Schizophrenia Bulletin Open
https://doi.org/10.1093/schizbullopen/sgab037

The Impact of Aberrant Data Variability on Drug–Placebo Separation and Drug/Placebo Response in an Acute Schizophrenia Clinical Trial

Alan Kott*[,1], Stephen Brannan[2], Xingmei Wang[3], and David Daniel[4]

[1]Signant Health, Prague, Czech Republic; [2]Karuna Therapeutics, Boston, MA, USA; [3]Signant Health, Blue Bell, PA, USA; [4]Signant Health, Mclean, VA, USA

*To whom correspondence should be addressed; Slezska 2127/13, Prague 2, 120 00, Czech Republic; tel: +441182067327, e-mail: dralankott@gmail.com

Objective: **In the current posthoc analyses, we evaluated the impact of markers of aberrant data variability on drug placebo separation and placebo and drug response in an acute schizophrenia clinical trial.*Methods:* Positive and negative syndrome scale data were obtained from a phase 2, randomized, double-blind, placebo controlled trial in hospitalized adults with schizophrenia experiencing an acute exacerbation. We assessed the impact of a total of six markers of aberrant data variability: erratic ratings, unusually large postbaseline improvement, high and low mean square successive difference (MSSD), identical and nearly identical ratings and compared the drug placebo difference, drug and treatment response at last visit in affected subjects vs those not affected. All analyses were conducted using generalized linear models.*Results:* In this posthoc analysis, drug placebo separation decreased with the presence of most markers of aberrant data variability. The only exception was high MSSD was associated with significant increase in the signal. In the affected subjects, the presence of indicators of increased data variability augmented the response to placebo, in the case of large postbaseline change and high MSSD, significantly. The presence of indicators of decreased variability numerically but not statistically decreased the response to placebo. Similar findings were observed in the drug treatment group with the exception of erratic ratings that numerically but not statistically decreased the response to the drug.*Discussion:* The presence of most indicators of aberrant data variability had a detrimental effect on drug-placebo separation and showed different effects on placebo and treatment response.**

Key words: placebo response/drug response/drug–placebo separation/data quality/data variability

Introduction

Nearly seven decades after the synthesis and commercial availability of chlorpromazine, schizophrenia remains one of the most debilitating mental disorders, associated with significantly decreased life expectancy[1] and severe socio-economic burden.[2] Even recently approved antipsychotic treatments are often associated with unwanted side effects, such as weight gain, sedation, and akathisia.

Successful development of new antipsychotics has been made more difficult by increasing placebo response and diminishing effect sizes over at least the last two decades.[3] These trends are widely acknowledged to be multifactorial and have been attributed to industry sponsorship, large number of sites, higher probability of receiving medication over placebo and a plethora of other causes.[3–6]

In addition to factors such as these, we have observed patterns of unexpected variability in psychopathology measurement that appear to be associated with increased placebo response and reduced separation between placebo and study drug. Among these patterns, erratic ratings emerged as a robust predictor of placebo response and drug-placebo separation in retrospective analyses of three clinical trials with bitopertin in schizophrenic subjects suffering from predominant negative symptoms.[7] Those trials did not separate from placebo. In the current retrospective analysis, we investigated which other patterns of unexpected variability would also impact placebo and drug response and drug-placebo separation in a phase IIb clinical trial of the effect of KarXT in acutely psychotic hospitalized schizophrenic subjects. We hypothesized that each of the individual measures of excessive and diminished variability tested would be

associated with diminished assay sensitivity of the clinical trial to detect differences in drug vs. placebo even in the context of a successful trial.

Methods

Intent-to-treat data was obtained from a phase 2, multicenter, 5 week, randomized, double-blind, placebo controlled trial of KarXT in hospitalized adults with DSM-5 schizophrenia in the United States experiencing an acute exacerbation or relapse of symptoms. (NCT03697252) The sponsor of the study employed an extensive set of procedures to address the reliability and accuracy of symptom measurement and modulation of placebo response. These procedures included: (1) site selection based on previous performance; (2) prestudy calibration of interview and symptom severity measurement technique; (3) placebo response mitigation training; (4) operationalization and monitoring of acuity criteria; (5) enhanced instructions and data quality checks embedded in eCOA; (6) recording and independent expert review of audio recorded PANSS interviews; (7) blinded analytic review of endpoint data for concerning patterns; (8) rapid remediation of rating and interview errors; and (9) site enrolment continually tied to data quality.

