Schizophrenia Bulletin vol. 46 no. 5 pp. 1027–1029, 2020
doi:10.1093/schbul/sbaa036
Advance Access publication 18 March 2020

INVITED EDITORIAL

Burying Our Heads in the Sand: The Neglected Importance of Reporting Inter-Rater Reliability in Antipsychotic Medication Trials

Steven Berendsen*[,1,2], Henricus L. Van[2], Loek M. A. Verdegaal[1], Mirjam J. van Tricht[1], Matthijs Blankers[2], and Lieuwe de Haan[1,2]

[1]University Medical Center, location Academic Medical Center, Department of Psychiatry, Amsterdam, the Netherlands; [2]Arkin Mental Health Care, Department of Research, Amsterdam, the Netherlands

*To whom correspondence should be addressed; Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands; tel: +31208913600, fax: +31208913701, e-mail: s.berendsen@amsterdamumc.nl

Key words: double-blind randomized trials/training procedures/consistency

The declining efficacy of antipsychotic medication in randomized clinical trials has led to major concern. Over the last decade, the number of failed phase II trials raised by 15%. In the search for the causes of the apparent declining potency of antipsychotic medication, it has been suggested that the explanation may be found in inadequate rating procedures.[1]

For several decades now, Helena Kraemer stressed the fundamental importance of inter-rater reliability (IRR) for randomized clinical trials,[2] in particular for the rating of psychotic symptoms since measurements are largely dependent on observational instruments that require acceptable reliability. In fact, reliability scores in the absence of training procedures are generally low (<0.6) for the observational instruments that are commonly applied in psychosis research.[3]

Unreliable assessments can have a major impact on the interpretation of study outcomes. Firstly, low reliability of data leads to underpowered studies, and therefore more false-negative findings and attenuated effect sizes.[4] Secondly, unsatisfactory training procedures, unreliable assessments combined with expectation biases of raters and time pressure to complete the inclusion may lead to inflated baseline severity scores. Then, after inclusion, rapid declines can be seen in severity scores as true severity scores of these participants are actually lower. These inflated baseline severity scores are associated with higher placebo responses, making it more difficult to identify real effects in the intervention condition.[5] Moreover, after the selection procedure and without controlling for reliability, rater drift might occur leading to increased measurement error with subsequent regression to the mean.

Although the value of training procedures and reliability assessment is abundantly clear, reporting in these areas is inconsistent. About 20 years ago, 2 papers found that only 9.5% of the included manuscripts reported training procedures and that only 19% or 35% of the included papers reported reliability measurements.[6,7] However, these reviews did not provide precise information about training procedures and IRR coefficients in antipsychotic medication trials, and we wondered whether there has been an improvement during the last 20 years.

We therefore conducted a new review to determine the proportion of papers with and without reported training procedures or IRR coefficients in double-blind randomized controlled trials (RCTs) with antipsychotic medication during the past 2 decades. To this end, we searched Medline for double-blind RCTs of antipsychotic medication for the treatment of schizophrenia spectrum disorders between January 2000 and January 2019. We also selected all double-blind RCTs of antipsychotic medication from 4 large meta-analyses published since 2000. Two authors (S.B. and L.V.) working independently retrieved the following coefficients from the published manuscripts and supplements: presence of an actual IRR coefficient: Intraclass correlation coefficient (ICC), Cohen's Kappa, Krippendorff's alpha or Agreement coefficient 1 (AC1). Further, we collected information about correlation coefficients or a minimum percentage agreement that were used as IRR coefficient, central rater, and any reported training of raters. The details of our approach can be found in the supplementary material, parts 1.1 to 1.5.

We identified 207 double-blind RCTs: 34.8% (N = 72) reported training for raters and 11.1% (N = 23) reported an actual IRR coefficient. Of the 23 RCTs reporting an

IRR coefficient, 78.3% (N = 18) used the ICC and 21.7% (N = 5) used Cohen's Kappa as a measure of IRR. In addition, 6.8% (N = 14) of all RCTs reported that the reliability of assessments was determined, but these studies did not report an IRR coefficient, 1.9% (N = 4) reported a correlation coefficient, 2.4% (N = 5) reported a percentage agreement and only 2.9% (N = 6) used central raters. We found no significant differences between studies sponsored by pharmaceutical companies or non-industry supported trials in the reporting of training variables or reliability measures.

Inappropriate measures of IRR, such as percentage agreement and correlation coefficients were reported in 4.3% of the RCTs. The latter analyses, as well as Cronbach's alpha, are unsuitable to evaluate IRR, as percentage agreement is not change corrected and correlation coefficients merely determines associations between raters without accounting for inter-individual agreement. These measures provide a false impression of sufficient IRR. In a correct analysis for IRR, the ICC, Cohen's kappa, or Krippendorff's alpha are applied.

