Journal of Clinical Epidemiology 66 (2013) 1061−1063

Journal of Clinical Epidemiology

EDITORIAL

The potential impact of unpublished results

Discussions on publication bias have taught the clinical community that selective (non)reporting of research results will generally distort both the available evidence picture and its clinical interpretation. Selective (non)reporting may occur deliberately or non-deliberately. The first can, for example, be the result of prejudiced decisions of researchers or pharmaceutical companies to only publish results that they consider favorable for their careers or products. The latter can happen when authors or editors think that only positive results are interesting for readers, or when the research has been done in a (e.g., educational) context or culture where publication of negative findings is not current practice. Also, it is possible that investigators or authorities, correctly or incorrectly, believe that full publication of results can be a threat to public safety [1]. All these variants have the same implications, in the sense that clinical decisions can be unfavorably affected, contributing to suboptimal health care, or that opportunities for scientific progress will be missed. Also, independent peer review and the open scientific debate on such research are eliminated, implying that the usual quality control by the scientific community cannot take place. Moreover, when research is not publicly reported, colleagues in the same field may never know about it and they may needlessly repeat the same experiments, while putting extra burdens on patients and, in the biomedical domain, on laboratory animals.

Obviously, selective (non)publication is a major problem for systematic reviews that aim to present a comprehensive and appropriate evidence picture. Although there have been developed approaches to identify and deal with this problem [2], selective publication may still go unnoticed. The use of international trial registers has done much good and may, when stringently applied, in the long run minimize the problem for experimental research. However, for non-experimental research, these registers do not (yet) exist [3] and are not easy to implement with sufficient coverage.

Of course, a key question is to what extent selective (non)publication does occur and what its impact may be. In this context, the article of McDonagh et al. provides very important information. They examined the DERP (Drug Effectiveness Review Project) systematic review reports published since 2003 for the use of FDA (US Food and Drug Administration) preapproval and post marketing documents to identify important unpublished evidence. The authors found that a minority of FDA documents contain unpublished evidence that can be highly useful in resolving publication bias and selective outcome and analysis reporting, and conclude that FDA documents can provide important unpublished evidence for systematic reviews. Recommendations for further research and for improving access to this important FDA information are presented. This important article should also stimulate active investigations of similar sources of unpublished evidence worldwide.

Additionally, Thaler and colleagues provide important insights in the way unpublished information can play a role. They systematically evaluated pooled-study publications (PSPs) that present statistical analyses of multiple randomized controlled trials (RCTs) without a systematic literature search or critical appraisal. These PSPs were excluded from a systematic review of second-generation antidepressants (SGAs). The authors report that PSPs of SGAs are almost exclusively funded by the pharmaceutical industry and often include unpublished data on secondary outcomes and subgroups that are not available in the primary publications. They conclude that guidance for reviewers and a system to assess susceptibility of PSPs to bias are required.

Guidance for reviewers is clearly a topic of more general interest. Berkman and her team examined the interrater reliability of applying guidance for grading strength of evidence in systematic reviews for the AHRQ (Agency for Health Research and Quality) Evidence-based Practice Center program. Based on data from two systematic reviews, the investigators found that the interrater reliability was highly variable for scoring strength of evidence domains, and low for combining scores to reach overall strength of evidence grades. Recommendations were made to support reviewers and for future research focused on improved methods in this field of evaluating complex bodies of evidence. It is interesting to connect Berkman's findings to the recent paper by Mustafa et al. [4] from the GRADE group, who found that trained individuals can reliably assess the quality of evidence using the GRADE system.

Speaking about more adequate reporting of findings that may be relevant to research and practice: case reports deserve more attention. According to Sun and co-workers, case reports are widely used to generate hypotheses, to

http://dx.doi.org/10.1016/j.jclinepi.2013.08.001

document rare or unusual phenomena, and to provide clinical stories for educational purposes. However, there is a need for more consistent quality in case reports. Therefore, these authors propose publication standards for case reports to improve manuscript quality and to enhance their usefulness for education, research, and clinical purposes. This paper precedes a broader initiative, entitled "Consensus-Based Case Reporting (CARA) guidelines," which will be published by the *Journal of Clinical Epidemiology*, along with other journals, later this year. We strongly encourage these groups to come together and agree on a single checklist to avoid confusion. Such collaboration between various initiatives on reporting standards has been highly successful before: a similar collaboration is how CONSORT was born. Also, correctly using and harvesting the wealth of information in routine administrative health care databases has a huge potential of enriching the arena of published data. Huber et al. provide an updated, pharmacy-based chronic disease score (CDS), based on a cohort study using medical claims data from insured persons. The CDS showed reasonable predictive validity of future health care utilization and medical expenditure. This may support health care decision makers in planning care delivery and allocating resources. Routine databases may also be used as a reference for validating self-reported health information. Parkinson's team examined the level of agreement between self-reported health and hospital administration records of arthritis-related surgeries in older women. Based on the good agreement between both sources, the authors support the use of self-reports surveys in epidemiological studies of joint procedures when adequate administrative data are not available or accessible. Also Teh and her group studied the agreement between self-reported health and medical records. They focused on octogenarians, using medical records on specific cardiovascular diagnoses. They concluded that the reliability of self-reported information on specific cardiovascular conditions is modest in octogenarians and is influenced by various factors, especially the number of comorbities. When deciding whether or not to use self-reporting of these conditions, participants' characteristics that may affect the level of agreement should be considered.