Excessive Data Variability

We retrospectively examined the impact of three markers of excessive data variability, erratic ratings, unusually large decrease in PANSS total score in the first postbaseline visit, and high mean square successive difference (MSSD) on placebo response and placebo-study drug separation of the PANSS total score. The MSSD as an average of squared differences between successive observations is a measure of temporal instability accounting for both variability and temporal dependency over time.[8]

Erratic ratings in the PANSS total score were originally operationally defined as at least one occurrence of a 15+ point change in opposite direction at two consecutive visits. We selected 15 points change on the PANSS as it was shown to represent the minimum clinically important difference in the PANSS.[9] Since the occurrence of these erratic ratings in the dataset was low and did not allow for a comparison, we lowered the cut-off to 10+ point changes in opposite direction at two consecutive visits.

Unusually large changes were operationally defined as decreases in the PANSS total score by week 2 (the first postbaseline visit where PANSS was administered) that were below the 10th percentile of the study. Thus, about 10% of subjects were expected to fulfill this criterion.

High mean square successive differences were operationally defined as those differences that exceeded the 90th percentile of the study. Again, approximately 10% of subjects were expected to fulfill this criterion.

Decreased Data Variability

Originally, we evaluated three markers of decreased variability: identical ratings across visits in the PANSS; nearly identical ratings across visits in the PANSS; and low mean square successive difference. Since the occurrence of identical ratings in the dataset was extremely low, we combined identical and nearly identical ratings into a single marker.

Nearly identical and identical ratings, respectively, were operationally defined as PANSS ratings where 27 or more of the total of 30 PANSS items or 30/30 of the PANSS items had exactly the same score at two or more consecutive visits.

Low mean squared successive difference was operationally defined as MSSD below 10th percentile of the study.

Statistical Analysis

For each of the data quality concerns, we compared the change in the PANSS from baseline to last visit in subjects affected by the respective data concerns vs. those not affected. Additionally, we estimated the drug placebo difference and the treatment effect size in the group of subjects not affected by the respective data quality concerns versus those in the group of subjects who were affected. A generalized linear model was fitted with fixed effect of subgroup, treatment, baseline PANSS total score and subgroup treatment two-way interaction as covariates. The analysis was carried out in SAS 9.4 (TS1M2).

Given the exploratory nature and the small number of planned analyses, no correction for multiple testing was applied.

Results

The dataset consisted of 165 subjects with evaluable data. In the current phase 2b data set, quality concerns were significantly decreased at alpha of 0.05 for the originally defined erratic ratings, identical and nearly identical ratings compared to historical data.[10] For those quality concerns defined as a percentage of affected visits (i.e. large postbaseline changes, high and low MSSD) frequency was at the expected levels of 10% based on historical data.[10]

The distribution of subjects with the examined variability indicators is summarized in Table 1. For four out of the five tested indicators there was no difference in distribution of the affected subjects between the placebo and the treatment groups. The only significant difference (P = .021) was observed in the increased presence of large postbaseline changes in the treatment group compared to the placebo group.

Table 1. Distribution of Variability Indices in the Dataset

	Placebo	KarXT	Total	*P*[a]
Erratic change	6/84 (7.1%)	4/81 (4.9%)	10/165 (6.1%)	0.553
Large postchange	3/84 (3.6%)	11/81 (13.6%)	14/165 (8.5%)	0.021
High mean squared successive difference	6/84 (7.1%)	9/81 (11.1%)	15/165 (9.1%)	0.375
Nearly identical ratings	11/84(13.1%)	8/81(9.9%)	19/165(11.5%)	0.517
Low mean squared successive difference	7/84 (8.3%)	3/81(3.7%)	10/165(6.1%)	0.213

Shown are the number and percentage of subjects by treatment arm and total.
[a] KarXT group compared to placebo group using Fisher's exact test.

