillance method from data accumulated from 5 distinct clinical trials of schizophrenia that included 3647 site-based PANSS or BPRS interviews that were paired with site-independent scores.

2. Material and methods

Data for this ratings reliability analysis was obtained from 5 phase II or III clinical trials conducted between 2011 and 2019 as part of vendor grants awarded to Clintara LLC (or Signant Health) to conduct quality assurance/surveillance monitoring programs for ratings reliability. The 5 selected studies were registered in Clinicaltrials.gov as: NCT 01939548, 01469039, 03697252, 01499563, and 02469155. The analysis was limited to clinical trials that had conducted double-blind, placebo-controlled trials involving subjects with schizophrenia experiencing an exacerbation of psychosis and had obtained paired ("dual") site-independent scores based upon audio-digital recordings of site-based structured interviews of the PANSS or BPRS as part of the study (Kay et al., 1987; Overall and Gorham, 1988; Ventura et al., 1993; Crippa et al., 2001; Targum et al., 2014, 2015). Four studies were conducted exclusively in the United States and one study was conducted in the United States and Eastern Europe. All enrolled subjects met DSM-IV or DSM-5 criteria for schizophrenia confirmed by either the Mini International Psychiatric Interview (MINI) or Structured Clinical Interview for DSM-IV (SCID-TR) depending on the study (APA, 1994; Sheehan et al., 1998; First et al., 2007; APA, 2013). The work described in each study was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans. All subjects provided written informed consent for study participation and recording of interviews approved by an independent review board prior to participation in the studies.

The site-based raters included physicians (mostly psychiatrists), nurses, doctorial and masters prepared clinicians who all had a minimum of 2 years research experience with the study specific instruments. All of the site-independent rater/reviewers were psychiatrists with extensive clinical and research experience using standardized instruments. All site-based raters and site-independent reviewers participated in a comprehensive rater training and certification program designed for each study that included didactic presentations, observation of video recordings of PANSS and/or BPRS interviews based upon the specific study requirements, and demonstration of PANSS and/or BPRS scoring competency via inter-rater reliability assessments of the video interviews (Muller and Szegedi, 2002; Kobak et al., 2004, 2007; Targum, 2006; West et al., 2014). In addition, site-based raters were required to administer either the PANSS or BPRS interviews with mock subjects to demonstrate their interviewing competency skills.

Paired ratings were obtained as part of a contracted quality assurance program for each study that was designed to identify discordant site-based rater outliers who might require interview and/or scoring remediation. As part of this quality assurance (surveillance) program, the site-based raters were trained to conduct the PANSS or BPRS interviews using an audio-digital recording pen or tablet. The digital notes completed by the site-based rater captured corroborative, informant and any other relevant information that was observed by the site-based rater to support their scores beyond the recorded interview. The recorded interviews with the accompanying digital notes were electronically forwarded to the surveillance vendor. The audio-digital recordings were administratively reviewed to ascertain the technical quality and completeness of the interview prior to random assignment to a site-independent reviewer/rater for scoring. A few interviews could not be independently reviewed due to technical issues, such as failure of electronic transmission. The site-independent rater/reviewers who received the recorded interviews were blinded to the site-based rater's scores. The site-independent rater generated his or her own independent score by listening to the audio recording and reading the site-based rater's accompanying digital notes. All BPRS items and 26 PANSS items were independently scored in this manner. The 4 PANSS items that required direct observation of subject behavior during the interview (blunted affect, tension, mannerisms and posturing, and motor retardation) could not be adequately assessed by listening to the recorded interview and were carried over from the site-based score. Markedly discordant scores were subjected to a second site-independent review in order to confirm the paired discordance. If the discordance was confirmed, the site-rater was contacted for discussion and remediation of the specific rating issues. The objective of rater remediation was to improve subsequent interview performance, although some raters failed to improve and were removed from the study after additional efforts at remediation.

The data from the 5 studies were examined to assess ratings reliability and scoring concordance or deviations between the individual paired site-based and site-independent ratings. Scoring deviations were defined as the difference between the site-based rater's total score minus the site-independent rater's score. Positive scoring deviations indicate that the site-based score was higher than the paired site-independent score whereas negative deviations indicate that the site-based score was lower than the site-independent score. We also examined the effect of the total PANSS or BPRS score severity and study visit on paired scoring deviations. The paired scores for each study were divided into sub-groups based upon the total site-based PANSS score (≥110, 100–109, 90–99, 80–89, 70–79, and <70) or total site-based BPRS score (≥ 70, 60–69, 50–59, and <50) and into separate visit groups based upon the available data.

