The Effects of Erratic Ratings on Placebo Response and Signal Detection in the Roche Bitopertin Phase 3 Negative Symptom Studies—A Post Hoc Analysis

Daniel Umbricht¹, Alan Kott*,², and David G. Daniel³

¹Neuroscience and Rare Diseases, Roche Pharma Research & Early Development, F. Hoffmann-La Roche Ltd, Basel, Switzerland; ²Signant Health, Prague, Czech Republic; ³Signant Health, Mclean, VA

*To whom correspondence should be addressed; Slezska 2127/13, Prague 2, 120 00, Czech Republic; tel: +1-610-225-5871, fax: +420-224-254267, e-mail: dralankott@gmail.com

Objective: In the current post hoc analyses, we assessed the impact of erratic ratings, a marker of questionable measurement quality, on placebo and drug response and drug-placebo separation in schizophrenia negative symptom trials. Methods: Data were obtained from three phase 3, multi-center, 24-week, randomized, double-blind, placebo-controlled trials with bitopertin in the treatment of negative symptoms of schizophrenia. Erratic ratings were operationally defined as at least one occurrence of at least a 20% change in negative symptom factor score in the opposite direction at consecutive visits. The effect of erratic ratings on placebo and drug response and drug-placebo separation was assessed by the protocol on a subject and site-level using a mixed model repeated measures analysis. Results: Placebo response was significantly increased in the presence of erratic ratings, both at the subject and site levels. Treatment response in the presence of erratic ratings was mixed and inconsistent across doses and protocols. In most cases removing data generated by subjects and sites with erratic ratings resulted in a numerical increase of drug-placebo difference favoring treatment. Additionally, in this post hoc analysis, 10 mg of bitopertin separated statistically significantly from placebo at the end of study in one of the protocols. Discussion: Erratic rating patterns appear to be consistently associated with increased response to placebo and a mixed response to treatment in the bitopertin negative symptoms trials and to have degraded drug-placebo separation. As a quality indicator, they may flag the need for enhanced quality control of affected sites in clinical trials.

Introduction

There are currently no drugs approved specifically for the treatment of negative symptoms with the exception of amisulpride in Europe. A large placebo response has been reported to be a significant complicating factor in detecting a signal in schizophrenia clinical trials and has been tied to industry sponsorship, a high probability of receiving drug vs placebo, larger numbers of sites, and several other factors.¹⁻³ There are other challenges to minimizing placebo responses that are unique to negative symptom trials. For example, subjects with negative symptoms who have been socially isolated may respond particularly positively to the attention from study staff and raters. Moreover, negative symptoms may be particularly difficult to rate reliably because assessments of avolition and apathy items cannot be based on immediate observation by the clinician but rely to a large extent on the reports of patients and caregivers, which are often of questionable reliability.⁴ In addition, there are subjective elements of the assessment of rapport, blunted affect, and psychomotor retardation.

In the current post hoc analyses, we address the question whether erratic ratings, a putative indicator of poor quality symptom measurement (ie, rating), affect placebo and drug response in negative symptom trials and if addressed may offer a possibility to improve signal detection.

Identification of markers of poor data quality would enable rapid remediation of poor measurement technique before it is replicated and potentially obscures drug-placebo differences after randomization. A straightforward approach to monitoring data quality is to assess for aberrant or illogical patterns of symptom change. The Marder negative symptom factor score (NSFS)⁵,⁶ of the Positive and Negative Syndrome Scale (PANSS)⁷ is a good substrate for such evaluation because it is the most commonly used primary outcome measure in trials in
patients with negative symptoms and is measured based on a synthesis of observations from the interview, informant data, and self-report.

In the current project, we performed post hoc analyses of three phase 3 negative symptom clinical trial data sets in which bitopertin or placebo was added to the antipsychotic regimen of stable schizophrenia patients with predominant, persistent negative symptoms. We evaluated the effects of unusual patterns of erratic changes in NSFS we have previously observed anecdotally in data quality monitoring that are not considered to be consistent with the expected course of stable patients with persistent, predominant negative symptoms.

We tested the hypotheses that the occurrence of erratic changes at the subject and the center level would be associated with increased placebo and drug response and decreased signal detection. To address the latter we performed sensitivity analyses to determine whether the trial outcomes would have been different without the subjects and sites affected by erratic ratings.

**Methods**

Intent-to-treat data was obtained from 3 similarly designed, phase 3, multi-center, 24-week, randomized, double-blind, placebo-controlled trials with bitopertin in the treatment of the negative symptoms of schizophrenia.\(^8\) In WN25308 NCT01192880, data from 596 subjects (199 on placebo) was analyzed; in WN25309 NCT01192906, data from 603 subjects (202 on placebo) was analyzed; and in NN25310 NCT0119286, data from 595 subjects (197 on placebo) was analyzed. The PANSS was conducted at baseline and at weeks 4, 8, 12, 16, 20, and 24 (visits 1–7).

