Editorial peer review for improving the quality of reports of biomedical studies (Review)

Jefferson T, Rudin M, Brodney Folse S, Davidoff F

THE COCHRANE COLLABORATION®

This is a reprint of a Cochrane review, prepared and maintained by The Cochrane Collaboration and published in *The Cochrane Library* 2008, Issue 2

http://www.thecochranelibrary.com

TABLE OF CONTENTS

[Methodology Review]

Editorial peer review for improving the quality of reports of biomedical studies

Tom Jefferson[1], Melanie Rudin[2], Suzanne Brodney Folse[3], Frank Davidoff[4]

[1]Vaccines Field, The Cochrane Collaboration, Roma, Italy. [2]Health Reviews Ltd, Roma, Italy. [3]Health and Wellness Division, Blue Cross Blue Shield of Rhode Island, Providence, USA. [4]Institute for Healthcare Improvement, Hartford, CT 06109, USA

Contact address: Tom Jefferson, Vaccines Field, The Cochrane Collaboration, Via Adige 28a, Anguillara Sabazia, Roma, 00061, Italy. jefferson.tom@gmail.com. jefferson@assr.it; jefferson.tom@gmail.com.

Editorial group: Cochrane Methodology Review Group.
Publication status and date: Edited (no change to conclusions), published in Issue 2, 2008.
Review content assessed as up-to-date: 19 February 2007.

Citation: Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews* 2007, Issue 2. Art. No.: MR000016. DOI: 10.1002/14651858.MR000016.pub3.

ABSTRACT

Background

Scientific findings must withstand critical review if they are to be accepted as valid, and editorial peer review (critique, effort to disprove) is an essential element of the scientific process. We review the evidence of the editorial peer-review process of original research studies submitted for paper or electronic publication in biomedical journals.

Objectives

To estimate the effect of processes in editorial peer review.

Search strategy

The following databases were searched to June 2004: CINAHL, Ovid, Cochrane Methodology Register, Dissertation abstracts, EMBASE, Evidence Based Medicine Reviews: ACP Journal Club, MEDLINE, PsycINFO, PubMed.

Selection criteria

We included prospective or retrospective comparative studies with two or more comparison groups, generated by random or other appropriate methods, and reporting original research, regardless of publication status. We hoped to find studies identifying good submissions on the basis of: importance of the topic dealt with, relevance of the topic to the journal, usefulness of the topic, soundness of methods, soundness of ethics, completeness and accuracy of reporting.

Data collection and analysis

Because of the diversity of study questions, viewpoints, methods, and outcomes, we carried out a descriptive review of included studies grouping them by broad study question.

Main results

We included 28 studies. We found no clear-cut evidence of effect of the well-researched practice of reviewer and/or author concealment on the outcome of the quality assessment process (9 studies). Checklists and other standardisation media have some evidence to support their use (2 studies). There is no evidence that referees' training has any effect on the quality of the outcome (1 study). Different methods of communicating with reviewers and means of dissemination do not appear to have an effect on quality (3 studies). On the

basis of one study, little can be said about the ability of the peer-review process to detect bias against unconventional drugs. Validity of peer review was tested by only one small study in a specialist area. Editorial peer review appears to make papers more readable and improve the general quality of reporting (2 studies), but the evidence for this has very limited generalisability.

Authors' conclusions

At present, little empirical evidence is available to support the use of editorial peer review as a mechanism to ensure quality of biomedical research. However, the methodological problems in studying peer review are many and complex. At present, the absence of evidence on efficacy and effectiveness cannot be interpreted as evidence of their absence. A large, well-funded programme of research on the effects of editorial peer review should be urgently launched.

PLAIN LANGUAGE SUMMARY

Editorial peer review for improving the quality of reports of biomedical studies

Editorial peer review is used world-wide as a tool to assess and improve the quality of submissions to paper and electronic biomedical journals. As the information revolution gathers pace, an empirically proven method of quality assurance is of paramount importance. The increasing availability of empirical research on the possible effects of peer review led us to carry out a review of current evidence on the efficacy of editorial peer review. We found few studies of reasonable quality, and most of these were concerned with the effects of blinding reviewers and/or authors to each others' identity. We could not identify any methodologically convincing studies assessing the core effects of peer review. Major research is urgently needed.

BACKGROUND

Learned societies and journal editors usually rely on the views of independent (outside) content experts in making decisions on publication of submitted manuscripts or presentation of reports at meetings. This system of appraisal is known as peer review. The use of peers to assess the work of fellow scientists goes back at least 200 years (Kronick 1990; Rennie 2003). It provides a crucial integrating social force in the scientific community, since "In science, the acceptance by scientific journals of contributed manuscripts establishes the donor's status as a scientist - indeed, status as a scientist can be achieved only by such gift-giving." (Hagstrom 1965, p. 13). As a consequence, "The organization of science consists of an exchange of social recognition for information" (Hagstrom 1965, p. 13). For intellectual purposes, the use of peer review is usually assumed to raise the quality of the end-product (i.e. the journal or scientific meeting) and to provide a mechanism for rational, fair and objective decision-making (Rennie 2003). However, these assumptions have rarely been tested in rigorous fashion.

Surveys indicate that clinicians give greater credence to findings published in peer-reviewed journals (Sievert 1996) yet little attention is paid to the definition of peer review or effects of variations in review practices between journals (Godlee 2003). However, the need to prioritise information sources is crucial, since over 20,000 biomedical journals are now published globally (Godlee 2003) while decision-makers such as clinicians have little time to read (Sackett 1997). Judgements about information sources are often based on the reputation of the publishing institution, or the editors of a book. However, this does not necessarily guarantee high quality information. For example, editorials and update articles written specifically for busy decision-makers sometimes contain misleading information on the effects of healthcare interventions (Antman 1992; Jefferson 1999). While clinicians often rely on peer review to filter and assess the work of others, many have criticised the current systems and identified defects. One editor has commented that "every scientist has a story to tell about the inequities of the peer-review system" (Rennie 1990).

