Adherence to Published Standards of Reporting

A Comparison of Placebo-Controlled Trials Published in English or German

Christoph A. Junker, MD, MSc

Context.—Although standards of reporting randomized controlled trials are well established internationally, essential study elements continue to be omitted, which hampers interpretation and systematic review of randomized controlled trials.

Objective.—To identify deficiencies in the quality of reporting of placebo-controlled randomized trials published in German or English.

Design.—Observational study comparing 32 German- and 89 English-language reports of placebo-controlled trials with parallel design, published by the same group of authors between 1985 and 1994.

Main Outcome Measure.—High reporting quality, defined as adherence to published standards and measured by an 18-item scale based on 2 standard guidelines.

Results.—The mean quality score was 8.4 (SD, 3.0; range, 1-16) of 18. The difference of the mean quality scores between English-language reports compared with German-language reports was small (0.27; 95% confidence interval, −0.97 to 1.52). More articles reported clinical aspects than trial methods or statistics. Moreover, US reviewers show a stronger preference for US papers than non-US reviewers (accept: 7.0% vs 3.6%; provisionally accept: 31.3% vs 30.5%; reject with resubmission: 26.1% vs 21.6%; reject: 35.6% vs 44.3%) (P = .001).

Data were analyzed to determine whether non-US reviewers evaluate non-US papers similarly. This process was also used to determine whether non-US reviewers evaluate the 2 sets of papers similarly. Non-US reviewers rank papers submitted from the United States more favorably (accept: 3.6% vs 3.2%; provisionally accept: 30.5% vs 24.7%; reject with resubmission: 21.6% vs 24.6%; reject: 44.3% vs 47.6%; P = .001). Moreover, US reviewers show a stronger preference for US papers than non-US reviewers (accept: 7.0% vs 3.6%; provisionally accept: 31.3% vs 30.5%; reject with resubmission: 26.1% vs 21.6%; reject: 35.6% vs 44.3%) (P = .001).

Based on logistic regression looking simultaneously at reviewer nationality and manuscript source, the data show that the manuscript source was significant (P = .01), with domestic papers having an odds ratio of 1.49 for background of review and background of acceptance, while the reviewer’s nationality was not significant (P = .22) (Table).

COMMENT

Several factors prevent conclusive findings on international bias in this study. First, non-US reviewers were pooled because there were not enough authors and reviewers from the same country to obtain statistically significant results. Second, nationality was based on the corresponding author’s location rather than actual nationality. Third, associate editors may choose non-US reviewers who have similar training and viewpoints; however, data regarding location of training were not available. Fourth, reviewers were not blinded. Most important, bias cannot be unequivocally inferred without measuring the quality of individual papers. Still, this study does provide insights into potential variability in manuscript evaluations across national lines and lends itself to further investigation and more complex studies.

I wish to extend a special thanks to WB Saunders for funding the presentation of this study at the Third International Congress on Peer Review in Biomedical Publication.

References

From the Department of Social and Preventive Medicine, University of Bern, Bern, Switzerland. Presented at the Third International Congress on Peer Review in Biomedical Publication, Prague, Czech Republic, September 19, 1997. Reprints: Christoph A. Junker, MD, MSc, ISPM, Finkenhubelweg 11, CH-3012 Bern, Switzerland (e-mail: junker@ispb.unibe.ch).
To create a homogeneous group of trials, all 131 reports of placebo-controlled trials with parallel design were selected from these 783 articles. Ten duplicates reporting the same study with the same outcomes were excluded, using a random procedure. This resulted in a final sample of 32 German and 89 English trial reports in 5 and 58 journals, respectively.

### Outcome Measures

We defined report quality as “adherence to published standards on trial reporting.” A scale for the assessment of reporting quality items was constructed from the checklist for statistical review of articles on clinical trials for BMJ and from the third edition of the Uniform Requirements. Items were interpreted according to references.  

We used 14 of the 24 items of the BMJ checklist and 11 of 12 aspects of the Uniform Requirements (Table). Items were not used when they were not applicable to the selected sample or when their assessment required clinical specialty knowledge. Three items in the BMJ checklist on blinding and completeness of follow-up reflected trial quality and could not be used in the scale of reporting quality. The 18 items used could be answered with a yes or no. Unclear situations were treated as not meeting the criteria. The quality score was calculated as the sum of met criteria. Data were analyzed with SAS for Windows 6.10, Proc Freq, and Proc Univariate (SAS Institute, Cary, NC).

### RESULTS

#### Reporting Deficiencies

The mean quality score was 8.4 (SD, 3.0; range 1-16). Fifty percent of the 121 reports met 9 or more of the 18 criteria (Table). Most frequently, information was missing in articles for confidence interval (96%), method of randomization (93%), reference of the statistical software used (90%), and sample size calculation (89%). However, a sufficient description of the experimental treatment was missing in only 7% of articles. Criteria were met more often for clinical aspects than for trial methodology and statistical analysis.

#### Differences by Language

The difference of the mean quality score between English- and German-language reports was 0.27 (95% confidence interval, −0.97 to 1.52). Seventeen of the items showed no difference between German-language and English-language reports ($P > .1$ in $\chi^2$ or Fisher exact tests). Side effects were reported in 63% of German and 39% of English reports ($P = .04$ in $\chi^2$ test).

#### Duplicate Publications

There were 11 sets with 2 to 4 papers in each set concerning the same study.