Despite strong recommendations in the literature concerning the relevance of the inclusion of IRR coefficients or training procedures, no improvement was observed during the past 2 decades. The finding that differences between antipsychotic medication and placebo have become smaller during recent decades may be attributed in part to the lack of training procedures and shortcomings in reliability.

The description of training procedures that we found in the reviewed RCTs varied strongly: from detailed descriptions of repeated training procedures to merely stating that raters were trained. The bottom line is that the value of high-quality training procedures for accurate signal detection is widely recognized for decades, and we still seem to bury our heads in the sand and ignore its vital importance.

The neglect of training procedures could be caused by the preconception that clinically experienced raters conduct reliable assessments and that training is not required. However, several studies have indicated that even experienced clinicians cannot make reliable assessments of at least one-third of the individual PANSS items.[3] Additionally, it is possible, albeit highly unlikely, that some authors actually did implement training procedures or reliability estimations without reporting them. In the more likely event that there was actually no training, this may have been due to the perception that rater training is too costly, time-consuming or difficult to implement in large multi-national trials.

Nevertheless, significant savings can be made by improving reliability since it improves power, meaning that smaller sample sizes are needed to demonstrate effectiveness. To illustrate: improvement of reliability from 0.7 to 0.9 will reduce the required sample size by 22%.[4]

Using central raters could result in major improvements in the areas discussed here: they are independent of study design, highly trained, and they have high IRR scores. It has been shown that using central raters results in significantly less baseline-score inflation in studies with antidepressant medication.[5]

Changes in the procedure could be considered. Firstly, training procedures should include a course on interview skills, video-taped interviews followed by reliability assessment. Independent interviews of the same patient by several raters would be ideal. However, we consider the feasibility of such a procedure in multicenter projects as problematic. Secondly, assessments during clinical trials could be recorded to be reevaluated for reliability and rater drift. Subsequently, inadequate observations can be adjusted and raters demonstrating insufficient observations may receive additional training. Ultimately, raters could even be removed from the trial if their assessments persistently fail. Thirdly, by reevaluating each assessment by several raters the average score can be used as an outcome measure. As a result, reliability would increase as well as power and effect sizes.

In conclusion, training procedures and IRR coefficients are still often neglected in double-blind RCTs with antipsychotic medication. Despite urgent recommendations, there has been no improvement in reporting on, and probably the implementation of, training procedures and reliability assessment in the last 2 decades. Editors of psychiatric journals could contribute to improvement in the future by imposing strict and detailed requirements for reporting on training procedures and IRR coefficients in manuscripts. Furthermore, the use of central raters could provide major benefits in terms of reliability, the prevention of baseline-score inflation and accurate study outcome.

Supplementary Material

Supplementary material is available at *Schizophrenia Bulletin* online.

Funding

No external funding or financial resources have been used for this project.

Acknowledgments

S.B. and L.M.A.V. contributed to the study design and proposal, literature search, data collection, analysis, and interpretation. S.B. drafted the manuscript and all other authors provided critical revisions. H.L.V., M.J.T., M.B., and L.H. supervised statistical analysis, study design, and writing of the manuscript. All authors contributed to and have approved the final manuscript. All authors declare

not to have any conflicts of interest that might be interpreted as influencing the content of the manuscript.

References

1. Kemp AS, Schooler NR, Kalali AH, et al. What is causing the reduced drug-placebo difference in recent schizophrenia clinical trials and what can be done about it? *Schizophr Bull.* 2010;36(3):504–509.
2. Kraemer HC, Thiemann S. A strategy to use soft data effectively in randomized controlled clinical trials. *J Consult Clin Psychol.* 1989;57(1):148–154.
3. Müller MJ, Rossbach W, Dannigkeit P, Müller-Siecheneder F, Szegedi A, Wetzel H. Evaluation of standardized rater training for the Positive and Negative Syndrome Scale (PANSS). *Schizophr Res.* 1998;32(3):151–160.
4. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry.* 2000;47(8):762–766.
5. Kobak KA, Leuchter A, DeBrota D, et al. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. *J Clin Psychopharmacol.* 2010;30(2):193–197.
6. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater reliability in clinical trials of depressive disorders. *Am J Psychiatry.* 2002;159(9):1598–1600.
7. Vacha-Haasem T, Ness C, Nillson J, Reetz D. Practices regarding reporting of reliability coefficients: a review of three journals. *J Exp Edu*. 1999;67:335–341.