BDA Payoffs and Pitfalls:
Examples from a Phase 3 Program

Colin Sauder, PhD

Disclosures

Employee, Bristol-Myers Squibb

Blinded Data Analytics (BDA): Overview

BDA vs Risk-based Monitoring

- The review of accumulating blinded data in ongoing trials:
  - Primary and key secondary endpoints
  - Identifying outlier performance at the site or rater level
  - Support study management decision making

BDA: Service vs. Sponsor

- Most eCOA vendors offer some form of blinded oversight
  - Rater performance, KPI/KRI (“Flags”), endpoint trajectory

JOURNAL ARTICLE

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

Alan Kott ✉, Stephen Brannan, Xingmei Wang, David Daniel

Schizophrenia Bulletin Open, Volume 2, Issue 1, January 2021, sgab037,
https://doi.org/10.1093/schizbullopen/sgab037
Published: 07 August 2021 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

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.

Using PANSS Score Profiles to Predict Early Termination in a Study on Acute Exacerbation of Schizophrenia

Mark Opler[1,2], Jonathan Lam[1], Jennifer Lord-Bessen[1], Elizabeth Hanson[3], Atul Mahableshwarkar[3], Xinxin Dong[3], Maggie McCue[3], Tom Macek[3]

[1]*ProPhase LLC, New York, NY,* [2]*New York University, School of Medicine, New York, NY,* [3]*Takeda Development Center Americas, Inc., Deerfield, IL*

Background

Early termination in clinical trials, especially in those involving schizophrenia, is a significant concern, with some studies showing early discontinuation rates of over 50% (Rabinowitz and Davidov, 2008). Missing data due to early dropouts can potentially compromise the results of a trial. While this area has been identified as an issue in antipsychotic clinical trials, there is little research on understanding the factors that lead to early termination (Mocks et al., 2002). One strategy that may be useful is to identify subjects who present with atypical symptom profiles. For instance, one would expect a patient who presents with high levels of hallucinatory behavior (represented by P3) to also show a correspondingly high level for conceptual disorganization (represented by P2). A patient who instead shows high scores on P3 but low scores on P2 is somewhat atypical. Atypical symptom profiles may represent subjects who are inappropriate for the clinical trial.

A recent Phase 2 study was completed in the US, evaluating the efficacy, safety and tolerability of treatment for 6 weeks with TAK-063 compared with placebo in subjects with acutely exacerbated schizophrenia. TAK-063 is a potent and selective inhibitor of the phosphodiesterase 10A (PDE10A) enzyme that is expressed primarily in the striatal medium spiny neurons (MSN) of the basal ganglia complex (Coskran et al., 2006), which receives extensive cortical (glutamatergic), thalamic, and nigral (dopaminergic) input. In phase 1 studies, TAK-063 has been shown to be safe and well tolerated at single doses up to 1000 mg in healthy subjects and following multiple dosing once daily (QD) for 7 days up to 100 mg in subjects with stable schizophrenia. A retrospective analysis of preliminary blinded data from this Phase 2 clinical trial presented an opportunity to investigate what strategies might be effective in identifying early terminators.

Objectives

Analyses were aimed to determine how different atypical PANSS score profiles (identified using algorithms) were able to predict early termination.

Results

- Forty-nine subjects (32.0%) were identified as early terminators from the study (see Table 1 for demographic information).
- Sensitivity and specificity analyses on the algorithms revealed a wide range of sensitivity and specificity values (Sensitivity: 0 to .94; Specificity: 0 to .99; see Table 2-6).
- Two algorithms had the most useful trade-off between sensitivity and specificity, one involving negative symptoms (labeled NEG: N1 [Blunted Affect] and N5 [Difficulty in Abstract Thinking]), and one involving two positive symptoms (labeled POS1: P2 [Conceptual Disorganization] and P1 [Delusions]).
- Use of recursive partitioning and regression trees also identified these two algorithms for early terminators (see Figure 1).
- Finally, logistic regression analyses revealed a significant interaction effect between these two algorithms ($z=2.1$, $p<.03$). The ROC curves show the use of NEG only (red) in a logistic regression (AUC = .55) while the blue curve shows the use of both POS1 and NEG as predictors (AUC = .61; see Figure 2).

Table 1. Demographic Characteristics of Sample

Mean Age (SD)	42.2 (10.5)
Mean PANSS Score at Baseline (SD)	98.2 (10.4)
Sex	29 (19.0%) Female 124 (81.0%) Male
Race	
Black/African-American	103 (67.3%)
White	46 (30.1%)
Other	4 (2.6%)

Table 4. Contingency Table for POS2 Algorithm

		Early Termination	
		No	Yes
POS2	No	13	8
	Yes	91	41

Table 5. Contingency Table for NEG Algorithm

		Early Termination	
		No	Yes
NEG	No	67	37
	Yes	37	12

Figure 1. CART Analysis of Algorithms

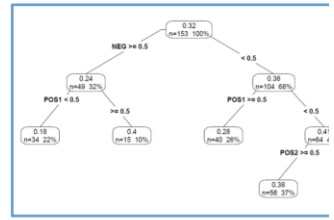


Table 6. Selection of Algorithms and Sensitivity/Specificity Analyses

Algorithm Label	Relevant PANSS Items	Sensitivity
ANX	G2 (Anxiety), G4(Tension)	.00
POS1	P2 (Conceptual Disorganization), P1 (Delusions)	.35
POS2	P3 (Hallucinatory Behavior), G15 (Preoccupation)	.94
NEG	N1 (Blunted Affect), N5 (Difficulty in Abstract Thinking)	.24

JOURNAL ARTICLE

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 Umbricht, Alan Kott ✉, David G Daniel

Schizophrenia Bulletin Open, Volume 1, Issue 1, January 2020, sgaa040,
https://doi.org/10.1093/schizbullopen/sgaa040
Published: 24 August 2020 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

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

The Rater Applied Performance Scale: Evaluating Clinical Interview Skill via Audio Recordings of MADRS Assessments in a Clinical Drug Trial

Engelhardt, N[1], Yavorsky, C[1], McNamara, C[1], Wolanski, K[1], Burger, F[1], Di Clemente, G[1]
[1]Cronos CCS, Inc., Lambertville NJ

Introduction

The quality of clinical interviews in CNS drug trials is frequently overlooked in favor of establishing interrater reliability with passive scoring tasks such as rating patients from a video recorded interview. Audio monitoring of primary outcome clinical interviews has been successfully applied to multi-center psychiatry trials as a way to monitor the quality of outcome data. However, optimal assessment of raters' applied skills requires systematic guidelines so that reviewers can reliably judge a rater's clinical interview skill.

The Rater Applied Performance Scale (RAPS)[1] was developed to provide a systematic and objective assessment of applied rater performance. The RAPS evaluates the clinical interview skill of raters as well as how reliably raters apply scoring criteria. It has been used in structured interview guide development, training, and active monitoring of rater performance in a clinical trial.[2]

We sought to evaluate the level of clinical interview skill of raters in a multicenter clinical drug trial evaluating a compound to treat depression, and to determine if clinical interview skill, as measured by the RAPS, is related to scoring accuracy. We also evaluated the relationship between the severity of illness, as measured by MADRS total score, and RAPS performance.

Methods

The RAPS measures six domains of clinical assessment: Adherence to scale administration guidelines, Follow-up questioning, Clarification of ambiguous responses, Neutrality, Rapport, and scoring Accuracy. Table 1 provides abbreviated

Each domain is rated as Excellent, Good, Fair, or Unsatisfactory. Table 2 provides abbreviated criteria for judging RAPS performance.

