Learned Publishing (2001)**14**, 257–263

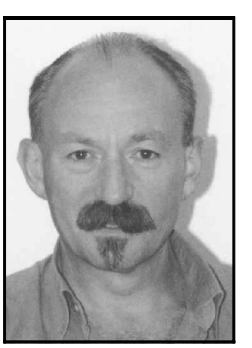
Shortcomings of peer review in biomedical journals

Elizabeth Wager
Sideview

Tom Jefferson
UK Cochrane Centre

ABSTRACT: Peer review is well established across most academic disciplines, and publishers, editors, and researchers devote considerable resources to it. This paper uses examples from biomedical journals to examine its shortcomings. Although mainly anecdotal, the evidence suggests that peer review is sometimes ineffective at identifying important research and even less effective at detecting fraud. Most reviewers identify only the minority of a paper's defects and they may be biased. Peer review plus other editorial processes are associated with improvements in papers between submission and publication, but published papers remain hard to read and a significant proportion contain errors or omissions. While it is hard to quantify the costs, peer review does not seem an efficient use of resources. Research into the outcomes of peer review, the establishment of sound methods for measuring the quality of the process and its outcomes, and comparisons with alternative methods are needed.

Elizabeth Wager



Tom Jefferson

Introduction

Peer review is embedded into the structures and processes of virtually all academic journals. The recent ALPSP survey[1] confirmed that, with only minor variations, peer-review processes are used across the range of disciplines encompassing the sciences, arts, and humanities. Virtually all academic institutions and many commercial organizations assess individuals for employment or promotion by their publication record, and the fate of whole departments or project teams may rest on their success as judged by the number and impact of their publications. An enormous amount of money and energy is therefore invested into peer review and it is almost impossible to imagine academic publishing without it. Yet, that does not mean that we should not examine it critically, consider its shortcomings and seek improvements.

To accept that a system has shortcomings, we must first agree on what we expect it to do, and on what a perfect process would look like. This is harder than it might seem, because peer review performs a number of functions and, despite its familiarity, measures of its quality have not been adequately defined.[2] However, most editors and authors would agree that peer review is designed to act as a filter and to improve submissions. Most would also say that the process should be fair, unbiased, and objective. It should make efficient use of resources and be better than alternative systems.

Peer review as a filter

An effective filter would have good sensitivity for important papers and high selectivity for poor research. It would take a major study to measure this function systematically, but there is anecdotal evidence that peer review sometimes fails to recognize important papers. Stephen Lock provides examples of the rejection of papers by two

Nobel Prize winners, one identifying the hepatitis B virus particle and the other describing the technique of radioimmunoassay for the first time. In the former case the reviewer suggested the particles were dirt on the microscope slide. In the latter case, the frustrated author was moved to comment that 'The truly imaginative are not being judged by their peers. They have none.'[3]

peer review fails to identify errors

There is more systematic evidence that peer review fails to identify errors in papers. Godlee and her colleagues at the *British Medical Journal* took a paper that had been accepted by the journal and, with the author's co-operation, used it to test reviewers' responses.[4] They modified the paper, introducing eight important methodological weaknesses, and sent it to people who review for the journal and were likely to have some knowledge of the subject area. Of the 221 reviewers who responded, the average number of weaknesses commented on was two out of the possible eight, 16% of reviewers failed to identify any weaknesses in the paper's methodology and only 10% identified four or more defects.

Peer review as an objective process

Lack of consistency and objectivity, and inability to spot papers that had previously been published were suggested by Peters and Ceci.[5] They took 12 articles originating from prestigious institutions and resubmitted them to the psychology journals that had published them 18–32 months previously. The only change they made was to replace the author's names and affiliations with fictitious ones so the submissions came from people and institutions that nobody recognized. Of the 12 articles, only three were identified as duplicates, one was accepted for publication (again) and the other eight were rejected. The stated reason for rejection in each case was methodological weakness rather than lack of originality. Peters and Ceci concluded that this demonstrated bias against less prestigious institutions. Although it is difficult to prove bias on the basis of single submissions to a range of journals, and alternative explanations for the results have been suggested,[6] this study certainly does not increase confidence in the effectiveness and objectivity of peer review.

Another study designed to determine if the peer-review process is biased suggests that reviewers are affected by the language in which a paper is written.[7] Magne Nylenna and colleagues sent reviewers a manuscript written either in their native languages (Danish, Norwegian, or Swedish) or in English. The manuscript contained methodological weaknesses such as inappropriate sampling units, statistical tests and control groups, and 'suboptimal standards of problem formulation, interpretation and language style'. Reviewers who had received the English-language version of the paper rated it, on average, more highly than those who had received the version in their native language. However, like Godlee *et al.*, the Scandinavian study found that only 25% of the 80 reviewers mentioned the incorrect sampling unit and only one commented on the inappropriate statistical test. On a four-point scale (where 4 indicated that all major flaws were identified), the average reviewer's score was 1.7.

