

7 Deadly Sins: Proposed Guidelines for Reporting Clinical Trials Methodology Research

Popp D¹, Williams JBW^{1,2}, Detke M^{1,3}

¹MedAvante, Inc. ²College of Physicians and Surgeons, Columbia University ³Indiana University School of Medicine

INTRODUCTION

In several disease areas, approximately 50 percent of clinical trials fail, even those powered at 80-90 percent¹. Many methodological approaches for increasing signal detection have been proposed, including increased rater training, patient-reported outcomes, computerized assessment tools, and centralized assessments. Questions remain about the efficacy (or lack thereof) of each of these methods in increasing signal detection. Variability in reporting results of studies aimed at determining the efficacy of these novel methods makes it difficult to interpret, evaluate and compare findings. Standardizing reporting across studies and methodologies would alleviate these limitations and reduce reporting bias.

METHOD

While the Consolidated Standards of Reporting Trials (CONSORT) developed a set of guidelines for standardizing reporting of findings of Randomized Clinical Trials (RCTs), no similar set of guidelines exist for the reporting of results from studies of the efficacy of clinical trial methodologies. We propose a set of seven such guidelines, explain the importance of each, and where appropriate, provide a detailed illustration of how misuse or omission can influence the interpretation of study results.

RESULTS

1. Report interrater reliability (IRR).

IRR (with a 95 percent confidence interval) should be reported in all studies with multiple raters and multiple observations.

Low IRR will reduce study power and the ability to detect drug-placebo separation². Therefore, it is important to accurately assess and report IRR prior to study start and throughout the course of a clinical trial.

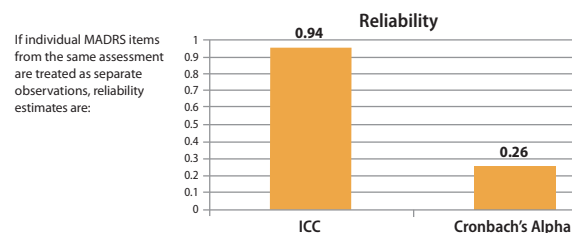
Typically, reliability estimates are obtained at a single point in time (i.e., prior to the start of the study) in artificial settings (i.e., at investigator meetings) and outside of the conduct of the clinical trial (i.e., scoring of videotapes, not actual in-study assessments). As such, these estimates may be inflated. For continuous variables (e.g., most severity assessments), the intraclass correlation coefficient (ICC) is required to accurately determine IRR³.

$$ICC = \frac{\text{variance due to subjects}}{\text{variance due to subjects} + \text{variance due to raters} + \text{residual variance}}$$

Correctly calculating the ICC requires more than one rater, and more than one subject, session or recording. One common error in calculating the ICC is to treat individual items from a single severity scale as separate observations, in order to compensate for a lack of multiple complete observations. However, ICCs calculated this way may be inversely related to the reliability of a construct.

For example, in the following sample, five raters scored the ten items of the Montgomery-Asberg Depression Rating Scale (MADRS) for a single assessment:

MADRS Item	Rater				
	A	B	C	D	E
1	2	2	1	1	2
2	5	6	5	6	5
3	4	5	4	5	4
4	0	0	1	1	0
5	3	3	4	3	3
6	1	1	1	2	1
7	5	6	6	5	5
8	2	1	0	1	1
9	5	6	5	6	6
10	5	5	5	6	5



By definition, an ICC calculated this way can only achieve a high value because the between-item mean squares are large in relation to the within-item mean square. That is, higher ICCs are actually inversely related to internal scale consistency, which may indicate that raters are not applying the scale correctly and additional observations may reveal that interrater reliability issues are present.

2. Use appropriate statistical tests.

Statistical tests should be appropriate for the type of variable (i.e., continuous, categorical, etc.) being analyzed.

Kappa (or percent agreement) does not capture concordance on continuous variables. Kappa is highly influenced by the criterion measure selected. At times, a fixed criterion (e.g., +/- 20 percent) is used to indicate rater agreement with a "gold standard" score.