---

**METHODOLOGY**

### Selection of Reports

All 224 trial reports published from 1985 through 1994 in 5 leading German-language general medical journals were identified in a manual search. Subsequently, 559 reports published in English by the same key authors (first, second, and last authors) during the same 10-year period were identified through MEDLINE. This sample of studies had been used previously to study aspects of language bias and of trial quality.

---

**Table:**

<table>
<thead>
<tr>
<th>Source†</th>
<th>Language, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>BMJ†</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Was the objective of the study sufficiently described?</td>
<td>1</td>
</tr>
<tr>
<td>Was there a satisfactory statement given of diagnostic criteria for entry into the trial?</td>
<td>2</td>
</tr>
<tr>
<td>Was there a satisfactory statement of source of subjects?</td>
<td>3</td>
</tr>
<tr>
<td>Were the treatments well defined?</td>
<td>5</td>
</tr>
<tr>
<td>Was the method of randomization described?</td>
<td>7</td>
</tr>
<tr>
<td>Was a prestudy calculation of sample size reported?</td>
<td>12</td>
</tr>
<tr>
<td>Was the main end point defined in advance, or was there only 1 end point given?†</td>
<td>...</td>
</tr>
<tr>
<td>Was a primary time point defined?‡</td>
<td>(13)</td>
</tr>
<tr>
<td>Were treatment and control group comparable in relevant measures?</td>
<td>14</td>
</tr>
<tr>
<td>Were the dropouts described by treatment and control groups, or was there a statement indicating no dropouts?</td>
<td>17</td>
</tr>
<tr>
<td>Were side effects of treatment reported?</td>
<td>18</td>
</tr>
<tr>
<td>Statistical analysis and presentation</td>
<td></td>
</tr>
<tr>
<td>Were all statistical procedures adequately described or referenced?</td>
<td>19</td>
</tr>
<tr>
<td>Were the statistical analyses used appropriate?</td>
<td>20</td>
</tr>
<tr>
<td>Were confidence intervals given?</td>
<td>23</td>
</tr>
<tr>
<td>Were $P$ values given together with effect estimates?</td>
<td>...</td>
</tr>
<tr>
<td>Were exact $P$ values given?</td>
<td>...</td>
</tr>
<tr>
<td>Was the statistical software package used named?</td>
<td>...</td>
</tr>
<tr>
<td>Was the conclusion drawn from the statistical analysis justified?</td>
<td>24</td>
</tr>
</tbody>
</table>
altogether 26 articles reporting on 12 studies (in 1 case, 2 studies were reported in 3 papers). Previous publication was handled and referenced to standard rules32 in only 4 cases. Four studies had been reported in identical or similar reports and in 1 case 4 papers had increasing numbers of patients. In another case, 2 papers reported on different subgroups of the same study without crossreference.

Three such sets were in English only, 1 in German only, and 7 with articles in both languages. In addition, 5 articles gave a reference to a previous publication or stated that they were preliminary reports, where the counterpart was not in the sample. Six English and 4 German papers were excluded from the analysis because they were identical with another paper.

**COMMENT**

Using a sample of reports from German-speaking Europe, the present study confirms the existence of large deficiencies in the reporting of placebo-controlled randomized trials with parallel designs. Missing sample size calculations and missing descriptions of the method of randomization are major deficiencies. These have been reported earlier.10,11,13 The results suggest that it may be easier for investigators to describe clinical features (inclusion criteria and therapeutic regimens) than to report trial methodology and statistical features of their studies. Yet, incomplete reporting of trial methodology impedes reading and systematic reviewing of articles.

No difference between German and English articles was found. The size of the sample would have allowed the discovery of a difference in the scores of 2 with a power of .9 on a 2-sided test at the α level of .05.

There are a number of limitations to the present study. The concept of quality poses some problems. The basic assumption underlying the measurement of quality in a single score is that items of quality are correlated with each other. This can be questioned because 1 fatal error can invalidate the results of otherwise excellent work. However, we are often not aware of the particular excluding criteria. Also, only papers already published were investigated. Within the present study, scoring of report quality was not duplicated and interrater reliability could therefore not be evaluated. External validity was aimed at by using a random sample. Finally, only articles of authors from German-speaking Europe were assessed. A further study is under way to measure the reliability of the scale used and to compare the present sample with articles from English language countries.

Standardized assessment of papers submitted for publication can improve the adherence of authors to these standards.14,15 Adherence could be promoted by better dissemination of content and meaning of the standards such as the Uniform Requirements. The Uniform Requirements are also known as “The Vancouver Style,” suggesting that the Uniform Requirements only refer to the style of manuscripts, but not to methodological and statistical reporting. MEDLINE references 58 publications of the 5 editions of the Uniform Requirements in 8 languages. A German translation of the fourth edition was published in 1993, but only by an ophthalmology journal.30

The Consolidated Standards of Reporting Trials (CONSORT) statement is an important step toward improving reporting of quality of RCTs since it requires that authors themselves complete the checklist on reporting quality.31 It might be recommended that all journals, including those not following the CONSORT statement strictly, request their authors to fill out a checklist on reporting quality.

This work was part of a wider project looking at language bias in meta-analysis (supported by the Swiss Academy of Medical Sciences, Basel, Switzerland). The author was partly funded by the Swiss National Science Foundation, Bern, Switzerland.

I thank Theodor Aeblin, MD, MPH, and Matthias Egger, MD, MSc, for their substantial support. Christoph E. Minder, PhD, for advice on the statistical methods, and the anonymous reviewers for their clarifying comments.

**References**