Common Rater Errors for Assessment in Pediatric Rare and Orphan Disease Trials

Sanchez, CE[1], Khan, A[1], Stern, L[1], Atkins, AS[1], Snyder, H[1] & Keefe, RSE[1, 2]

[1]Verasci, Inc. (Durham, NC), [2]Duke University (Durham, NC)

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

Mucopolysaccharidoses (MPS) are a group of rare, inherited lysosomal storage disorders that can lead to significant neurological manifestations, particularly in rapidly progressing MSP I, II, III, and VII. Patients with MPS are missing an enzyme that recycles cellular waste. While babies may show little signs of the disease, as the cellular waste begins to accumulate, symptoms start to appear.

Neurological symptoms include impaired cognition, difficulties in language and speech, behavioral abnormalities, sleep problems and/or seizures (Neufeld et al., 2001). As such, **neurocognitive impairment acts as a sensitive indicator of disease progression**. In order to obtain reliable and valid neurocognitive data, it is critical to select reliable and valid tests appropriate for MPS patients, and, in addition, to use them in a standardized manner.

One important measurement challenge involves accurate administration and scoring of assessments with multiple items and complex scoring rules in MPS in international clinical trials. Given the potentially large impact of erroneous, invalid and missing data on treatment outcomes in clinical trials with small samples, identification of administration and scoring errors is critical. Specifically, prompt identification and correction of rater administration and scoring errors ensures accurate calculation of study endpoints.

AIMS

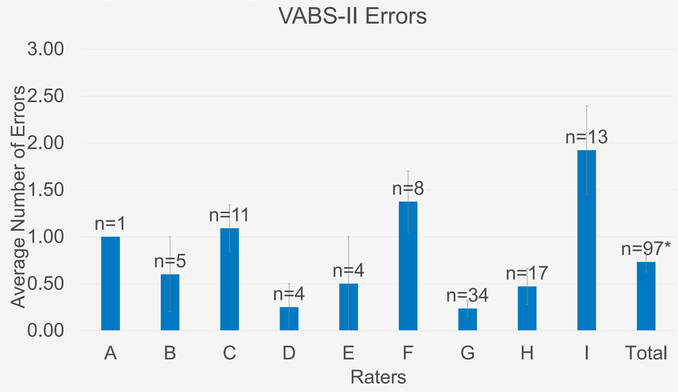
To investigate the impact of data quality review, we quantified the number of **scoring and administration errors** typically found in paper-based review of cognitive and adaptive behavior assessments, particularly tools that are sensitive to neurocognitive change in MPS disorders (Shapiro et al, 2017).

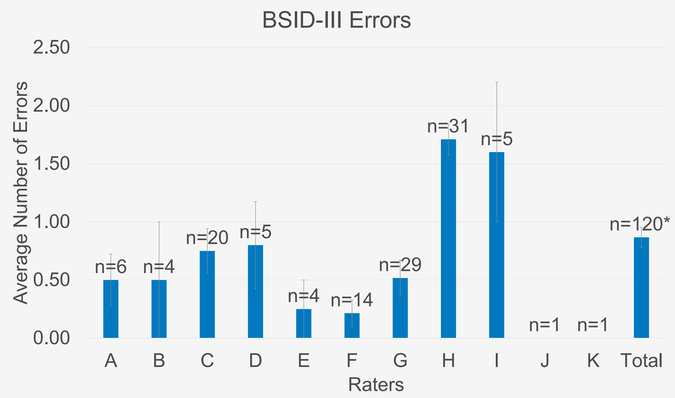
METHODS

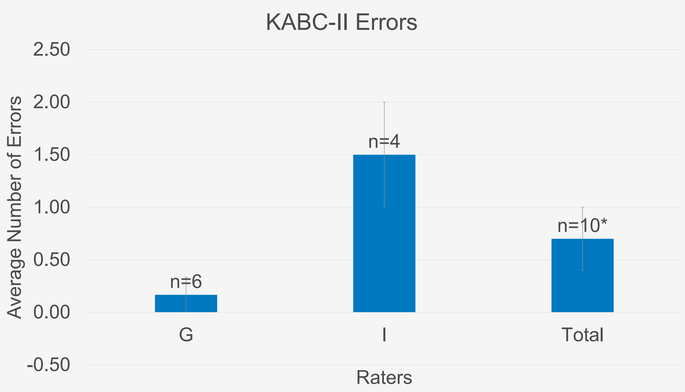
- A Ph.D.-level psychologist reviewed assessments from a total of 135 visits (originating from 37 participants) for administration and scoring errors.
- 9 previously-trained Ph.D-level neuropsychologists served as raters for the international clinical trials.
- On average, NCT feedback was initiated after an average of 9.11 (82/9) visits had been completed. After initiation of feedback, an average of 5.88 (53/9) visits were completed.
- Assessments reviewed included:
  - Vineland Adaptive Behavior Scale, 2nd Edition, Extended Interview Form (VABS-II; n = 101) and
  - Bayley Scales of Infant Development, 3rd Edition (BSID-III; n = 121) **or**
  - Kaufman Assessment Battery for Children, 2nd Edition (KABC-II; n = 14).
- **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