The predictive value of paired site-independent scores to match site-based response/nonresponse treatment outcomes was evaluated in the two studies that had paired PANSS data at the baseline and study endpoint visits. We used ≥30% improvement from the baseline visit as the criteria for treatment response (Leucht et al., 2007; Correll et al., 2011).

Statistical analyses used Students' *t*-test, Chi square analysis with Yates correction for continuity, and intra-class correlation (ICC) as appropriate to compare the site-based and site-independent scores. The significance level was set at 5% for all tests in these analyses. In addition, we plotted Bland-Altman scatterplots to examine the limits of agreement (LOA) between each site-based and site-independent rating pair and calculated 95% confidence intervals as the mean difference between paired ratings ±1.96 standard deviations (Bland and Altman, 1986).

3. Results

Every PANSS interviews at every study visit was recorded and submitted for possible independent review. Although all of the PANSS interview recordings were submitted, only a pre-defined number of post randomization PANSS interviews were independently scored. Paired screen and baseline BPRS interviews were examined in Study 03 and only at the screen interviews in Studies 04 and 05. The interviews were conducted by 218 certified site-based raters and reviewed by 33 site-independent raters. Some of the raters participated in more than one study; hence, there were 166 unique site-based raters and 23 unique site-independent rater/reviewers.

Across the 5 studies, 3647 paired scores were obtained that included 1810 PANSS scores and 1837 BPRS scores.

3.1. Comparison of paired total PANSS and BPRS scores

The intra-class correlation (ICC) between the 1810 paired site-based and site-independent total PANSS scores was r = 0.801 (r = 0.818 and 0.765 for the 2 studies using the PANSS) and r = 0.897 for the 1837 paired BPRS scores (r = 0.821, 0.937, and 0.896 respectively for the 3 studies using the BPRS). Study 01 included 653 U.S. PANSS pairs and 604 Eastern European pairs with ICC of r = 0.765 and 0.839 respectively.

Figs. 1 and 2 display the distribution of scoring deviations generated by the paired PANSS and BPRS ratings for all studies. There was a normal distribution with some marked outliers. After ruling out electronic transmission errors as the source of discordance and getting secondary, independent confirmation of the markedly discordant scores, ratings remediation was provided to the identified site-based rater outliers. The most common reasons for marked paired scoring deviations were due to resistant or uninformative subjects, poor interview quality, short interviews that lacked sufficient information to allow independent scoring replication, and/or a misunderstanding or failure to apply the conventional scoring anchors. Subsequent review of site-based rater performance following remediation revealed greater paired scoring concordance in almost every case. In 3 instances, the site-based rater was removed from the study.

Figs. 3 and 4 display Bland-Altman scatterplots that examine the individual paired scoring differences between the site-based and site-independent scores relative to the average score symptom severity (Bland and Altman, 1986). As shown, there was a high level of agreement across the range of total scores for both the PANSS and BPRS. There were 62 PANSS score pairs (3.4%) above, and 36 pairs (2.0%) below the 95% confidence level. 94.6% of the U.S. paired PANSS scores and 93.7% of the European scores were within the 95% confidence interval. Similarly, 80 BPRS pairs (4.4%) were above, and 45 BPRS pairs (2.5%) were below the confidence interval. 93.9% of the paired BPRS scores were within the calculated 95% confidence interval.

Table 1 displays the mean (±SD) site-based and site-independent total PANSS and BPRS scores for each study by visit. Higher positive paired scoring deviations were noted primarily at the screen and baseline visits. As shown in Studies 01 and 02 where interim visits were scored, there was substantially less paired positive scoring deviations after the baseline visit. As shown below in Table 2, the symptom severity (magnitude) of the site-based PANSS or BPRS scores had much greater impact on paired scoring deviations than the study visit itself.

3.2. Effect of total PANSS or BPRS severity score on paired scoring deviations

The severity (magnitude) of symptoms as measured by the total site-based PANSS and BPRS scores affected the extent and directionality of paired scoring deviations in each of the 5 studies examined. There was a significant positive correlation between the 1810 total site-based PANSS scores and paired scoring deviations (r = 0.246; r^2 = 0.060; df = 1808; t = 10.8; p < 0.0001). Similarly, there was a significant positive correlation between the 1837 total site-based BPRS scores and paired BPRS scoring deviations (r = 0.176; r^2 = 0.031; df = 1835; t = 7.68; p < 0.0001).

As shown in Table 2 and Figs. 5 and 6, the magnitude of total PANSS and BPRS symptom severity scores affected the extent and proportional directionality of paired scoring deviations. Higher site-based scores yielded more positive paired scoring deviations whereas lower scores generated more negative scoring deviations across all studies.