Table 2: RAPS Ratings

UNSATISFACTORY	consistently poor performance or any systematic deviation that would compromise the validity of the assessment, e.g., skipping an item, repeatedly failing to follow up and/or clarify critical information, consistently failing to listen to information that would alter a rating, rushing the subject or repeatedly challenging negative answers.
FAIR	several marked deviations or omissions such as paraphrasing required probes rather than asking verbatim, changing the wording on several questions so that the meaning of the item is slightly altered, making reassuring comments around the subject's hope for treatment, failing to clarify ambiguous responses.
GOOD	less than optimal performance on several items. Taken together, errors do not result in more than 1 or 2 item scores being difficult to verify due to insufficient information. Good is distinguished from Excellent by the frequency and/or significance of errors.
EXCELLENT	a high level of performance throughout, e.g., thorough and consistent clarification of ambiguous information; rapport is neither too therapeutic nor too rigid; asks all necessary follow up questions. An Excellent is warranted if the Rater makes a couple of minor, inconsequential, or insignificant errors.

Using domain ratings and pass/fail criteria developed for this analysis, we analyzed individual domain and total RAPS scores of 585 audio recorded interviews of the Structured Interview Guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA) in a randomized, double blind placebo-controlled depression trial. Audio recorded interviews were conducted by remote blinded raters over the telephone. The RAPS was conducted by clinical specialists who were trained and calibrated on the scale by one of the scale authors (Engelhardt).

We also evaluated the relationship of depression severity with RAPS performance in a sample of 550 audio recorded interviews with non-missing value MADRS assessments. Depression severity was represented by total MADRS scores. In one analysis, we split the sample according to the median MADRS total score of 23. In a second analysis we split the sample according to clinically established cutoffs for severity. The sample was split into a low severity range (≤ 12) and a moderate-severe range (≥ 28).

72% probability that raters who were Good or Excellent in Follow Up met criteria for passing Accuracy (n = 512). However, for those raters who received Fair or Unsatisfactory in Follow Up (n = 72), the probability of passing Accuracy was 44% and not substantially different than the probability of failing Accuracy (56%).

Relationship between severity of illness and RAPS performance was evaluated in two separate analyses using different MADRS total score cutoffs. RAPS Pass/Fail rates by MADRS total score are show in Table 3. The RAPS pass rate for low severity (≤ 12) was 91.4% while the moderate-severe group (≥ 28) was 77.3%. The difference between the two groups was statistically significant, with $z = 3.29$, $p<.001$.

Table 3: RAPS Pass/Fail Rate by MADRS Total Score Severity

MADRS total score	Pass	Fail	n
≤ 23	84.1%	15.9%	277
≥ 23	77.0%	23.0%	287
≤ 12	91.4%	8.6%	116
≥ 28	77.3%	22.7%	172

Conclusion

The majority of remote raters in this sample demonstrated adequate clinical interview skill and scoring accuracy. Remote raters, in addition to being blind to study visit and protocol, received extensive training and regular, ongoing calibration, which may differentiate their rating performance from site raters. Another study found that lower interview quality, as measured by the RAPS, was associated with greater scoring discrepancies between site and remote raters.[3] Raters who engage in appropriate use of follow-up questions to elicit sufficient information tend to score more accurately than raters who do not. Rapport, thought to be critical in mitigating placebo response, was not related to RAPS

Blinded Data Analytics (BDA): Overview

BDA vs Risk-based Monitoring

- Here, BDA refers to the review of accumulating blinded data in ongoing trials:
  - Primary and key secondary endpoints
  - Identifying outlier performance at the site or rater level
  - Support study management decision making

BDA: Service vs. Sponsor

- Most eCOA vendors offer some form of blinded oversight
  - Rater performance, KPI/KRI ("Flags"), endpoint trajectory
- The sponsor remains the decision-maker
  - Integrate data from multiple sources
  - Balance risk/benefit of action

BDA: Factors Influencing Sponsor Action

Blinded Data is Incomplete

- Looking for outliers in blinded data → site or study outcomes
  - Significant outliers may invalidate statistical assumptions
  - Established risk factors such as erratic raters, low baseline acuity
- Once you've taken the BDA out of the box you can't put it back
  - Inaction becomes an action

BDA: Factors Influencing Sponsor Action

Blinded Data is Incomplete

- Looking for outliers in blinded data → site or study outcomes
  - Significant outliers may invalidate statistical assumptions
  - Established risk factors such as erratic raters, low baseline acuity
- Once you've taken the BDA out of the box you can't put it back
  - Inaction becomes an action

Decisions Not Made in a Vacuum

- Pressure to meet study timelines
  - Stopping sites → reduced enrollment
  - Go slower (GASP!) or reallocate
- Operational limitations
  - Disqualifying raters
  - Managing site relationships
- Limited opportunities to intervene
  - GET IT RIGHT, but without all the data

BDA: Pragmatic Examples

A Phase 3 Acute Schizophrenia Program

- How we monitored blinded endpoint data in real time
- Examples of decisions made
  - Were we right or not (answer: sometimes)?

Trial Details

- Acute 5-week inpatient study in adults with schizophrenia
  - Minimum PANSS score of 80
- Highly statistically significant and clinically meaningful effect of drug vs placebo

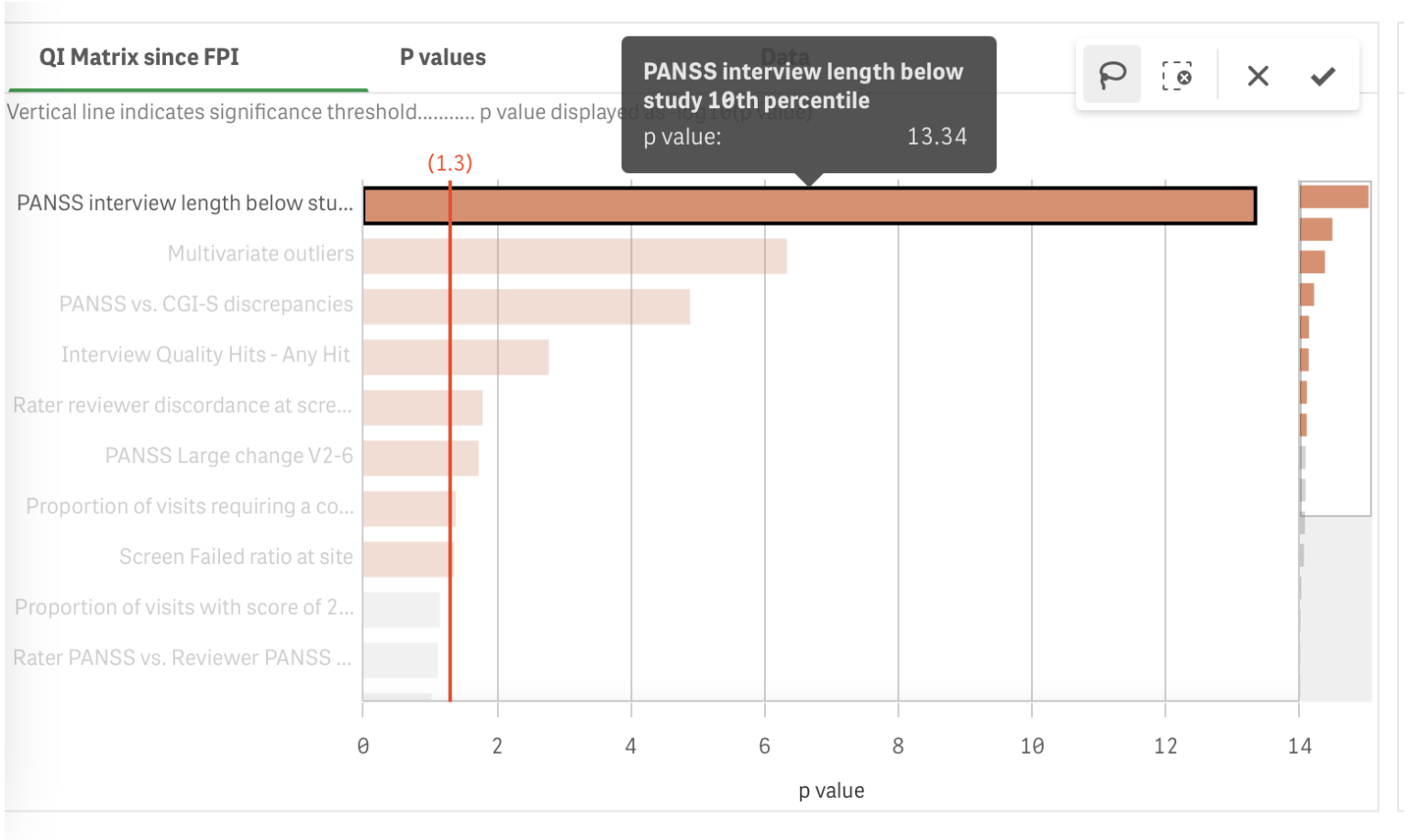
BDA: Pragmatic Example (Three Sites)

