Initial Findings of the Rater Quality Questionnaire (RQQ): A New Tool for Evaluating Quality of Ratings in Global Clinical Trials

David G. Daniel, MD¹ and Alan Kott, MD²

¹Bracket, McLean, VA, USA; ²Bracket, Prague, CZ
INTRODUCTION

As drug placebo differences have diminished over time in global schizophrenia clinical trials, the quality of interview and rating procedures performed by sites has come under increasing scrutiny (1-4). The Ratings Quality Questionnaire (RQQ) is a culturally neutral two item tool for evaluation of the quality of site ratings and diagnostic procedures. We describe the initial posted results of use of the RQQ in evaluation of interview and ratings quality in ten ongoing global or international schizophrenia clinical trials.

METHOD

Many of the studies comprising the pooled dataset are ongoing. Thus, the number of interviews evaluated with the RQQ has grown from 906 at the time the ECNP abstract was submitted to 1187 currently. The descriptive statistics have been updated accordingly.

A proprietary video/audio recording system was utilized to record PANSS rating procedures in ten schizophrenia clinical trials conducted in North America, Europe, South America and Asia. Sites uploaded video or audio recorded ratings assessments for review by calibrated external reviewers. External reviewers provided feedback on ratings quality on an ongoing basis to the site and Bracket. Prior to study initiation, raters were trained at investigator meetings by highly interactive procedures, including slide presentations, rating of videotaped patient interviews, and, in some cases, interview and rating of live actors trained to portray schizophrenia symptoms. Instruments used to assess ratings and interview quality included a new tool, the Rater Quality Questionnaire (RQQ) (5) and the Research Interview Assessment Scale (RIAS) (6).

The RQQ is a global measure of interview and ratings quality consisting of two domains addressing: 1) the overall quality of the patient and/or informant material collected; and 2) the overall quality of the ratings or diagnostic evaluation (including proper application of the rules and anchor points of the rating scale or structured interview) for the current subject visit. Each domain is rated on a Likert-like scale (1-3). A score of 1 represents no deficiencies in information gathering and/or scoring. A score of 2 represents deficiencies that have only a minimal impact on the scoring. A score of 3 reflects serious deficiencies in either collecting the information or rating the symptoms observed.

RESULTS

SUMMARY OF RESULTS I: For the RQQ analysis, the current pooled dataset consists of 1187 evaluated interviews. The mean RQQ scores for sufficiency of the data collected by interview and competency in use of the rating scales were 1.35±.51 and 1.31±.50, respectively.

SUMMARY OF RESULTS II: In 67.3% (n=799) of visits the patient and/or informant material required to rate the patient was classified as fully adequate; 30.6% (n=366) as some deficiencies but overall adequate; and 1.9% (n=23) as inadequate to rate the patient.

SUMMARY OF RESULTS III: In 71.1% (n=844) of visits the ratings were fall to reflect proper application of the anchor points and other instructions of the measurement tool; in 27.2% (n=323) the ratings generally, but not uniformly reflected proper application of the anchor points and other instructions of the measurement tool. 3.0% (n=36) of interviews had minor deficiencies in information gathering and/or instruction, which was scored using proper techniques, and 19.5% of interviews had minor deficiencies in both information gathering and scoring.

SUMMARY OF RESULTS IV: 97.5% of the interviews were graded as adequate with no more than minimal deficiencies in any area. Only 1.4% of interviews had serious deficiencies in one of the domains and only 1.1% of interviews had serious deficiencies in both domains. Most (80.2%) interviews were considered as fully proficient in both information gathering and adherence to PANSS scale instructions. The remainder had at least minor deficiencies. 7.2% of interviews had good quality of information gathering but had minor difficulties in scoring this information. 10.9% of interview had minor deficiencies in information gathering but the information was scored using proper techniques. 19.5% of interviews had minor deficiencies in both information gathering and scoring.

SUMMARY OF RESULTS V: 1442 videotaped PANSS administrations at sites were graded for interview quality by external reviewers using the RISA. 64.6% were regarded as excellent (RISA Score 28-30), 28.9% as acceptable (RISA Score 24-27) and only 6.5% as poor or unacceptable.

DISCUSSION

The initial data indicate that in global clinical trial settings with surveillance of ratings and ongoing feedback to investigators the quality of interview data and proficiency of ratings as assessed by the RQQ were judged to be adequate or better by external reviewers in the large majority of cases. The results of interview assessment using the RISA in the same 10 studies (but with a modestly larger sample size) are included for reference. Both the RQQ and total score of the RISA, categorized the large majority of interviews as adequate or better. The RISA is a 16 item scale with 4 domains, providing a comprehensive assessment of interview quality. The RQQ is a global impression-like scale. It addresses two separate items: the overall quality of the patient and/or informant material collected, as well as the quality of the ratings at a given subject visit.

This study’s findings differed from those of Jagl et al (2007) and Engerdahl et al (2006) in which the majority of interviews were judged to be deficient. We speculate that rapid feedback to the raters from the external reviewers may have been useful in maintaining interview and ratings quality in the current studies.

These results are preliminary and additional data will be reported from ongoing studies. Future presentations will address the psychometric qualities of the RQQ, including inter-rater reliability, test-retest reliability, content validity and convergent validity. Regional variation in interview quality will also be addressed.