Peer review and fraud

Another area in which peer review has not performed particularly well is the detection of fraud. Since it is very hard to measure the extent of misconduct, it is almost impossible to find a denominator against which to measure the effectiveness of peer review. So the evidence is, once again, anecdotal, but nonetheless thought-provoking. Stephen Lock provides a useful review of major cases that came to light in the 1980s.[6] These included John Darsee who published 44 invalid papers based on falsified results, Elias Alsabti who published at least 60 plagiarized papers, and Joseph Cort who published two papers on a molecule that had not even been synthesized. More recently, the Annual Reports of the US Office of Research Integrity (ORI) provide details of the cases that resulted in findings of scientific misconduct.[8–10] Those from 1997–9 are summarized in Table 1. All the published cases involved data fabrication or falsification. These cases resulted in the retraction or correction of over 30 papers.

Table 1 Cases of research misconduct reported by the Office for Research Integrity 1997–9[8–10]

	Total no. cases	No. published	Total no. publications affected	No. journals affected
1997	14	5	15	14
1998	9	4	>5[a]	>5[a]
1999	12	5	10	9
Total	25	14	>30	>28

[a]One case reported 'several publications'.

They had been published in a wide range of journals including respected titles such as the *American Journal of Psychiatry*, *Blood*, *Nature*, and the *Proceedings of the National Academy of Science*. In only one of the cases had the misconduct been detected during a journal peer-review process, but only after false data had already been published in five research papers and two review articles.

In two cases reported recently by ORI, misconduct arose as a direct result of the abuse of the peer-review system. In 1997, ORI demonstrated that a scientist had plagiarized the aims, background, experimental design, research plan, and most of the references from a grant application he had reviewed, and used these to submit his own application. In 1999, ORI reported that a member of an NIH review committee retained a grant application and an unpublished manuscript and used these to plan similar experiments. The former case was held to constitute plagiarism, and resulted in a finding of research misconduct. However, the 1999 case did not result in a positive finding against the scientist, since it was deemed 'a serious deviation from commonly accepted practices within the scientific community for conducting research, but was not serious plagiarism' because the experiments were 'directed at a different biological specimen'.

The extent of deliberate abuse of the peer-review system is, like other forms of scientific misconduct, hard to determine, but Rennie cites several examples from the 1980s.[11] In 1989, the NIH determined that David Bridges had abused his position as a reviewer. He had been asked to review a manuscript about rhodopsin regeneration for the *Proceedings of the National Academy of Sciences*. He retained the paper for several weeks then returned it, explaining that he was unable to review it fully because he was working on a similar study. However, he enclosed a handwritten note saying that the paper was poorly written and lacked primary data. Nevertheless, *PNAS* accepted the paper and published it in April 1987. In the meantime, Bridges submitted a paper on the same subject to *Science*, where it appeared in June 1987. An NIH investigation found that Bridges had not been working on this topic before he received the paper from *PNAS*, that he could not substantiate his research records, and that his own paper had 'internal inconsistencies, incomplete data and misrepresentation'.

Rennie also cites an instance where abuse of the peer-review system led to costly litigation.[11] In 1983, scientists working for Cistron Biotechnology Inc. submitted a manuscript describing the DNA sequence for interleukin-1 to *Nature*. It was reviewed by Steven Gillis who worked for a rival company, Immunex. He recommended rejection and enclosed a letter to the editor asserting that he was in possession of the correct sequence and implying that the one in the manuscript must therefore be incorrect. *Nature* declined to publish the paper from Cistron but it was accepted by *PNAS*. In the meantime, both biotech companies obtained patents on elements of interleukin-1. In 1992, Cistron showed that the Immunex patent contained errors that also appeared in the manuscript submitted to *Nature* and claimed that the sequence must have been appropriated during the peer-review process. In 1996, after a lengthy legal case, Immunex paid Cistron $21million in an out-of-court settlement.