Erratic ratings in the NSFS total score were operationally defined as at least one occurrence of a 20% change in the opposite direction at consecutive visits.

For each protocol separately we conducted the following statistical analyses to explore the effect of the presence of erratic ratings on placebo and drug response.

1. We separately compared the response to placebo and individual treatment arms in subjects affected by the presence of erratic ratings vs subjects not affected by erratic ratings.

2. We separately compared the response to placebo and individual treatment arms at sites affected by the presence of at least one instance of an erratic rating vs sites not affected by any erratic ratings.

Additionally, we removed all data coming from subjects and sites affected by erratic ratings identified under steps 1 and 2 above from the analytical dataset and calculated the drug-placebo difference for each protocol separately to assess whether such a step would change the outcomes of the study.

All analyses were conducted using a mixed model repeated measures (MMRM) analysis with fixed effects of treatment, visit, treatment visit interaction, baseline value, subgroup treatment visit 3-way interactions, subgroup treatment visit 2-way interactions and baseline visit interaction as covariates, and with a AR(1) matrix structure. Least square mean difference between subgroups for each dose at each visit was computed.

**Results**

The dataset consisted of 1794 randomized subjects. Erratic ratings affected a total of 171 (9.5%) subjects across all 3 protocols, 50/596 (8.4%) in protocol WN25308, 48/603 (8.0%) in protocol WN25309, and 73/595 (12.3%) in protocol NN25310. Fifty-seven subjects terminated before week 8 (the second post-baseline visit) and, therefore, could not be affected by an erratic pattern. For the purposes of the analyses, these subjects were considered not affected by erratic ratings. Table 1 summarizes the distribution of erratic ratings across treatment arms in each protocol. No statistical differences were identified in the distribution of erratic ratings across the individual treatment arms within each protocol. Erratic ratings were significantly more frequent in protocol NN25310 compared to the other 2 protocols ($P < .05$).

**Results 1: Effect of Erratic Ratings on Placebo and Drug Response in Subjects Affected vs Subjects not Affected by Erratic Ratings**

As shown in figure 1, the LS mean change from baseline in the placebo arm was significantly greater ($P < .05$, 2-tailed) at all but one visit in subjects that had at least one erratic rating compared to the remaining subjects. In contrast, the LS mean change from baseline in the treatment arms was not consistently different in subjects affected by erratic

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Placebo</th>
<th>Bitopertin 5 mg</th>
<th>Bitopertin 10 mg</th>
<th>Bitopertin 20 mg</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN25308</td>
<td>16/199 (8.0%)</td>
<td>18/199 (9.0%)</td>
<td>16/198 (8.1%)</td>
<td>50/596 (8.4%)</td>
<td></td>
</tr>
<tr>
<td>WN25309</td>
<td>17/202 (8.4%)</td>
<td>19/204 (9.3%)</td>
<td>12/197 (6.1%)</td>
<td>48/603 (8.0%)</td>
<td></td>
</tr>
<tr>
<td>NN25310</td>
<td>20/197 (10.2%)</td>
<td>30/200 (14.9%)</td>
<td>23/198 (11.6%)</td>
<td>73/595 (12.3%)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53/598 (8.9%)</td>
<td>19/204 (9.3%)</td>
<td>60/596 (10.1%)</td>
<td>171/1794 (9.5%)</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Shown are the number and percentage of subjects with erratic ratings randomized in each treatment arm.
Fig. 1. Least square mean change from baseline in negative symptom factor score in subjects affected by erratic ratings (dashed lines), and subjects not affected by erratic ratings (full lines). P values represent the comparison of subjects affected by erratic ratings vs those not affected within each treatment arm. A) Protocol WN25308; B) Protocol WN25309; C) Protocol NN25310. *P < .05; **P < .01; ***P < .001.
ratings vs those who were not. For example, in the 10 mg bitopertin arm, the change from baseline in subjects impacted by erratic ratings was significantly greater at 2 time points in protocol WN25308, at all 6 time points in protocol WN25309 and did not differ at any time point in protocol NN25310 (For analysis of data combined from all 3 protocols see Supplementary Figure 1).

As shown in figure 2, at week 24 the apparent effect of erratic ratings on change from baseline in protocol WN25308 was significantly stronger in the placebo arm than in either treatment arm. At week 24 in protocol WN25309 erratic ratings did not appear to impact the placebo and treatment arms differently. In protocol WN25310 erratic ratings were associated with significantly greater change from baseline in the placebo arm vs 10 mg of bitopertin but not vs 20 mg of bitopertin (For analysis of data combined from all 3 protocols see Supplementary Figure 2).