	Study Average	Site A	Site B	Site C
Screened	**407**	9	17	23
SCF Rate	**38%**	33%	41%	30%
ET Rate	**23%**	0%	20%	19%
Age (Years)	**46**	40	48	42
PANSS Total	**98**	106	100	98

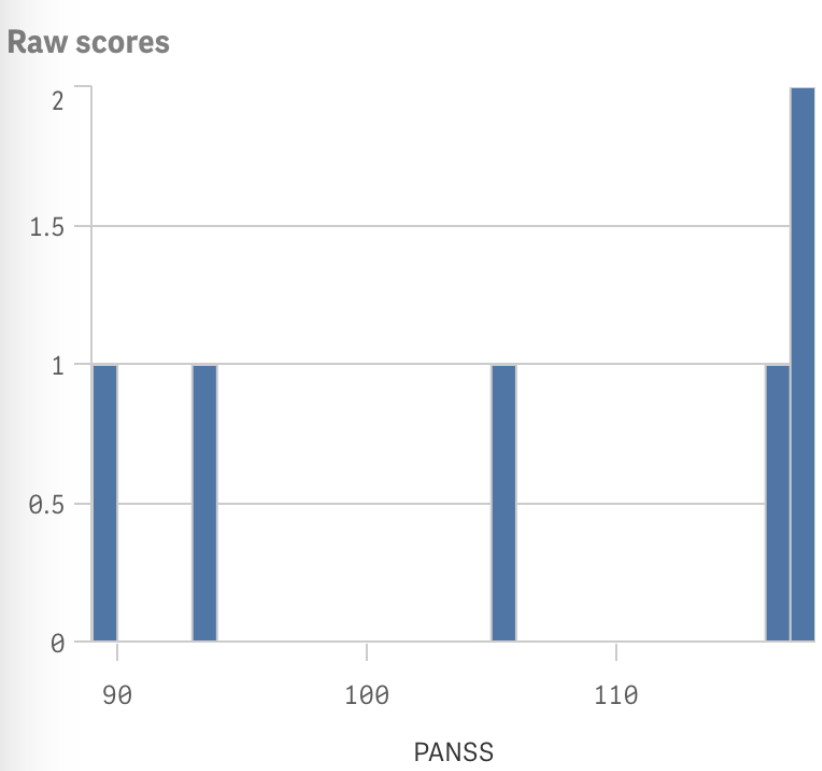
BDA: Pragmatic Example (Three Sites)

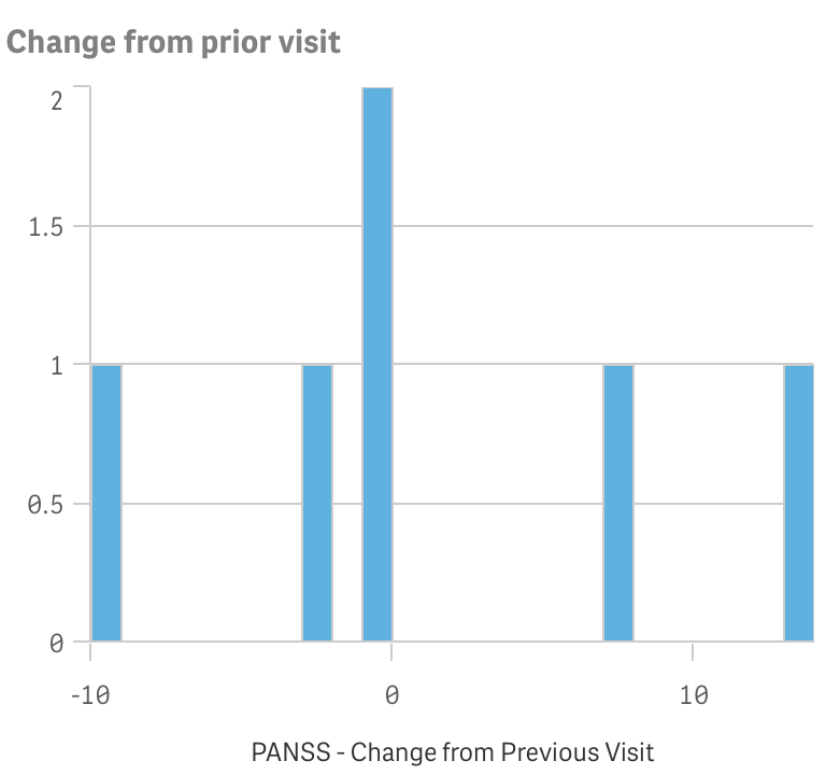
	Study Average	Site A	Site B	Site C
Screened	**407**	9	17	23
SCF Rate	**38%**	33%	41%	30%
ET Rate	**23%**	0%	20%	19%
Age (Years)	**46**	40	48	42
PANSS Total	**98**	106	100	98

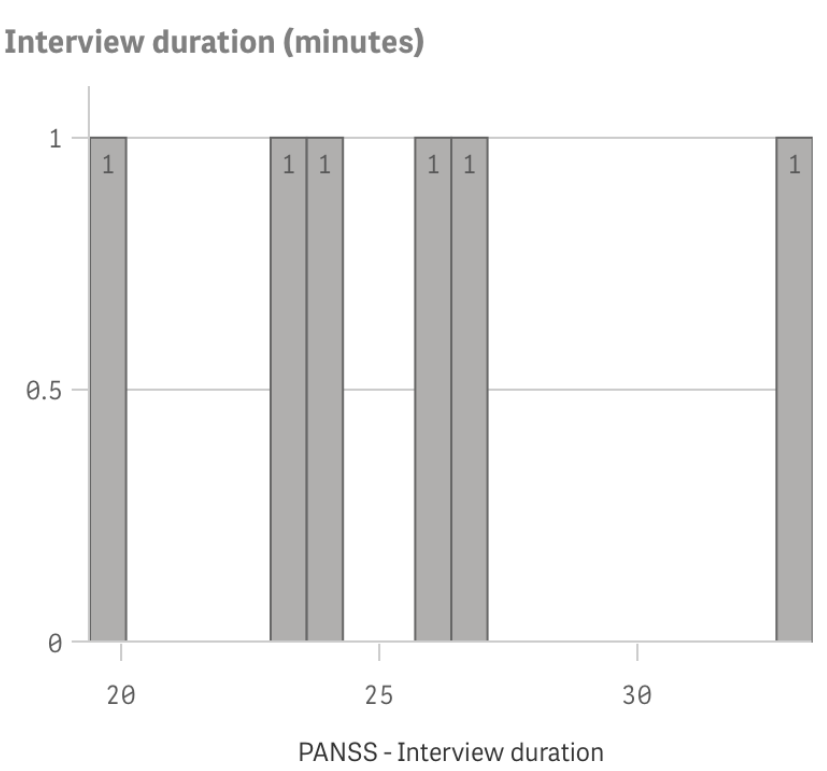
Site A: Flagged by eCOA Vendor



Site A: Baseline eCOA Performance







Site A: Additional Context

Study screening opened January 15th

- On that day, five subjects were screened by Site A

4. Subject is experiencing an acute exacerbation or relapse of psychotic symptoms, with onset less than 2 months before screening.
   a. The subject requires hospitalization for this acute exacerbation or relapse of psychotic symptoms.

