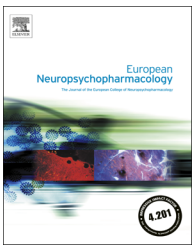
European Neuropsychopharmacology (2015) 25, 312–318



www.elsevier.com/locate/euroneuro

Impact of BPRS interview length on ratings reliability in a schizophrenia trial

Steven D. Targum[a,*], J. Cara Pendergrass[a], Chelsea Toner[a], Laura Zumpano[b], Philip Rauh[a], Nicolas DeMartinis[b]

[a]*Clintara, LLC, Boston, MA, United States*
[b]*Pfizer Inc., Cambridge, MA, United States*

Received 30 August 2014; received in revised form 26 October 2014; accepted 26 November 2014

KEYWORDS
Clinical trials;
BPRS;
Interview length;
Schizophrenia;
Ratings precision;
Inter-rater reliability

Abstract
Signal detection in clinical trials relies on ratings reliability. We conducted a reliability analysis of site-independent rater scores derived from audio-digital recordings of site-based rater interviews of the structured Brief Psychiatric Rating Scale (BPRS) in a schizophrenia study. "Dual" ratings assessments were conducted as part of a quality assurance program in a 12-week, double-blind, parallel-group study of PF-02545920 compared to placebo in patients with sub-optimally controlled symptoms of schizophrenia (ClinicalTrials.gov identifier NCT01939548). Blinded, site-independent raters scored the recorded site-based BPRS interviews that were administered in relatively stable patients during two visits prior to the randomization visit. We analyzed the impact of BPRS interview length on "dual" scoring variance and discordance between trained and certified site-based raters and the paired scores of the independent raters.
Mean total BPRS scores for 392 interviews conducted at the screen and stabilization visits were 50.4±7.2 (SD) for site-based raters and 49.2±7.2 for site-independent raters (t=2.34; p=0.025). "Dual" rated total BPRS scores were highly correlated (r=0.812). Mean BPRS interview length was 21:05±7:47 min ranging from 7 to 59 min. 89 interviews (23%) were conducted in less than 15 min. These shorter interviews had significantly greater "dual" scoring variability (p=0.0016) and absolute discordance (p=0.0037) between site-based and site-independent raters than longer interviews.
In-study ratings reliability cannot be guaranteed by pre-study rater certification. Our findings reveal marked variability of BPRS interview length and that shorter interviews are often incomplete yielding greater "dual" scoring discordance that may affect ratings precision.

1. Introduction

Randomized clinical trials conducted for central nervous system (CNS) indications depend on scoring reliability and ratings consistency to evaluate the efficacy of a candidate

*Correspondence to: 11 Beacon Street Suite 600 Boston MA 02108. Tel.: +1 617 824 0800.
E-mail address: sdtargum@yahoo.com (S.D. Targum).

http://dx.doi.org/10.1016/j.euroneuro.2014.11.023
0924-977X

drug (Kobak et al., 2005; Targum, 2006; Targum et al., 2013). Although the research instruments used to score CNS symptoms are usually structured and standardized and the raters are trained and certified beforehand, each interview is likely to differ because of the nature of the disease under study. In schizophrenia studies, patients may be more or less informative and/or cooperative from visit to visit such that raters may be challenged to conduct complete and reliable interviews. The core features of the illness itself (suspiciousness, negative symptoms, conceptual disorganization) may complicate the reliability and consistency of ratings. Beyond patient-related factors, the competency of the clinician-rater to conduct research interviews may also affect ratings reliability (Kobak et al., 2005; Targum, 2006).