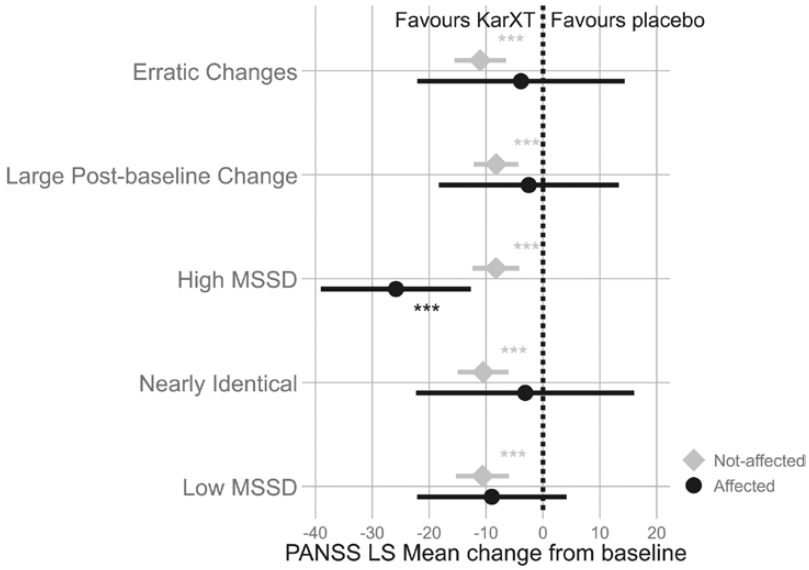

Fig. 1. Last visit drug−placebo difference in least square mean change from baseline for affected and not affected subjects. *P* values represent comparison with the placebo group. **P* < .05; ***P* < .01; ****P* < .001.

Results 1. *Effect of Variability Indicators on Drug Placebo Separation*

The objective of the first set of analyses was to assess whether the variability indicators impacted signal detection by comparing drug placebo separation for the affected versus the not affected subjects. Figure 1 shows the drug placebo difference at end of treatment in subjects not affected by the indicators and in affected subjects. KarXT separated from placebo in all tested groups in the subjects not affected by the indicators, while in the affected group in four out of the five tested indicators the drug did not separate from placebo. The only exception was the presence of high mean squared successive difference where the KarXT treatment arm significantly separated from placebo (P < 0.001) (further details in table 2)

Results 2. *Effect of Variability Indicators on Placebo and Drug Response*

The objective of the second set of analyses was to assess the impact of the variability indicators on the end of treatment change in the placebo and treatment groups, respectively. The presence of indicators of increased variability augmented the response to placebo in the affected subjects. As shown in figure 2 and table 2, the LS mean change from baseline in the placebo arm was significantly increased (*P* < .05, two-tailed) in subjects affected by high mean squared successive difference and in subjects affected by large postbaseline change. The presence of indicators of reduced variability then reduced the response to placebo in the affected group; however, the differences were only numerical and not statistically significant. A very similar pattern was observed in the case of subjects randomized to KarXT, with the exception of erratic ratings, which presence unlike in the placebo arm decreased numerically the response to treatment in the affected group (figure 2 and table 2).

Discussion

Previously, we observed that erratic ratings increased placebo response and diminished drug-placebo separation in a very large, phase 3 global negative symptom clinical program.[7] In the current posthoc analysis, we examined the effect of erratic ratings as well as other measures of

Table 2. Summary of PANSS Changes at Last Visit for Placebo and KarXT Treatment Arms in Groups not Affected and Affected by Examined Indicators in NCT03697252 Trial