3.3. Predictive value of paired site-independent ratings for study outcome

BPRS interviews were only administered and recorded during pre-randomization visits in the 3 studies evaluated in this report, whereas The PANSS interviews were administered and recorded at every study visit. By design, the surveillance program collected but did not independently score all of the recorded interviews at every visit. There were 171 paired PANSS scores available that assessed both the baseline and endpoint visits from the same subject in the 2 studies employing the PANSS. The paired site-independent scores were made by raters who were blind to study site, study visit, and to any adverse event data that might have been known at the trial site. Merging data from the 2 studies, 144 of the 171 paired site-independent PANSS scores (84.2%) correctly matched the response/non-response treatment outcomes of the site-based raters at the study endpoint with a combined sensitivity of 76.7% and specificity of 85.9%.

4. Discussion

We examined the utility of audio-digital recordings and blinded, site-independent scoring as a surveillance strategy for quality assurance of site-based interviews from data obtained from 5 different clinical studies of schizophrenia. The site-independent raters were blind to the study site, study visit, and to any adverse event data. Site-independent scoring of the audio-digital recordings of 1810 site-based

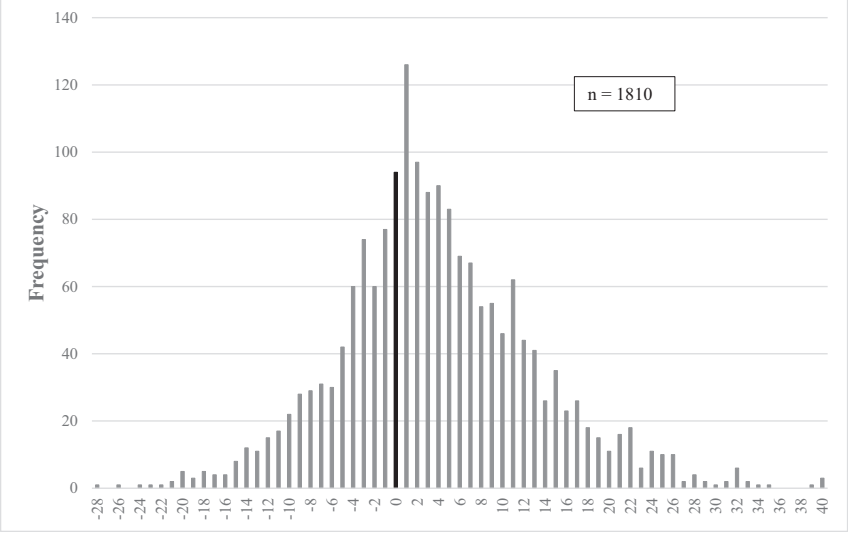

Fig. 1. Distribution of paired site-based and site-independent PANSS scoring deviations
NOTE: Figure displays scoring deviations (discordance) between paired site-based and site-independent total PANSS scores. Positive scores indicate that paired site-based scores are higher whereas negative scores indicate that site-independent scores are higher.

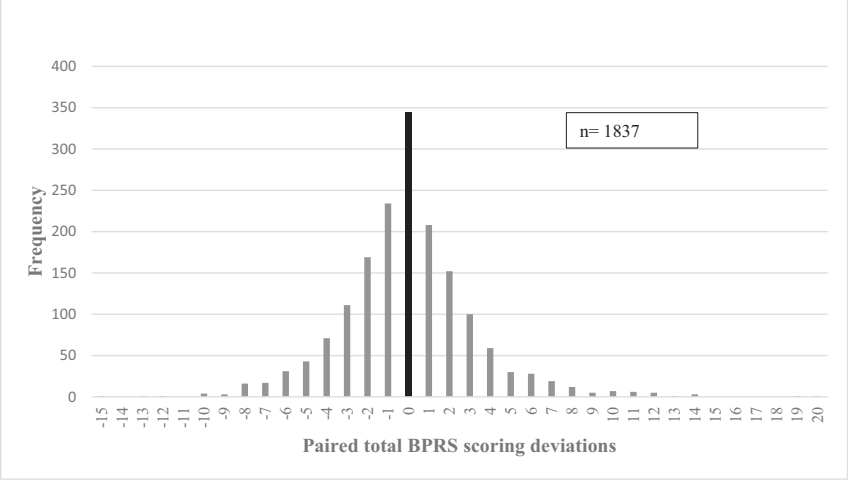

Fig. 2. Distribution of paired site-based and site-independent BPRS scoring deviations
NOTE: Figure displays scoring deviations (discordance) between paired site-based and site-independent total BPRS scores. Positive scores indicate that paired site-based scores are higher whereas negative scores indicate that site-independent scores are higher.

