

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

Brown, B<sup>1</sup> De Santi, S<sup>2,3</sup> Detke, M<sup>1,4,5</sup> Brown, J<sup>1</sup> Williams, JBW<sup>1,5</sup>

MedAvante, Inc.<sup>1</sup>, NYU Langone Medical Center<sup>2</sup>, Bayer Healthcare Pharmaceuticals<sup>3</sup>, Indiana University School of Medicine<sup>4</sup>, McLean Hospital, Harvard Medical School<sup>5</sup>, College of Physicians and Surgeons, Columbia University<sup>6</sup>

## 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 common<sup>1</sup>, and superior interview performance as measured by the Rater Applied Performance Scale (RAPS) is associated with drug-placebo separation<sup>2</sup>. We report the first findings using Continuous Quality Control (CQC), a new approach to monitoring and remediating the administration and scoring of clinical outcome measures.

**Methods:** 26 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. Data from two on-going clinical trials were pooled. Site raters audio recorded all MADRS, and HAM-A/HAM-D (SIGH-AD) administrations and uploaded the recordings to a central server. A priori scoring accuracy and RAPS interview quality criteria were established. Calibrated quality reviewers independently scored 797 site raters' assessments and rated interview quality using the RAPS. Only after scores and RAPS were submitted was the calibrated quality reviewer given access to the site raters' scores. Feedback was provided to the site raters on both interview quality and scoring accuracy before their next reviewed assessment.

**Results:** 797 assessments were reviewed. At the first review of 129 site ratings, 56% met the a priori criteria for scoring accuracy, 64% for interview quality and 44% met both criteria. By review nine or later (n=136) there were substantial improvements: 74% met criteria for scoring accuracy, 84% for interview quality and 68% met both criteria. Improvements generally occurred by month four of the study. Analysis of RAPS domains showed that inadequate interview follow-up was the most common contributor to poor interview quality.

**Conclusions:** QA/QC of clinical assessments identified significant scale administration and scoring issues. Repeated feedback improved rater performance substantially. 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) recently showed that 51-52% of clinical trials failed with known effective antidepressants and anxiolytics.

• 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. Raters drift occurs in scale administration interviewing techniques and scoring.

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

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

**We report the first 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:

- 26 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.<sup>4,5</sup>

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

### Study Design:

- 129 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.
- 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; and Neutrality) with a pre-specified definition of "meets criteria."
- The calibrated quality reviewer enters his/her RAPS scores of the site rater interview into the system.
- After all of the reviewer's scale item scores and RAPS scores are entered, the reviewer is given access to the site rater scores for comparison.
- Scoring feedback as well as interview quality feedback are then prepared and sent to rater and sponsor.
- Scale-specific criteria defining the necessary level of scoring agreement between the site rater and the reviewer were pre-specified.

## CONCLUSION

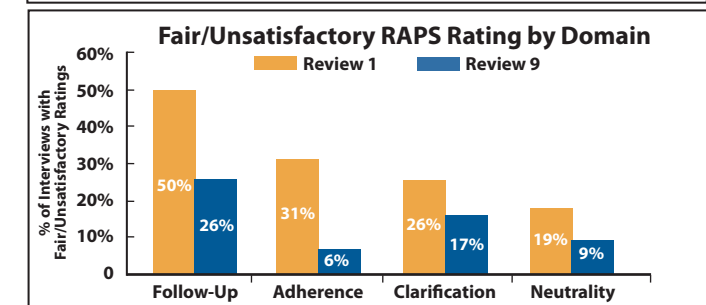
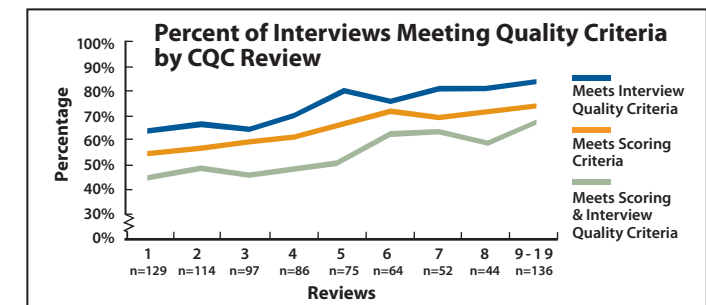
- Multi-site trials may pose special challenges in standardizing administration and scoring of clinical outcome measures across many raters.
- Scoring and interview quality may require ongoing monitoring and training to achieve and maintain an acceptable standard.

### 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.

## RESULTS

- At first review, CQC of clinical assessments identified significant scale administration and scoring issues. Only 44% of raters met both scoring and interview quality criteria at review 1 - a potentially substantial risk to the trial.
- Interview quality was most impacted by the Follow-up domain of the RAPS.
- Rater performance improved substantially in both scoring and 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 seen in well-documented rater drift.



### 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.

This study was supported by MedAvante, Inc. Authors report no conflict of interest.