As shown in figure 3, in the sensitivity analyses, the removal of subjects affected by erratic ratings resulted in 10 mg of bitopertin significantly outperforming placebo at weeks 20 and 24 in the WN25309 protocol ($P < .05$, ES = 0.20 and 0.15, respectively). No other statistically significant effects were observed (For analysis of data combined from all 3 protocols see Supplementary Figure 3).

Results 2: Effect of Erratic Ratings on Placebo and Drug Response at Sites With At Least One Instance of Erratic Ratings vs Sites With No Erratic Ratings

In protocol WN25308 we identified 27 out of 92 sites, in protocol WN25309 28 out of 96 sites, and in protocol NN25310 40 out of 105 sites affected by at least one set of erratic ratings.

As shown in figure 4, starting week 8 for all 3 protocols the LS mean change from baseline in the placebo arms was significantly greater ($P < .05$, 2-tailed) at those sites that had at least one erratic rating compared to the remaining sites. In contrast, in the treatment arms, the difference in LS mean change from baseline at the affected sites compared to the not affected sites did not exhibit a clear pattern. In protocol WN25308, the LS mean change from baseline in the 10 mg and the 20 mg bitopertin arm did not differ at the affected sites from the not affected sites (figure 4a). In protocol WN25309, the LS mean change from baseline was statistically greater ($P < .05$) at the affected sites at one time point in the 10 mg bitopertin arm and 4 time points in the 5 mg arm, respectively (figure 4b). In protocol NN25310, we observed a significantly larger LS mean change from baseline at the affected sites compared to the not affected ones in the 20 mg but not 10 mg bitopertin arm (figure 4c) (For analysis of data combined from all 3 protocols see Supplementary Figure 4).

As shown in figure 5, at week 24, the difference in LS mean change from baseline between the affected and not affected sites did not differ between placebo and the treatment arms in protocols WN25308 and WN25309. In protocol NN25310, the effect of erratic ratings was significantly larger in the placebo group compared to 10 mg but not 20 mg of bitopertin (For analysis of data combined from all 3 protocols see Supplementary Figure 5).

As shown in figure 6, in protocol WN25308 the removal of sites affected by at least one erratic rating resulted in
Effects of Erratic Ratings on Placebo Response

Numerical decreased change from baseline in all 3 study arms at all post-randomization visits. Starting week 12, both active treatment arms numerically outperformed placebo, however at no time point was this difference statistically significant. In contrast, in protocol WN25309 removing sites with at least one set of erratic ratings resulted in the 10 mg dose of bitopertin statistically significantly outperforming placebo at weeks 20 and 24 ($P < .05$, ES = 0.30 at both visits). In protocol NN25310 the removal of sites affected by at least one set of erratic ratings resulted in 10 mg but not 20 mg of bitopertin numerically (but not statistically significantly) outperforming placebo from week 4 visit onwards (For analysis of data combined from all 3 protocols see Supplementary Figure 6).

Discussion

In the current post hoc analyses, we found that erratic ratings were associated with sustained increased placebo response lasting until the 24-week endpoint. The findings were consistent across 3 separate clinical trials of adjunctive treatment of stable patients with negative symptoms of schizophrenia. This effect was observed not only at the subject level but as well at the level of sites affected by at least one occurrence of erratic ratings. In contrast, in the treatment arms, the impact of erratic ratings at both the subject and site level was inconsistent. In the sensitivity analyses, removing the data generated by subjects affected by erratic ratings resulted in a numerical increase in drug-placebo differences at the end of treatment favoring the actual treatment in all 3 protocols, with the separation reaching statistical significance for 10 mg of bitopertin in WN25309. Similarly, the removal of sites affected by at least one erratic rating resulted in all treatment arms with the exception of 20 mg bitopertin arm in protocol NN25310 numerically outperforming placebo. As in the case of removing the affected subjects, the removal of affected sites in protocol WN25309 resulted in this post hoc analysis in 10 mg of bitopertin separating from placebo at weeks 20 and 24. Moreover, the effect size would have been twice as large as in the case of removing only the affected subjects. Taken together these results are consistent with the notion that stringent data quality monitoring procedures, including limiting enrollment of poorly performing sites, have the capacity to enhance detection of any drug signal.