It is known that inter-rater reliability directly affects the power of the test to achieve signal detection (Muller and Szegedi, 2002). Despite the best efforts of pre-study rater training and inter-rater reliability assessments, the quality of in-study interviews still varies from rater to rater and may also vary from visit to visit. It is conceivable that the pressure to conduct a complete research interview during a busy clinic day may result in shorter interviews that might compromise ratings reliability. The same rater may conduct some good and some less good interviews based upon the daily circumstances. Recently, surveillance strategies using site-independent raters have been used in clinical trials in an attempt to assure in-study ratings quality and enhance precision throughout the trial (Shen et al., 2008; Schoemaker et al., 2009; Sharp et al., 2011; Targum and Pendergrass, 2014; Targum et al., 2014). Site-independent scoring, including video-conferencing and secondary telephone interviews have been used in the assessment of patients with schizophrenia as well (Zarate et al., 1997; Shen et al., 2008; Schoemaker et al., 2009; Sharp et al., 2011; Targum et al., 2012). Each surveillance strategy has potential benefits and inherent limitations (Sharp et al., 2011; Targum and Pendergrass, 2014). Given the nature of the illness of schizophrenia and the inherent complexity of research study visits, it would be desirable to minimize the extra burden of a second, independent interview in patients with schizophrenia.

We have explored the utility of audio-digital recording of site-based interviews in randomized clinical trials (Targum et al., 2012; Targum and Pendergrass, 2014; Targum et al., 2014). A review of the primary, site-based rater's interview by a paired independent rating of the same interview ("dual" scoring) can be a reasonable metric to examine ratings reliability (precision), to explore "dual" scoring discordance, and to identify individual site-based rater's who reveal excessive scoring discordance ("outlier's"). In the current study, we assessed ratings reliability of the structured Brief Psychiatric Rating Scale (BPRS) in the pre-randomization phase of a schizophrenia trial using the audio-digital recording method (Overall and Gorham, 1988; Targum and Pendergrass, 2014). The recording method allowed us to examine the impact of BPRS interview length as it related to interviewing competency and scoring concordance. Our findings reveal that independent scoring based upon audio-digital recordings can usually confirm the scores from most BPRS interviews but that some shorter interviews are insufficient and generate significantly greater "dual" scoring discordance. The in-study identification of poor ratings performance and early remediation of rater "outliers" may enhance ratings precision for a clinical trial.

2. Experimental methods

This study was conducted as part of a quality assurance (QA) surveillance strategy in patients with relatively stable symptoms of schizophrenia. Ratings reliability was evaluated during a 12-week, double blind, parallel-group study of PF-02545920 compared to placebo in patients with sub-optimally controlled symptoms of schizophrenia (ClinicalTrials.gov identifier NCT01939548). The study was conducted in compliance with the informed consent regulations of an Institutional Review Board (IRB) and International Conference on Harmonization (ICH) for Good Clinical Practice (GCP) Guidelines at 26 clinical trial sites in the United States. All study eligible patients needed to meet DSM-IV criteria for schizophrenia based upon diagnostic confirmation with the Mini-International Psychiatric Interview (M.I.N.I.) and have a minimum total Brief Psychiatric Rating Scale (BPRS) score ≥ 39 at the screening, stabilization, and randomization visits (Lukoff et al., 1986; Overall and Gorham, 1988; Sheehan et al., 1998). The BPRS score was used to confirm symptom stability between visits and was part of the eligibility criteria for subject selection prior to randomization (visit 3) for this particular study. The stabilization visit (visit 2) was conducted a minimum of 14 days after the screen visit to re-assess BPRS scores and confirm relative symptom stability between these visits. The BPRS was chosen as the symptom severity measure to ascertain subject eligibility in order to un-couple the eligibility criteria from the primary efficacy measure. The Positive and Negative Syndrome scale (PANSS) was the primary efficacy measure and was not administered until the randomization visit (Kay et al., 1987). Consequently, we used the BPRS scores to examine symptom stability and "dual" ratings precision prior to randomization.

All raters participated in a comprehensive rater training and certification program that was conducted at the study initiation for all rating instruments, and included the BPRS and PANSS (Overall and Gorham, 1988; Kay et al., 1987). All qualified site-based and site-independent raters had at least two years experience with the BPRS and primary measure (PANSS) and needed to demonstrate scoring accuracy on sample video interviews of the BPRS and PANSS (inter-rater reliability) as well as interviewing competency via mock interviews. The rater-training program used structured versions of the BPRS (version 4.0) and PANSS (Lukoff et al., 1986; Kay et al., 1987; Ventura et al., 1993; Crippa et al., 2001). Light's weighted kappa scores demonstrated inter-rater reliability for the BPRS ratings (κ=0.809) and PANSS (κ=0.615) amongst all raters.