PANSS interviews and 1837 BPRS interviews yielded highly reliable paired scores (r = 0.801 and 0.897 respectively). Overall, 93.9% of paired PANSS and BPRS scores were within the 95% confidence intervals calculated by Bland-Altman scatterplots (Figs. 3 and 4). The high correlation and limits of agreement found been site-based and paired, blinded site-independent scores based on audio-digital recordings is consistent with previously published analyses of depression rating instruments in major depressive disorder (Targum and Catania, 2019). Paired scoring reliability was observed across all visits. Further, a subgroup of paired PANSS scores (n = 171) collected at the baseline and study endpoint visits yielded a high predictive value (84.2%) in which the site-independent scores matched the response/nonresponse site-based treatment outcomes.

The high correlation and limits of agreement observed in each of these 5 different schizophrenia studies are consistent and affirm the utility of audio-digital recording of both the PANSS and BPRS interviews as a quality assurance (surveillance) method to assess and optimize site-based ratings reliability. Ratings reliability is contingent upon competent site-based interviews, particularly with psychotic subjects. The audio-digital surveillance strategy reinforces sustained, competent ratings performance because raters are aware that their recorded interviews are subject to independent monitoring. Nonetheless, some raters in this analysis conducted shorter, incomplete interviews, obtained insufficient information, or failed to correctly apply ratings conventions. We found that rater remediation of rater "outliers" improved ratings performance on subsequent interviews in most instances.

As shown in Table 1, the screen and baseline visits were the most likely visits to reveal positive paired scoring deviations. Post-randomization visits revealed substantially fewer paired scoring deviations. It is possible that the positive paired scoring deviations uncovered some site-based score inflation that aimed to meet study eligibility severity criteria. However, the primary driver of the paired scoring deviations across all 5 studies was symptom severity (Table 2 and Figs. 5 and 6).

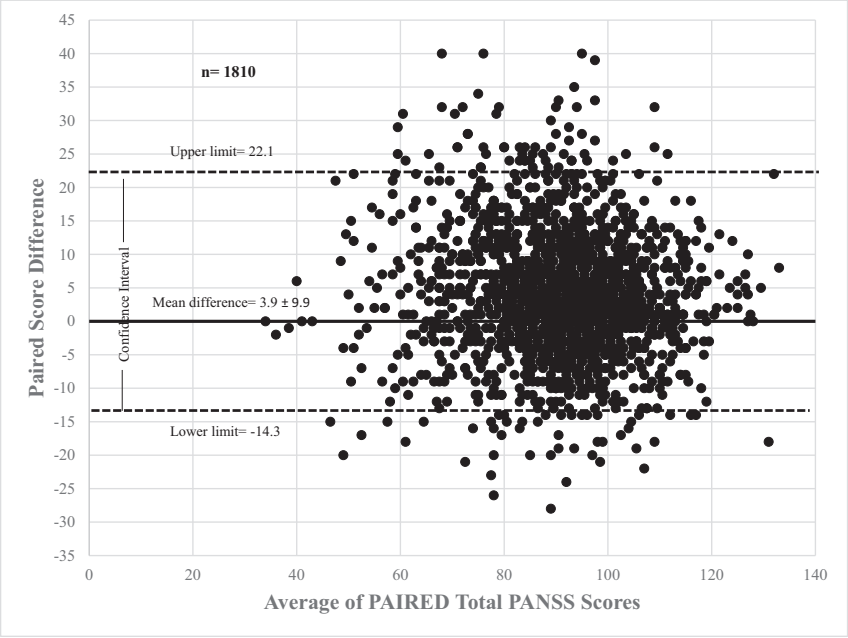

Fig. 3. Bland Altman Scatterplot for paired total PANSS scores.