Site A: Additional Context

Study screening opened January 15th

- On that day, five subjects were screened by Site A

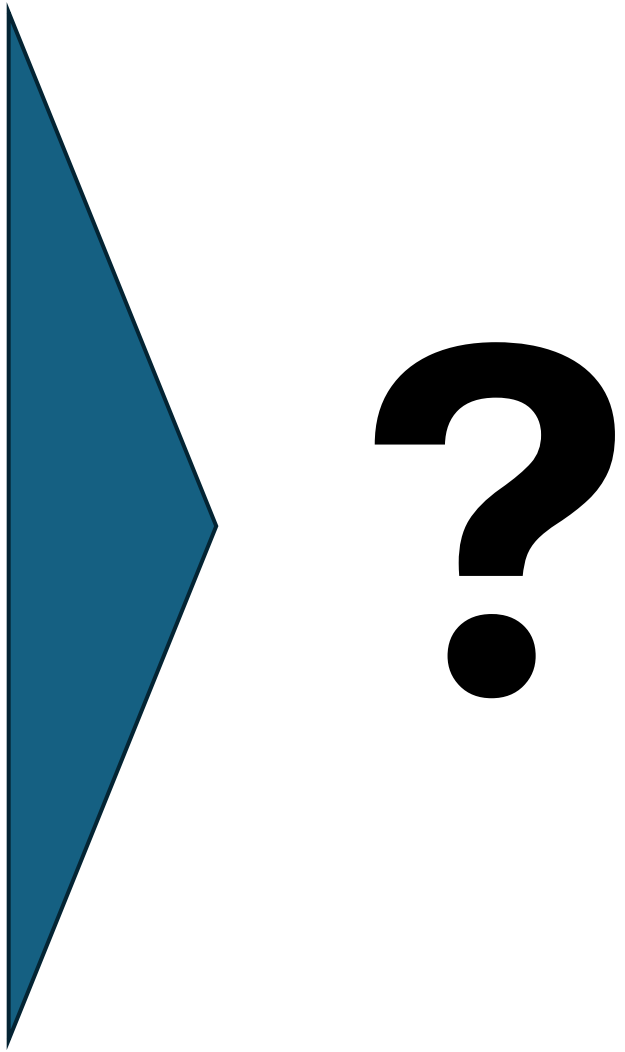
4. Subject is experiencing an acute exacerbation or relapse of psychotic symptoms, with onset less than 2 months before screening.
   a. The subject requires hospitalization for this acute exacerbation or relapse of psychotic symptoms.

Abnormalities in eCOA recordings

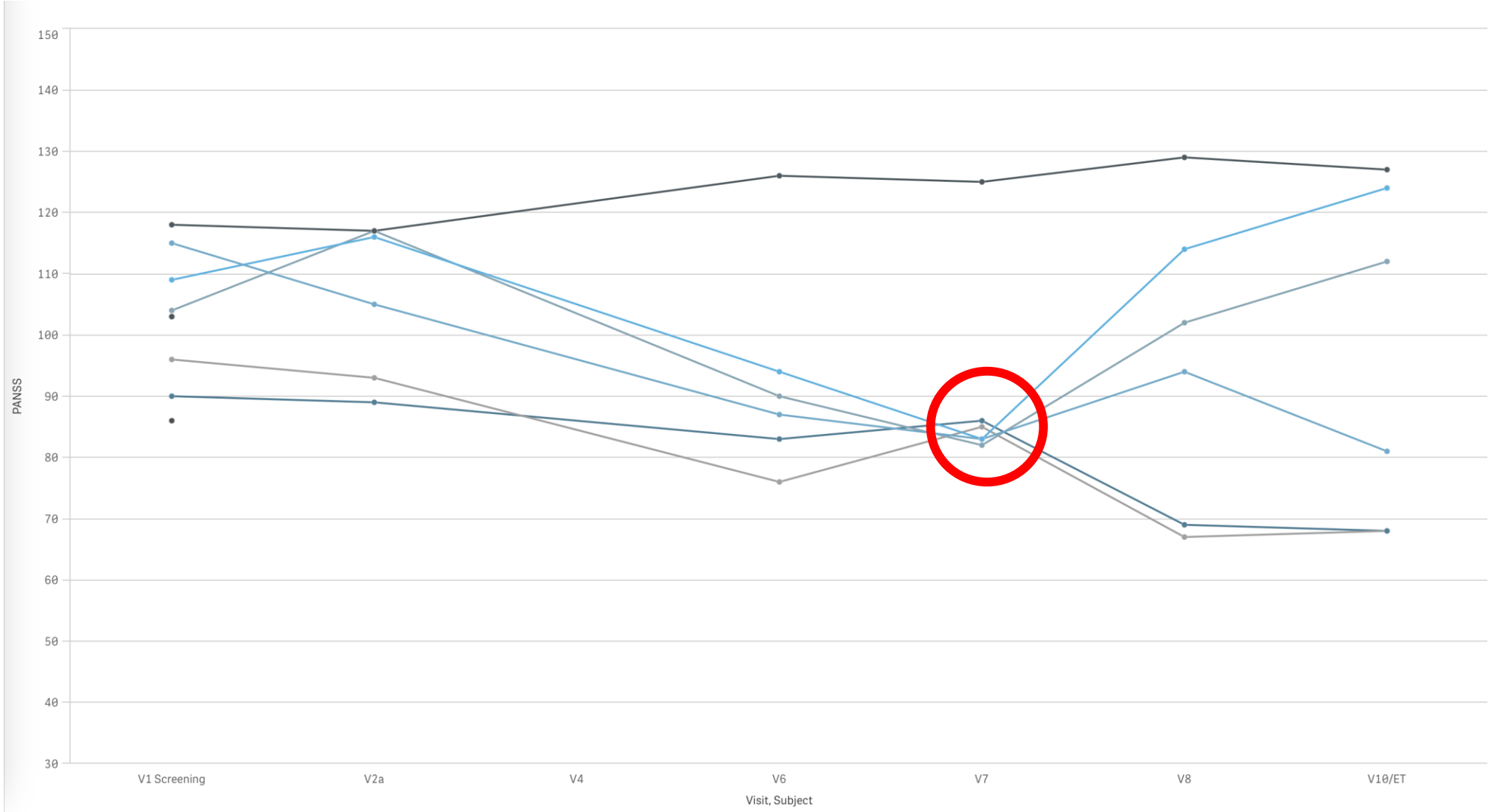
- Stopping and restarting interview at key moments

Previously worked with the site…

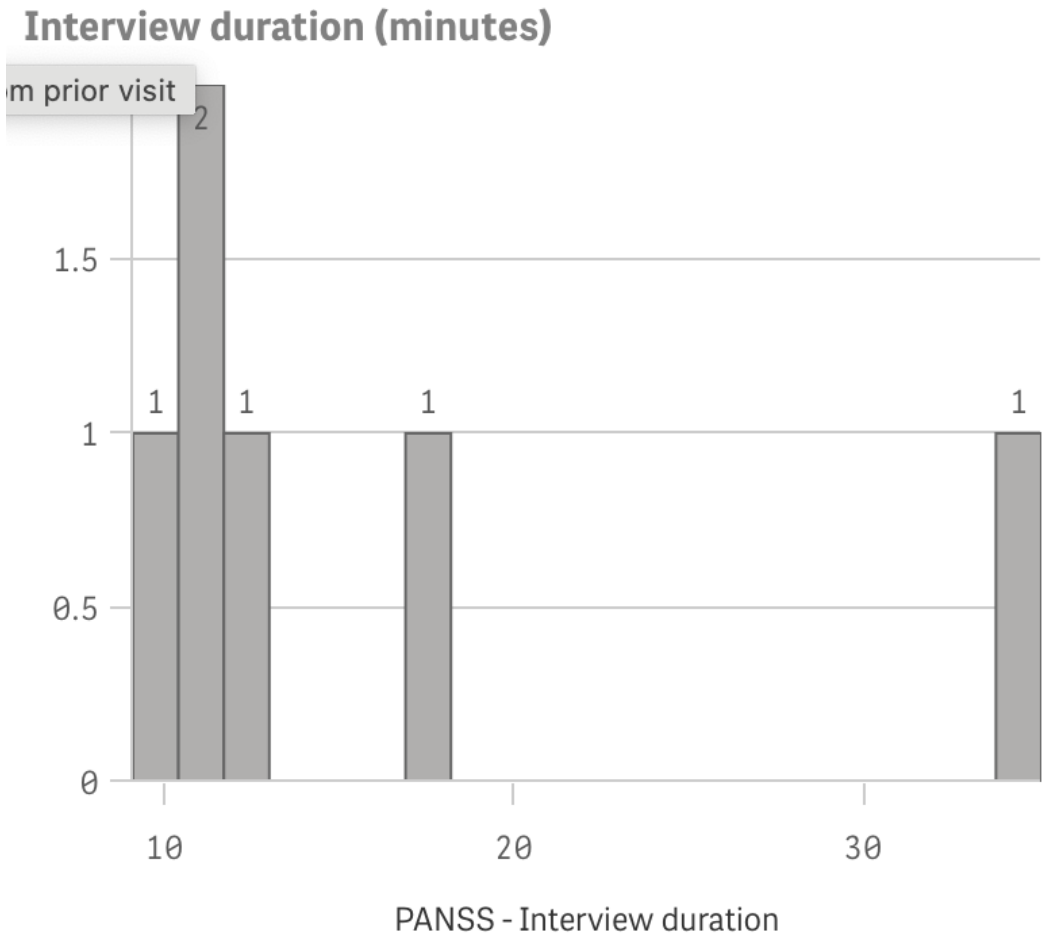
- But insisted on a PI change coming from Phase 2 to Phase 3