Site-based raters were also trained to use audio-digital recording pens for the BPRS interview (Targum and Pendergrass, 2014). All patients consented to audio-digital recording of site-based BPRS interviews as part of their consent to participate in the study. The audio-digital pens simultaneously record the site-based interview and digitally capture accompanying written notes on specially manufactured source books. The recorded BPRS interviews were electronically forwarded to Clintara LLC (Boston, MA) for random assignment to the site-independent reviewers. The

site-independent raters met the same ratings qualification standards as the site-based raters. Site-independent raters were blinded to the study visit and trial site and generated their "dual" scores by listening to the audio recording and reading the site-based rater's corroborative digital notes. Independent raters were blinded to the site-based rater's scores because the digital notes sent to the blinded rater did not include these scores.

As part of the QA surveillance program, we examined BPRS scoring variance and absolute discordance between the "dual" site-based rater scores and the paired site-independent rater scores. Absolute discordance reflects the deviation of site-independent scores from the paired site-based BPRS scores in either direction. These "dual" scoring analyses were performed in "real-time" on 100% of the screen and stabilization visits in order to identify site-based rater's who revealed excessive "dual" scoring deviations ("outliers") early in the study. Rater remediation was initiated when a second, independent rater confirmed the discordant total BPRS scores. Remediation included telephone adjudication to discuss the specific score differences, individual item deviations, and interview style. These raters were followed to ascertain their subsequent performance and were subject to removal from the study if it did not improve. In addition to these analyses, we timed the length of the full BPRS interview conducted by site-based raters in order to determine whether length had an impact on the quality of interviews and "dual" scoring concordance.

We examined visits 1 and 2 separately and as a pooled group of paired interviews independent of the rater who did the interview or that actual patient assessed. Statistical analysis of the data included ICC, χ^2, Student's *t* test, kappa statistics, and one-way analysis of variance (ANOVA) using the online VassarStats program.

3. Results

3.1. Site-based total BPRS scores

Forty-three site-based raters recorded 228 BPRS screen visits and 164 BPRS stabilization visits ($n=392$). The mean symptom severity scores as measured by the total BPRS were relatively stable without any statistically significant difference between the screen and stabilization visits (Table 1). The mean site-based total BPRS score was 50.4 ± 7.2 (SD) at the screen visit and 50.5 ± 7.0 (SD) at the stabilization visit ($t=0.129$; $p=0.90$). The scores ranged from 39 to 81 (screen) and 39 to 75 (stabilization visit). The pooled total BPRS for both visits combined was 50.4 ± 7.2 (SD) for all 392 paired interviews available for this analysis.

3.2. Comparison of site-based and site-independent total BPRS scores

We compared the total BPRS "dual" scores between site-based and site-independent raters (Table 1). The paired "dual" scores were highly correlated at both visits ($r=0.812$ for both visits combined; $t=35.31$; $p<0.0001$). Similar to the stability of the site-based scores, the "dual" site-independent scores were also stable between visits 1 and 2 ($t=0.13$; $p=0.90$). However, the site-independent raters mean total BPRS scores were significantly lower than site-based raters at each visit ($t=2.24$; $p=0.025$ for both visits combined).

There were individual scoring differences (deviations) between the site-based and paired site-independent total BPRS scores. The site-based scores deviated between 14 points lower and 20 points higher than the site-independent "dual" scores (Figure 1).

3.3. BPRS interview length

The audio-digital recording method made it possible to time the length of each BPRS interview. The interviews ranged in length from 7:25 to 59:44 min across the two visits. The mean length of all 392 interviews combined was $21:05\pm7:47$ min (median interview length=21 min). There was no significant difference in interview length between the two interviews. The mean screen visit interview length (v1) was $20:34\pm7:41$ min and the mean stabilization visit interview length (v2) was $21:50\pm7:52$ min ($t=-1.58$; $p=0.11$).

Eighty nine BPRS interviews (23%) were conducted in less than 15 min whereas the other 303 interviews (77%) took at least 15 min to complete. The shorter interviews had significantly lower mean total BPRS scores than the longer site-based interviews that took at least 15 min ($t=-2.34$; $p=0.020$). Alternatively, as shown in Table 2, the mean site-based BPRS scores from the shorter interviews were actually significantly higher than the "paired" site-independent scores ($t=2.71$; $p=0.007$).