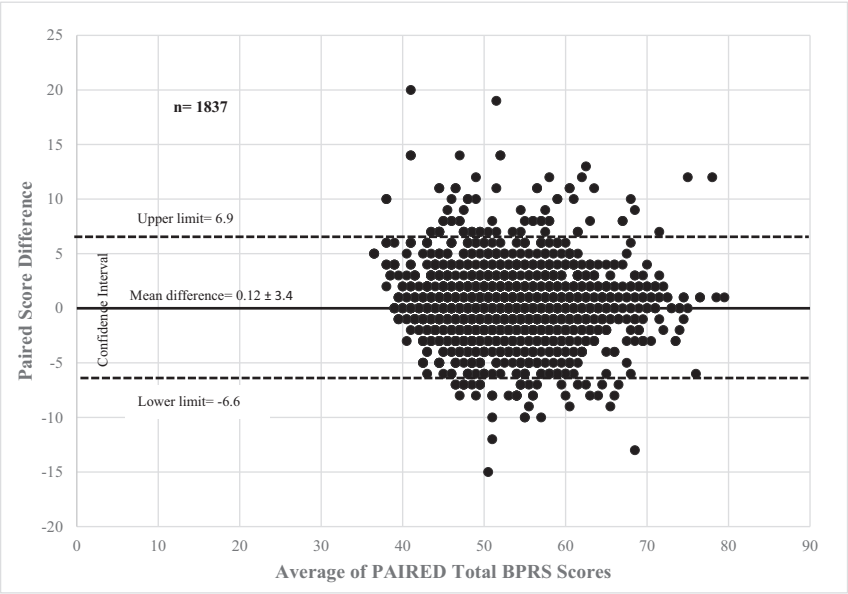

Fig. 4. Bland Altman Scatterplot for paired total BPRS scores.

There was a significant positive correlation between the site-based PANSS and BPRS scores and paired scoring deviations (r = 0.246 and r = 0.176 respectively). We found that the magnitude (severity) of the total score influenced the directionality of the paired scoring deviations. Site-based total PANSS or BPRS scores were often higher than the paired site-independent scores for the most symptomatic (severely ill) subjects. Alternatively, the lower (less severe) site-based total PANSS or BPRS scores were often lower than the paired site-independent scores and yielded greater negative deviations (Figs. 5 and 6). The finding of bi-directionality is consistent with the analyses of the MADRS we previously reported in 5 clinical studies of major depressive disorder and appears to be an inherent characteristic of this method of ascertainment (Targum and Catania, 2019). The observed paired scoring differences may be due to a non-quantifiable clinical nuance that is possible during a live interview that cannot be matched by simply listening to an audio recording of the same interview. It is also possible that blinded, independent raters listening to an audio recording are less likely to use the extreme scoring ranges of an instrument than a live rater who observes the subject. Although site-independent raters do use a 1 or 6 on the PANSS in many cases, they are often a point higher on the low side in symptomatically improved subjects and a point lower on the high side in the most severely ill subjects who they can only hear but not observe.

There are some limitations in this report that must be noted. First, the majority of the paired scores examined in this analysis were derived from clinical trial sites in the United States. Although the correlations, Bland-Altman plot, and directional trends noted in the U.S. pairs was similar in the Eastern European sample of Study 01, we cannot assert that these findings are universal and would be present in all regions in

Table 1
Comparison of site-based and site-independent PANSS and BPRS scores per visit.

	n	Mean site-based score (±SD)	Mean site-independent score (±SD)	Difference	t	p
Study 01	**PANSS**					
Screen visit	433	96.7 ± 14.1	91.6 ± 16.0	5.1	14.33	<0.0001
Baseline visit	347	93.7 ± 14.6	85.6 ± 15.5	8.1	16.70	< 0.0001
week 1	162	91.8 ± 15.0	86.4 ± 14.8	5.3	3.22	<0.002
week 2	92	89.4 ± 15.3	86.2 ± 15.4	3.2	1.43	0.15
week 3	87	85.1 ± 12.0	83.0 ± 14.4	2.1	1.06	ns
week 4	73	82.7 ± 14.3	81.0 ± 16.7	1.7	0.66	ns
week 5	33	91.7 ± 16.7	92.6 ± 16.7	−0.8	−0.20	ns
week 6/ET	30	84.7 ± 14.1	84.8 ± 18.4	−0.1	−0.06	ns
All visits	1257	92.7 ± 15.0	87.5 ± 16.0	5.1	8.31	<0.0001
Study 02	**PANSS**					
Screen visit	215	95.7 ± 8.6	95.3 ± 9.4	0.4	0.98	ns
Baseline visit	169	97.2 ± 8.9	95.7 ± 10.0	1.5	1.44	0.15
week 2	10	86.2 ± 13.4	86.1 ± 11.7	0.1	0.02	ns
week 4	9	91.1 ± 15.0	90.6 ± 12.2	0.6	0.09	ns
week 5/ET	150	85.1 ± 16.7	83.6 ± 13.1	1.5	0.85	ns
All visits	553	93.0 ± 12.7	92.0 ± 12.0	1.0	2.91	<0.004
Study 03	**BPRS**					
Screen visit	362	53.9 ± 6.1	51.9 ± 7.2	2.0	4.16	<0.0001
Baseline visit	154	50.7 ± 7.2	49.1 ± 7.5	1.6	1.86	0.06
All visits	516	50.6 ± 6.1	49.2 ± 6.9	1.4	3.02	<0.003
Study 04	**BPRS**					
Screen visit	905	54.8 ± 7.2	55.4 ± 7.0	−0.6	−1.60	0.11
Study 05	**BPRS**					
Screen visit	416	53.4 ± 7.6	53.4 ± 7.4	0.0	0.00	ns