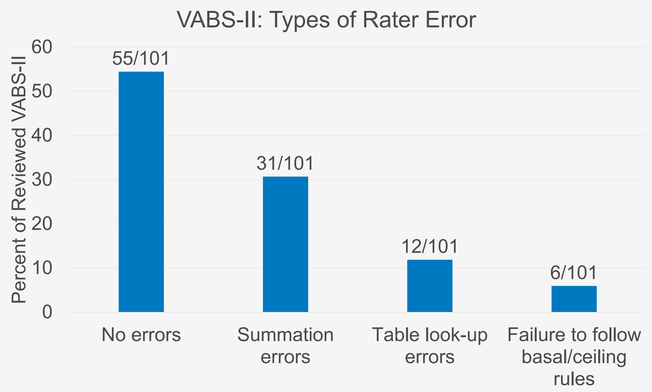
Average Errors by Rater

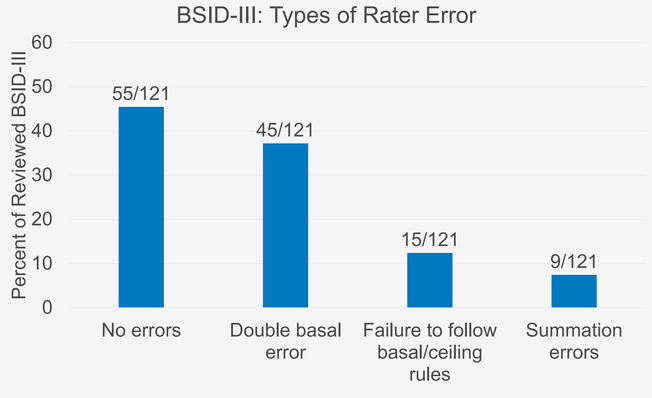


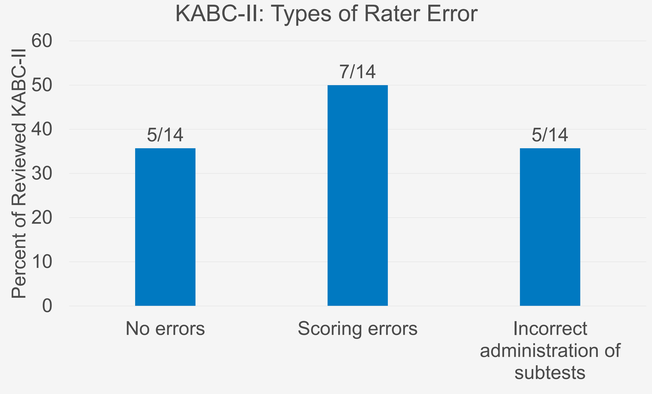




Qualitative Review of Errors







*VABS-II was not scored by raters for 4 visits. BSID-III was not scored by a rater for 1 visit. KABC-II was not scored by a rater for 4 visits.

CONCLUSIONS

- Expert data review of cognitive and adaptive behavior assessments demonstrated the high frequency of administration and scoring errors committed by raters in R&O disease clinical trials.
- The impact of errors is considerable, leading to inaccurate calculation of endpoints across the subtest and domain levels.
- A central rater serves a critical role in R&O disease trials to identify errors of scoring and correct them for accuracy.
- Common errors point to areas rater training could address. Specifically,
  - VABS-II--The use of computer-based scoring programs to assist in score summation
  - BSID-III--Explicit training on the double basal guideline and its application to scoring
  - KABC-II--Provision of clear guidelines for scoring
- Data quality review and remediation can have large positive effects on signal detection in R&O trials.

ACKNOWLEDGEMENTS AND DISCLOSURES

We would like to thank Armagen and Lysogene for the use of their data. We would also like to express our appreciation to all of the children and families that contributed to this research.

RSE Keefe currently or in the past 3 years has received investigator-initiated research funding support from the Department of Veteran's Affairs, Feinstein Institute for Medical Research, GlaxoSmithKline, National Institute of Mental Health, Novartis, Psychogenics, Research Foundation for Mental Hygiene, Inc., and the Singapore National Medical Research Council. He currently or in the past 3 years has received honoraria, served as a consultant, or advisory board member for Abbvie, Akebia, Amgen, Astellas, Asubio, AviNeuro/ChemRar, BiolineRx, Biomarin, Boehringer-Ingelheim, Bristol-Myers Squibb, Eli Lilly, EnVivo, Helicon, Lundbeck, Merck, Mitsubishi, Otsuka, Pfizer, Roche, Shire, Sunovion, Takeda, Targacept. He is also a shareholder in VeraSci, Inc.