misconduct arose as a direct result of the abuse of the peer-review system

Since journal peer review was not designed as a policing system to detect fraud it is, perhaps, asking too much to expect it to perform well in this area. Furthermore, if research misconduct rarely occurs and is hard to detect it may not be 'cost-effective' to expect peer review to protect readers from fraudulent data. However, most editors state that peer review and subsequent editorial processes perform an important function in refining papers between submission and publication. Several studies have

shown that the readability of papers is improved, but the authors of one commented that 'the degree of overall improvement was modest, and there was still substantial room for improvement in quality scores after manuscript revision'[12] and the authors of another note that 'The peer review and editorial processes slightly improved readability of original articles and their abstracts, but both remained difficult to read at publication.' [13]

Peer review and quality

Another way to assess peer review and other editorial processes is to consider the quality of published papers. One drawback of the studies showing that reviewers comment on only a minority of a manuscript's defects is that this could be because they do not bother to highlight further weaknesses after encountering a fatal flaw. One might hope that papers recommended for acceptance would undergo a more thorough review. One might also hope that the processes of revision by the authors and technical editing by the journal would guarantee consistently high standards of published papers. But this does not appear to be the case.

the scientific community pays at least part of the bill

David Moher and the group responsible for the CONSORT statement reviewing the quality of reports of randomized controlled trials reported that 89% of publications did not include details of the sample size calculation, 86% did not use confidence intervals, 60% provided inadequate detail on randomization, and 77% gave insufficient details of methods of treatment allocation concealment.[14]

Roy Pitkin and colleagues studied the accuracy of abstracts in papers published in major medical journals.[15] They found deficiencies (i.e. data in the abstract that were inconsistent with, or absent from, the main body of the text) in 18–68% of abstracts, despite the fact that 'these journals have full-time professional staffs who can be presumed to devote a good deal of time and energy to the editorial and production processes'.

We have recently conducted a systematic review of studies in which we looked at the accuracy of bibliographic references.[16] Like abstract accuracy, we found that this varied from journal to journal but the median proportion of inaccurate references (across 64 journals) was 36% (range 4–67%). The median proportion of errors serious enough to make retrieval of references impossible was 8% (range 0–38%). Similarly, our review showed that the median percentage of inaccurate quotations (i.e. those that were not a true representation of the cited paper) was 20% (range 0–44%).

Peer review and efficiency

A more general measure of the quality of a process is its efficiency, i.e. whether it makes good use of resources. This aspect should be examined at the level of individual journals and that of the universe of scientific publishing. In order to determine whether a process offers good value for money, one needs some idea of how much it costs. The full cost of peer review to the scientific community is difficult to measure since it relies to such an extent on the willingness of reviewers to perform their task unpaid. However, the cost to a journal has been estimated. Relman came up with a figure of $40 per article for the *New England Journal of Medicine* in the late 1970s and Lock considered the cost to the *BMJ* to be at least £48 per paper in 1984.[6] Suzanne and Robert Fletcher reported that peer review accounted for 2.7–7.5% of a journal's total expenditure.[17] These costs seem relatively modest although, for a journal with a high rejection rate, the cost per published article is more significant. (For a journal that rejected 85% of the manuscripts it received, the cost, based on Lock's calculation, per published paper would be £320). More recently, Donovan carried out an informal survey of editors and reported that the cost per accepted paper in medical journals was around £200.[18] He concluded that 'the system is labour intensive and expensive' and that 'a quality control mechanism is essential'.

High rejection rates may ensure a title's prestige, but the scientific community pays at least part of the bill, since most papers get published eventually. Independent investigations at the *NEJM* and *Journal of Clinical*

Investigation indicated that 85% of papers rejected by these journals were published elsewhere, following a second and sometimes third cycle of peer review. A more recent study using different methods indicated that at least 69% of papers rejected by *Annals of Internal Medicine* were published elsewhere.[19] Now, it could be argued, that the peer-review system adds value by causing articles to be refined and improved at each round of review. However, the evidence suggests that the majority of papers, perhaps as many as 80%, are not revised before resubmission, so the benefits of the first reviewers' comments are lost, to both the authors and the second set of reviewers. This cycle of review and rejection also carries a cost in terms of time.

Given the evidence that most papers rejected by general journals are published in other peer-reviewed journals, it has been suggested that peer review should be regarded less as a filter and more as a 'traffic policeman' directing articles to the most appropriate journal. Before the days of electronic databases and sophisticated search engines this, albeit less exalted, function still performed a useful service to readers, since it meant they could focus their attention on the titles most relevant to their interests. However, electronic retrieval makes this role increasingly obsolete. It is no longer possible to keep abreast of a subject by reading only one or two prestigious journals, and a literature review that restricted its contents in this way would not be taken seriously.