Site A: Treatment Progression



Site A: Visit 7 PANSS Duration



Site A: Decision

Decision: Site was closed to further enrollment

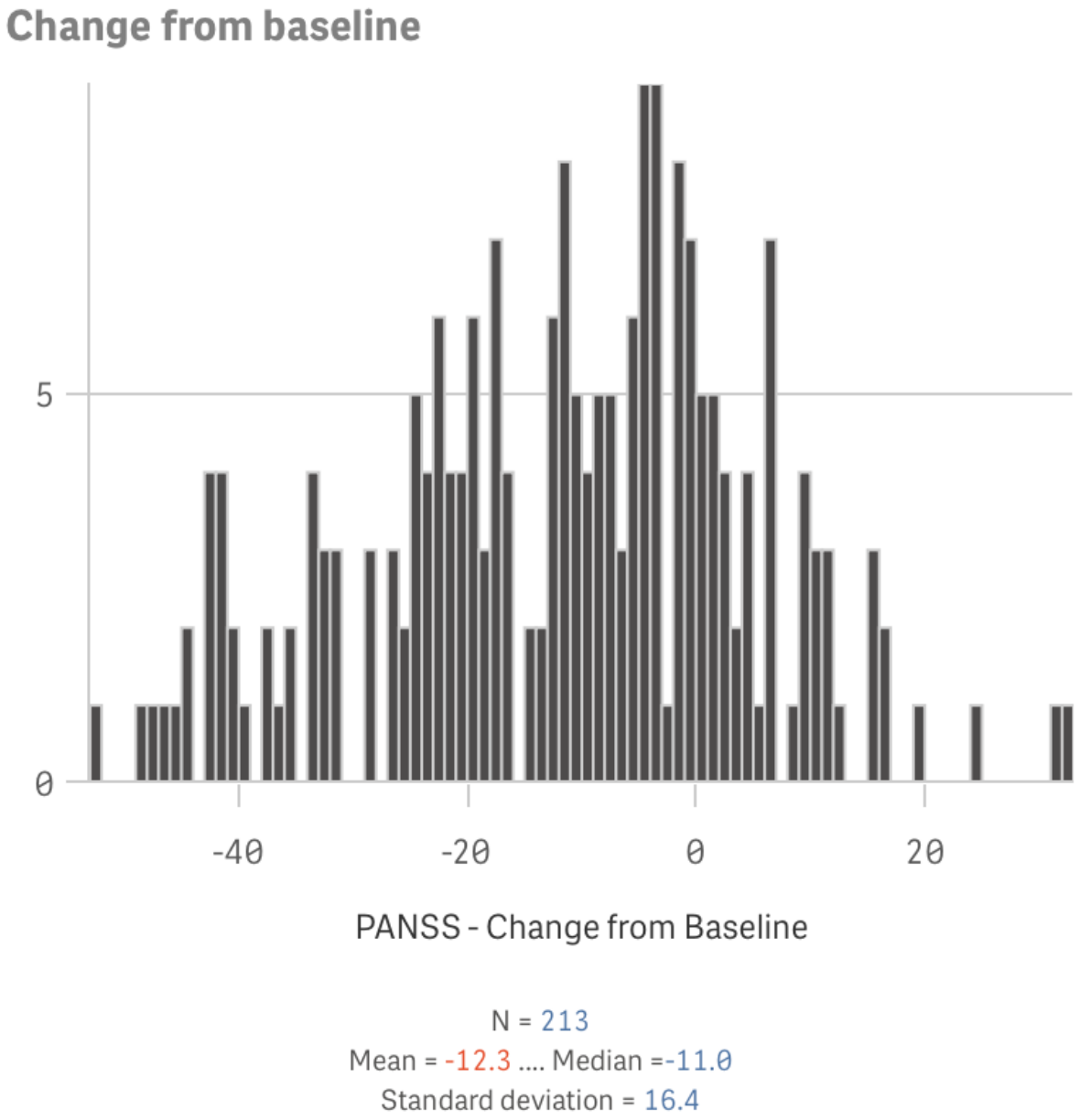
Factors impacting decision

- Site enrollment vs study enrollment
- Previous experience with the site
- Multiple indicators of potential poor quality

Alternative options considered

- Disqualify the rater
- Pace enrollment and continue to monitor

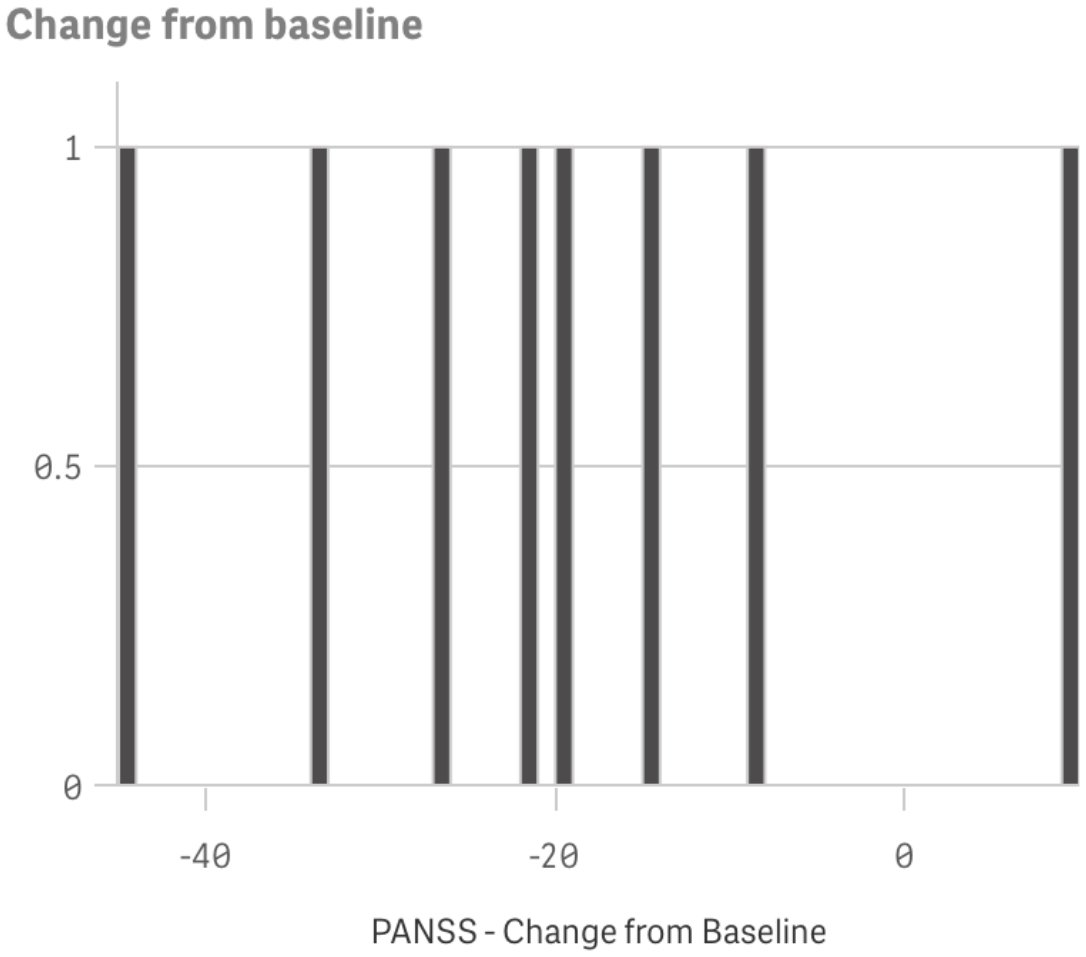
Study-wide CFB at Endpoint



BDA: Pragmatic Example (Three Sites)

