High fidelity: What is the real impact of a data monitoring program on data quality?

Yavorsky C, Di Clemente G, Wolanski K, Burger F*

*all authors are employees of Cronos CCS, Lambertville, NJ, USA

Introduction

- Despite the emphasis on rater training at the outset of CNS clinical trials and, in many cases re-current training, there are still significant concerns about the reliability of in-study data[1].
- The well-documented phenomenon of rater drift suggests that there is degradation of training impact over time. This drift can cause problems for both reliability and validity and impact sample size and power calculations.
- Further variability comes from the sheer numbers of raters participating in a multi-site clinical trial. When multiple raters are used in a clinical trial, differences between raters in terms of interviewing technique and scoring criteria introduce variability that can distort the outcome measures. Idiosyncratic rating practice can emerge, and if not dealt with, can introduce systematic error into study results[2,3].
- Risk-based data-monitoring is widely used in clinical trials to help manage risk implicit in subjectively derived outcome measures. Few studies to date have critically examined the impact of such programs. In this study our intention was to determine how many raters contributed data that contained confirmed significant error, and to forward estimate the result of non-intervention (had we not provided remediation and support), then estimate the potential impact of this non-intervention on statistical power using the original study protocol parameters.

Objectives

- To estimate the impact of an active data-monitoring program on a randomized, double-blind, placebo-controlled, Phase III trial for the treatment of adults with schizophrenia in terms of statistical power and sample size estimates.

Methods

- The present study investigated a sample of subjects enrolled in a completed schizophrenia trial using the PANSS as the primary outcome measure. The data was processed daily and, if risks to data quality were detected, contact with the site rater was initiated. Feedback was provided as necessary when problems in scale use were identified. A forward analysis estimating the impact of non-intervention with raters contributing poor quality data was conducted with this data partitioned by subject.
- The protocol specified parameters to calculate sample size to achieve 90% power (nominal alpha of 0.05, 2-sided, though alpha for final analysis was 0.0452).
- The power was recalculated using G*Power[4] software with the original protocol parameters but with the theoretically reduced sample size resulting from the partitioning of poor quality subject data.

Results

- The first analysis allowed for summarization of the overall efficacy of the method while identifying those raters whose error would have adversely impacted trial outcomes to re-estimate sample size. We theorized a non-intervention scenario with patients continuing to be assessed incorrectly and subtracting these. Using this estimated reduction in sample size we recalculated power based on original parameters and found a reduction from .90 to .78. Note also the increase in β in Table 1.1 wherein the reduced sample size has resulted in increased likelihood of committing a Type II error.

Analysis: **Original protocol parameters**

Input:
Tail(s)= Two
Effect size |ρ| = .1840
α err prob = .0452
Total sample size = 310
Output:
Noncentrality parameter δ = 3.2960
Critical t = 2.0110
Df = 308
Power (1-β err prob) = 0.9001

Analysis: Estimated power recalculation

Input:
Tail(s) = Two
Effect size |ρ| = .1840
α err prob = .0452
Total sample size = 222
Output:
Noncentrality parameter δ = 2.7892
Critical t = 2.0143
Df = 220
Power (1-β err prob) = 0.7804

Figure 1.0 (*n* = 310)

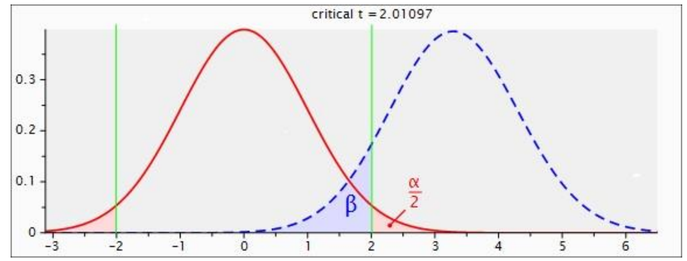
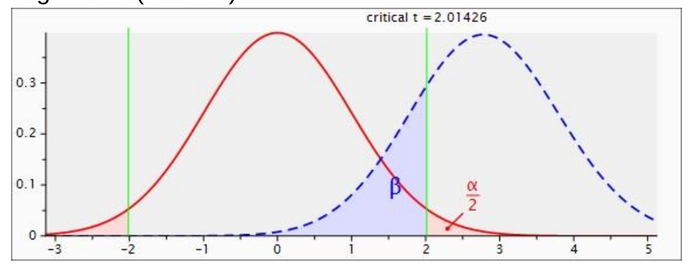


Figure 1.1 (*n* = 222)



Conclusion

- Risk-based data-monitoring can identify and correct error in-study and ensure protocol fidelity. This analysis was designed to estimate the impact of allowing data contributed by problematic raters to persist, i.e., assuming that after significant error was found and confirmed, there was no intervention.
- If sample size can be assumed to be reduced by the presence of incorrect data (or data that is sufficiently flawed as to be partitioned), then power is reduced and the probability of Type II error increases.

References

- [1] Müller MJ, Szegedi, A. Effects of Interrater Reliability of Psychopathologic Assessment on Power and Sample Size Calculations in Clinical Trials. J Clin Psychopharm. 2002; 22(3):318-325.
- [2] Demitrack MA, Faries D, Herrera JM, DeBrota DJ, Potter WZ. The problem of measurement error in multicenter clinical trials. Psychopharmacol Bull. 1998; 34: 19-24.
- [3] Brown B, DeSanti S, Detke M, Brown, J, Williams JBW. Assessing interview quality and scoring accuracy in clinical trials with continuous quality control. Poster presented at the New Clinical Drug Evaluation Unit Annual Meeting, Boca Raton, FL; 2010.
- [4] Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Meth. 2009; 41, 1149-1160.

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

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