Electronic publishing has also altered the economics of production and distribution and removed the space constraints of paper journals. This has led some peer-reviewed journals to move away from the traditional model of selecting only the most relevant and interesting articles and to replace it with a 'bias to publish' (Fiona Godlee, personal communication). In this model, peer review serves to weed out reports of scientifically unsound or unethical research, retains its role of providing constructive criticism with an aim of improving submissions, but does not serve to select the very 'best' papers for the journal. Instead of scanning a highly selective and focused journal, readers are alerted when relevant papers are published, and can search the journal electronically. So far, very few journals have adopted this principle, so it is too early to judge its impact, but if accepted by authors and readers, it could fundamentally alter the economics and efficiency of peer review.

Virtually no journals have tested their peer-review systems rigorously against alternatives but Lock has, to date, got closest to this. He compared the assessments and recommendations of external reviewers with those of two senior in-house editors and found 65% agreement.[6] He also attempted to measure the efficacy of peer review to select important findings by examining the fate of rejected papers. This revealed that 68% of articles rejected by the *BMJ* were published elsewhere, 15% in high impact factor specialist journals and 10% in high impact factor general journals. Papers accepted by the *BMJ* had higher average citation rates than those it rejected. He concluded that 'the loss of 83 papers (5%) to general journals with high impact factors did not seem too disturbing out of a total cohort of 1551'.

Virtually no journals have tested their peer-review systems rigorously against alternatives

One variation in the peer-review process that would be expected to affect the cost and efficiency of the system, is the proportion of papers that are sent out for external review. Many specialty journals, run by an editor and editorial board who remain in full-time academic employment, send virtually all submissions out for review. However, some of the general journals that employ in-house editors use them as an initial filter and may reject as many as 50% of papers on their recommendation (Frank Davidoff, personal communication). Journals operating this system obtain external opinions on all papers that are published, but this has not always been the case. A former editor of the *Lancet* wrote 'I am a convinced opponent of routine peer review of articles . . . I believe that general journals should normally review articles internally.'[20]

One of Douglas-Wilson's chief concerns was the time that external peer review takes. He suggested that if more widespread use were made of internal review then

> The authors can have some confidence that their papers have been considered, if

> not knowledgeably, at least without scientific bigotry; they will receive an answer, whether yea or nay, with little delay.[20]

It is impossible to put a value on the time cost of peer review since this depends so much on the context. To the scientific community, there may be a cost if the publication of important findings is delayed, since new results may shape the course of other research. For some clinical papers a delay may, arguably, harm patients, either by exposing them unnecessarily to a treatment known to be harmful or by preventing them from receiving the most effective treatment. However, few single papers advance either science or medicine to such an extent that a few months' delay in publication has a significant impact.

However, from the author's perspective, long periods waiting for the decision of reviewers incur a 'human cost'; Suzanne and Robert Fletcher also used this phrase to describe the effects of receiving insensitive or discourteous reviews.[17] The cost of delay to a commercial company may be more readily quantified. Since the profitability of a medicine declines substantially when its patent expires and it is exposed to generic competition, manufacturers are keen to maximize sales during the early part of a product's lifecycle. Therefore, in the case of clinical studies that demonstrate the benefits of marketed products, the promoters of the medicine might measure a delayed publication in terms of lost revenue. However, delays are not necessarily synonymous with peer review. Some journals offer expedited review for important papers and others guarantee rapid review. The Fletchers have also pointed out that the time taken to review a publication usually represents only a small proportion of the total time from the conception, planning, execution, and reporting of a study.[17]

delays are not necessarily synonymous with peer review

While electronic publishing has the ability virtually to eliminate delay between acceptance and publication, which has been an important component in the time taken by traditional paper journals to publish research, it probably has little effect on the time taken to reach a decision. The use of fax, email, or online reviewing may trim a few days off either end of a schedule, but they do not improve the speed of response of unpaid reviewers who continue to be busy people with competing demands on their time.

Conclusions

So, we have seen that peer review is costly in terms of time and resources, and, especially when papers are rejected by a number of journals, does not appear to be particularly efficient. It has the power to improve the quality of submitted articles, yet, in many cases, suggestions from journals that reject a paper are ignored. It does not always detect important work, nor does it reliably protect readers from fraudulent reports. Reviewers do not reliably comment on all the weaknesses of a paper, and a significant proportion of the end-product (i.e. the papers that are published) contains deficiencies that make it hard for the reader to evaluate the research thoroughly or to retrieve cited references. Reviewers may also exhibit bias in relation to authors' identities, professional affiliations, and publication language. In a few cases, reviewers are known to have abused the peer-review system.