		Placebo			KarXT				
Indicator	Group	*N*	LS Mean	SEM	*N*	LS Mean	SEM	*P*[a]	Cohen's *d*[a]
Erratic changes	NON-ID[b]	78	−5.63	1.63	77	−16.65	1.64	<.0001	−0.76
	ID[c]	6	−8.44	5.88	4	−12.31	7.2	0.6772	−0.28
Large postbaseline change	NON-ID	81	−4.69	1.37	70	−12.92	1.47	<.0001	0.65
	ID	3	−36.37	7.13	11	−38.87	3.73	0.7573	0.3
High MSSD	NON-ID	78	−4.98	1.45	72	−13.25	1.51	<.0001	−0.66
	ID	6	−16.42	5.23	9	−42.27	4.27	0.0001	−1.49
Nearly identical	NON-ID	73	−6.43	1.68	73	−17.06	1.68	<.0001	−0.73
	ID	11	−1.83	4.35	8	−10.82	5.07	0.1798	−0.67
Low MSSD	NON-ID	77	−6.53	1.62	78	−17.03	1.61	<.0001	−0.72
	ID	7	1.94	5.37	3	−1.17	8.19	0.7502	−1.04

[a] KarXT group compared to placebo group.
[b] Not-affected subjects.
[c] Affected subjects.

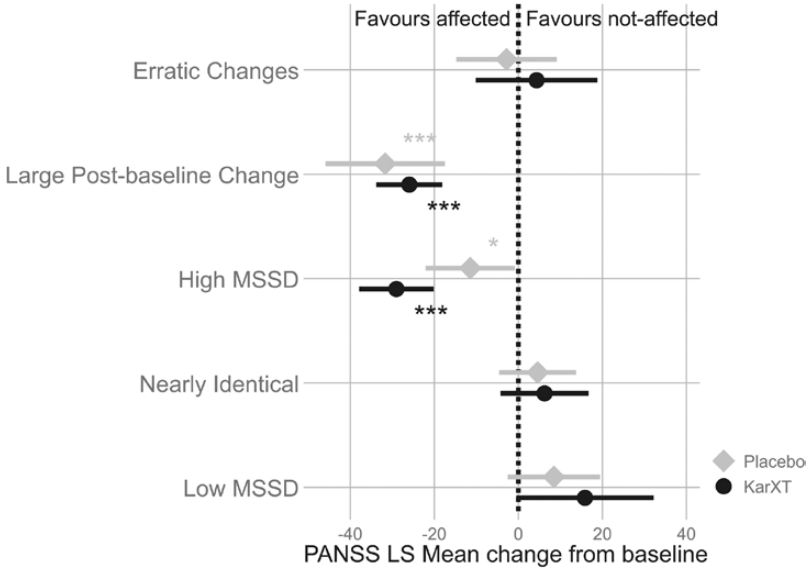

Fig. 2. Last visit difference in least square mean change from baseline between affected and not affected subjects. *P* values represent the comparison of subjects affected vs those not affected within each treatment arm. **P* < .05; ***P* < .01; ****P* < .001.

increased and decreased data variability on placebo response, drug response and drug placebo separation in a phase 2 acute schizophrenia trial. The current findings appear to confirm and extend the previously reported detrimental impact of erratic ratings on drug-placebo separation in a qualitatively different schizophrenia clinical trial population. Finding that erratic ratings have essentially the same impact in both successful trials and unsuccessful trials and in acute schizophrenia trials as well as negative symptom trials extends the importance of paying attention to erratic ratings. In examining this acutely psychotic population, we broadened the retrospective analyses to include additional measures of data variability. In addition to confirming the impact of erratic ratings, we observed significantly increased response to placebo in those subjects who were affected by other measures of increased variability, that is, large postbaseline change and high MSSD. Numerically lower response to placebo was observed in the case of decreased data variability in affected subjects; however, compared to the nonaffected subjects the difference did not reach statistical significance, possibly due to the relatively low number of affected subjects. In the affected group, the drug placebo difference was severely impacted by the presence of all examined data quality concerns, and with the exception of high MSSD drug would not separate from placebo in the group of affected subjects.