At the same time, others have found clinicians' views to be that "More of an effort should be made to have critical review of the literature done by experts and screened for the rest of us," particularly since the exploding pace of medical knowledge means that clinicians necessarily (and should) rely "on journal editors to ensure scientific rigor in statistics, conclusions, etc. It is not important that [clinicians] spend their limited neurons on these

technical issues; it's hard enough to retain the results and employ them in clinical care" (Saint 2000). Similarly, Vandenbroucke has suggested that if all edited journals were suddenly abolished by decree, and all scientific information were simply posted directly by the authors on the Internet, "within a few days...we will be receiving commercial e-mail messages from some young unknowns who tell us that they have the time as well as the technical resources to retrieve the information for us. They will add that, of course, since so much information on the Internet is nonsensical, they also will obligingly sieve out the nonsense...Because they are not conversant with all aspects of medicine, they will invoke the help of some trusted friends to read the retrieved information and judge its quality. Within a week, my best guess, our edited journal system will be born again, inclusive of peer review" (Vandenbroucke 1998).

Given the amount of biomedical data generated, it is impossible to envisage a situation where each item of information can be assessed for its quality by a single authority or individual. An alternative is to group information by the way in which it has been managed before reaching the public domain. This would require an understanding of the effects of peer review and other processes applied to scientific information before publication.

The processes applied to paper or electronic articles on science-related topics can be divided into the following activities:

- Procedures aimed at assessing and ensuring the scientific quality of output. These are usually known as editorial peer review.
- Technical processes, such as copy-editing. These aim to produce output of good technical quality (for example, articles that are free of spelling mistakes and in the book's or journal's house style). They may also aim to ensure or increase the readability and/or comprehensibility of the content by the target audience.
- Administrative processes, aimed at ensuring the regular production of a journal (for example, the assignment of a unique registration number or tracking device to each submission).

This review assesses the effects of processes undertaken as part of editorial peer review of original research studies submitted for paper or electronic publication in biomedical journals. Technical editing is the topic of a separate review (Wager 2007). Both are complex procedures, but editors believe they ensure the quality of the end-product and most medical journals invest considerable time and/or money in them.

Despite the fact that peer review has such a long history and is so well established, rigorous research into its effects is a recent phenomenon. The first international congress on peer review was held in 1989 and even then one contributor commented that "the systematic study of peer review is even more deficient in theory than it is in data" (Knoll 1990). Since then, the body of original research on the effects of peer review has been growing and secondary research (i.e. systematic review and synthesis) may now be possible (Overbeke 2003).

The original review was published on The Cochrane Library in 2003 and in paper format in 2002 (Jefferson 2002a; Jefferson 2002).

OBJECTIVES

To estimate the effect of processes, or combinations of processes, in editorial peer review on the outcomes studied in the research on this topic. These processes are grouped as:

- different ways of screening submissions (i.e. carrying out a preliminary assessment of the submission)
- different ways of assigning submissions (i.e. choosing and assigning assessors to the submission)
- different ways of masking submissions (i.e. concealing the identity and background of the authors and/or the assessors)
- different ways of eliciting internal opinions (i.e. opinions on the scientific quality of the submission from those within the publishing organisation)
- different ways of eliciting external opinions (i.e. opinions on the scientific quality of the submissions from those outside the publishing organisation)
- different (group or single person) decision-making procedures (i.e. deciding whether to publish the submission)
- different types of feedback to author(s) and subsequent revision of submissions
- a combination of the above
- any other process that can be considered as editorial peer review

METHODS

Criteria for considering studies for this review

Types of studies

The following types of comparative studies were included:

- randomised controlled trials
- quasi-randomised controlled trials
- interrupted time series
- before and after studies
- other observational studies where there was some attempt to control for confounding

The following were excluded:

- Surveys comparing editorial practice or editorial outcomes with characteristics of journals or reviewers
- Review articles

Types of data

Reports of original research submitted to biomedical journals (not reviews or articles commenting on the work of others), published or unpublished, or people acting as peer reviewers.

Types of methods

The studies should compare two or more interventions or an intervention against do-nothing from within one of the following categories:

- different ways of screening submissions (i.e. carrying out a preliminary assessment of the submission)
- different ways of assigning submissions (i.e. choosing and assigning assessors to the submission)
- different ways of masking submissions (i.e. concealing the identity and background of the authors and/or the assessors)
- different ways of eliciting internal opinions (i.e. opinions on the scientific quality of the submission from those within the publishing organisation)
- different ways of eliciting external opinions (i.e. opinions on the scientific quality of the submissions from those outside the publishing organisation)
- different (group or single person) decision-making procedures (i.e. deciding whether to publish the submission)
- different types of feedback to author(s) and subsequent revision of submissions
- a combination of the above
- any other process that can be considered as editorial peer review

Types of outcome measures

The main outcome we hoped to find was the quality of the published report, however measured. The measurement of quality of the editorial process should reflect the aims of the process. Editorial peer review aims to identify good submissions on the basis of:

- importance of the topic dealt with (the impact on health and healthcare of the findings of the study)
- relevance of the topic to the journal or broad coherence with the aims (and record) and the aspirations of readers
- usefulness of the topic (the paper's or proposal's contribution to the scientific debate or body of knowledge on the subject)
- soundness of methods (the ability of the methods used to answer the study question)
- soundness of ethics (honesty in carrying out and reporting the study and its limitations and the avoidance of unnecessary harm if humans or animals are involved)
- completeness and accuracy of reporting (ensuring that all relevant information is clearly presented)

A gold standard editorial process is one that produces studies that are important, appropriate to the publication medium, useful, original, methodologically sound, ethical, complete and accurate. How should these characteristics be measured? There are two possible levels at which measurement should be made, an ideal one (in which true outcomes are used as indicators of quality) and a 'pragmatic' one (in which surrogate outcomes are used). We have summarised outcomes, their standards, indicators by rank and methods to assess them in the table (Table 1).