There was a modest, but significant correlation (ICC) between interview length and symptom severity as measured by the total BPRS at these two visits ($r=0.253$; $t=6.14$; $p<0.0001$). This correlation was present in the 303 interviews that took at least 15 min ($r=0.263$; $t=4.56$; $p<0.0001$) but not in the shorter interviews ($r=0.057$; $t=0.4$; $p=0.690$).

Table 1 Comparison of total BPRS "dual" scores between site-based and site-independent raters.

	n	Site-based BPRS	Site-independent BPRS	*r*	*t*	*p*
Screen (visit 1)	228	50.4 ± 7.2	49.5 ± 7.1	0.809	1.34	0.182
Stabilization (visit 2)[a]	164	50.5 ± 7.0	49.0 ± 7.3	0.82	1.88	0.061
All BPRS interviews	392	50.4 ± 7.2	49.2 ± 7.2	0.812	2.24	0.025

[a]A stabilization visit was conducted a minimum of 14 days after the screen visit to assess and confirm symptom stability between visits as part of the study eligibility criteria.

3.4. Impact of interview length on "dual" scoring variance and discordance

The least "dual" scoring variance between site-based and site-independent BPRS scores occurred in interviews that were conducted between 20 and 24 min (Figure 2). The variance increased as the interviews became shorter or longer.

Table 3 displays the mean BPRS scoring variance, mean absolute discordance, and discordance rates between the "dual" BPRS scores at the screen and stabilization visits.

We established a cut-off range of ≥ 8 BPRS points as an arbitrary discordance range for "dual" scoring deviations based upon >1SD of the mean total BPRS score (± 7.2) for all BPRS interviews. The cut-off range is useful to identify

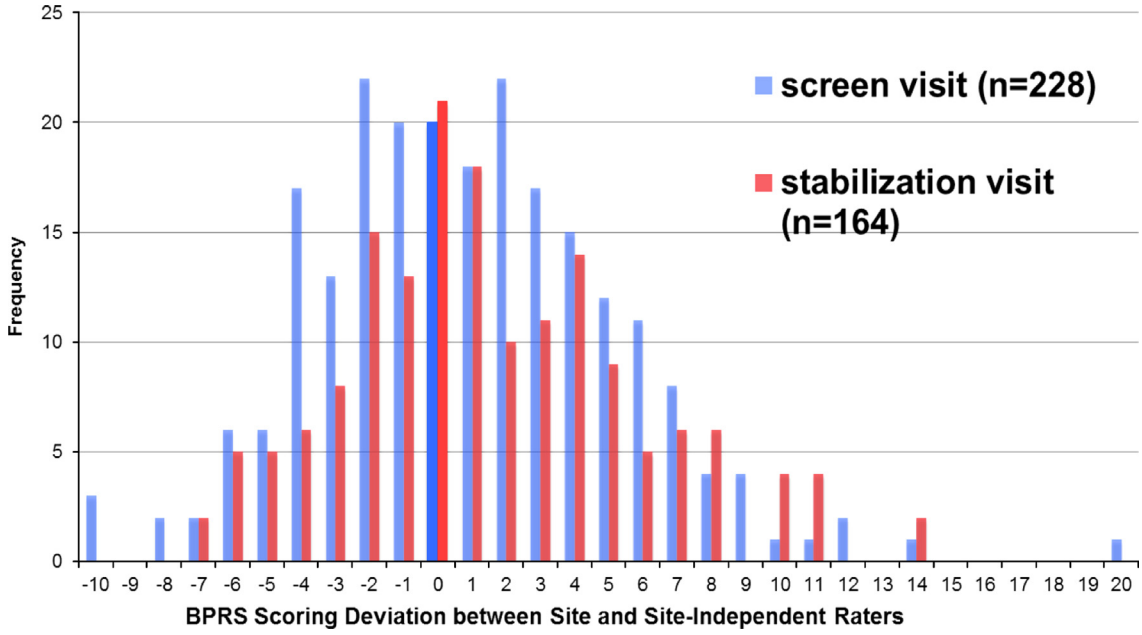


Figure 1 Total BPRS scoring deviations between site-based and site-independent raters*.