Table 2
Effect of total PANSS and BPRS severity scores on paired scoring deviations.

	n	Mean site-based score ± SD	Mean site-independent score ± SD	Difference	t	p
Study 01	**PANSS**					
≥110	146	116.7 ± 6.4	108.1 ± 10.8	8.6	8.26	<0.0001
100–109	255	104.4 ± 2.9	97.8 ± 9.0	6.6	11.1	<0.0001
90–99	357	94.4 ± 2.9	89.8 ± 8.7	4.6	9.46	<0.0001
80–89	290	85.0 ± 2.9	80.6 ± 9.8	4.4	7.44	<0.0001
70–79	128	74.7 ± 2.8	70.9 ± 10.8	3.8	3.83	<0.0002
<70	81	60.7 ± 8.9	59.3 ± 11.5	1.4	0.86	ns
All visits	1257	92.7 ± 15.0	87.5 ± 16.0	5.1	8.31	<0.0001
Study 02	**PANSS**					
≥110	34	114.4 ± 3.6	109.5 ± 10.6	4.9	2.83	<0.01
100–109	136	103.7 ± 2.6	98.7 ± 8.1	5.0	6.84	<0.0001
90–99	205	94.6 ± 2.9	93.4 ± 7.6	1.2	2.03	<0.05
80–89	123	84.9 ± 2.8	85.8 ± 7.7	−0.9	−1.21	ns
70–79	22	75.0 ± 3.1	80.4 ± 7.7	−5.3	−3.00	<0.005
<70	33	59.8 ± 7.3	68.2 ± 9.8	−8.5	−3.97	<0.0002
All visits	553	93.0 ± 12.7	92.0 ± 12.0	1.0	2.91	<0.004
Study 03	**BPRS**					
≥70	7	73.9 ± 3.6	67.7 ± 5.0	6.1	2.65	<0.02
60–69	67	62.2 ± 1.9	59.3 ± 4.5	2.9	4.83	<0.0001
50–59	185	53.6 ± 2.8	51.4 ± 5.6	2.1	4.59	<0.0001
<50	257	44.9 ± 2.9	44.6 ± 4.5	0.3	0.92	ns
All visits	516	50.6 ± 7.2	49.2 ± 7.4	1.4	3.02	0.003
Study 04	**BPRS**					
≥70	38	72.2 ± 2.2	71.1 ± 3.3	1.1	1.66	ns
60–69	172	63.4 ± 2.7	63.1 ± 3.6	0.4	1.08	ns
50–59	487	54.2 ± 2.8	54.9 ± 3.5	−0.7	−3.66	<0.0003
<50	208	46.0 ± 2.4	47.1 ± 3.5	−1.1	−3.8	<0.0002
All visits	905	54.8 ± 7.2	55.4 ± 7.0	−0.6	−1.6	ns
Study 05	**BPRS**					
≥70	11	74.0 ± 4.2	72.0 ± 4.2	2.0	1.17	ns
60–69	76	63.0 ± 2.7	61.5 ± 4.6	1.5	2.49	<0.02
50–59	201	54.0 ± 2.9	54.1 ± 4.3	−0.1	−0.3	ns
<50	128	45.1 ± 2.5	46.0 ± 3.3	−0.9	−2.43	<0.02
All visits	416	53.4 ± 7.6	53.4 ± 7.4	0.0	0	ns

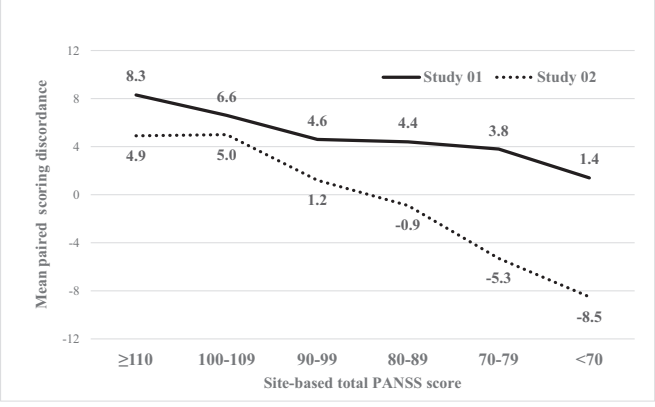

Fig. 5. Effect of PANSS score severity on paired scoring deviations
NOTE: Mean scoring deviations (discordance) between site-based and site-independent total PANSS scores. Positive mean scores indicate that the site-based scores are higher whereas negative scores indicate that independent scores are higher: n = 1257 (Study 01) and 553 (Study 02).

