Rev Saúde Pública 2006;40(N Esp):30-5

Moyses Szklo

Quality of scientific articles

Qualidade de artigos científicos

ABSTRACT

The paper discusses the difficulties in judging the quality of scientific manuscripts and describes some common pitfalls that should be avoided when preparing a paper for submission to a peer-reviewed journal. Peer review is an imperfect system, with less than optimal reliability and uncertain validity. However, as it is likely that it will remain as the principal process of screening papers for publication, authors should avoid some common mistakes when preparing a report based on empirical findings of human research. Among these are: excessively long abstracts, extensive use of abbreviations, failure to report results of parsimonious data analyses, and misinterpretation of statistical associations identified in observational studies as causal. Another common problem in many manuscripts is their excessive length, which makes them more difficult to be evaluated or read by the intended readers, if published. The evaluation of papers after their publication with a view towards their inclusion in a systematic review is also discussed. The limitations of the impact factor as a criterion to judge the quality of a paper are reviewed.

KEYWORDS: Editorial policies. Peer review, research. Evaluation, standards. Publications, standards. Scientific publications.

RESUMO

O artigo discute as dificuldades em julgar a qualidade de manuscritos científicos e descreve algumas das armadilhas comuns que devem ser evitadas ao preparar um artigo para submissão a um periódico. A revisão por pares é um sistema imperfeito, com confiabilidade abaixo da ideal e validade duvidosa. Entretanto, como provavelmente continuará sendo o principal processo de seleção de artigos para publicação, os autores devem evitar alguns erros ao preparar um manuscrito baseado em achados empíricos da pesquisa humana. Entre eles estão: resumos excessivamente longos, amplo uso de abreviações, falha ao relatar resultados de análise de dados parcimoniosa, e interpretação equivocada de associações estatísticas encontradas em estudos observacionais como sendo causais. Outro problema comum em muitos manuscritos é sua excessiva extensão, o que os torna mais difíceis de serem avaliados ou lidos por leitores-alvo, se publicados. Discute-se a avaliação de artigos após sua publicação e inclusão em revisão sistemática. São revisadas as limitações do fator de impacto como um critério para julgar a qualidade de um artigo.

DESCRITORES: Políticas editoriais. Pesquisa avaliada pelos pares. Avaliação, normas. Publicações, normas. Publicações científicas.

Department of Epidemiology. Bloomberg School of Public Health. The Johns Hopkins Medical Institutions. Baltimore, MD, USA

Correspondence:
Moyses Szklo
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street; Room W6009
21205 Baltimore, MD, USA
E-mail: mszklo@jhsph.edu

Received: 6/12/2006

INTRODUCTION

In 1731, the Royal Society of Edinburgh stated that "Memoirs sent by correspondence are distributed... to those members who are most versed in these matters... Nothing is printed in the review which is not stamped with the mark of utility."[11] Thus, the peer review process – which represents the main approach to evaluate the quality of scientific publications – is not recent, even though it was fairly rare until after the Second World War, when the increasing number of papers submitted to journals and the increasing specialization of science made it virtually mandatory.[5]

According to Lock,[10] a former editor of the British Medical Journal, peer review has three main objectives: prevent publication of bad work, improve scholarship, and improve language and data presentation. That these objectives are probably not being thoroughly met has been known for some time. For example, Yankauer,[21] a former editor-in-chief of the American Journal of Public Health (AJPH) found out that 61 of 72 papers submitted to that journal had been submitted initially to other journals; however, only 1/3 of these had been revised substantially prior to submission to the AJPH. This finding was consistent with that of other study showing that, of the 70-85% of papers rejected by the British Medical Journal, only 20% were changed substantially before submission to a different journal.[11] In addition, it appears that a large proportion of published articles are never used as reference after publication.[11,20] Thus, the extent to which peer review is achieving an improvement in scholarship is open to question.

The objectives of peer review proposed by Lock[10] suggest a conceptual framework within which quality of articles might be judged. Thus, the two main questions that outside experts and editors must ask when reviewing a scientific paper are: (a) Does the research on which the manuscript is based advance scientific knowledge? and, (b) Is the report clearly written? Not always the answers to these questions converge. When the study seems to advance knowledge notwithstanding problems with the report's data presentation and language, it may well be salvageable, in which case the editor defers the decision on acceptability pending the necessary revisions. On the other hand, a negative answer to the first question makes the second moot, as it implies that the paper should be rejected (Table).

