Gary Sachs,1, 2 Susan Edman,1 Dan DeBonis,1 Douglas Vanderburg3 Why Do Clinical Trials Fail? Learning From Computer Administered Assessments

1Bracket, Lexington, MA, USA; 2Massachusetts General Hospital, Boston, MA, USA; 3Pfizer Inc., New York, NY, USA.
Why Do Clinical Trials Fail? Learning From Computer Administered Assessments

Gary Sachs1,2; Susan Edman1; Dan DeBonis1; Douglas Vanderburg3
1Bracket, Lexington, MA, USA; 2Massachusetts General Hospital, Boston, MA, USA; 3Pfizer Inc., New York, New York, USA

PURPOSE

The high failure rate of randomized controlled trials (RCT) is a well recognized obstacle to drug development, but remains poorly understood [1]. We report exploratory analyses utilizing data collected by interactive computer interviews to examine the impact of protocol specified eligibility criteria and rating reliability on signal detection in a Bipolar Depression RCT.

METHODS

These clinical trial methodology (CTM) analyses were undertaken as part of a failed randomized controlled trial. The efficacy study was a double-blind placebo controlled trial designed to test adjunctive ziprasdone as treatment for acute depression in 265 bipolar I subjects. Based on assessments administered by the site-based rater (SBR), subjects met all requirements for randomization including DSM IV diagnostic criteria and baseline HAM-D scores ≥20. In addition to the efficacy study assessments, subjects completed independent computer administered assessments for diagnostic confidence and symptom severity (HAM-D_COMP, MADRS_COMP, YMRS_COMP) prior to randomization and at least one postrandomization assessment on the MADRS.

The CTM study compared drug-placebo differences on change from baseline MADRS_COMP and MADRS_SBR (primary outcome measure) for subgroups created based on the computer assessments including protocol specified baseline severity criteria (HAM-D_COMP ≥20, Diagnostic Confidence, Presence of Mixed episode, Baseline Inflation, and Overzealous subject reporting). Diagnostic confidence was measured using the Bipolarity index [2]. A mixed episode was diagnosed, if at least 3 items on the computer administered YMRS were rated to be of clinically significant severity (defined as a score of ≥4 on items 5, 6, 8, and 9 or a score ≥3 on the other YMRS items.). Rating reliability was defined as poor and attributed to Baseline inflation by the SBR if (MADRS_COMP − MADRS_SBR) ≤ -10, or attributed to Overzealous Reporting by the Subject if (MADRS_COMP − MADRS_SBR) ≥ +10.

CTM signal detection analyses used Stata version 11.0 statistical software. Since the efficacy study was powered to detect a drug-placebo difference of ≥4.0 points on change from baseline MADRS, but was not powered for these exploratory analyses, criteria which resulted in more than 1.0 difference in the drug-placebo signal on both the SBR and Comp ratings were arbitrarily defined as “impactful” prior to the data analysis.

RESULTS

Figure 1: Impact of Key Eligibility Criteria: Active-Placebo Difference in Mean Improvement from Baseline MADRS

Figure 2: Rater Performance at Baseline: A Quantitative Quality Metric

Figure 3: Impact of Mixed Episodes

CONCLUSIONS

The main findings of the exploratory CTM analyses, was the observation of “impactful” but nonsignificant trends favoring placebo in the subgroups with mixed episodes rather than bipolar depression and among two subgroups with large discordance between MADRS_COMP and MADRS_SBR at baseline. This is consistent with the suggestion that unreliable ratings by subjects as well as SBRs contribute to study failure. These results suggest more stringent subject selection processes may improve RCT signal detection, but do not support the practice of qualifying subjects based on a threshold score obtained on a different depression scale than that used for the primary outcome.

REFERENCES
