Impact of Centralized Over-Read on Outcomes in Depression and Schizophrenia Trials

Echevarria B[1], Negash S[1], Ropacki MT[1]
[1] MedAvante, Inc.

The Methodological Question Being Addressed

Centralized over-read (Central Review) of clinical assessments in depression and schizophrenia trials administered at key visits (e.g., screening, baseline and end of study) is intended to improve inter-rater reliability and reduce scoring variability. We investigated the impact of performing Central Review on two widely used efficacy measures administered at key visits in schizophrenia and depression trials, the Positive and Negative Syndrome Scale (PANSS) and Hamilton Depression Rating Scale (HAM-D).

Introduction (Aims)

Scoring variability and poor inter-rater reliability have been reported to contribute to high placebo response rates and inconclusive results in clinical trials[1], and rater training has been demonstrated to be insufficient to remedy these challenges[2]. Quality assurance measures such as Central Review methodology have been implemented to avoid rater drift and improve inter-rater reliability, but the impact of this approach on scoring variability and rater agreement in PANSS and HAM-D has not been adequately explored.

Methods

Data from PANSS and HAM-D assessments in two randomized, double-blind, multisite schizophrenia and depression clinical trials were analyzed. The scales were completed by raters rigorously trained on administration and scoring conventions, then qualified by successful completion of a rater precision exercise prior to conducting in-study assessments. A cohort of expert calibrated Central Reviewer clinicians examined video/audio recordings and source documents to identify raters' administration and scoring errors. When scoring discrepancies were identified, raters were given an option to either agree with the reviewer's feedback and change the score, or provide rationale for rejecting the feedback and maintain the original score. Number of discrepancies and score changes were examined for screening, baseline and end of study visits.

Results

A total of 2057 PANSS and 358 HAM-D assessments were reviewed, and the average number of scoring discrepancies for each of the key study visits analyzed. Scoring discrepancies declined from screening (27%) to end of study visits (16%) for the PANSS, as well as HAM-D (40% to 23%, screening to week 8, respectively). Central Review resulted in a mean total score change on average of three points (PANSS) and two points (HAM-D), and total score change also decreased as raters began to score more accurately as the trial progressed.

Conclusions

Initial rater training is necessary, but not sufficient to ensure inter-rater reliability and prevent rater drift in schizophrenia and depression trials. Additional in-study oversight through Central Review by expert and highly calibrated external clinicians resulted in a positive impact on individual item scores and total scale score for key study visits. Reduced error variance translates to increase study power and a higher probability of signal detection in interventional trials.

References:

1. Khan A, Yavorsky WC, Liechti S, DiClemente G, Rothman B. Assessing the sources of unreliability (rater, subject, time-point) in a failed clinical trial using items of the Positive and Negative Syndrome Scale (PANSS). *J Clin Psychopharmacol*. 2013 Feb; 33(1):109-17.
2. Targum SD, Evaluating Rater Competency for CNS Clinical Trials. J Clin Psychopharmacol 2006; 26:308-310.