HOW TO JUDGE QUALITY

How can the quality of a paper be judged, and which are the "risk factors" for a paper lacking in quality? Assessment of quality remains elusive. In the American Journal of Epidemiology, for example, quality items typically include originality, design, conclusions, importance/interest, presentation, and documentation. However, the reviewer is not given instructions as to how these should be assessed. Each of these items is assigned to an "excellent", "very good", "satisfactory" or "poor" category. On the basis of this classification, which also includes an overall category ("overall quality"), the editor decides to accept with routine copy editing, accept with only minor revisions, defer the decision because the paper is currently unacceptable (with an indication as to whether or not the paper is likely to become acceptable after major revisions), or reject the paper.

Although in most peer reviewed journals in the public health area the instructions to reviewers are as relatively nonspecific as those from the American Journal of Epidemiology, some authors have tried to quantify quality in a more systematic fashion. An example is Chalmers et al[6] scoring system for evaluating the quality of a clinical trial, which is based on an assessment of both internal validity items addressing comparability (e.g., analysis of efficacy of randomization, blinding in evaluating outcomes, compliance), and external validity items related to the generalizability of findings (e.g., study population characteristics, list of eligible but not enrolled individuals, discussion of side effects). Basically, however, peer review remains a process with imperfect reliability and undetermined validity.[2,8,14,16]

Systematic reviews are noteworthy – as well as its quantitative expression, meta-analysis – as instruments for evidence-based medicine and public health,[4] given their increasing popularity and because their analytic unit is *study* (rather than *individual*). Thus, it is also important to judge the quality of papers *after they are published*. One of the ways in which quality of papers should *not* be judged after publication is by use of the citation indexes of the journals

Table - Usual editorial decisions according to characteristics of an article submitted to a reputable peer-reviewed journal.

Study improves scholarship	Report is clearly and accurately written	Usual editorial decision
No	No	Rejection
No	Yes	Rejection
Yes	No	Deferral
Yes	Yes	Acceptance

in which they were published. In a recent study by Berghmans et al,[3] for example, a very weak association was found between the quality of lung cancer randomized clinical trials and these indexes, which included the impact factor.* A similar conclusion was reached by Barbui et al[1] with regard to trials related to an antidepressant. Because it is a common practice to use the impact factor as a marker of quality of articles in certain academic settings, particularly in some European countries (where it can be used when considering the promotion of a faculty member to a higher rank), it is useful to list some important conceptual and technical flaws inherent to this index:[20]

1. the number of journals in the Institute for Scientific Information, which estimates the impact factor, is a small proportion of those published;
2. literature reviews are cited more often than results of single empirical studies, thus favoring journals that publish reviews frequently or exclusively;
3. the impact factor does not take self-citations into consideration;
4. the accuracy of the assumption of a positive link between citations and quality is questionable, as papers can, for example, be cited because they are based on flawed studies.

"RISK FACTORS" FOR POOR QUALITY OF EPIDEMIOLOGIC MANUSCRIPTS

It is possible to follow some simple rules and avoid some pitfalls that may decrease the likelihood of acceptance of a paper. The discussion that follows is not meant to convey rigid guidelines, but merely point out some ways to improve a paper's quality based on this author's experience with common avoidable mistakes detected over the years in papers submitted to the American Journal of Epidemiology.

Assuming that the research findings being reported are sound, one of the key objectives of a scientific paper is that the rationale, design, findings and inferences from the study results be clearly conveyed and understood by the intended readers. Therefore, one of the quality issues that authors should keep in mind when writing their papers is, What is the target audience for whom the publication is intended? Epidemiologists? Other public health practitioners? Policy makers? Practicing physicians? Each type of reader requires a different approach to reporting study findings. As a general strategy, however, it is highly recommended that the language used be both simple and unambiguous. Technical terms should be explained and abbreviations avoided, with a few possible exceptions represented by well known abbreviations, such as HIV or EKG. Consider, for example, the following sentence, which is difficult to understand in view of the large number of abbreviations: *The US-determined IMT of participants with AP showed a strong correlation with ABI. Patients with low ABI also had a high CRP.* When properly modified, the sentence should read, *The B-mode ultrasound-determined intimal-medial thickness of the carotid arteries in participants with angina pectoris showed a strong correlation with ankle-brachial index. Patients with low ankle-brachial index also had a high level of C-reactive protein.*