Table 2 BPRS interview length and ratings precision: impact of "short" interviews.

	n	Length (minutes)	Site-based BPRS	Site-independent BPRS	*t*	*p*
All BPRS interviews[a]	392	21.1±7.8	50.4±7.2	49.2±7.2	2.24	0.025
Interview length						
<15 min	89	11.7±2.0	48.8±6.1	46.4±5.8	2.71	0.007
≥15 min	303	23.4±6.9	50.8±7.4	50.1±7.4	1.29	0.200

[a]Includes all BPRS visits from the screen and stabilization visits.

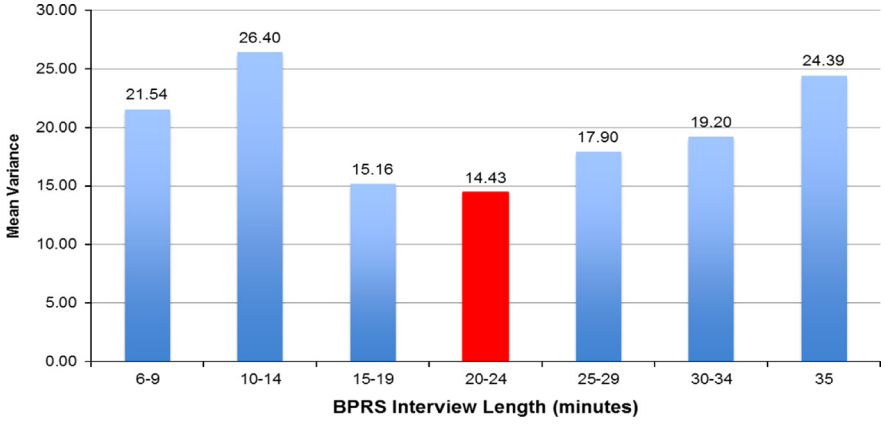


Figure 2 "Dual" scoring variance relative to BPRS interview length ($n=392$).

raters who might need ratings remediation. Using this range, 35 of the 392 interviews (8.9%) were discordant and were evenly distributed between the two visits (Table 3). Fifteen of these discordant "dual" scores (42.8%) were conducted in less than 15 min ("short" interviews). Shorter interviews were significantly more likely to be discordant than interviews that took at least 15 min to conduct ($\chi^2=8.89$; df=1; $p=0.003$).

There was a significantly greater "dual" scoring variance ($p=0.0016$) and absolute mean scoring discordance ($p=0.0037$) amongst the "short" BPRS interviews that were conducted in less than 15 min compared to longer interviews (Table 4). In contrast to the "short" interviews, longer BPRS interviews that took more than 35 min to complete did not reveal significant scoring variance or absolute discordance.

Many shorter interviews (less than 15 min) were characterized by fewer questions asked and less information collected. Blinded raters indicated that they were often uncertain about the best "short" interview scores because of poor rater performance and not because of poor patient cooperation. On the other hand, longer interviews were mixed between difficult patients who were hard to examine (providing tangential or evasive responses) and inexperienced interviewers who took longer to get the necessary information.

3.5. Individual raters and "dual" scoring variability

At part of the QA surveillance program, the impact of BPRS interview length on "dual" scoring discordance was reviewed throughout the enrollment process. At the approximate midpoint (after 208 "dual" BPRS reviews), three of 43 site-based raters were identified who accounted for nearly 40% of the short interviews that had been conducted at the screen visit at that time (visit 1). Of note, these three "outlier" raters had more scoring discordance when they conducted "short" interviews than on their own longer interviews. The raters were remediated, one was removed from the study, and the overall performance of the other two improved.

In the final review of the 392 available BPRS interviews reported in this analysis, the "short" interviews were now distributed amongst all of the site-based raters and no individual clinician-rater stood out as an "outlier" rater.

