ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 241−242

Journal of Clinical Epidemiology

EDITORIAL

Are we doing enough to ensure quality of trials?

The Journal of Clinical Epidemiology has an ongoing interest in study quality including publication bias [1,2]. A full 30% of trials published in high-impact journals still fail to measure patient-important outcomes and use safeguards against bias. Bala and colleagues compared the methodological characteristics of randomized controlled trials (RCTs) published in 219 higher-impact and 250 lower-impact core clinical journals. They found that higher-impact journals enroll substantially more patients, more frequently measure patient-important primary outcomes, conduct prespecified subgroup analyses, report a test of interaction for subgroup effects, and report greater safeguards against the risk of bias. Despite these journals having more credibility than lower-impact journals, there were major omissions in all journals studied. The authors warn that readers of these journals should be aware that publication in a prestigious journal does not ensure a low risk of bias.

Citation bias is a variant of publication bias where studies with statistically significant results get cited more often than published studies with statistically nonsignificant results. Janno et al conducted a cohort study of all therapeutic intervention studies included in meta-analyses published between January and March 2010 in the Cochrane database. They identified 89 research questions addressed in 458 eligible articles. Significant studies were cited twice as often as nonsignificant studies. The implication of this is that authors looking for references in articles to explore the literature on a specific question should be cautious because treatments may thus seem more effective to the readers of medical literature than they really are. Similarly, an increasing amount of research and guidelines has been published on search methodology and the reporting of search strategies in systematic reviews; Golder et al assessed whether this has led to any improvements in the reporting and quality of searching in systematic reviews of adverse effects. They found that although some improvements are apparent, poor reporting of search strategies does continue to be a significant obstacle.

Do web-like representations such as spydergrams, radial plots, or stargrams help readers appreciate small differences in the results of substantive trials that are not so obvious on bar graphs? In this issue of the JCE, an interesting Variance and Dissent is presented on the controversy about the utility of spydergrams. Boers asks for a debate on methodologic issues raised by an article that was recently published using spydergrams to present and interpret SF-36 health-related quality of life data within rheumatic diseases [3]. Spydergrams are an attempt to graphically depict more than two dimensions on a flat surface. He begins by presenting some of the challenges that spydergrams present. These include possible surface area distortion, comparison of surface area as opposed to rings, and correlation of the sub-dimensions. He then presents two alternatives to spydergrams, the transparent radar plot and the stargram. However, he maintains that a simple bar graph is usually the best option. Strand and Loftis present a countering view based on the tenet that graphs should be judged on how they are actually used and presented. They also state that guidance for the interpretation of graphs and particularly spydergrams should accompany the graphic. Based on these two tenets, the latter authors argue that spydergrams are a useful means with which to interpret data. We welcome the views of our readers.

One paper reports on patient preferences in decision support. Valdes and colleagues generated and validated a scale to measure the informed choice of contraceptive methods in a middle-income country setting. Informed choice is a multidimensional construct made up of orientation, information, communication, and the quality of treatment. The scale was deemed to be culturally appropriate for women attending a family health care service in Chile.

A different aspect of preferences are reviewed by Prady and colleagues. Preference trials are recommended for situations in which participants are likely to want one particular treatment that results in differential dropout caused by baseline differences in preferences between groups. They focused on acupuncture trials and found 31 acupuncture trials (RCTs and CCTs) that fulfilled the inclusion criteria. Six were comprehensive cohorts (one also elicited postrandomization treatment preferences), seven were preference designs, and 18 elicited treatment. The results from this review indicate that around three quarters of potential participants turned down the offer of being randomized into an open-label acupuncture trial, which makes it a great case study. However, the details are poorly reported, which permit the benefits of the different preference designs to be assessed.

Recently, there have been major advances in statistical techniques for assessing central tendency and measures of association. Wilcox et al claim that, during the last 25 years, many new and improved methods for comparing

0895-4356 © 2013 Published by Elsevier Inc. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.jclinepi.2012.12.007

groups and studying associations have been derived that have the potential of documenting important statistical relationships that are likely to be missed when using standard techniques. They maintain that, although the practical utility of modern methods has been documented extensively in the statistics literature, they remain underused and relatively unknown in clinical trials. They address this issue by reviewing common problems with standard methodologies by summarizing alternative methodologies and by illustrating the practicality of those methodologies in a randomized control trial using one of the most important software developments during the last 30 years: the free software R.

The body of evidence for computer-adaptive testing (CAT) improving measurement precision and efficiency in assessing physical function and fatigue, with an impact on reducing sample size requirements, is added to by the study of Petersen et al from the European Organisation for Research and Treatment of Cancer Quality of Life Group. Geere et al assessed more traditional instruments to estimate whether the predefined variables in study design, instrument type, and patient characteristics account for variance in reported retest reliability for the Oswestry Disability Index and Roland Morris Questionnaire. They found that study design and population influence the reliability of a given instrument. However, a greater difference in reliability exists between instruments with the Oswestry Disability Index, was more reliable than the Roland Morris Questionnaire after adjusting for confounding.

Koopman et al investigated whether a prenotification letter instead of a second reminder and varying senders of the questionnaires would improve response rates of medical and health surveys. They found that the prenotification groups returned their questionnaires faster. However, no significant differences were found for response speed, respondent characteristics, item nonresponse, or mean scores. They conclude that a prenotification letter should be considered when quick response is desirable, but that this will not increase overall response rates.

Risk prediction models are increasingly being called for as we move from individual care to systems, as well as having to make choices in the allocation of scarce resources. The methods are often chaotic as shown in a systematic review by Collins et al, who found 14 different models just for chronic kidney disease. These were often developed using inappropriate methods and were generally poorly reported. They indicate that using poor methods can affect the predictive ability of the models, whereas inadequate reporting hinders an objective evaluation of the potential usefulness of the model. In a similar vein, Hudson and colleagues conducted a systematic review to determine the validity of the diagnostic algorithms for osteoporosis and fractures using administrative data. Following a review of 12 studies, they found that administrative data can be used to identify hip fractures. However, existing diagnostic algorithms to identify osteoporosis and vertebral fractures in administrative data were found to be suboptimal.

Peter Tugwell
André Knottnerus
Leanne Idzerda
Editors
E-mail address: lidzerda@uottawa.ca (L. Idzerda)

References

[1] Guyatt G, Oxman A, Vist G, Kunz R. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.
[2] Moreno SG, Sutton AJ, Ades AE, Cooper NJ. Adjusting for publication biases across similar interventions performed well when compared with gold standard data. J Clin Epidemiol 2011;64:1230−41.
[3] Strand V, Crawford B, Singh J, Choy E, Smolen JS, Khanna D. Use of ''spydergrams'' to present and interpret SF-36 health-related quality of life data across rheumatic diseases. Ann Rheum Dis 2009;68:1800e4.