Table 1. Outcome measures

Outcome	Standard	Ideal indicator	Surrogate - rank 1	Surrogate - rank 2	Surrogate - rank 3
Importance	The study findings have a potential high impact on health and healthcare.	Changes in health status. Changes in healthcare delivery.	Citation rates. Media coverage.	Tone and content of assessors' and editors' reports. Correspondence.	Process centred (e.g speed or cost of reviewing).
Usefulness	The study contributes significantly to the	Contributes significantly within a systematic	Contributes significantly within a re-	Assessors' and editors' reports. Corre-	Process centred (e.g speed or cost of re-

Table 1. Outcome measures (*Continued*)

	scientific debate or knowledge on the subject.	review of the topic. Narrows CIs around estimates of effect.	view of the topic.	spondence.	viewing).
Relevant	The study topic is relevant to the journal, its aims and to the readership.	Is relevant and consistent with the aims and readership of the journal confirmed by survey.	Citation rates.	Editors' and assessors' reports. Correspondence. Internet hit rates.	Process centred (e.g speed or cost of reviewing).
Methodologically sound	The methods used can answer the study question.	Study findings are replicated several times across settings, where studies appear to have been well-conducted.	Closeness of fit between methods and 'evidence-based' methodological checklist.	Editors' and assessors' reports. Correspondence.	Process centred (e.g speed or cost of reviewing).
Ethical	The study has been carried out and reported honestly. Its limitations are explicit and unnecessary harm to humans or animals has been avoided.	No divergence between reality and the report. The rights of humans and animals are safeguarded. Privacy and informed consent are maintained throughout.	No duplicate publication. No complaints from participants. The study has been given ethical clearance.	Editors' and assessors' reports. Correspondence. The study was carried out in a highly reputable institution by highly reputable investigators.	Process centred (e.g speed or cost of reviewing).
Complete	All relevant information is presented.	There is no selective presentation of data. Raw data match aggregate data.	The text is complete.	Correspondence.	Process centred (e.g speed or cost of reviewing).
Accurate	All relevant information is a true reflection of what went on.	Measurements truly reflect magnitude of findings.	The figures add up.	The figures add up. Correspondence.	Process centred (e.g speed or cost of reviewing).

As one moves through the columns from left to right, the importance of the chosen indicator decreases and the proxy or surrogate nature increases. On the far right, indicators of the quality of the process (such as speed or cost of providing opinions) replace measures of the quality of the outcome. All measures of quality of the process were regarded as third rank indicators. On the basis of these expected outcomes, we planned to synthesise findings from different studies if homogeneous outcome measures and similar study designs were found.

An additional aim of the peer-review process is to improve the quality of submissions going through the editorial process.

Search methods for identification of studies

The following databases were searched to June 2004: CINAHL, Ovid, Cochrane Methodology Register, Dissertation abstracts, EMBASE, Evidence Based Medicine Reviews: ACP Journal Club, MEDLINE, PsycINFO, PubMed. For the full search strategy, see appendices (Appendix 1; Appendix 2).

Data collection and analysis

For the 2005 update, two reviewers working in pairs (TJ and MR) and one reviewer working alone (SB) examined abstracts of retrieved citations for possible inclusion. Studies for possible in-

clusion were retrieved in full. The same reviewers then examined studies independently against the selection criteria. During preparation of the 2003 review there was disagreement on one study (Cho 1998). This was resolved, and the study excluded, after re-reading the study in question and discussing its design.
The same reviewers extracted information on study design and outcomes. We collected descriptive information on study quality, as reported below. Once we had read and classified the reports by the study questions they were attempting to answer and by their design, we considered pooling the results of similar studies into a formal meta-analysis. We decided against this course of action as no two studies were alike and all were asking a slightly different study question, or using a different design, or unit of randomisation, or outcome measure.

RESULTS

Description of studies

See: Characteristics of included studies; Characteristics of excluded studies; Characteristics of ongoing studies.
Twenty-eight studies fulfilled our criteria. We decided to group studies by the broad questions or issues they were addressing, thus carrying out a descriptive review.

Studies assessing the effect of blinding/masking on the quality of external opinions

- Mc Nutt (McNutt 1990) reports the conduct and results of a double-blind randomised trial designed to answer the question: "Does blinding of external reviewers improve quality of reviews?" Quality of reviews was assessed by an editor blinded to authors' and reviewers' identity, and compared to assessments by an editor who was not blinded, taking the perspective of both editor (4 items) and author (5 items) on five-point scales. Authors also judged quality of review (6 items including a summary grade) on a five-point scale. 127 consecutive submissions to the Journal of General Internal Medicine were sent to two external reviewers one of whom was randomly chosen to be blinded as to the authors' and institution's names. Review signing was optional. No association between signing and review quality was detected, but the authors found that blinding the reviewer resulted in a statistically significantly higher quality review, as measured by the editor

- Fisher (Fisher 1994) reports the conduct and results of a randomised controlled trial designed to answer the question: "Does knowledge of authors' identity bias reviewers towards authors with a high number of previous publications?" Two reviewers blinded to the authors' identity and two non-blinded reviewers were randomly assigned to 57 consecutive submissions to the Journal of Developmental and Behavioural Paediatrics. Editors assessed quality of blinded and non-blinded reviewers using a five-point scale. Items of judgment were: accept the submission, accept with optional revision, accept with mandatory revision, reject with additional review and reject. In addition, a survey questionnaire asked reviewers to guess the identity of the author(s). Results showed that in 50 (46%) cases reviewers guessed authors' identity correctly and that blinded reviewers gave better scores to the work of authors with longer publication records. As this finding persisted even after adjustment for correct guessing of identity, the study authors interpret it as suggesting that blinded peer review is effective in producing less biased reviews. They argue that non-blinded reviewers are more likely to be biased against the work of authors with long publication record

- Jadad (Jadad 1996) reports the conduct and results of a randomised controlled trial comparing the quality of 36 reports of RCTs in the field of pain control. Seven reviewers were randomly allocated to carry out the assessment under blind and seven reviewers under open conditions. Quality was assessed with a score on 3 and 6 -item lists. Blinded assessment produced lower quality scores and more consistent results than open assessment.