4. Discussion

Audio-digital recordings of site-based BPRS interviews were used as part of a quality assurance surveillance program that allowed us to examine ratings reliability (precision) and its relationship to BPRS interview length. In this analysis, "dual" scoring of the site-based BPRS interviews by paired blinded, site-independent ratings revealed a high correlation ($r=0.812$) and minimal discordance at the screen and stabilization visits of a schizophrenia trial. This finding demonstrates that blinded, site-independent raters can confirm the total BPRS scores of most, but not all audio-digital recorded interviews.

Interview length had a significant impact on "dual" scoring concordance. Shorter interviews (<15 min) revealed significantly more absolute discordance than longer interviews ($p=0.0037$) and had significantly greater "dual" scoring variance ($p=0.0016$) as well.

In this study, the "sweet" spot for the least discordance were BPRS interviews that were conducted between 20 and 24 min. Obviously, this finding does not mean that all competent BPRS interviews must be conducted within 20–24 min. There are some interviews that require more time to complete and others that might be sufficient when conducted in less than 20 min. The key point is that a research interview must seek to obtain complete information at each visit and not rely on previous interviews or presumptions to achieve reliable scores. Although the shorter interviews yielded significantly less severe total BPRS scores than interviews that took at least 15 min ($p=0.020$), this difference cannot explain the variance observed between site-based and site-independent ratings of

Table 3 Variance and absolute discordance comparison of BPRS "dual" scores between site-based and site-independent raters.

	n	Mean variance of "Dual" scoring deviations	Mean absolute discordance	Discordance rate[a]
Screen (visit 1)	228	19.72±4.44	3.50±2.86	19 (8.3%)
Stabilization (visit 2)	164	18.53±4.31	3.50±2.86	16 (9.8%)

[a]The discordance rate reflects the number of paired "dual" scores at each visit that exceed one standard deviation from the mean total BPRS score (50.4±7.2 points).

Table 4 Impact of "short" BPRS interviews on scoring variance and absolute discordance.

	n	BPRS "dual" scoring variance Mean±SD	Absolute discordance Mean±SD
All BPRS interviews[a]	392	19.3±4.4	3.49±2.9
Interviews <15 min	89	26.2±5.1	4.27±3.7
Interviews ≥15 min	303	16.7±4.1	3.26±2.6
ANOVA "short" vs. longer interviews		$F=10.05$	$F=8.49$
		$p=0.0016$	$p=0.0037$

[a]Includes all BPRS visits from the screen and stabilization visits.

the same interview. The BPRS instrument used in this study was a structured version that provides prompt questions for each item and tends to increase inter-rater reliability (Lukoff et al., 1986; Ventura et al., 1993; Crippa et al., 2001). A review of many of the shorter interviews revealed that the interviewer asked fewer questions and thus less information (data) was collected. Therefore, it is understandable that blinded raters were not able to confirm the scores as well as they did with longer, more complete interviews. Some of the longer BPRS interviews (>35 min) revealed more discordance than the mid-range length interviews but did not yield significant "dual" scoring differences. Longer interviews were often a consequence of more difficult patients (tangentiality, evasiveness) or extra time taken by a less experienced rater. Other factors that might affect interview quality and ratings reliability such as rater experience with the BPRS, patient cooperation, the day of the interview, or even the time of the day were not systematically examined in this study.

The analysis in this study used the BPRS interviews that were conducted at two pre-randomization visits. In other schizophrenia studies, we have found that shorter PANSS interviews (less than 15 min) will also yield significantly greater absolute discordance and "dual" scoring variability than longer interviews between site-based and site-independent raters (Targum et al., 2014).

The audio-digital recording method used in this study serves as a constant reminder to site-based raters to conduct complete interviews. By turning on the recording pen the site-based raters obviously knew that they were being "observed" during the interview. It is reasonable to assume that the mere awareness of being observed would alter site-based ratings behavior in most raters (e.g. Hawthorne effect) and therefore assure data integrity (McCarney et al., 2007). Yet, despite the obvious fact of recording their own in-study BPRS interviews, some raters still conducted short, insufficient interviews. Clearly, pre-study rater training and certification did not guarantee in-study ratings performance. In fact, some site-based raters revealed greater "dual" scoring discordance when they did "short" interviews relative to their own longer interviews. Apparently, these raters were able to conduct scorable, complete interviews when they took the time to do it. Clearly, time is an issue in a busy research or clinic setting and may, in some instances compromise the quality (competency) of the interview. It is noteworthy that in-study rater remediation abetted subsequent ratings performance and these raters later acknowledged that some of their interviews had been rushed.