It should be also strongly emphasized that, even within the length allowed by the journal to which the paper is submitted, the paper should be as short as possible, as a short paper not only makes the journal more reader-friendly, but also eases the burden on reviewers and editors.[18] The preparation of a short (rather than a long) report is a difficult task that may require a considerable time investment, as highlighted by a reputed letter from Mark Twain to a friend: *Dear friend: I am sorry I had to write you such a long letter. I did not have time to write you a short one…*

Using the format favored by most epidemiologists and empirical researchers in the biomedical field since the 1970s[15] (IMRAD, or – in addition to the abstract – "Introduction", "Methods", "Results" and "Discussion"), the following pitfalls should be avoided:

Abstract

The abstract is often the only part that is read. It should be used as a way to encourage the potential reader to read the whole paper. Abstracts should state findings and conclusions, not only the purpose of the study (as done often in abstracts sent in advance of a meeting's oral presentation). Thus, "Results will be presented…" is not allowed in abstracts of manuscripts submitted to a peer-reviewed journal for consideration for publication. Abstracts are of poor quality if they are long, use abbreviations and acronyms, and include redundant information, such as both confidence limits and standard errors.

Introduction

The Introduction is the section where the study's justification and statement of hypotheses belong. It should contain an extensive review of the literature, but rather a review of scientific findings that are relevant to the study hypotheses. In a biomedical and public health

*Defined by Garfield[7] as the number of citations to the articles published in a given journal during the two preceding years, divided by the number of articles published in that journal within the same period.

paper, it is often necessary to state the plausibility of the hypothesis (biologic, sociologic, and others). It should be also clearly stated whether an observed interaction was an *a priori* hypothesis, rather than a result of subgroup analysis. Authors should in addition describe the main reason why the study was carried out: to test a new hypothesis (or hypotheses); to confirm a previously tested hypothesis when replication is necessary; or to examine a previously tested hypothesis using a new design or methodology that is superior to those used in previous studies.

Methods

Methods is a key section, as it allows the reviewers and the editor – and, if the paper is published, the readers – to judge the soundness of the study design and the analytic strategies. In order to evaluate the level of the study's external validity (generalizability), the study population should be described in detail; inclusion and exclusion criteria should be specified. A description of the study variables and confounders, including the categories chosen and the criteria to select them should be provided (e.g., based on traditional categorization such as ages 45-49, 50-54 and so on, or for continuous variables, use of percentile values, such as the median or quartiles). The approach to identify confounders and effect modifiers should also be described. Whereas the former is usually based on either previous literature or assessment of study data, the latter is a function of the biologic (or other) plausibility of the hypothesis. As mentioned previously here, if subgroup analyses are carried out and interaction(s) were not part of the *a priori* hypothesis, specification of these analyses as exploratory is necessary.

Data collection procedures must be described in detail. If the study was questionnaire or inventory-based, the authors should offer to publish the instrument online (provided that there are no copyright issues). Quality assurance – which includes training of data collectors – and quality control measures should be described. The latter include observation monitoring, studies of replicate samples using phantoms, and estimation of validity and reliability indexes, such as sensitivity, specificity, kappa and intra-class correlation coefficient.

Analytic approaches should be as parsimonious as possible. For example, for case-control data, if only a few variables have to be adjusted for, it may not be necessary to use logistic regression (unless these variables are continuous). A simpler method, such as the Mantel-Haenzsel technique to adjust for odds ratios, may be sufficient.

When assessing interactions, most authors choose the interaction model as a function of the regression model used for adjustment. Because these models are often ratio-based (e.g., logistic or Cox multiple regression), reference in the paper to interaction without specification of the interaction model often means multiplicative interaction. However, regardless of the model chosen for adjustment, additive interaction is important for public health purposes,[13,17] and can be readily assessed even in the context of ratio-based models.[19]

A simple technique that is becoming increasingly popular, and allows establishing the limits of the accuracy of the inferences drawn by the authors is sensitivity analysis, which is generally defined as "the quantitative assessment of systematic error on an effect estimate".[9,*] It is a tool to examine changes in the output (results) of a given model resulting from varying certain model parameters (or assumptions) over a reasonable range.[17] Use of this technique reflects the authors' concerns with validity of their findings, and thus it is looked upon favorably by journal reviewers and editors.

