

Assessing Interview Quality and Scoring Accuracy in Clinical Trials with Continuous Quality Control (CQC)

Brown, B¹ De Santi, S^{2,3} Detke, M^{1,4} Williams, JBW^{1,5}

MedAvante, Inc.¹, MedAvante, Inc. (former)², Bayer Healthcare Pharmaceuticals, Inc.³, Indiana University School of Medicine⁴, College of Physicians and Surgeons, Columbia University⁵

ABSTRACT

Introduction: CNS clinical trials fail more often than their *a priori* powering indicates they should. Quality assurance/quality control (QA/QC) safeguards for clinical (including primary) outcome measures have rarely been utilized. The large number of raters performing assessments in multi-site trials increases the probability of variability in ratings. Rater drift over time is well-documented and ubiquitous¹, and superior interviews as measured by the Rater Applied Performance Scale (RAPS), were associated with drug-placebo separation². The RAPS is a scale used to rate interviewing skills on the domains of **adherence** to the interview guide, use of **follow-up** questions, **clarification**, **neutrality**, and research **rapport**. We report early findings using Continuous Quality Control (CQC), a new approach to monitoring and remediating the administration and scoring of clinical outcome measures via the RAPS.

Methods: 18 calibrated quality reviewers were rigorously trained and continuously calibrated on scale scoring and interview quality. This cohort was tightly calibrated on the MADRS, HAM-A and HAM-D, with ICCs = .91-.94 on observed interviews. The reviewers were also calibrated on assessing interview quality via the RAPS and on delivery of feedback. Data from two on-going clinical trials were pooled. 128 Site raters audio recorded all MADRS, and HAM-A/HAM-D (SIGH-AD) administrations and uploaded the recordings to a central server. A subset of these interviews was selected for review by calibrated quality reviewers based upon a pre-determined algorithm. *A priori* scoring accuracy and RAPS interview quality criteria were established. The reviewers listened to the recorded interviews in full and independently scored 1158 site raters' assessments and rated interview quality using the RAPS. Only after the reviewers' scores and RAPS ratings were locked was the reviewer given access to the site raters' scores. Detailed feedback was provided to the site raters on both interview quality (covering each domain of the RAPS) and scoring before their next reviewed assessment.

Results: 1158 assessments were reviewed. At the first review of each of the 128 site raters, 60% met the *a priori* criteria for scoring accuracy, 72% for interview quality and 48% met both criteria. By review nineteen or later (n=100) there were improvements: 68% met criteria for scoring accuracy, 81% for interview quality and 61% met both criteria. Analysis of RAPS domains showed that inadequate follow-up questioning was the most common contributor to poor interview quality. Notable improvements were shown in 4 out of the 5 RAPS domains from review 1 to review 19+, with ratings on Follow-up and Neutrality progressing the most over time from Fair/Unsatisfactory ratings to Good/Excellent ratings. Additional data are currently being collected.

Conclusions: QA/QC of clinical assessments identified significant scale administration and scoring issues. Repeated feedback improved rater performance over the course of the trial. This is in contrast to the well-documented phenomenon of rater drift seen in trials without QA/QC safeguards. Scoring and interview quality may require ongoing monitoring and training to achieve and maintain an acceptable standard. Study outcomes will be evaluated to determine if continuous QA/QC of study assessments assists sponsors in identifying risks that contribute to CNS trial failures.

INTRODUCTION

Khan (2005) showed that 51-52% of clinical trials failed with known effective antidepressants and anxiolytics³.

Possible reasons for failed trials include:

- Variability in ratings of clinical scales due in part to the sheer number of raters performing assessments in multi-site trials.
- Rater drift over time: study start-up rater standardization does not persist and rater calibration drops off after a short time. Rater drift occurs in scale administration interviewing techniques and scoring.

Fair to unsatisfactory interview performance, as measured by the Rater Applied Performance Scale (RAPS) (Lipsitz et al, 2004), has been associated with a failure in drug-placebo separation (Kobak et al, 2007).

