Title: Common Rater Errors for Assessments in Pediatric Rare and Orphan Disease Trials

Authors[1]: Carmen E. Sanchez, Anzalee Khan, Leah Stern, Alexandra Atkins, Heather Snyder & Richard Keefe

[1]All authors are with NeuroCog Trials

Methodological Question

For many rare diseases, well-characterized efficacy endpoints appropriate for the disease are not available. The small size of the target population, coupled with the challenges of utilizing cognitive and adaptive behavior instruments that are not designed specifically for these vulnerable populations, makes quality review of assessment skills and scoring critical. Does the application of methodological and measurement considerations to review neurocognitive ratings in Rare and Orphan (R&O) trials have the potential to enhance the quality of clinical trial data?

Background

In accordance with the FDA's goal to expedite the development of novel drugs in R&O diseases, an accompanying obligation arises to assess the measurement challenges in these small patient populations. One critical measurement challenge involves accurate administration and scoring of assessments with multiple items and complex scoring rules. Since errors in administration and scoring, and any missing or excluded endpoints in small samples, significantly affect treatment outcome, identification of these errors is critical. Specifically, prompt identification and correction of rater administration and scoring errors ensures accurate calculation of study endpoints.

Aims

To investigate the impact of data quality review, we quantified the number of scoring and administration errors typically found in paper-based review of cognitive and adaptive behavior assessments, particularly tools that are sensitive to neurocognitive change in Mucopolysaccharidosis (MPS) disorders (Shapiro et al, 2017).

Methods

Assessments for 37 participants' visits (administered by 11 previously trained Ph.D-level raters worldwide) were reviewed by a psychologist for administration and scoring errors, for a total of 126 visits. NCT feedback was initiated after an average of 2.1 visits per rater had been completed. Assessments reviewed included the Bayley Scales of Infant Development, 3rd Edition (BSID-III; n = 112) or Kaufman Assessment Battery for Children, 2nd Edition (KABC-II; n = 14) and Vineland Adaptive Behavior Scale, 2nd Edition, Extended Interview Form (VABS-II; n = 91). Errors made on a subtest were only counted once (i.e., a single error on one subtest can result in multiple score changes to subtest and domain-level scores).

Results

The most frequently observed scoring error across all visits was incomplete scoring of the entire assessment, with only 8.73% of visit assessments being completely scored by the rater. Within tests, errors were as follows: 1) VABS-II: Summation errors for subtests (23.08%, 21/91), and referencing normative tables in the manual (18.68%, 17/91); 2) BSID-III: Incorrect application of the double basal

rule (37.50%, 42/112) and violation of the basal and ceiling rules (7.14%, 8/112); 3) KABC-II: Incorrect administration of subtests or items (21.43%, 3/14) and incorrect summing of scores or no scoring (28.57%, 4/14). In response to corrective feedback, 5 out of 8 raters showed improvement on the VABS-II and BSID-III.

Conclusions

The high frequency of administration and scoring errors in cognitive and adaptive behavior assessments underscores the importance of expert data review. Furthermore, the impact of errors is considerable, leading to inaccurate calculation of endpoints across the subtest and domain levels. Common errors point to areas rater training could address. Data quality review and remediation can have large positive effects on signal detection in R&O trials.