- Van Rooyen (van Rooyen 1998) reports the conduct and results of a randomised controlled trial of 527 consecutive submissions to the BMJ. The trial was designed to answer the question: "Does knowledge of reviewer's identity affect quality of the review?" The study design comprised three arms in which submissions were randomly assigned to an uninformed (to assess any Hawthorne effect), masked (reviewers masked to co-reviewers' identities) or non-masked arm (reviewers not masked to co-reviewers' identities). Reviewers refusing consent to non-masking after randomisation were transferred to a preference arm, and remained masked. In each arm, reviewers were then randomly assigned to be blinded or unblinded to authors' identities. Each submission was managed by a randomly assigned editor. Review quality was assessed on a seven-item five-point scale and on time taken to review. The items were: importance of research question, originality, method, presentation, constructiveness of comments, substantiation of comments, interpretation of results and politeness of tone. Results show that knowledge of one or both reviewers' or authors' identity made no difference to the quality of reviews, recommendations or time taken in reviewing.

- Godlee (Godlee 1998) reports the conduct and results of a randomised controlled trial of 221 reviews of submissions to the BMJ. The objective of the study was to assess the effects of open peer review compared with authors' details removed. Quality of reviews was defined as the ability of reviewers to identify eight deliberate errors inserted in a manuscript already accepted for

publication, whose authors consented. The trial had four arms all with different types of disclosure of reviewers and authors and a fifth or control arm. Out of 420 reviewers asked to return their opinions only 221 (53%) did so. Results failed to find any effect of concealment on the quality of reviews.

- Justice (Justice 1998) reports the conduct and results of a randomised controlled trial designed to answer the question: "Does ignorance of authors' identity affect quality of reviews?" Quality of reviews was judged on a four-item five-point scale. Items are: importance of the research question, methodological soundness, courteousness, and reviewer's production of evidence to support their views. In addition, reviewers were asked whether they recognised author identity. 118 submissions to Annals of Emergency Medicine, Annals of Internal Medicine, JAMA, Obstetrics and Gynaecology and Ophthalmology were randomised to be assigned to pairs of reviewers either using usual journal blinding practice, or to one reviewer blind to the author's identity and the other unblinded. Seventy-seven review pairs were judged to contain sufficient data to be entered into the study. Results did not show that blinding had an effect on review quality. Submissions of well-known authors were more difficult to blind successfully (blinding success was 68%).

- Van Rooyen (van Rooyen 1999) reports the conduct and results of a randomised controlled trial of 113 consecutive submissions to the BMJ to assess whether open peer review led to lower quality reviews and whether reviewers would refuse to review openly. The identity of consenting reviewers was revealed when sending the review to the authors. The quality assessment instrument was a modified version of that used in Van Rooyen 1998 (van Rooyen 1998). Results failed to show that asking reviewers' consent to identification affects quality of reviews, time taken to review or recommendations to publish, however a significant number of reviewers are likely to refuse to give their opinions.

- Das Sinha (Das Sinha 1999) reports the conduct and results of a single-blind randomised trial of 78 consecutive submissions to The National Medical Journal of India. Each submission was randomly assigned to an Indian and a non-Indian reviewer with the aim of comparing reviewers' performance by nationality. Although study design included 100 submissions, for 22 the review process was incomplete. Evaluation of reviews was carried out by two editors, blinded to reviewers' identity and background using an evaluation form with five main items (importance of question, target key issues, methods, presentation and general). Reviewers' evaluation was scored from 1 to 100. The authors concluded that their study showed than non-Indian peer reviewers produced better quality reviews than Indians. In addition, exchanging reviews among reviewers was not found to make a difference to their quality.

- Walsh (Walsh 2000) reports the conduct and results of a randomised controlled trial designed to assess the feasibility of running an open peer-review system (in which both reviewer and author were aware of each other's identity). The study was set in the British Journal of Psychiatry and lasted 18 months, involving 408 reviews (194 signed returned and 164 unsigned returned) of consecutive submissions to the journal. Fifty-seven reviewers refused to participate in the trial and were not randomised. Outcomes were assessed using the Black instrument, consisting of seven items scored on a five-point scale. The items were: importance of research question, originality of topic, methods, organisation, writing style and tone of the comments. Additional assessment of the tone of each review was made by the trainee editors on a five-point scale. All reviewers were also asked to estimate the time spent reviewing and give recommendations on whether to publish the submission or not. Review quality was significantly higher and of more courteous tone in the signed group. Time spent on the review was higher in the signed group, although response rate for this item was low (54%). Signed reviewers were less likely to recommend rejection. The authors conclude that an open system is feasible and that signed reviews are possibly of better quality although they take longer to complete.

Studies assessing the effects of submission checklists on the outcome

- Jefferson (Jefferson 1998) attempts to answer the question: "Did publication of the BMJ guidelines on peer review of economic submissions affect the quality of the editorial process?" The study has a before and after (publication of the guidelines) comparative design, in which quality of 192 economic submissions and of the editorial process is compared in the BMJ and Lancet settings. Quality of economic submissions was assessed using the BMJ's 36-item checklist and an ad hoc questionnaire to elicit editors' views. The 36 points were based on the fundamentals of economic analysis defined by an expert panel. Results showed that the guidelines had no apparent impact on the quality of submissions but helped editors in managing submissions.