a multinational study. Second, the paired scores were derived from audio recordings and digital corroborative rater notes but without video observation. The scores from the 4 PANSS items that require direct observation of subject behavior during the interview were carried over from the site-based score. Consequently only 26 of the 30 PANSS items were truly site-independent scores. From a quality assurance perspective, this slight difference is not a material issue and the use of audio without video is clearly less invasive for the subject and less expensive to conduct.

In summary, the current analysis of 3647 paired PANSS and BPRS scores from 5 clinical studies of schizophrenia affirms the utility of audio-digital recording and site-independent scoring of site-based interviews as a strategy for quality assurance of ratings performance. The use of site-independent ratings as a primary measure beyond its utility for quality assurance was not investigated in this analysis and still needs further exploration. The audio-digital surveillance method can reinforce ratings reliability and allay concerns about deceptive ratings practices (Lee et al., 2018). In addition, the high predictive value of blinded site-independent ratings to replicate site-based treatment outcomes may be useful to affirm primary site-based results when there is a potential of functional unblinding.

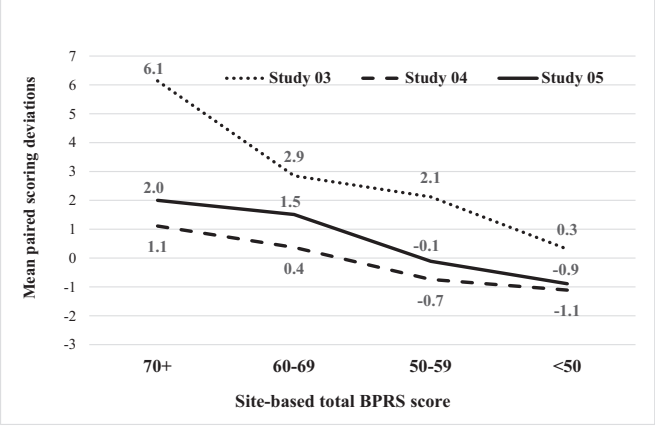

Fig. 6. Effect of BPRS score severity on paired scoring deviations
NOTE: Mean scoring deviations (discordance) between site-based and site-independent total BPRS scores. Positive mean scores indicate that the site-based scores are higher whereas negative scores indicate that independent scores are higher: n = 516 (Study 03), 905 (Study 04), 416 (Study 05).

Role of the funding source

Support for the analyses of the paired ratings data came from Signant Health. The data for each study was obtained from vendor grants to conduct quality assurance programs during clinical trials sponsored by Alkermes, Karuna Therapeutics, and Intra-Cellular Therapies Inc. Neither Signant Health nor any sponsor had any role in the analysis and/or interpretation of the data in this report, the writing of this report, or the decision to submit the manuscript in its current form.

CRediT authorship contribution statement

Dr. Targum participated in the design, implementation, and analysis of the original studies and conceived, analyzed, and wrote the current analysis reported in this manuscript. Dr.'s Pendergrass and Murphy assisted with the execution of these studies, supervision of site-independent raters, collection, collation, and analysis of the data and reviewed and approved the final manuscript.

Declaration of competing interest

Dr. Targum is an employee of Signant Health and has received vendor grants or consulted with Acadia Pharmaceuticals, Alkermes Inc., BioXcel, EMA Wellness LLC., Denovo Biopharma, Epiodyne, Frequency Therapeutics, Functional Neuromodulation, Intra-Cellular Therapies, Johnson and Johnson PRD, Karuna Therapeutics, Merck Inc., Methylation Sciences Inc., Navitor Pharmaceuticals Inc., Neurocrine Biosciences Inc., Pfizer Inc., and Sunovion Inc. during the past 3 years.

J. Cara Pendergrass is currently employed by William James College (Newton, Massachusetts) but was an employee of Clintara LLC and Signant Health when these studies were conducted and has no other disclosures. Dr. Murphy is an employee of Signant Health.

Acknowledgement

None.