A peripheral, but vitally important issue emerges about the requisite length of a reliable rating instrument and the necessary length of a research study visit. Recently, some researchers have explored using shorter forms of rating instruments in an effort to sustain ratings reliability without the burden of long interviews (Levine, 2013; Levine and Leucht, 2013; Velthorst et al., 2013). Clearly, long interviews and long clinic visits are burdensome to both the patient and staff and can potentially saturate the patient such that data accuracy is compromised. Further, long study visits extend patient exposure to study staff and potentially introduce placebo effects related to familiarity and other non-drug benefits of study participation. Although a briefer interview and study visit is certainly desirable, our findings reinforce the need for sufficient interview length to obtain enough data to generate reliable ratings.

In summary, our findings in this study suggest that some, but not all "short" BPRS interviews are insufficient because they are incomplete. We have also found that shorter PANSS interviews yield greater "dual" scoring discordance between site-based and site-independent (Targum et al., 2014).

"Dual" scoring of site-based interviews using the audio-digital recording method is a relative straightforward way to reinforce ratings competency and a direct way to affirm that site-based raters have really conducted complete interviews.

Each surveillance strategy used in randomized clinical trials has advantages and limitations. Audio recording is a less intrusive method of surveillance than video observations and "dual" ratings of site-based interviews are definitely less burdensome than entirely separate, second interviews. However, a limitation of the audio-digital recording method is that the blinded, independent scores are entirely dependent on the quality of the site-based interview and the responses of the patient to the site-based interviewer. Further, "dual" concordance can provide scoring confirmation but cannot demonstrate that either the site-based or independent score is truly accurate. Of course, the scoring accuracy of psychiatric symptoms is an elusive target for any ratings method. Alternative surveillance methods, like separate telephone interviews or live video-fed interviews conducted by an independent rater offer a more distinct "second" opinion but have inherent limitations as well (Shen et al., 2008; Schoemaker et al., 2009; Targum et al., 2013). Separate ("second") telephone-based interviews are time consuming to arrange, add cost and burden to the study visits, require a greater level of patient cooperation and participation than a recorded first interview, and introduce temporal and informational variance about symptom severity that will likely generate different scores. Each of these surveillance methods, as well as patient-reported outcomes and site-based ratings represent different data "sourcing" that will influence the scores.

The current study examines ratings precision but does not examine the impact of "dual" ratings on treatment outcome. Additional studies that examine other rating instruments, the relevance of interview length and "dual" scoring concordance throughout a clinical trial, and the impact of "dual" ratings on trial outcome are needed. Meanwhile, the audio-digital recording method applied in this study is a relatively unobtrusive surveillance strategy whose use in multi-site randomized clinical trials may assure rater performance and enhance ratings precision.

Role of funding source

Partial support for this research came from Pfizer Inc who provided a vendor grant for audio-digital recording of study interviews.

Contributors

Steven D. Targum helped to design the "dual" ratings component of the study, assisted in the execution of the study, performed the data analysis and wrote the manuscript.

J. Cara Pendergrass helped to design the study and assisted with study execution and data analysis.

Chelsea Toner assisted in the study execution and research coordination and reviewed and approved the manuscript.

Laura Zumpano helped to design the study and participated in study execution and reviewed and approved the manuscript.

Philip Rauh assisted with study coordination and data analysis for this study and approved of the manuscript.

Nicholas DeMartinis designed the study, assisted with study execution, data analysis, and reviewed and approved the manuscript.