Results to be achieved in real practice may not always be well predicted by research based on randomized trials. Among the reasons for this can be methodological challenges such as nonadherence and contamination. Using model calculations, Brenner c.s. analyzed the potential impact of these phenomena in RCTs evaluating endoscopic screening for colorectal cancer. The authors conclude that, in the era of widespread endoscopy use even outside screening programs, RCTs may strongly underestimate the effects of colorectal cancer screening and that additional analyses are crucial for disclosing the true screening effects. Study participation was also evaluated in pediatric clinical research. Based on a cross-sectional survey on all clinical studies conducted in six pediatric clinical investigation centers, Kaguelidou and co-workers evaluated refusal rates. It was found that the refusal rate was low and that it was influenced by characteristics of the studies and the recruiting physicians. The authors make recommendations on how to further improve recruitment in pediatric clinical research.

In a systematic review, which is an update of a similar study in 2004 [5], Whiting et al summarize the current evidence on the sources of bias and variation in studies of the accuracy of diagnostic tests. Consistent evidence was found for the effects of a number of sources of bias and variation, having generally more impact on sensitivity than specificity. Accordingly, decades after the appearance of early work stressing the need for better diagnostic research [6−8], there is still substantial room for improvement in minimizing the potential for bias in primary diagnostic accuracy studies.

Van Hoorde c.s. considered the well-known observation that prediction models may perform poorly in other settings. Using case studies on testicular and ovarian tumors, they studied which model updating methods should be applied for models of polytomous outcomes. Simple dichotomous methods behaved well when applied to polytomous models. The results suggest that recalibration is preferred, but when larger validation sets are available, revision or even redevelopment can be a valid alternative.

A number of articles address the validity and performance of instruments. Because there were limited data on health-related quality of life (HRQL) in chronic obstructive pulmonary disease (COPD) patients with chronic hypercapnic respiratory failure, Struik et al. assessed and compared the performance of four instruments among patients with severe COPD. The Severe Respiratory Insufficiency (SRI) questionnaire performed slightly better than the other three, the authors suggest the SRI to be the preferred HRQL instrument in patients with very severe COPD. Nikolaus and co-authors developed an item pool based on the patients' perspective to construct a future computerized adaptive test for fatigue in rheumatoid arthritis. Using data from a large group of patients with rheumatoid arthritis, they provided an initially calibrated item bank and showed which dimensions and items can be used for the development of a multidimensional computerized adaptive test for fatigue. Also starting from the patient's perceptions, Carlesso c.s. performed a cross-sectional survey of patients receiving manual physiotherapy, focusing on the identification and occurrence of adverse responses related to manual therapy and predictors of such responses. Based on their findings, the authors conclude that for developing a comprehensive framework for defining adverse responses in manual therapies, the patient perspective is indeed important.

While summary scores from methodological quality scales (such as the PEDro scale) are often used, da Costa et al. argued in an earlier paper that the use of such scores should be discouraged and that the PEDro database be restricted to presenting the scores for individual items of the scale [9]. In a letter to the editors, Costa et al. comment on this paper, and suggest further research on this topic.

We have received many positive responses to the series by Cals and Kotz on effective writing and publishing scientific papers. We therefore welcome the new one-pager in this series, addressing 'the discussion.'

J. André Knottnerus
Peter Tugwell
Editors
E-mail address: anneke.germeraad@maastrichtuniversity.nl (J.A. Knottnerus)

References

[1] Fouchier RA, García-Sastre A, Kawaoka Y. The pause on Avian H5N1 influenza virus transmission research should be ended. MBio 2012;3(5). pii: e00358-12. doi: 10.1128/mBio.00358-12.

[2] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence−publication bias. J Clin Epidemiol 2011;64:1277−82.

[3] Swaen GM, Carmichael N, Doe J. Strengthening the reliability and credibility of observational epidemiology studies by creating an Observational Studies Register. J Clin Epidemiol 2011;64:481−6.

[4] Mustafa RA, Santesso N, Brozek J, Akl EA, Walter SD, Norman G, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. J Clin Epidemiol 2013;66:736−42.

[5] Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189−202.

[6] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926−30.

[7] Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411−23.

[8] Knottnerus JA. Interpretation of diagnostic data: an unexplored field in general practice. J R Coll Gen Pract 1985;35:270−4.

[9] da Costa BR, Hilfiker R, Egger M. PEDro's bias: summary quality scores should not be used in meta-analysis. J Clin Epidemiol 2013;66:75−7.