- Gardner (Gardner 1990) is a before and after study asking a similar question: "Is statistical assessment using a checklist beneficial to the quality of submitted manuscript?" Gardner and colleagues used a 12 and 24-item referee checklist addressing study design, features, analysis and presentation of statistical data and recommendations to detect methodological problems in the statistical sections of 45 papers submitted to the BMJ. Only 5 (11%) were methodologically acceptable upon submission, but this number increased to 38 (84%) after publication. The authors concluded that statistical assessment with the use of checklists is beneficial to increasing the statistical quality of published studies.

Studies assessing the effects of communication media on the outcome

- Bingham (Bingham 1998) reports the conduct of an open study carried out within the Medical Journal of Australia to assess whether open peer review on the Internet improved the quality of peer review. Fifty-six papers accepted for publication were posted on the Journal's website together with 150 referees' reports and comments were invited. Seven papers were changed by the authors in response to the comments received. No change in quality of commissioned reviews was noted before or during the study. Electronic comments from readers were few.

- Neuhauser (Neuhauser 1989) reports the conduct of a randomised controlled trial of 177 submissions to Medical Care. Ninety submissions were randomised to be assigned to reviewers warned of the impending review by phone and 87 who received the submission by post without a warning. The aim of the study was to assess whether a warning would speed review turnaround. Results showed that total turnaround time was significantly longer in the group with warned reviewers and calling reviewers increases costs.

- Pitkin 2002 is a randomised controlled trial comparing the effects of sending the same manuscript for review either with no prior warning ('justsend') or with a preliminary fax asking for reviewer availability to review ('askfirst'). Two hundred and eighty-three consecutive manuscripts submitted to the Journal of Obstetrics & Gynaecology between September 1999 and May 2000 were sent to 566 reviewers. Reviewers were chosen by the editor on the basis of availability of address and fax number. Two hundred and forty-seven (87%) of the 'justsend' referees and 177 (63%) of the 'askfirst' referees produced a review. Twenty-two of the 'justsend' declined at several stages to review and the manuscripts were sent on to other referees (all 22 of first refusals were eventually refereed). One hundred and eighty-one of the 'askfirst' referees agreed to review, whereas 102 of the initial referees either did not respond (n=59) or declined (n=43). All 102 manuscripts were eventually refereed by 'substitute' referees. Although differences in frequency of specific declines were significant (8% vs 15%), the production rate of reviews by original 'justsend' referees who did not optout and 'askfirst' reviewers who agreed to referee were not significant (247/261 vs 177/181). Review quality in the subset of 151 referee reports agreed to review and returned reviews was not significantly different. Blinded quality assessment of a review was carried out using a not described instrument on a 5-point scale. The authors concluded that advance warning of review did not affect quality, but elicited 36% of turndowns or actual reviews. However, the extra work involved in finding 'substitute' reviewers was made up by an 'askfirst' speedier review turnaround.

Studies assessing the effects of training, feedback and correspondence on the outcome

- Callaham (Callaham 1998) reports a study conducted to address the question of whether attendance at voluntary training workshops improves quality of peer reviewers of medical journals. The authors compared the quality of output of 39 reviewers who had completed at least two reviews for Annals of Emergency Medicine (AEM) and had agreed to attend a 4-hour training workshop with 39 reviewers matched for review quality and number of reviews before the training session and 220 unmatched controls. Assessment of quality was carried out using a 5-point in-house Likert scale by journal editors. There was no significant change in any performance measure after the workshop.

- Callaham 2002 reports two randomised controlled trials to assess if peer reviewers from AEM attending a formal interactive training session produced better reviews. In study 1, 25 of the 173 invited AEM reviewers who volunteered attendance were randomised to 25 control reviewers who were invited but did not attend, using the same inclusion criteria - average review quality score (median < 4). Study 2 (intense recruitment) was conducted to reduce self selection bias (to address the limitations of study 1, as attendees and controls were randomised by invitation and not attendance). Participants had the same average review quality score. However, non-responders were aggressively followed up for a response by e-mail, etc until a confirmed response of the invitees was obtained. Twenty-nine reviewers accepted the invitation and 12 actually attended a formal 4-hour interactive workshop on peer review. Outcomes for study 1 and 2 were the same: number and mean rating of reviews in preceding 2 years were compared with the number and mean rating of reviews in the 2 years after the workshop. The difference was expressed as mean rating change. In study 1, the mean change in rating after the workshop was 0.11 (95% confidence interval (CI) - 0.25 to 0.48) for control reviewers and 0.10 (95% CI - 0.20 to 0.39) for attendees. In study 2, of 75 reviewers intensively recruited, only 12 (41%) of those who said they would attend did. All of the participants thought the workshop would improve their performance ratings. Test scores at the end of the workshop improved in 73% of participants compared with scores on pretest. The control reviewers' average rating changed by - 0.10 (95% CI - 0.49 to 0.29) versus 0.06 (95% CI - 0.34 to 0.23) for attendees. The authors concluded that attendance was low and did not improve the quality of subsequent reviews, contrary to the predictions of attendees. Efforts to aggressively recruit average reviewers to a second workshop (study 2) were time consuming, had low success rates, and showed a similar lack of effect on ratings, despite improvement in scores on a post-test instrument assessing knowledge of peer review.

- Callaham 2002a reports two randomised controlled trials assessing whether written feedback to reviewers of AEM improved subsequent reviews. Authors, editors and reviewers were blinded to each other's identity and reviewers to editor's rating of their reviews. Reviewers were selected from an eligible pool. Study 1 included 35 reviewers with a median quality score of 3 or lower (low-volume, low-quality reviewers) and study 2 included 95 reviewers with median score of 4 or lower (low-volume, average quality reviewers). For study 1, 51 reviewers were eligible and randomized and 35 (182 reviews) had sufficient data for analysis. For study 2, 127 reviewers were eligible and randomized, and 95 (324 reviews) had sufficient data for analysis. Reviewers in both studies were randomized to intervention or control (standard procedure). The intervention in study 1 consisted of feedback with other reviewer's assessment of the same submission and a copy of the editor's letter to the authors. Study 2 had the same intervention as study 1 but with the addition of the editor's ratings of reviewers' performance and a sample high grade review of another submission to the journal. Follow-up was two further reviews for each study. For study 1, the mean individual reviewer rating change was 0.16 (95% Cl -0.26 to 0.58) for control and -0.13 (-0.49 to 0.23) for intervention. For study 2, controls had a mean individual rating change of 0.12 (95% Cl, -0.20 to 0.26) and intervention reviewers, 0.06 (-0.19 to 0.31). The authors concluded that in study 1 minimal feedback from editors on review quality had no effect on subsequent performance of poor-quality reviewers, and the authors thought there might have been a negative effect. In study 2, feedback to average reviewers, despite being more extensive and supportive, produced no improvement in reviewer performance. In summary, simple written feedback to reviewers seems to be an ineffective educational tool.

