Psychometric data monitoring in clinical trials: Can disparities between patient and clinician reported outcomes predict measurement error?

Christian Yavorsky[1], Nina Engelhardt[2], Cynthia McNamara[2], Guillermo DiClemente[2]
Cronos CCS[1] Berkeley, CA, Cronos CCS[2] Lambertville, NJ

Introduction

- A randomized, placebo-controlled trial of a compound used to treat patients with Major Depressive Disorder used psychometric data-monitoring to minimize risk associated with human error in study measurement. Psychometric data monitoring is a risk-mitigation strategy that has been implemented across industry, academic and governmental organizations. It consists of computational algorithms to identify risk based on predictive analytics (accumulated trial data) alongside scale dynamics (e.g., do items agree), alongside active identification and remediation of raters at higher risk for contributing non-informative data.
- In this study we reviewed data generated by a patient self-report HAMD using Interactive Voice Response (IVR) and the clinician administered MADRS. We used the correlation between the MADRS administered by a clinician and the patient-rated HAMD using IVR as a potential proxy for risk of assessment error.
- The literature suggests, moderate to strong correlations have been reported between validated instruments in cross-clinician comparison. The Montgomery-Åsberg Depression Rating Scale (MADRS) and the Hamilton Depression rating scale (HAMD) are the most widely used assessments in clinical trials for depression. These scales are well-validated and many studies (e.g., Jiang & Ahmed, 2009) indicate moderate (r= 0.62) to strong (r= 0.92) correlations when these scales are administered by trained clinicians. Strong correlation (r=0.96) between patient and clinician ratings has also been reported for the HAMD (Kobak et al, 1999). Whether these instruments remain correlated when one is administered by a clinician and the other is patient report, has been shown in academic studies but has not been investigated sufficiently in clinical trials with much larger sample sizes. This result has implications for alternative study design considerations which may be implemented to improve signal detection in depression trials.

Methods

- Data from 575 patient visits was analyzed using SAS 9.3 and correlations obtained across visits and between MADRS and IVR HAMD scores. The change in individual item scores across visits was also calculated to examine similarity (a potential risk predictor) and magnitude of change captured across scales.

Results

- There was very weak correlation with poor significance between MADRS item Reported Sadness and HAMD item Depressed Mood (Spearman's r = 0.055, p = 0.590) at the baseline visit. MADRS Inner Tension and HAMD Psychic Anxiety had similarly weak correlations at baseline and visit 2 though items across scales assessing sleep and appetite had moderate to strong correlations across visits. MADRS and HAMD total scores by visit were moderately correlated at baseline (Spearman's r = 0.454, p < .0001) with weak correlations at visit 2 (randomization visit; Spearman's r = 0.383, p < .0001) and moderate to strong correlations at visits 4, 5 and 6 (scales were not performed at visit 3). The mean MADRS score at baseline was 30.94 (SD: 4.52) and mean HAMD at baseline was 24.48 (SD: 5.12). Individual item correlations across visits were also computed. There was very weak correlation with poor significance between MADRS item "reported sadness" and HAMD item "depressed mood" (Spearman's r = 0.055, p = 0.590) at the baseline visit. The two items were more closely correlated by visit 4 (Spearman's r = .518, p < .0001). MADRS item "inner tension" and HAMD item "psychic anxiety" had similarly weak correlations at baseline and Visit 2 though the items across scales assessing both sleep and appetite had moderate to strong correlations across all visits.

Conclusion

- We found that the two scales under consideration did not agree strongly at key visits and individual item correlations measuring similar constructs across scales had very weak or no correlation. This correlation improved as the study continued.
- Some researchers (e.g., Kobak, 2000) indicate that IVR use of the scale is essentially equivalent to clinician administration. In this analysis we found that the two scales under consideration did not appear to agree strongly and individual item correlations thought to measure similar constructs across scales had very weak or no correlation.
- There is also some evidence for agreement with our findings from this study (Kunugi et al, 2013) noting that patients "tended to overestimate depression severity and have limited agreement with the clinician reported version of the scale". In this case it appeared that the raters estimated depression severity as consistently higher; baseline and Visit 2 time points seemed especially discrepant between patient and rater assessments.

References

- Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960; 23: 56–62.
- Jiang Q, Ahmed S. An analysis of correlations among four outcome scales employed in clinical trials of patients with major depressive disorder. Ann Gen Psychiatry. 2009; 8(4): 1-6.
- Kobak KA, Mundt JC, Greist JH, Katzelnick DJ, Jefferson JW. Computer assessment of depression: Automating the Hamilton Depression Rating Scale. Drug Inf J. 34: 145-156, 2000
- Kunugi H, Koga N, Hashikura M, Noda T, Shimizu Y, Kobayashi T, Yamanaka J, Kanemoto N, Higuchi T. Validation of computer-administered clinical rating scale: Hamilton Depression Rating Scale assessment with Interactive Voice Response technology – Japanese version. Psychiat Clin Neuros. 2013; 67: 253-258.
- Montgomery S, Åsberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 2004; 161: 2163-77.
- Moore HK, Wohlreich MM, Wilson MG, Mundt JC, Fava M, Mallinckrodt CH, Griest JH. Using daily Interactive Voice Response Assessments. Psychiatry (Edgmont). 2007; 4(3): 30-38.

Corresponding Author:
Christian Yavorsky, PhD wcy@cronosccs.com

Disclosure: The authors report no conflicts of interest for this work.