QA/QC safeguards for clinical (including primary) outcome measures have rarely been utilized in clinical trials.

We report findings from two randomized clinical trials of major depression using Continuous Quality Control (CQC), a new approach to monitoring and remediating the administration and scoring of clinical outcome measures.

METHODS

Calibrated Quality Reviewers:

- 18 calibrated quality reviewers were extensively trained and calibrated on the MADRS, HAM-A, and HAM-D
- Reviewers were calibrated to clinical scale scoring, assessing interview quality and delivery of feedback.
- Reviewers were calibrated prior to study start and quarterly throughout the study.
- MedAvante Clinicians have historically achieved high levels of interrater reliability on independent interviews.^{4,5}

MADRS Scoring ICC = .93 | HAM-D Scoring ICC = .93 | HAM-A Scoring ICC = .91

Study Design:

- 128 site raters were selected by the sponsors to interview patients in these studies.
- All site raters were trained and qualified MedAvante prior to study start through completion of didactic and applied training. Applied training required raters to meet a pre-specified definition of "meets criteria" pertaining to both interviewing skills and scoring.
- All study assessments are audio recorded.
- Assessments are uploaded to a central server where they can be accessed by the calibrated quality reviewer.
- A study-specific algorithm determines which assessments are reviewed.
- The calibrated quality reviewer listens to the assessment, scores all items, assesses interview quality with the RAPS (Adherence, Follow-up; Clarification; Neutrality; and Rapport) with a pre-specified definition of "meets criteria" identical to that used in initial training.
- The calibrated quality reviewer enters his/her RAPS ratings and item scores of the site rater interview into the system.
- After all of the reviewer's scale item scores and RAPS ratings are entered, the reviewer is given access to the site rater scores for comparison.
- Scale-specific criteria defining the necessary level of scoring agreement between the site rater and the reviewer were pre-specified.
- Scoring feedback as well as interview quality feedback (covering each domain of the RAPS) are then prepared and sent to rater and sponsor.

RESULTS

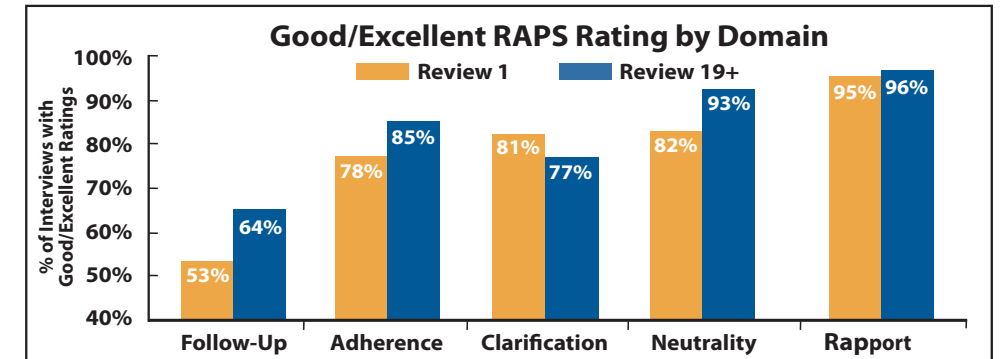
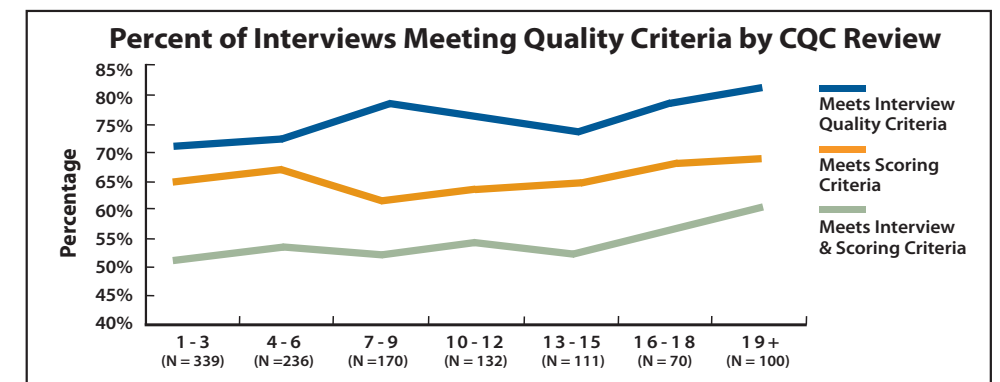
- At first review, CQC of clinical assessments identified significant scale administration and scoring issues. Only 48% of the qualified raters met both scoring and interview quality criteria at review 1- a potential risk to the trial success.
 - At the initial review interview quality was most impacted by the Follow-up domain of the RAPS (47% rated Fair or Unsatisfactory).