Results

The presentation of results should go from the most to the least parsimonious approaches. Thus, it is strongly recommended that authors begin this section with a presentation of frequency distributions, means, medians. In observational studies the likelihood of finding a real association is, among other factors, a function of the variability of the factors under study. Thus, authors should keep in mind the possibility of systematically reporting coefficients of variation (mean values divided by their standard deviations) and/or inter-quartile ranges for key variables. The Results section should, in addition, include a comparison between individuals included in the study and those who, although eligible, refused inclusion. For prospective studies, at a minimum, a comparison between persons lost and those not lost to observation with regard to baseline characteristics should be provided.

Stratification allows both (partial) assessment of confounding and examination of heterogeneity (interaction), and should generally follow the data presentation described in the previous paragraph. Following stratification, results of regression methods should be presented, if necessary. However, when reporting these results in a table format, the authors should be as reader-friendly as possible by:

1. presenting the most easily understandable figu-

*Sensitivity analysis can be also used to assess the impact of random errors. See, for example, Peppard et al.[12]

res – e.g., odds ratios instead of regression coefficient;
2. using stratification for reporting interactions, and limiting the use of interaction terms to statistical testing, as outside the realm of prediction these terms are not readily interpretable;
3. making sure that the titles of tables and figures are so informative that their interpretation does not require that the reader consult the text;
4. specifying units for each variable entered in a regression model (e.g., age in individual years or age groups, smoking as current, former and never smokers);
5. avoiding redundancy, such as for example reporting both p values and confidence intervals for a given measure of association.

For figures, the most common mistake in papers submitted to the American Journal of Epidemiology is the use of an arithmetic scale in the ordinate with a reference of zero for plotting odds ratios or relative risks. Because these measures of association are ratios, a semi-log scale with a baseline of 1.0 should be always used. However, the arithmetic scale can be used when plotting the logarithms of these ratio-based association measures.

Discussion

The Discussion section is where the authors have the greatest latitude to speculate about the findings of their study. The quality of a Discussion section depends on how persuasive the authors' arguments are with regard to having proved or disproved the hypothesis. There is no standard format for this section, but in most published papers, the following sequence is used by authors:

1. a brief review of the study findings (e.g., "In our study, a relationship was found between x and y, which could not be explained by the effects of known confounders";
2. a review of the pertinent literature, in which the similarities and discrepancies between the findings of the present study and those of previous studies are highlighted;
3. a discussion of the strengths and limitations of the present study, and how future studies can improve the methodology and design needed to reach firmer conclusions about the hypothesis;
4. if appropriate, provide a discussion of policy implications. With regard to this issue, caution is necessary because translation of epidemiologic findings into policy decisions – particularly as they relate to primary prevention – requires that a cause-effect relationship be established, which is impossible in the context of a single observational study. Thus, in reports of observational studies, the word "effect" should be avoided (such as when stating that "the effect of x on y was not changed by adjustment"); instead, these reports should refer to associations (which may or may not be causal).

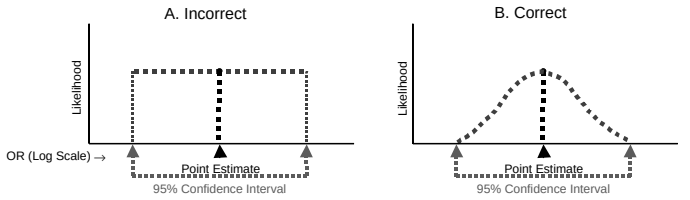

Figure 1 - Incorrect (A) and correct (B) interpretations of the 95% confidence interval of a relative risk or an odds ratio.

The interpretation of p-values and 95% confidence limits, which assist the authors in concluding whether or not an association is a chance finding, deserves special caution. Although informative, these values should not be construed as rigid rules on which to base definitive inferences. A hypothetical example would be a manuscript in which the authors concluded that excessive alcohol consumption is not associated with risk of causes other than liver cirrhosis, on the basis of relative risk (RR) of 1.5 and a 95% confidence interval for the RR of 0.9, 2.6. This inference is not warranted, as the authors overlooked the fact that the confidence interval function of the relative risk is not a straight line that is parallel to the abscissa, but rather follows a lognormal distribution (Figure 1). For this example, most of the area under the curve is to the right of the null hypothesis (Figure 2), thus suggesting that most of the range of possible values support the presence of an association (assuming no bias).