For example, with a criterion of +/- 20 percent of the gold standard, 85 percent of raters may "meet criteria". However, if the criterion is narrowed to within +/- 10 percent of the gold standard, the number of raters meeting criteria may drop to 45 percent. Selecting a broader criterion range can artificially inflate Kappa.

As such, Kappa is not an appropriate measure of the level of rater agreement on severity scales. ICCs use raw scores instead of artificially created criterion values; therefore they are not influenced by subjective definitions of agreement.

3. Include effect size measures.

Measures of effect size (e.g., Cohen's d with a 95 percent confidence interval) should be reported for all means comparisons regardless of statistical significance.

Effect sizes permit readers to decide if findings are clinically relevant without regard to sample size.

$$Cohen's d = \frac{\bar{X}^1 - \bar{X}^2}{S}$$

Reporting effect size allows comparable evaluations both within and across studies with different outcome variables, although these must be interpreted cautiously relative to effect sizes for existing methods.

4. Identify a priori and post-hoc analyses.

Methodological comparisons should be identified a priori in a statistical analysis plan, much like efficacy analyses.

Ideally, a primary analysis should be explicated. If not a priori, analyses should be identified as post-hoc when reported. Typically, post-hoc analyses are performed on small subsets of the sample.

Failing to report these details can result in over-interpretation of exploratory analyses performed on small subsets of data.

5. Acknowledge and correct for multiple comparisons.

If multiple comparisons are performed on a single sample, all analyses should be reported, whether or not they are published.

Appropriate multiplicity adjustments must be made (e.g., Bonferroni or Hochberg) in order to avoid inflating false positives. Reporting a significant result on a subset of data without indicating the total number of comparisons made across the entire data set may lead to over-interpretation of false positives.

For example, a researcher might report in isolation a statistically significant (p = .04) t-test comparing drug-placebo separation between two methodologies. This would appear to indicate that one methodology was statistically significantly superior to the other.

If these results represent data from only one country, and the same t-test was performed separately for each of the 20 countries in the study, then at least one of these comparisons would likely reach significance at <.05 by random chance. In such a case, the critical p would now be .0025, leading to an interpretation that one methodology failed to demonstrate superiority to the other. Adding analyses by rater education, site enrollment levels, country, etc. may additionally inflate the likelihood of obtaining at least one false positive.

PROPOSED CHECKLIST FOR REPORTING CLINICAL TRIALS METHODOLOGY RESEARCH

1. Interrater reliability

- When assessed?
- How many raters?
- How many observations?
- How were ratings obtained (i.e., scoring videos, joint interviews, independent interviews)?
- What statistic was calculated?

2. Inferential statistics

- What inferential statistics were calculated?
- What significance tests are appropriate?

3. Analyses a priori vs. post-hoc

- What was primary analysis?
- Were analyses reported post-hoc?

4. Multiple comparisons

- How many comparisons were made across entire sample?
- What multiplicity adjustments were made?
- What is adjusted critical p?

5. Appropriateness of statistical test

- Test appropriate for variable type?

6. Effect sizes for all analyses regardless of significance

- Effect size statistic reported?
- Interpretation of magnitude of effect?

7. Interpretation of null hypothesis testing (NHT)

- What does p value indicate?
- Is interpretation of significance affected by magnitude of effect size?

6. Include inferential statistics for means comparisons.

Statements concerning differences or patterns in means should be substantiated with inferential statistics.

As an example, when citing one measurement as being "numerically larger" than another, researchers should qualify such statements by including appropriate inferential measures, such as a t-test or ANOVA. Doing so allows readers to conclude whether or not the observed difference is likely to have occurred by chance.

7. Correct interpretation of null hypothesis testing (NHT).

Expression of NHT should be carefully checked to avoid commonly made errors.