Erratic ratings represent clinically unexpected symptom fluctuations over a short period of time. The originally proposed cutoff of 15 points could not be used

for the purposes of our analysis since only two subjects meeting those cutoffs were identified in the dataset. The cutoff of 10 points utilized identified a small subset of affected subjects almost equally represented in the placebo and treatment groups. Erratic ratings likely represent a number of measurement errors that translate into the zig-zag pattern of disease severity fluctuation over time. In the current data, erratic ratings were in 33% of cases associated with rater change that could be, at least in part, responsible for the data variability observed. This raises the possibility that in the context of rater change erratic ratings occurred because the two raters were not calibrated sufficiently in their interview styles and/or symptom measurement technique. In other cases, erratic ratings might occur in the context of inconsistencies or deficiencies in a single rater's interview and measurement/scoring technique. For example responsible raters could be more inclined to use extreme scores on individual item levels in combination with relative ratings compared to previous visit, rather than rating the actual symptom severity using the appropriate anchors. Both scenarios could be prevented by enhanced training. The observation that erratic ratings appear to enhance placebo response while reducing drug response is puzzling but in alignment with our prior findings in the three bitopertin studies in subjects suffering from negative symptoms.[7] A plausible explanation is that in some cases raters are accurately identifying an unusual subgroup of subjects, for example subjects possibly exhibiting lability associated with the untreated, placebo state, albeit the amplitude of the changes may be incorrectly pronounced. In any case, attempted replication in a larger data set may be informative.

Unusually large postbaseline improvements are sometimes interpreted as indicative of intentional inflation of a subject's severity at baseline that would allow the subject to meet study inclusion criteria. If so, the large drop in scores immediately following baseline could be more reflective of the true subject severity at the time of entry into the study rather than a real improvement. Another explanation for unusually large improvement immediately after baseline is expectation bias by the subject and/or rater that the test medication will be highly effective. This could potentially explain the increased frequency of this indicator in the KarXT treatment arm where expectation bias may have amplified the amount of actual improvement in the affected subjects. Intentional baseline inflation seems unlikely in the current study because all subjects identified with these large changes had scores well above the minimal threshold and independent reviews of audio-recorded baseline interviews may have deterred possible baseline inflation. Although unexpectedly large early changes from baseline were infrequent in this study, they could represent a subgroup of rapid responder. In addition more intensive training efforts to modulate expectation bias originating from subjects, informants and investigative sites might further increase confidence in the veracity of these large changes.

Identical ratings represent a highly concerning data finding. Clinically, individual symptom severities are expected to oscillate over time even in stable subjects; additionally, given the reliability properties of the PANSS,[11] it is exceedingly difficult for an expert rater to rate the same subject exactly the same even under the hypothetical situation of absolutely no symptom change. We have as well previously demonstrated that two random raters rating the same videotaped interview are expected to agree on all 30/30 PANSS items in 0.016% of cases,[12] Identical ratings may be indicative of expectation bias of no change, failure to conduct an independent interview, copy pasting of scores obtained during prior interview(s) or data fabrication. Visits with identical ratings were substantially rarer (<1%) in the current data set compared to our database of acute schizophrenia studies (2.5%) and affected in total only three subjects, one randomized to KarXT and two to placebo. Less is known about the significance of nearly identical PANSS ratings. In the current study, the combined identical and nearly identical PANSS rating groups did numerically decrease response to both, investigational drug and placebo, and affected subjects failed to separate drug from placebo.

Mean squared successive difference (MSSD) assesses unusual levels of variability over the entire course of the subject's participation in the study. Thus, subtle anomalous data patterns can be identified that might not have been detected by more point in time quality indicators. For the purposes of the analysis, thresholds were set at 10% of data at either end of distribution, but more stringent cut-offs could be used in larger studies to increase specificity. Placebo response was significantly increased in subjects affected by high MSSD; subjects affected by low MSSD showed on average worsening on placebo but did not differ from the remaining subjects in the placebo arm. While subjects affected by low MSSD did not separate from placebo in the current study possibly due to a small number of cases, subjects affected by high MSSD showed a clear separation between drug and placebo with an effect size larger than the effect size in the nonaffected group. It is important to note that the magnitude of improvement in both arms seems exaggerated, the placebo response in the affected subjects was numerically larger than the response to KarXT in the not affected group. The data supports the intuitively obvious notion that outlying measurement patterns have the potential to enhance as well as attenuate drug-placebo separation.

