Assessing the sources of unreliability (rater, subject, time-point) of placebo responders using items of the Positive and Negative syndrome Scale (PANSS)

Authors: Khan A1,5, Yavorsky C1,2, Opler M1,3

Affiliations: 1ProPhase, LLC, 2CROnos CCS, 3New York University School of Medicine, 4Johns Hopkins University Bloomberg School of Public Health, 5Nathan S. Kline Institute for Psychiatric Research.

Abstract

Background: Failed trials are an acknowledged problem for the development of pharmacological treatments. A trial is considered failed when the active treatment does not differentiate from placebo. Contributing factors may include: escalating placebo response rates, dosing regimens, low sensitivity in the clinical scores used to evaluate efficacy and inconsistency in rating scores, especially in the case of multicenter trials. Interim monitoring of assessment tools, clinical outcomes, and measurement errors is recognized as an important tool for early decision-making.

Objectives: Using the principles of generalizability theory, the aim was to examine the extent to which each facet (i.e., raters, subjects, and time-points (visits)) contributed to the variability (and inconsistency) in scores of the PANSS for subjects identified as placebo responders.

Methods: This study was designed as a sub-study from a failed Phase II double-blind, placebo-controlled evaluation into the safety and efficacy of a new antipsychotic compound for the acute exacerbation of schizophrenia. The larger study consisted of 392 subjects with schizophrenia ages 18-65 years across 30 centers (United States). PANSS was completed 1 week apart for 8 weeks/time-point. Of the 130 placebo subjects, 71 completed all visits. There were 59 raters used in the entire study. Based on the 20% improvement, in PANSS scores, 15 patients were identified as placebo responders.

STATISTICAL ANALYSIS: Generalizability analysis was selected to examine the dependability of the performance of measurements and raters across time periods (Visit 1 to Visit 8). The reliability of symptom ratings was assessed by raters, time-points (visits), and subjects (individual items by subscale). The ultimate goal is to map out as many potential sources of variability as possible.

Results: INTER-RATER RELIABILITY: At Visit 1, ICC = 0.68 for the placebo group; ICC = 0.71 for the active treatment group. At endpoint, ICC = 0.75 for the placebo sample; ICC = 0.79 for the active treatment group. SOURCES OF MEASUREMENT ERROR: POSITIVE SYMPTOMS - Raters accounted for the most variability, ranging from 33.00% (P7. Hostility) to 72.00% (P1. Delusions and P4. Excitement), with an average of 60.71%. The next largest source of variance was attributed to the interaction between the rater and visit, with an average of 12.28%. SOURCES OF MEASUREMENT ERROR: NEGATIVE SYMPTOMS - Raters accounted for the most variability, ranging from 54.00% (N6. Lack of Spontaneity/Flow of Conversation) to 69.00% (N1. Blunted Affect and N5. Difficulty in Abstract Thinking), with an average of 62.29%. The next largest source of variance was attributed to the interaction between the rater and visit, with an average of 11.00%.

Conclusions: This study uniquely applied generalizability theory to assess the extent to which each facet (i.e., raters, subjects, and time-points (visits)) contributed to the variability and inconsistency in subscale and item scores of the PANSS for subjects receiving placebo treatment. The results provide the first demonstrable relationship between rater reliability and subsequent placebo response using clinical trial data. The significance of rater training for reliability and validity in clinical trials is recognized as vital. Findings can be used to guide data monitoring, rater training and identification of PANSS items which may require supplementary training.