References

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). 4th edition. American Psychiatric Association, Washington DC.
American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders. Fifth edition. American Psychiatric Press, Arlington VA.
Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.
Correll, C.U., Kishimoto, T., Nielsen, J., Kane, J.M., 2011. Quantifying clinical relevance in the treatment of schizophrenia. Clin. Ther. 33 (12), B16–B39.
Crippa, J.A., Sanches, R.F., Hallak, J.E., Loureiro, S.R., Zuardi, A.W., 2001. A structured interview guide increases Brief Psychiatric Rating Scale reliability in raters with low clinical experience. Acta Psychiatr. Scand. 103 (6), 465–470.
First, M.B., Williams, J.B.W., Spitzer, R.L., Gibbon, M., 2007. Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Clinical Trials Version (SCID-CT). Biometrics Research. New York State Psychiatric Institute, New York.
Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13 (2), 261–276.
Kobak, K.A., Engelhardt, N., Williams, J.B., Lipsitz, J.D., 2004. Rater training in multicenter clinical trials: issues and recommendations. J. Clin. Psychopharmacol. 24 (2), 113–117.
Kobak, K.A., Feiger, A.D., Lipsitz, J.D., 2005. Interview quality and signal detection in clinical trials. Am. J. Psychiatr. 162 (3), 628.
Kobak, K.A., Kane, J.M., Thase, M.E., Nierenberg, A.A., 2007. Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms? J. Clin. Psychopharmacol. 27, 1–5.
Lee, C.P., Holmes, T., Neri, E., Kushida, C.A., 2018. Deception in clinical trials and its impact on recruitment and adherence of study participants. Contemp. Clin. Trials https://doi.org/10.1016/j.cct.2018.08.002.
Leucht, S., Davis, J.M., Engel, R.R., Kane, J.M., Wagenpfeil, S., 2007. Defining 'response' in antipsychotic drug trials: recommendations for the use of scale-derived cutoffs. Neuropsychopharmacology 32, 1903–1910.
Muller, M.J., Szegedi, A., 2002. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. J. Clin. Psychopharmacol. 22, 318–325.
Muller, M.J., Wetzel, H., 1998. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. Acta Psychiatr. Scand. 98 (2), 135–139.
Mulsant, B.H., Kastango, K.B., Rosen, J., Stone, R.A., Mazumdar, S., Pollock, B.G., 2002. Interrater reliability in clinical trials of depressive disorders. Am. J. Psychiatry 159, 1598–1600.
Overall, J.E., Gorham, D.R., 1988. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling. Psychopharmacol. Bull. 24, 97–98.
Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiat. 59 (20), 22–33.
Targum, S.D., 2006. Evaluating rater competency for CNS clinical trials. J. Clin Psychopharm. 26 (3), 308–310.
Targum, S.D., Catania, C., 2019. Audio-digital recordings for surveillance in clinical trials of major depressive disorder. Contemp. Clin. Trials Commun. https://doi.org/10.1016/j.conctc.2019.100317.
Targum, S.D., Pendergrass, J.C., 2014. Site-independent confirmation of subject selection for CNS trials: "dual" review using audio-digital recordings. Annals Gen. Psychiatry 13, 21.
Targum, S.D., Pendergrass, J.C., Toner, C., Asgharneshad, M., Burch, D.J., 2014. Audio-digital recordings used for independent confirmation of site-based MADRS interview scores. Eur. Neuropsychopharmacol. 24, 1760–1766.
Targum, S.D., Pendergrass, J.C., Toner, C., Zumpano, L., Rauh, P., DeMartinis, N., 2015. Impact of interview length on ratings reliability in a schizophrenia trial. Eur. Neuropsychopharmacol. 25 (3), 312–318.
Targum, S.D., Brannan, S., Murphy, C., Daniel, D., Breier, A., 2020. Site ratings versus site-independent ratings of PANSS interviews in a schizophrenia study. Presented at the Annual Meeting of ISCTM. September 23, 2020.
Ventura, J., Lukoff, D., Nuechterlein, K.H., Liberman, R.P., Green, M., Shaner, A., 1993. Appendix 1: Brief Psychiatric Rating Scale (BPRS) expanded version (4.0) scales, anchor points and administration manual. Int. J. Methods Psychiatr. Res. 3, 227–244.
West, M.D., Daniel, D.G., Opler, M., Wise-Rankovic, A., Kalali, A., 2014. Consensus recommendations on rater training and certification. Innov. Clin. Neurosci. 11 (11—12), 10–13.
Zarate Jr., C.A., Weinstock, L., Cukor, P., Morabito, C., Leahy, L., Burns, C., Baer, L., 1997. Applicability of telemedicine for assessing patients with schizophrenia: acceptance and reliability. J. Clin. Psychiatry 58, 22–25.