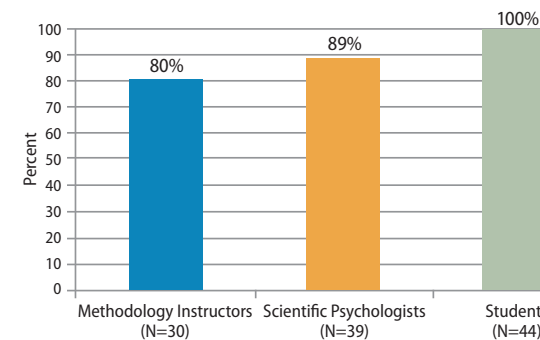
NHT is commonly misinterpreted in clinical trial methodology.

$$p = \text{the probability of the data if the null hypothesis is true}$$

As Cohen notes in his classic article, common misinterpretations of NHT include concluding that smaller p values indicate more important effects, or that a non-significant p value represents a finding of no difference⁴.

More recently, Haller and Krause asked methodology instructors, scientific psychologists, and students to interpret the meaning of p < .01. The following data presents the percentages of participants endorsing various incorrect interpretations of p⁵.

Fallacy	Methodology Instructors (N = 30)	Scientific Psychologists (N = 39)	Students (N = 44)
HO is absolutely disproved.	10	15	34
Probability of HO is found.	17	26	32
HO is absolutely proved.	10	13	20
Probability of H1 is found.	33	33	59
Probability of Type 1 error.	73	67	68
Probability of replication.	37	49	41



Recent research suggests that pharmaceutical statisticians are not immune from misconceptions similar to those identified by Haller and Krause⁶.

While a significant p-value can be used to reject the null hypothesis as false, the reverse is not true. A non-significant p-value does not allow one to conclude that the null hypothesis is true. Since both statistical significance and non-significance are highly influenced by sample size, results of NHT should be interpreted cautiously in the context of sample size, multiplicity adjustments, and effect size estimates. Careful interpretation of findings can avoid over-interpretation of both false positives and findings of non-significance.

CONCLUSION

We demonstrate how adherence to proposed guidelines for standardized statistical reporting on outcomes of studies examining clinical trial methodologies can reduce reporting bias. Empirical research evaluating the effectiveness of methods holds important consequences, not only for clinical trial methodology but also for future drug development decisions facing sponsors and regulators.

References

- Khan A, Kolts RL, Rappaport MH, Krishnan KR, Brodhead AE, Browns WA. Magnitude of placebo response and drug-placebo differences across psychiatric disorders. *Psychol Med.* 2005; 35(5): 743-749.
- Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *J Pharmacol Pharmacother.* 2010; 1(2):100-107.
- Muller, M. J., & Szegedi, A. Effect of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharm.* 2002; 22: 318-325.
- Shroff PE, Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull.* 1979; 86(2): 420-428.
- James LR, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. *J Appl Psychol.* 1984; 69(1): 85-98.
- Cohen J. Things I have learned (so far). *Am Psychol.* 1990; 45(12): 1304-1312.
- Haller H, Krause S. Misinterpretation of significance: A problem students share with their teachers? *Methods of Psychological Research Online.* 2002; 7(1), 1-20.
- Leconte MF, Poitevineau J, Lecoutre B. Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *Int J Psychol.* 2003; 38(1): 37-45.

Disclosures

Danielle Popp: Part 1: MedAvante, Inc.; Part 2: MedAvante, Inc.; Part 3: MedAvante, Inc.; Part 4: None; Part 5: MedAvante, Inc.
Janet B.W. Williams: Part 1: MedAvante, Inc.; Part 2: MedAvante, Inc.; Part 3: MedAvante, Inc.; Part 4: None; Part 5: MedAvante, Inc.
Mike Detke: Part 1: MedAvante, Inc.; Eli Lilly, Inc.; Sonkei, Inc.; Part 2: MedAvante, Inc.; Eli Lilly, Inc.; Part 3: MedAvante, Inc.; Eli Lilly, Inc.; Part 4: None; Part 5: MedAvante, Inc.