It is always easier to criticize than to improve a system. It would, indeed, be reassuring to think that these anecdotes of the failings of peer review could be set against a solid body of evidence about its benefits and efficacy. Yet, most research on peer review has focused on peripheral or surrogate measures,[2] and there have been virtually no attempts to compare peer review with alternative methods of assessing scientific publications to see whether they do a better job. Considering the resources devoted to peer review and its important place in academic science and medicine, such research is long overdue. We must therefore conclude that peer review is not only an imperfect process but also a largely untested one. The anecdotes and evidence presented in this paper do not prove that it is ineffective or positively harmful, yet, equally, there is no convincing evidence that it actually works.[21] This state of affairs is unacceptable for a process that shapes indi-

vidual careers and the direction of research itself.

Note

Based on a presentation given at the ALPSP seminar on peer review held in London, January 2001. It was given in the context of a day spent examining peer review and was balanced by a number of other presentations demonstrating the benefits of peer review. It therefore does not attempt to present a full evaluation of peer review, but instead concentrates on its shortcomings.

References

1. ALPSP/EASE Peer Review Survey. www.alpsp.org/peerev.pdf (12 Mar. 2001).
2. Jefferson, T., Wager, E. and Davidoff, F. Measuring the quality of editorial peer review. Accepted for presentation at: 4th International Congress on Peer Review in Biomedical Publication. Barcelona, Sep. 2001.
3. Lock, S. The grossest failures of peer review. *BMJ* 1993:307, 382.
4. Godlee, F., Gale, C.R. and Martyn, C.N. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. *JAMA* 1998:280, 237–40.
5. Peters, D.P. and Ceci, S.J. Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behavioral and Brain Sciences* 1982:5, 187–95.
6. Lock, S. *A Difficult Balance: Editorial Peer Review in Medicine*. London: BMJ, 1991.
7. Nylenna, M., Riis, P. and Karlsson, Y. Multiple blinded reviews of the same two manuscripts: effects of referee characteristics and publication language. *JAMA* 1994:272, 149–51.
8. Office of Research Integrity Annual Report 1997. Washington, DC: Department of Health and Human Services, 1998.
9. Office of Research Integrity Annual Report 1998. Washington, DC: Department of Health and Human Services, 1999.
10. Office of Research Integrity Annual Report 1999. Washington, DC: Department of Health and Human Services, 2000.
11. Rennie, D. Misconduct and peer review. In F. Godlee and T. Jefferson (eds), *Peer Review in Health Sciences*. London: BMJ Books, 1999, 90–9.
12. Goodman, S.N., Berlin, J., Fletcher, S.W. and Fletcher, R.H. Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. *Annals of Internal Medicine* 1994:121, 11–21.
13. Roberts, J.C., Fletcher, R.H. and Fletcher, S.W. Effects of peer review and editing on the readability of articles published in *Annals of Internal Medicine*. *JAMA* 1994:272, 119–21.
14. Moher, D. CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. *JAMA* 1998:279, 1489–91.
15. Pitkin, R., Branagan, M.A. and Burmeister, L.F. Accuracy of data in abstracts of published research articles. *JAMA* 1999:281, 1110–11.
16. Wager, E. and Middleton, P. Reference accuracy in peer-reviewed journals: a systematic review. Accepted for presentation at 4th International Congress on Peer Review in Biomedical Publication. Barcelona, Sep. 2001.
17. Fletcher, R.H. and Fletcher, S.W. The effectiveness of editorial peer review. In F. Godlee and T. Jefferson (eds), *Peer Review in Health Sciences*. London: BMJ Books, 1999, 45–56.
18. Donovan, B. The truth about peer review. *Learned Publishing* 1998:11, 179–84.
19. Ray, J., Berkwits, M. and Davidoff, F. The fate of manuscripts rejected by a general medical journal. *American Journal of Medicine* 2000:109, 131–5.
20. Douglas-Wilson, I. Editorial review: peerless pronouncements. *NEJM* 1977:296, 877.
21. Alderson, P.A., Davidoff, F., Jefferson, T., Middleton, P. and Wager, E. Editorial peer review for improving the quality of reports of biomedical studies. Accepted for presentation at 4th International Congress on Peer Review in Biomedical Publication. Barcelona, Sep. 2001.

Elizabeth Wager
Publication Consultant
Sideview Cottage, Station Road
Princes Risborough
Bucks HP27 9DE, UK
Email: liz@sideview.demon.co.uk
Website: www.lizwager.com

Tom Jefferson
UK Cochrane Centre
Summertown Pavilion
Middle Way
Oxford OX2 7LG, UK
Email: toj1@aol.com