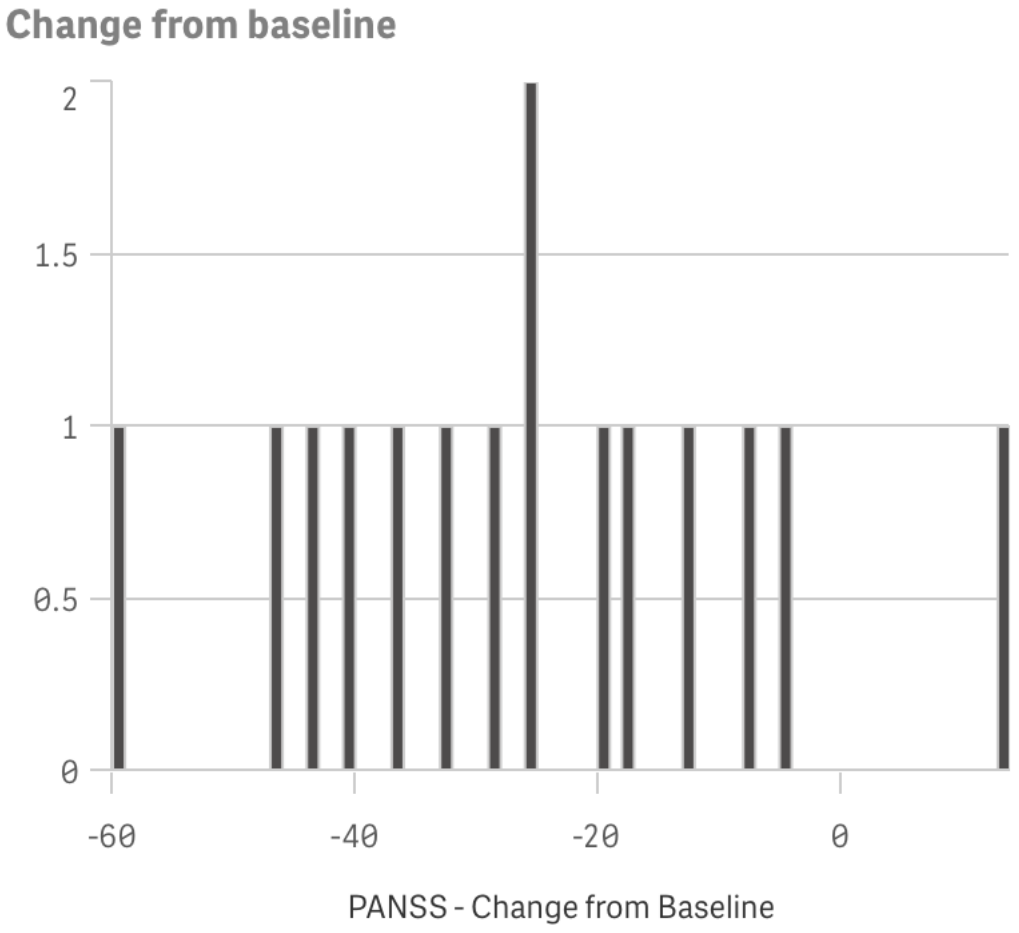
	Study Average	Site A	Site B	Site C
Screened	**407**	9	17	23
SCF Rate	**38%**	33%	41%	30%
ET Rate	**23%**	0%	20%	19%
Age (Years)	**46**	40	48	42
PANSS Total	**98**	106	100	98

Site B



N = 8
Mean = -20.4 .... Median =-21.0
Standard deviation = 16.3

Site C



N = 15
Mean = -26.3 .... Median =-26.0
Standard deviation = 18.7

Site B/C: Context

Site B & C were generally good performers

- eCOA BDA flags were, if present, not overly concerning
- Rater performance was adequate to good
- Baseline data was consistent with study average

Both sites show mean change larger than study average

- Risk is potential placebo response
- Site B is new to the program, whereas Site C participated in Phase 2

Site B/C Decisions

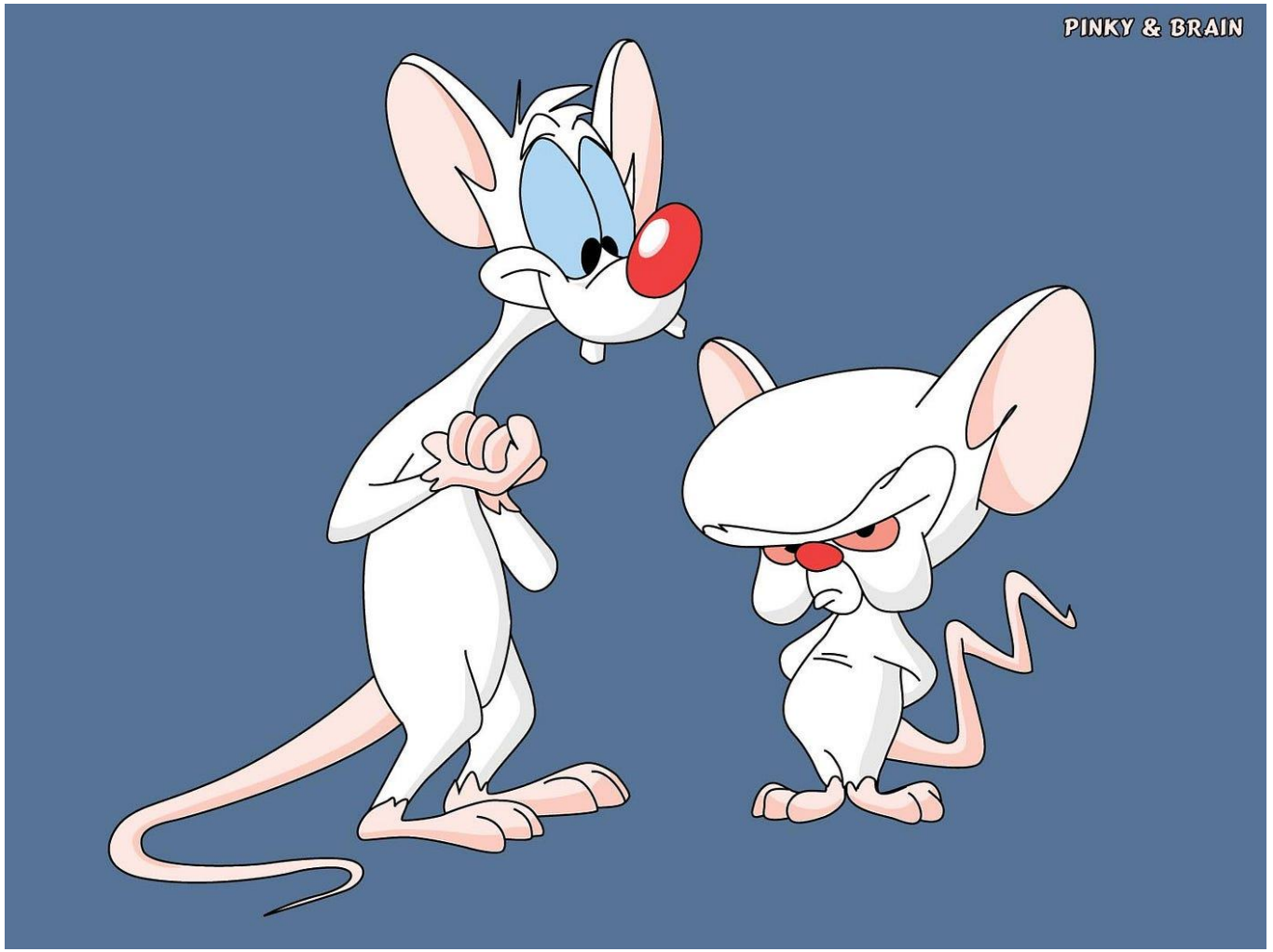
Site B closed to enrollment

- New site showing a pattern of response that may be indicative of placebo response
- Intervention would introduce bias toward non-response
- Study-wide enrollment was ahead of projections