It would appear that the outcome assessment instrument for both Callaham 2002 and Callaham 2002a was the AEM editor's routine quality rating (on a scale of 1 to 5) tool.

- Schroter 2004 reports a single-blind randomised trial assessing the effects of training on the quality of peer review. Consenting BMJ reviewers were randomised (609) to the control arm (arm 1) or to a self taught group who received a training package and CD Rom (based on the material used for arm 3) and techniques of critical appraisal for randomised controlled trials (arm 2) or to receive a full days' training and the CD Rom (arm 3). Reviewers were asked to rate three previously published papers, each describing an RCT of alternative generic ways of organising and managing clinical work. The trials included 14 deliberate errors. Review quality assessment was carried out using the eight-item instrument used in van Rooyen 1999. Reviewers in the self taught group scored higher in review quality after training than did the control group (score 2.85 versus 2.56; difference 0.29, 95% confidence interval 0.14 to 0.44; P = 0.001), but the difference was not judged to be of editorial significance and was not maintained at 6 months. Both intervention groups identified significantly more major errors after training than did the control group (3.14 and 2.96 versus 2.13; P < 0.001), and this remained significant after the reviewers' performance at baseline assessment was taken into account. Training had no impact on the time taken to review the papers but was associated with an increased likelihood of recommending rejection (92% and 84% versus 76%; P = 0.002). The authors concluded that short peer-review training packages have only a slight and non-sustained impact on the review quality and the value of longer interventions needs to be assessed.

- Strayhorn (Strayhorn 1993) reports the conduct and results of a before and after study carried out at the Journal of the American Academy of Child and Adolescent Psychiatry aimed at assessing the effects of training manuals for reviewers and a structured checklist. Two hundred and ninety-six pairs of reviewers' ratings were compared with 272 pairs after the introduction of new multi-item evaluation scales with more separate discrete items and training manuals. Results showed that inter-rater reliability increased after the new scales were introduced.

Studies assessing the presence and effects of reviewer bias on the outcome

- Ernst (Ernst 1999) reports the conduct and results of a double-blind randomised trial to test the hypothesis that there is no reviewer bias against an 'unconventional' drug. One hundred thirty out of 291 medical practitioners invited to review for a fictitious journal were randomly assigned to review either of two versions of the same letter to the editor. One version assessed a mainstream drug (metoprolol) and the other an unconventional drug (beef spleen cell extract). Apart from the study drug, all details were identical. Quality assessment was carried out using a 10-item visual analogue scale with 3-point scale with two overall scores. Items were: relevance, hypothesis-formulation, randomisation, inclusion and exclusion criteria, sample size, statistics, outcome choice, follow-up, clarity, linguistic quality. Results showed no evidence of reviewer bias towards the 'unconventional' drug but showed poor inter-rater reliability.

- Resch 2000 is a double-blind randomised trial to assess the hypothesis that peer review favours an orthodox form of treatment over 'alternative' therapy. The authors produced a fictitious short report of a placebo controlled trial of appetite suppressants. Manuscript A described testing an orthodox compound, whereas manuscript B described testing a homeopathic equivalent. Manuscripts originated from a fictitious institution and were identical except for the name of the drug. Participants were 369 reviewers (only 35.4% suitable for evaluation) unaware that the short report was fictitious.

Outcomes were assessed by scores of importance of the manuscripts and a visual analogue scale recommending to reject or accept the paper. The score sheet developed by the authors for peer-review assessment consisted of dichotomous questions, summarizing questions on importance (1= trivial even if true, 5= major contribution to knowledge in the field), and a visual analogue scale recommending to reject or accept the paper. After dichotomization of the rating scale, a significant difference in favour of the orthodox version with an odds ratio of 3.01 (95% confidence interval, 1.03 to 8.25) was found. This observation mirrored that of the visual analogue scale for which the respective medians and interquartile ranges were 67% (51% to 78.5%) for version A and 57% (29.7% to 72.6%) for version B. The authors concluded that reviewers showed a wide range of responses to both versions of the paper, with a significant bias in favour of the orthodox version.

Studies assessing the effects of peer review on study validity

- Arnau 2003 reports a double-blind randomised trial of consecutive manuscripts of original research received by Medicina Clinica (Barcelona), a weekly Spanish journal of internal medicine. The study aimed to assess the effect of joint clinician-statistician (arm 1) review compared to clinician only review (standard review, arm 2). The study took place between May 2000 and February 2001. Eighty-two manuscripts were assessed. Quality of final manuscript compared to its original version was assessed by two statisticians (blinded to authors' and reviewers' identity). Quality was assessed using a modified version of the 9-point Goodman scale, assessing quality of manuscript by sections scoring from 0 to 10. Analysis of referees' reports revealed four deviations involving 'co-opting' of statisticians as reviewers into arm 2, or presence of clinical reviewers who were also statistically qualified. The effect (median score) of additional statistical review (ITT analysis) was not significant:1.35 (- 0.45 to 3.16). However when the arm 2 'co-opted' statistical input is taken into account the difference is significant (1.96, 95% CI 0.25-3.67). The authors concluded that adding a statistical reviewer improves the process of peer review.

