The Impact of Site Characteristics on Efficacy Measures

Bethanne Friedmann, PsyD[1] , Erin B. Kornsey, MS[1], Kathryn Dawson, PhD[1], Henry J. Riordan, PhD[1], Michael F. Murphy, MD, PhD[1], Neal R. Cutler, MD[2]

[1] WORLDWIDE CLINICAL TRIALS, King of Prussia, PA, [2] WORLDWIDE CLINICAL TRIALS, Beverly Hills, CA

Abstract

Introduction

The success of psychiatric trials hinges on the quality of data collected by sites. Data errors and discrepancies between the key efficacy measures of a double-blind, randomized, placebo-controlled, multicenter, phase II clinical trial designed for adult attention deficit/hyperactivity disorder (ADHD), were closely monitored by a panel of clinical experts.

Methods

Data "flags" were based on various scales including the Clinical Global Impression of Severity (CGI-S), Clinical Global Impression of Improvement (CGI-I), Conners' Adult ADHD Rating Scale – Observer: Screening Version (CAARS-O:SV), and Conners' Adult ADHD Rating Scale –Self Report: Short Version (CAARS-S:S). Reports were generated twice a week based on information sites electronically entered since the last report was sent. These reports contained discrepancies between the scales, data entry errors, and rater errors. The clinical manager contacted the sites to gather more information about these data flags via telephone and email.

Results

Sixteen sites with randomized subjects for the trial and were included for analysis (mean number of flags 11.38, standard deviation 12. 47). The overall number of data flags per site was negatively correlated with both the number of subjects screened (r = - 0.45) and number of subjects randomized per site (r = -0.19). Prior studies have suggested that the number of flags per site were proportional to the number of subjects enrolled per site. Because there is evidence this was not the case, sites were divided into 2 categories: sites with a high rate of flags per randomized subject and sites with a low rate of flags per randomized subject. A significant difference in the mean number of subjects randomized by the data flag rates was determined (t = 2.43, p = 0.03) with a higher mean number of randomized subjects for site with lower flag rates. The severity of the data flags were also assigned numeric ratings from 1 (least severe, e.g. data entry error, missing data needs to be entered; n = 116), 2 (moderately severe, e.g. possible incorrect rater completing assessment; n= 59), to 3 (most severe, possible scale discordance; n = 7). This has important implications for the appraisal of sites with lower patient screen and randomization numbers.

Conclusion

By investigating and tracking the frequency and severity of the various flags over the course of the study it is possible to enhance the overall quality of data and ultimately lead to increased effect sizes.

Background

- Data errors and discrepancies between the key efficacy measures in double-blind, randomized, placebo-controlled trials detract from the quality of data and ability to detect separation from placebo.
- It was hypothesized that high enrolling sites could be susceptible to high error rates.
- Alternatively, a low number of data related errors may reflect the expertise gained from repeatedly administering the assessments at the high enrolling sites.

Methods

- Reports were generated twice a week based on new electronically entered data. These reports listed discrepancies between the scales, data entry errors, and rater errors (data "flags").
- Flags were based on efficacy measures including the CGI-S, CGI-I, CAARS-O:SV, and CAARS-S:S.
- The clinical manager contacted the sites to gather more information about the data flags and when necessary to re-educate the site.
- The severity of flags was divided into three levels.

Table 1. Description of Flag Severity

Least Severe	Moderately Severe	Most Severe
• Basic data entry errors, missing data • For example: CGI-S and CGI-I were not entered; one CAARS-O:SV subject left blank .	• Possible un-blinding of rater • For example: rater's initials on CAARS-O:SV and CAARS-S:S were GMJ. The CAARS INV rater was blinded to all other measures and therefore, should not have conducted the CAARS –S:S.	• Possible scale discordance; rater inflation • For example: from baseline to week 14 CAARS-INV decreased from 68-31; CAARS –S:S decreased from 58-30. CGI-S remained 4 from baseline.

Results

- Analysis was conducted on sites with randomized subjects (n=16). The mean number of flags for the sites was 11.38 (sd = 12. 47) with a total number of 182 flags for the current study.

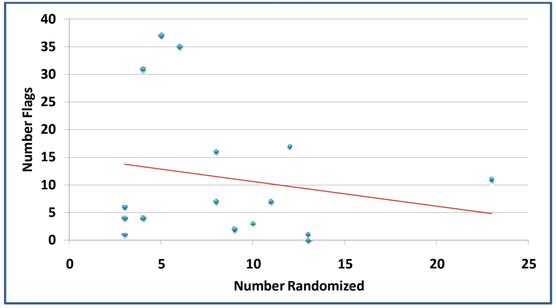
Table 2. Frequency of Randomized Subjects and Flags per Site

Site	Number Randomized	Number of Flags	Low Severity Flags	Moderate Severity Flags	Severe Flags
1	8	16	3	13	0
2	3	1	1	0	0
3	6	35	28	4	3
4	13	1	0	0	1
5	3	4	4	0	0
6	23	11	11	0	0
7	10	3	3	0	0
8	8	7	2	5	0
9	9	2	1	1	0
10	3	6	0	6	0
11	12	17	17	0	0
12	5	37	21	16	0
13	13	0	0	0	0
14	4	31	19	11	1
15	11	7	3	3	1
16	4	4	3	0	1

Results continued

- The overall number of data flags per site was negatively correlated with both the number of subjects screened (r = - 0.45) and number of subjects randomized per site (r = -0.19).

Figure 1. Number of observed flags by randomized subjects



- Sites were divided into 2 categories: high incidents of flags (as defined by ≥ 2 flags/randomized subject) and low incidents of flags per randomized subject.
- There was a significantly larger mean number of randomized subjects associated with 11 sites with a low flag rate (9.9±5.75) when compared to the 5 sites with the high flag rate (5.2±1.92) (t = 2.43, p = 0.03).
- There was no difference in the distribution of the severity of the flags by the number of randomized subjects (p = 0.871).

Table 3. Severity of Flags by Number of Randomized Subjects

	Least Severe	Moderately Severe	Most Severe
< 8 Randomized	6 (46%)	4 (31%)	3 (23%)
≥ 8 Randomized	7 (54%)	4 (31%)	2 (15%)

Conclusions

- Sites with the highest enrollment had the least number of flags which is counter to some past findings and our expectations.
- The distribution of severity of these flags did not differ by number of randomized subjects using a median split.
- Importantly, the number of flags can be decreased during the course of trial with expert rater feedback.
- More research is needed to determine if sites with the lowest number of flags, in particular the lowest number of severe flags, were better able to detect drug-placebo differences than those with higher number of flags.

No conflicts of interest exist in the research and development of this poster