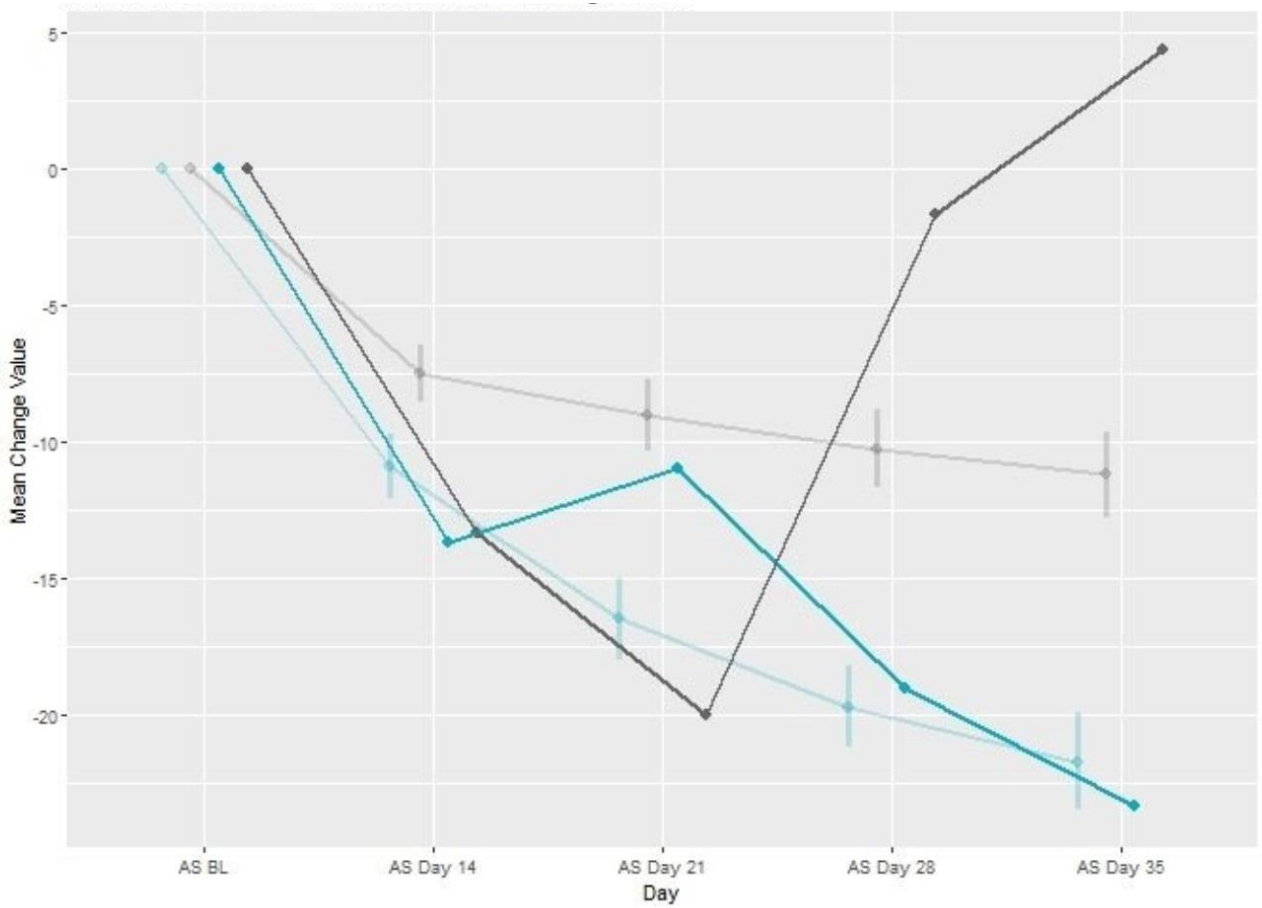
Site C continued enrollment

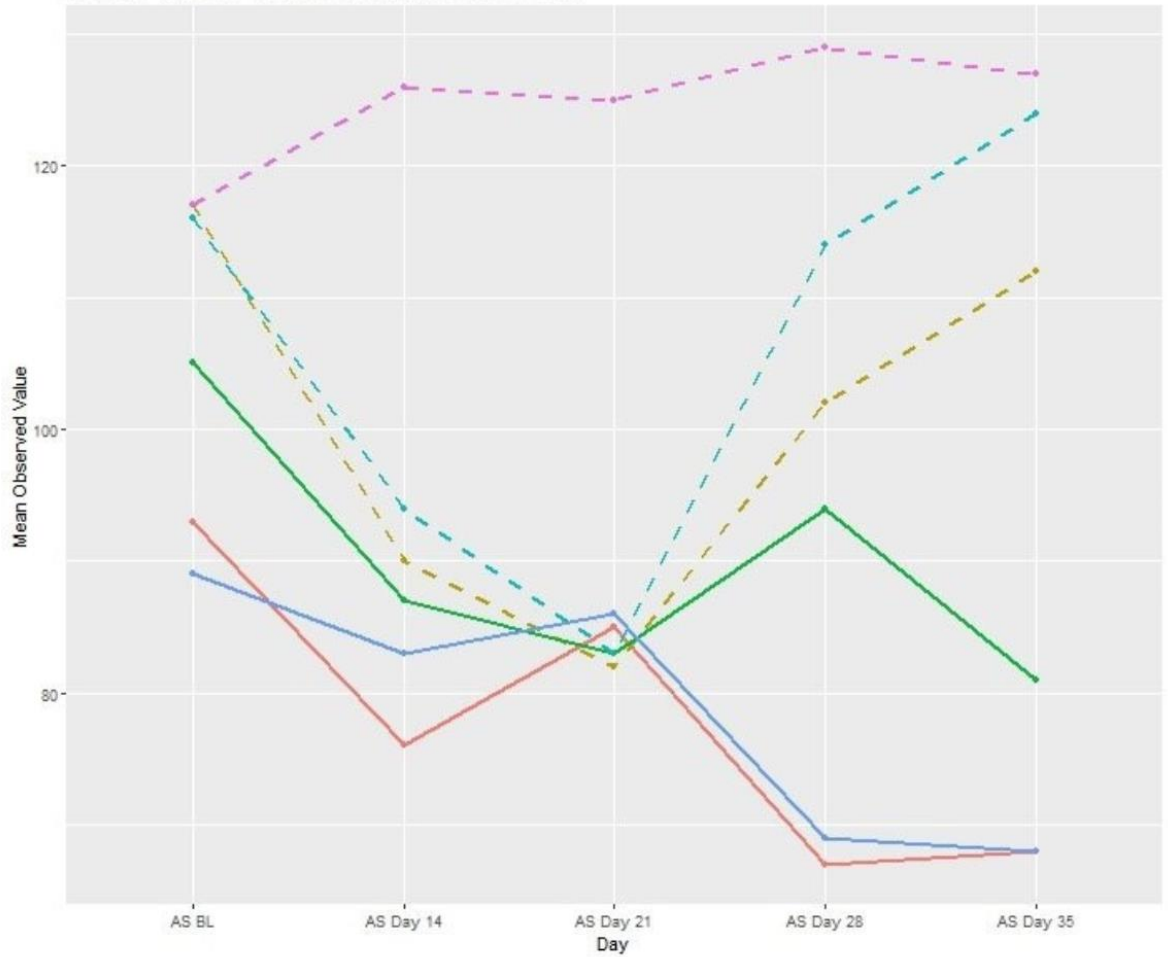
- Phase 2 showed a similar pattern of response, but good drug/placebo separation
- No other risk indicators at this site

So... Genius (Brain) or Idiot (Pinky)?