- Day 2002 is a cohort study assessing the use of dedicated methodology and statistical reviewers for peer review. Staff at AEM randomly selected reviews from 1998 and 1999 written by methodology reviewers (two reviewers) and blinded their identity. Thirty studies (15 in each arm) were reviewed by one of two methodology and statistical reviewers versus non specialist reviewers randomly selected from all original research articles sent to AEM in 1997. The authors developed a checklist using existing taxonomies (99 items in eight categories for classification of comments made in manuscript reviews). The checklist had a bias towards methodology. The authors concluded that this small study provides evidence that two dedicated methodology and statistical reviewers provided reviews that were generally consistent and emphasised methodology issues that were distinct from those raised by regular reviewers. Although these findings are insufficient to establish the value of dedicated methodology review, they highlight the potential of such reviews to improve the methodological quality of manuscripts.

- Elvik (Elvik 1998) reports the results of an open comparison of the quality of 44 road safety evaluation studies published in peer-reviewed journals compared to that of 79 similar reports published in non-peer-reviewed journals. The author asked the question whether peer-reviewed journals publish more valid reports than non-peer-reviewed journals. The author assessed validity against seven criteria addressing sampling techniques, total and mean sample size, specification of accident or injury severity, study design, number of confounding factors controlled and number of moderator variables specified. The only finding of note was that studies of authors whose affiliation to universities was published in peer-reviewed journals scored higher on validity.

- Rochon 2002 is a cohort study comparing review articles in published peer-reviewed versus throwaway journals focusing on diagnosis or treatment of medical conditions. Three hundred and ninety-four review articles published in 1998 either in the top five leading peer-reviewed journals (Annals of Internal Medicine, BMJ, JAMA, The Lancet and New England Journal of Medicine) or in highest circulation throwaway journals (Consultant, Hospital Practice, Patient Care and Postgraduate Medicine) were included. Of the 394 articles in the sample, 16 (4.1 %) were peer-reviewed systematic reviews, 135 (34.3%) were peer-reviewed non-systematic reviews, and 243 (61.7%) were non-systematic reviews published in throwaway journals. The mean [SD] quality scores were highest for peer-reviewed articles (0.94 [0.09] for systematic reviews and 0.30 [0.19] for non-systematic reviews) compared with throwaway journal articles (0.23 [0.03]). Throwaway journal articles used significantly more tables, figures, photographs, colour, and larger font sizes compared with peer-reviewed articles. Readability scores were significantly higher for throwaway journal articles (104 [77.0%] versus 156 [64.2%]; P=.01). Peer-reviewed article titles were judged less relevant to clinical practice than throwaway journal article titles (P<.001). The authors concluded that although lower in methodological and reporting quality, review articles published in throwaway journals have characteristics that appeal to physician readers.

Studies assessing the effects of peer review on study report quality

- Goodman (Goodman 1994) reports the conduct of a before and after study carried out in the editorial offices of Annals of Internal Medicine on 111 consecutive original manuscripts. The study aim was to evaluate the effects of peer review by expert assessors and editing on manuscript quality, between the submitted and 'ready for publication' version. Forty-four assessors reviewed the 111 manuscripts. Effects were assessed using a checklist instrument with 34 items grouped in 7 topics (introduction, methods - subjects and design, results - subjects and analysis, discussion and conclusions and general). The authors commented that the reliability of this checklist was low, with an intraclass correlation coefficient of 0.12 for average score. The authors concluded that peer review and editing improve the quality of medical research reporting, particularly in those areas that readers rely on most heavily to decide on the importance and generalisability of the findings and in methodologically weakest elements. However, the authors were cautious about the generalisability of these findings, because only a single journal was studied, and Annals of Internal Medicine may be atypical of journals.

- Pierie (Pierie 1996) reports the conduct of a non-randomised study to assess whether articles accepted by the Nederlands Tijdschrift voor Geneeskunde, (the Dutch Journal of Medicine), were improved after peer review (comparing quality of submitted versus accepted articles) and editing (comparing quality of accepted versus published articles). The first comparison was carried out by a 25-item questionnaire, the second with 17-item version. The questions were answered on five-point scales. Volunteer journal readers were asked to make these assessments of quality. Each volunteer assessor was sent a pack containing a set of identically appearing typescripts (i.e. blinding) of the submitted, accepted, and published versions of 50 articles that had been published in Nederlands Tijdschrift voor Geneeskunde . Each evaluator received two of the sets of versions, and each set was evaluated by one person from each group. The authors found that after peer review, the quality in 14 of 23 questions (61%) was significantly improved (p =0.03 or smaller). In particular, the overall score and general medical value were significantly improved (p =0.00001 for each). Editing led to significant improvement in 11 of 16 questions (69%, p =0.017 or smaller), and especially in style and readability (p = 0.001 and p=0.004). These indicated statistically significant improvement of published articles after both peer review and editing. It is difficult to draw conclusions about the importance of the absolute increases in quality detected.

Studies looking at authors' opinions

- Weller 1996 is part of a larger study on editorial peer review, investigating resubmission after rejection. The relevant part of the report compared the experiences of authors who published in two groups of indexed US medical journals, using a 'survey instrument' (questionnaire). The articles were randomly selected using a criteria developed by Weller. Participants were divided into two groups. Group one articles were taken from 17 'large prestigious medical journals'. Group two consisted of 742 'small speciality journals'. All of the journals were published in the USA and indexed on MEDLINE. The author reported that many authors made very positive comments about the review process, stating that editorial peer review, while imperfect, is the best process available. "Peer review had the greatest impact on the final presentation of the manuscript. Fifteen percent of all authors felt that the statistics were improved. It is through the statistical analysis of data that conclusions are drawn and clinical decisions are made. Peer review served to fine-tune the manuscript, and, for a small but improved percentage of the articles, it significantly improved them". According to the author, the group one review process did a slightly better job of making substantive changes to manuscripts than that of group two. Between 34 and 38% of authors felt review had improved content of the manuscripts.