Conflict of interest

Disclosures

S.D. Targum: Acadia Pharmaceuticals, Acumen, Alcobra, Alkermes Inc., AstraZeneca, BioMarin, BrainCells Inc., CeNeRx Pharmaceuticals, Clintara, LLC, Civitas, Eli Lilly and Company, EnVivo (Forum) Pharmaceuticals, Euthymics, Functional Neuromodulation Inc, Johnson & Johnson PRD, Intracellular Therapies Inc., Methylation Sciences Inc., Mitsubishi Tanabe, Neurophage, Nupathe, Pfizer Inc., Prana Biotechnology Ltd., ReViva Pharmaceuticals, Roche Labs, Sophiris, Sunovion, Targacept, Theravance, Transcept.

J.C. Pendergrass, C. Toner, P. Rauh: are all employees of Clintara, LLC (Boston MA).

L. Zumpano, N. DeMartinis: employees of Pfizer Inc (Cambridge MA).

Acknowledgments

We would like to acknowledge Pfizer Inc (No. 8500345121/1400) who provided partial support for this research via a vendor grant to provide training and quality assurance services for a clinical trial that included “dual” ratings and the staff of Clintara LLC for the coordination of this project.

References

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.

Kay, S.R., Opler, L.A., Fiszbein, A., 1987. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276.

Kobak, K.A., Feiger, A.D., Lipsitz, J.D., 2005. Interview quality and signal detection in clinical trials. Am. J. Psychiatry 162 (3), 628.

Levine, S.Z., Leucht, S., 2013. Psychometric analysis in support of shortening the scale for the assessment of negative symptoms. Eur. Neuropsychopharmacol. 23 (9), 1051–1056.

Levine, S.Z., 2013. Evaluating the seven-item center for epidemiologic studies depression scale short-form: a longitudinal US community study. Soc. Psychiatry Psychiatr. Epidemiol. 48 (9), 1519–1526.

Lukoff, D., Nuechterlein, K.H., Ventura, J., 1986. Appendix A: manual for expanded Brief Psychiatric Rating Scale (BPRS). Schizophr. Bull. 12 (4), 594–602.

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., Fisher, P., 2007. The hawthorne effect: a randomised, controlled trial. BMC Med. Res. Methodol. 7, 30. http://dx.doi.org/10.1186/1471-2288-7-30.

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.

Overall, J.E., Gorham, D.R., 1988. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling. Psychopharmacol. Bull. 24, 97–98.

Schoemaker J., Gaur R., Chawla V., Jansen W., Szegedi A. Expert Rater Assisted Score Evaluation (ERASE): a new method to enhance signal detection in randomized, placebo-controlled clinical trials. In: Proceedings of the 49th Annual NCDEU, Hollywood, FL, June 29-July 2, 2009.

Sharp, I.R., Kobak, K.A., Osman, D.A., 2011. The use of videoconferencing with patients with psychosis: a review of the literature. Ann. Gen. Psychiatry 10, 14.

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. Psychiatry 59 (S20), S22–S33.

Shen, J., Kobak, K.A., Zhao, Y., Alexander, M., Kane, J., 2008. Use of remote centralized raters via live 2-way video in a multicenter clinical trial for schizophrenia. J. Clin. Psychopharmacol. 28 (6), 691–693.

Targum, S.D., 2006. Evaluating rater competency for CNS clinical trials. J. Clin. Psychopharmacol. 26 (3), 308–310.

Targum, S.D., Little, J.A., Lopez, E., DeMartinis, N., Rapaport, M., Ereshefsky, L., 2012. Application of external review for subject selection in a schizophrenia trial. J. Clin. Psychopharmacol. 32 (2), 825–826.

Targum, S.D., Wedel, P.C., Bleicher, L.S., Busner, J., Daniel, D.S., Robinson, J., Rauh, P., Barlow, C., 2013. A comparative analysis of centralized, site-based, and patient ratings in a clinical trial of major depressive disorder. J. Psychiatr. Res. 47, 944–954.

Targum, S.D., Pendergrass, J.C., 2014. Site-independent confirmation of subject selection for CNS trials: “dual” review using audio-digital recordings. Ann. 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.

Velthorst, E., Levine, S.Z., Henquet, C., de Haan, L., van Os, J., Myin-Germeys, I., Reichenberg, A., 2013. To cut a short test even shorter: reliability and validity of a brief assessment of intellectual ability in schizophrenia-a control-case family study. Cogn. Neuropsychiatry 18 (6), 574–593.

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.

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.