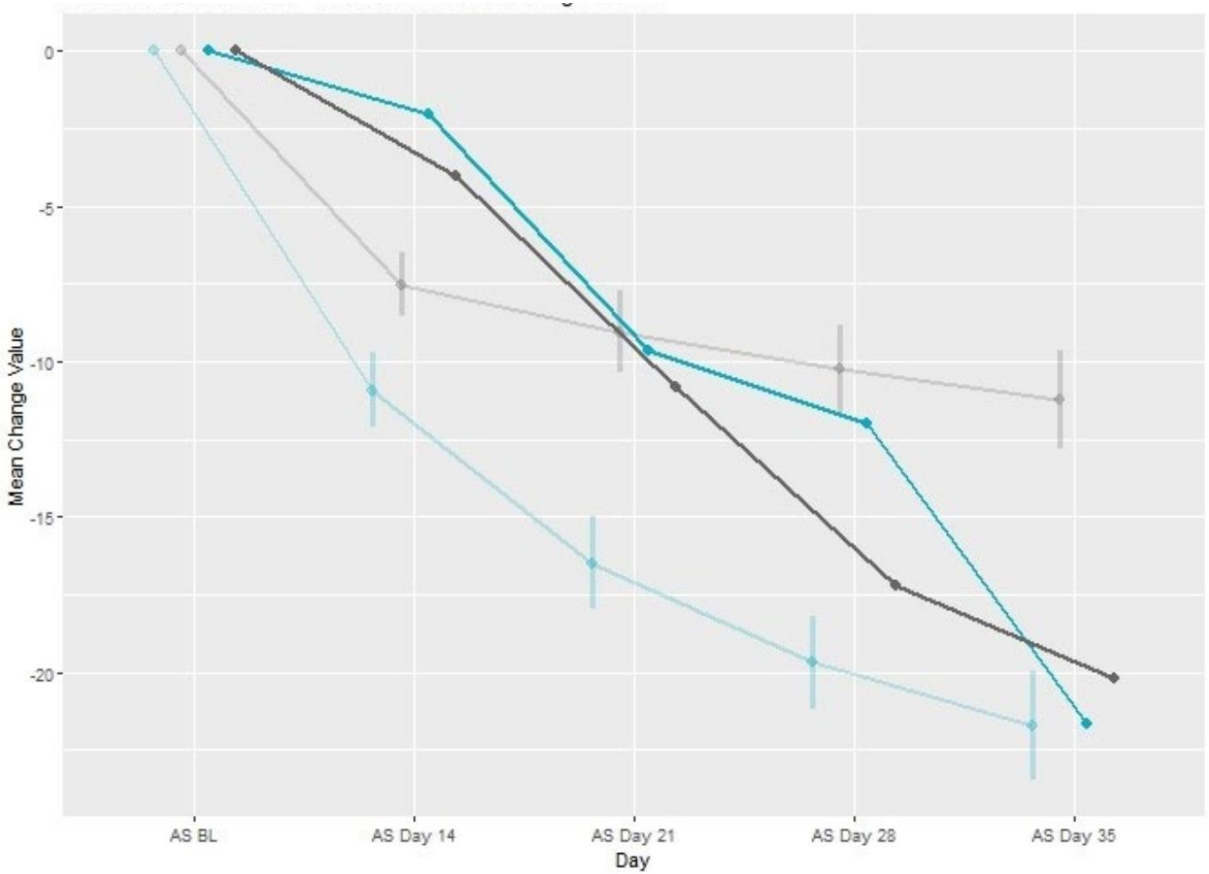


Site A: Outcome

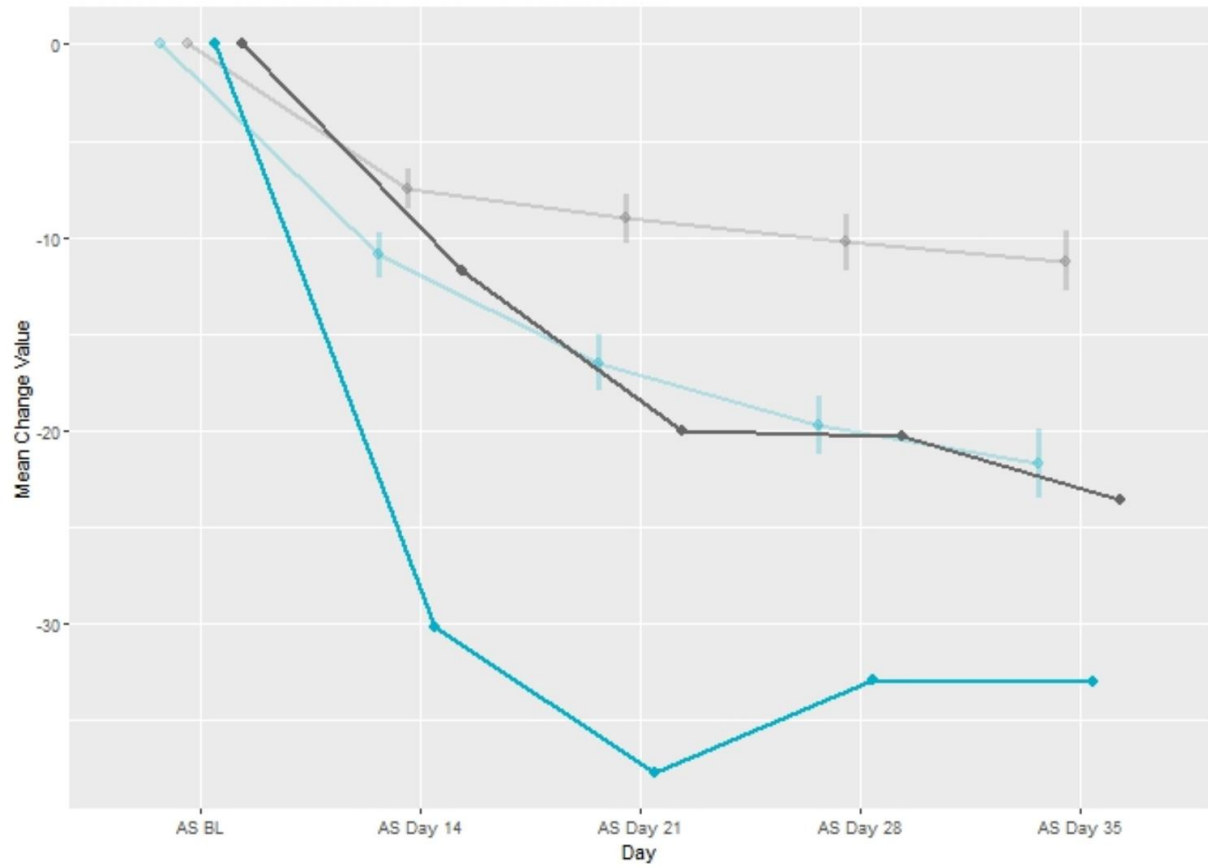




Site B


0
-5
-10
-15
-20
Mean Change Value
AS BL
AS Day 14
AS Day 21
AS Day 28
AS Day 35
Day

Site C


0
-10
-20
-30
Mean Change Value
AS BL
AS Day 14
AS Day 21
AS Day 28
AS Day 35
Day

In Summary

Making decisions based on BDA is <u>HARD</u>

- Once you begin reviewing blinded data, inaction is an action…

- Decisions to pause or end enrollment often conflict with pressure to complete studies on time
  - Replacement sites may not be better performers

- Decisions based on BDA are **<u>BEST GUESSES</u>**