P02-88

QUANTIFYING RATER DRIFT ON THE HAM-D IN A SAMPLE OF STANDARDIZED RATER TRAINING EVENTS: IMPLICATIONS FOR RELIABILITY AND SAMPLE SIZE CALCULATIONS

B. Rothman[1], C. Yavorsky[1], A. De Fries[1], J. Gordon[1], M. Opler[1,2]
[1]ProPhase LLC, [2]New York University, New York, NY, USA

Introduction/objectives/aims: Though rater drift in clinical trials has long been understood to negatively impact trial results, few studies have systematically quantified this. We examined training data for the HAM-D (Hamilton Depression Scale, 17-item version) at two time points to measure the impact.
Methods: Raters participating in a standardized training scored the HAM-D based on two videotaped interviews of depressed patients. To assess drift, data from an initial, post-online training session was compared to data obtained 12 months later. Intra-class correlation coefficients (Shrout & Fliess, 1979) and concordance with expert ratings were compared.
Results: Intra-class correlation coefficients (ICC) for raters (n=167) following initial training were good to excellent for individual raters (.695-.976, $p < .0001$) and good for the overall cohort (.752, $p< .0001$). Concordance with expert ratings was excellent at 99.3%. The overall ICC fell to .730 at the second assessment and although the upper bound of individual performance remained in the good to excellent range, the frequency of scores in the poor to fair range (< .65) increased. Concordance also fell slightly to 87%..
Conclusions: Rater drift occurred over 12 months, as gauged by the metrics of reliability and concordance. Drift was apparent in a limited portion of the cohort but resulted in a lower overall ICC at the second time point. Because studies are generally powered assuming that the ICC remains stable, there are implications for both this power calculation and the required sample size.

https://doi.org/10.1016/S0924-9338(11)72389-6 Published online by Cambridge University Press