The sponsor implemented a robust and continuous program to assure the validity of the data. It consisted of rigorous selection of sites based on prior performance, rater calibration at the beginning of the trial, placebo modulation training, as well as ongoing monitoring of subject eligibility and quality of ratings by audio recording of all PANSS assessments and their independent verification by

highly calibrated group of central reviewers. Additionally, all collected efficacy data were continuously scrutinized by central analytical program and sites were allowed to screen subjects based on ongoing performance. These implemented measures resulted in rapid corrective actions when data concerns were identified and minimization of spread of these concerns in the data. Thus, it was not surprising that the frequency of data quality concerns in the current study was modest.

In summary, the current posthoc analysis in an acutely psychotic schizophrenic population replicated an earlier finding in a predominantly negative symptom schizophrenia population that erratic ratings are associated with diminished drug-placebo separation. In addition, other measures of increased and decreased variability appeared to impact drug-placebo separation. Interpretation of the analyses are limited by their posthoc nature and the relatively small size of the affected groups. Despite these limitations, the findings are consistent with the notion that markers can be identified in blinded data sets that are associated with diminished signal detection. Raters prone to such patterns can be remediated or limited from rating new subjects, potentially minimizing their impact on data quality. This type of intervention could be added to other elements of a systematic approach to minimizing risks to endpoint data quality such as careful calibration of raters prior to the study, placebo response modulation training, blinded data analytics to identify aberrant rating patterns and external expert review of recorded interviews. A well designed prospective study randomizing sites and raters to this systematic approach or no intervention would provide a more definitive answer whether and how much these measures impact on data quality.

Conflicts of Interest: This posthoc analysis was funded by Signant Health. Dr Kott, Dr Daniel, and Ms. Wang are employees of Signant Health. Dr Brannan is an employee of Karuna Therapeutics.

References

1. Hjorthøj C, Stürup AE, McGrath JJ, Nordentoft M. Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. *Lancet Psychiatry.* 2017;4(4):295–301.
2. Vos T, Barber RM, Bell B, *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015;386(9995):743–800.
3. Leucht S, Leucht C, Huhn M, *et al.* Sixty years of placebo-controlled antipsychotic drug trials in acute schizophrenia: systematic review, Bayesian meta-analysis, and meta-regression of efficacy predictors. *Am J Psychiatry.* 2017;174(10):927–942.
4. Kinon BJ, Zhang L, Millen BA, *et al.*; HBBI Study Group. A multicenter, inpatient, phase 2, double-blind, placebo-controlled dose-ranging study of LY2140023 monohydrate in patients with DSM-IV schizophrenia. *J Clin Psychopharmacol.* 2011;31(3):349–355.
5. Alphs L, Benedetti F, Fleischhacker WW, Kane JM. Placebo-related effects in clinical trials in schizophrenia: what is driving this phenomenon and what can be done to minimize it? *Int J Neuropsychopharmacol.* 2012;15(7):1003–1014.
6. Leucht S, Chaimani A, Mavridis D, *et al.* Disconnection of drug-response and placebo-response in acute-phase antipsychotic drug trials on schizophrenia? Meta-regression analysis. *Neuropsychopharmacology.* 2019;44(11):1955–1966.
7. Umbricht D, Kott A, Daniel DG. The effects of erratic ratings on placebo response and signal detection in the Roche bitopertin phase 3 negative symptom studies—a post hoc analysis. *Schizophr Bull Open.* 2020;1(1):203.
8. Jahng S, Wood PK, Trull TJ. Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychol Methods.* 2008;13(4):354–375.
9. Hermes ED, Sokoloff D, Stroup TS, Rosenheck RA. Minimum clinically important difference in the Positive and Negative Syndrome Scale with data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE). *J Clin Psychiatry.* 2012;73(4):526–532.
10. Kott A, Brannan SK, Wang X, Murphy C, Targum SD, Daniel DG. Procedures to optimize endpoint data quality in an acute schizophrenia study. Poster presentation at the ISCTM 2020 Autumn Virtual Conference, September 21–25, 2020.
11. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261–276.
12. Kott A, Daniel DG. Rater change associated with identical scoring of the PANSS as a marker of poor data quality. Poster presentation at the 2014 CNS Summit held in Boca Raton, Florida, 13–16 November 2014.