Psychiatry clinical trials have considerably lower success rates (24% and 56%, respectively, in phases 2 and 3) than medical disorders such as hematology (57% and 75%, respectively, in phases 2 and 3), which utilize more objective endpoints. Psychiatry clinical trials are particularly vulnerable to failure because they utilize relatively subjective endpoints and drug-placebo differences are often modest. A recent meta-analysis of antipsychotic clinical trials conducted from 1960 to 2013 indicates decreasing magnitude of separation between placebo and medication treatment. Similar findings were observed in a meta-analysis conducted by Leucht and colleagues that included 167 double-blind randomized controlled trials published between years 1955 and 2016. The extent

![Fig. 3. Least square mean change from baseline in negative symptom factor score in subjects not affected by erratic ratings. P values represent comparison with the placebo group. *$P < .05$; **$P < .01$; ***$P < .001$.](https://academic.oup.com/schizbullopen/article/1/1/sgaa040/5896582)
to which rating error has contributed to the decline in separation between placebo and medication treatment in schizophrenia clinical trials is unknown. However, it appears that anomalous measurement patterns such as erratic ratings may be associated with significant degradation of the signal in clinical trials.

For bitopertin and other compounds that demonstrated statistically significant drug-placebo differences in phase 2 but not phase 3, the observations emphasize the difficulty in sorting out the impact of measurement error and excessive placebo response from drug activity or lack thereof. In the case of the phase 3 bitopertin schizophrenia negative symptom trials, to the extent degradation of drug signal might have occurred from measurement error, it is possible that data quality assurance methodologies such as the use of highly calibrated centralized raters might have been useful. Centralized rating methodologies have been shown in some instances to decrease baseline inflation and increase signal detection.10,11

The mechanism through which erratic ratings as a potential index of poor data quality affect placebo and drug response is unclear. In the bitopertin schizophrenia protocols, erratic ratings represent large changes in opposite directions at consecutive visits that were clinically unexpected in a population of subjects maintained on their original antipsychotic treatment regimens, selected for stability and predominant negative symptoms. It is unclear why erratic ratings impacted the placebo arm in the bitopertin studies more than the treatment arm. It is possible that bitopertin exerted a stabilizing effect on psychopathology. Another speculative view is that there may be a ceiling beyond which even the most “optimistic” low-quality rater cannot push a rating. Such a ceiling would be closer to the true effect of an efficacious drug than of placebo, it should, therefore, distort the placebo effect more.

Our results also point to additional factors. As our analysis removed not only data from affected subjects but in a separate step as well data from all patients randomized at sites affected by the presence of at least one erratic rating, the occurrence of erratic ratings as flag for questionable data quality may also point to other data quality issues present at affected sites caused by nonspecific increase of data variability and measurement error, such as inaccurate assessment of baseline overall and individual symptom severity or unidirectional inflation of symptom changes after randomization, ultimately affecting drug-placebo separation. Indeed, our prior analyses of data pooled from 14 double-blind schizophrenia clinical trials showed a significant association of erratic ratings with outlying PANSS changes and inaccurate PANSS scoring in the screening period.

In summary, the current post hoc findings in 3 separate protocols are consistent with the notion that erratic ratings in erstwhile stable patients with prominent, predominant negative symptoms are predictive of sustained placebo response. Such rating patterns appear to have
Effects of Erratic Ratings on Placebo Response

degraded the signal in the bitopertin phase 3 negative symptom trials addressed in these analyses. As indicators of questionable data quality, erratic ratings may flag the need for enhanced quality control of affected sites in clinical trials.

Novel methodologies that detect measurement errors and markers of placebo response early have the potential to become an important component of enhancing measurement accuracy and reliability, modulating placebo response and, ideally, improving signal. The types

Fig. 5. Week 24 difference in least square mean change from baseline between sites affected and not affected by erratic ratings. *P < .05; **P < .01; ***P < .001.

Fig. 6. Least square mean change from baseline in negative symptom factor score at sites not affected by erratic ratings. *P < .05; **P < .01; ***P < .001.
of analyses described in this presentation may be automated to produce rapid feedback to investigators and external data quality monitors when incorporated into
electronic clinical outcome assessments (eCOA). Rapid identification of such patterns allows for rater assessment and remediation to avoid propagation of errors that may decreases the chance of signal detection and also flag sites with potentially poor data quality. With eCOA systems, feedback to raters may be instantaneous, allowing reconsideration of errors prior to data submission. Our findings provide a strong incentive for the use of highly calibrated and well trained centralized assessments of symptom and ongoing reviews of study data to detect and remediate data quality issues as early as possible.

Funding

This post hoc analysis was funded by Signant Health.

Acknowledgments

The authors thank Xingmei Wang, M.S. for her support in performing the statistical analyses. A.K. and D.G.D. are full-time employees and stock option holders at Signant Health. D.U. is a full time employee of F. Hoffmann-La Roche Ltd.

References