References

1. Kobak K, Kane JM, Thase ME, Nierenberg AA. Why do clinical trials fail? The problem of measurement error in clinical trials: Time to test new paradigms. *Journal of Clinical Psychopharmacology* 2007; 27: 534-535.
2. Lipsitz J, Kobak K, Feiger A, Sikich D, Moroz BAG, Engelhardt N. The Rater Applied Performance Scale: Development and reliability. *Psychiatry Research* 2004; 127: 147-155.
3. Khan A, Kolts RL, Rapaport MH, Krishnan KR, Brodhead AE, Browns WA. Magnitude of placebo response and drug-placebo differences across psychiatric disorders. *Psychol Med* 2005; 35(5):743-9.
4. Williams JBW, Kobak KA. Development and reliability of the SIGMA: A structured interview guide for the Montgomery-Asberg Depression Rating Scale (MADRS). *British Journal of Psychiatry* 2008; 192: 52-58.
5. American Psychiatric Association, 159th Annual Meeting, Toronto, CA (May, 2006). Janet B.W. Williams, DSW and Kenneth A. Kobak, Ph.D. Special acknowledgment to Danielle Popp, Ph.D., MedAvante, Inc.

RESULTS *continued*

- Rater performance improved considerably in interview quality with application of the CQC of clinical assessments by a closely and continuously calibrated cohort of quality reviewers. This stands in stark contrast to the *decline* of rater performance seen in well-documented rater drift.
 - At review 1&2, 70% met interview quality criteria; however at review 19+, 81% met interview quality criteria.
 - RAPS performance improved on 4 out of 5 RAPS domains from review 1 to review 19+. Most notable improvement was in the domains of Follow-up and Neutrality.
- Scoring performance improved somewhat more modestly with application of CQC, but again this is in contrast to the documented decline in scoring agreement over time without intervention.
 - At review 1, 60% met scoring criteria; however at review 19+, 68% met scoring criteria.



CONCLUSION

- Multi-site trials may pose special challenges in standardizing administration and scoring of clinical outcome measures across many raters.
- Scoring and interview quality may require ongoing monitoring and training to achieve and maintain an acceptable standard.

Indicated Next Step:

- Rater performance throughout the remainder of these studies will continue to be monitored to determine the degree to which continued monitoring and training maintains the acceptable standards of interview quality and scoring.
- When the studies close, the effect of CQC on study outcomes will be evaluated.

Disclosures

B. Brown, MedAvante, Part 1; MedAvante, Part 2; MedAvante, Part 3; MedAvante, Part 5; S. De Santi, MedAvante, The Cognition Group, Bayer Healthcare Pharmaceuticals, Part 1; MedAvante, The Cognition Group, Part 2; MedAvante, The Cognition Group, Bayer Healthcare Pharmaceuticals, Part 3; Bayer Healthcare Pharmaceuticals, Part 5; M. Detke, MedAvante, Eli Lilly, Part 1; MedAvante, Eli Lilly, Part 2; MedAvante, Eli Lilly, Part 3; MedAvante, Part 5; J. Williams, MedAvante, Part 1; MedAvante, Part 2; MedAvante, Part 3; MedAvante, Part 5