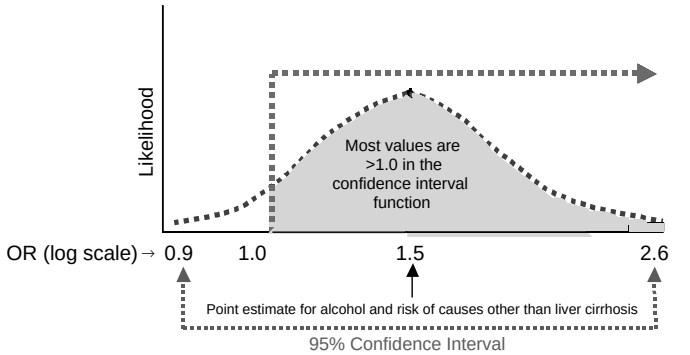

Figure 2 - Correct inference when the relative risk is 1.5 and the 95% confidence interval is 0.9, 2.6: there is an association, although the 95% confidence interval overlaps the null value (OR=1.0).

CONCLUSIONS

It is often difficult to judge the quality of a scientific article. The peer review process has been the bedrock of science, notwithstanding its flaws reflected by a less-than-perfect reliability and an unknown validity. However, as it is likely that the peer review process will continue to exist for years to come, authors should be well advised to follow a few simple rules in order to avoid a number of pitfalls when preparing a manuscript. This paper tried to summarize these rules, in the hopes that they could serve as guideposts for potential authors. It should, however, be emphasized that it is impossible to comment on all possible types of papers, and consequently this article was limited to a discussion of the kinds of challenges its author has faced as editor-in-chief of the American Journal of Epidemiology.

REFERENCES

1. Barbui C, Cipriani A, Malvini L, Tansella M. Validity of the impact factor of journals as a measure of randomized controlled trial quality. *J Clin Psychiatry.* 2006;67:37-40.

2. Berger E. Peer review: a castle built on sand or the bedrock of scientific publishing? *Ann Emerg Med.* 2006;47:157-9.

3. Berghmans T, Meert AP, Mascaux C, Paesmans M, Lafitte JJ, Sculier JP. Citation indexes do not reflect methodological quality in lung cancer randomized trials. *Ann Oncol.* 2003;14:715-21.

4. Bigby M, Szklo M. Evidence-based dermatology. In: Fitzpatrick's dermatology in general medicine. New York: McGraw-Hill; 2003. vol. 2, p. 2301-11.

5. Burnham JC. The evolution of editorial peer review. *JAMA.* 1990;263:1323-9.

6. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized clinical trial. *Control Clin Trials.* 1981;2:31-49.

7. Garfield E. Journal impact factor: a brief review. *CMAJ.* 1999;161:979-80.

8. Jefferson T, Wager E, Davidoff F. Measuring the quality of editorial peer review. *JAMA.* 2002;287:2786-90.

9. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology.* 2003;14:451-8.

10. Lock S. A difficult balance: editorial peer review in medicine. London: Nuffield Provincials Hospital Trust; 1985.

11. Lock S. Information overload: solution by quality? *Br Med J (Clin Res Ed).* 1982;284:1289-90.

12. Peppard PE, Young T, Palta M, Skatrud J. Prospective study of the association between sleep-disorder breathing and hypertension. *New Eng J Med.* 2000;342:1378-84.

13. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol* 1980;112:467-70.

14. Smith R. Peer review: a flawed process at the heart of science and journals. *J R Soc Med.* 2006;99:178-82.

15. Sollaci LB, Pereira MG. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J Med Libr Assoc.* 2004;92:364-7.

16. Strayhorn J Jr, McDermott JF Jr, Tanguay P. An intervention to improve the reliability of manuscript reviews for the Journal of the American Academy of Child and Adolescent Psychiatry. *Am J Psychiatry.* 1993;150:947-52.

17. Szklo M, Nieto FJ. Epidemiologic issues in the interface with public health policy. In: Epidemiology beyond the basics. 2nd ed. Boston: Jones & Bartlett Publishers; 2006. [in press]

18. The Editors. Rapid communications. *Am J Epidemiol.* 2006;163:299.

19. Thompson WD. Statistical analysis of case-control studies. *Epidemiol Rev.* 1994;16:33-50.

20. Walter G, Bloch S, Hunt G, Fisher K. Counting on citations: a flawed way to measure quality. *Med J Aust.* 2003;178:280-1.

21. Yankauer A. Peering at peer review. *CBE Views.* 1985;8:7-10.

*M. Szklo is also Editor-in-Chief of the American Journal of Epidemiology.