Risk of bias in included studies

We decided to assess study quality separately for randomised and non-randomised studies.

Randomised studies

- The quality of Neuhauser 1989 is difficult to assess, as the process of randomisation, reasons for drop-outs and allocation concealment are not described in any detail.

- In McNutt 1990, although no details of methods used to randomise reviewers are reported, the robustness of blinding was tested by asking reviewers and editors to guess the identity of the authors and their institutions.

- Fisher 1994 is a small randomised controlled trial with a randomisation process using a random number table. Despite the authors' stratification by correctly guessed author's identity, the study failed to show differences in results. The weakness of the blinding process calls into question the validity of carrying out studies in small specialty settings.

- Jadad 1996 is a small randomised controlled trial with a randomisation sequence generated by a random number table. No mention is made of assessment of success of blinding.

- van Rooyen 1998 is a well designed and conducted study. The authors report a clear randomisation plan, but blinding appears to have been difficult with 58% concealment success.

Again, reasons for concealment failure were a small research field and referencing one's own work in the text.

- Although Godlee 1998 is a well-designed study, the 53% reviewers' response rate raises the issue of the introduction of bias in the study. A subsequent analysis of the characteristics of respondents and non-respondents showed no significant difference.

- Justice 1998 is a well-designed study in which randomisation took place using a random number table. Yet again, the relatively low blinding success rate cannot exclude the possibility of bias.

- van Rooyen 1999 is well-designed and conducted with an apparently robust randomisation process. Eleven submissions (9%) were excluded from the study as the time taken to obtain two suitable reviews would have imposed an unacceptable delay on the editorial process.

- Although randomisation appears to be robustly applied in Das Sinha 1999, the report fails to mention any attempt at assessing allocation concealment and no attempt is made at analysing any differences between the 22 drop-outs and the submissions included in the study.

- The design of Ernst 1999 included a computer-generated random allocation schedule, but the text does not describe blinding and allocation in detail sufficient to exclude the play of biases in the results. The response rate was comparatively low (61%).

- Methods of randomisation employed in Walsh 2000 are unclear as intervention and control groups are uneven (222 and 186 respectively including drop-outs). In addition, no effort was made to investigate the reason for dropping out of the trial. The Editor was unaware of which reviewers had consented to participate in the trial and throughout the trial there were no changes in the editorial decision-making process. Review quality assessment was carried out by two trainee editors blinded as to the review author's identity, an apparently robust process. Intention to review analysis was not carried out.

- Resch 2000 is a clearly written report with randomisation carried out in blocks of four. However, the poor response rate (35.4% of the total to evaluate) was partly due to the impossibility of sending out a reminder as this would have jeopardised confidentiality and increased the authors' ethical dilemma (the study did not have ethical approval).

- In Callaham 2002 randomisation is not described and the text of the report may engender some confusion (for example the use of the words 'matched controls'). Attrition bias makes the results very difficult to interpret.

- In Callaham 2002a the description of the studies is very brief. Randomisation and the possible effects of attrition selection bias are not described. Given these problems, generalisability of the results may be difficult.

- In Pitkin 2002 randomisation is not described and there is no description of the rating scale used in the study.

- In Arnau 2003 randomisation took place in blocks of four and was centralised in the editorial office. It is a well written study, but the tiny denominator and setting-specific issues may limit the generalisability of its conclusions.

- Schroter 2004 may have numerous biases such as a high, unexplained and differential drop out rate (77% versus 55% versus 74%) . There is no description of randomisation and reviewers may have under performed or over performed, knowing they were taking part in a trial. Some reviewers may not have persisted in detecting all the errors after identifying enough to condemn a paper. The authors believe that these factors are likely to have affected each of the three randomised groups of reviewers equally.

Non-randomised studies

- Generalisation from the results of Gardner 1990, a small open study based on the BMJ, may be limited. The potential for bias and a Hawthorne effect of the study is possible but its impact unknown in the absence of a control arm.

- In Strayhorn 1993,the criteria for selection of reviewers to undertake training are unclear. Additionally, the open nature of the design makes the interpretation of its findings difficult. The presence of several possible biases (such as the selection of reviewers) could account for the study findings.

- The small sample and before and after design are the greatest limitations of Goodman 1994. The study conclusions therefore have limited generalisability.

- Pierie 1996 also has a small denominator and a design in which bias could have affected results. For instance, volunteer assessors (selected on the basis of their order of reply to an advertisement) are very unlikely to be representative of all readers. Other potential biases are discussed in the text but their possible impact is played down with weak arguments.

- The open design of Bingham 1998 makes it almost a descriptive or feasibility study of uncertain generalisability.

- There are several methodological problems which make the interpretation of the results from Jefferson 1998 difficult. A number of submissions were untraceable, and the authors could not assess their relationship to the study population. In addition, the open design and low-key publication of the guidelines made the assessment of their impact difficult.

- In Callaham 1998 the self-selected nature of attendees at the training workshop and the absence of blinding raises the issue of the presence of bias in the study results. Study findings can be explained as a regression to the mean of editors and reviewers views after the face-to-face session.

- Elvik 1998 is the only study which we were able to identify addressing the crucial issue of the effects of peer review on validity. Despite such importance, it remains a small effort by a single author. The choice of checklist criteria are open to debate, but the author did attempt to adjust his results for potential confounders.

- The results of Day 2002 are difficult to interpret given the possibility of bias in the random extraction of submissions and the tiny denominator.

- Rochon 2002 is a well-written report. The study may have introduced bias (as also discussed by the authors). The scoring tables used had a bias towards systematic reviews, the reviewers were not truly representative of the population as they were new graduates and there may be a heavy selection bias as the top five peer-reviewed journals. Finally journals and throwaway journals have entirely different purposes and the authors may not have been comparing like with like.

- In Weller 1996 methods of random extraction of papers for both groups were different. Group one used a random number table. As group two had too many articles indexed from the 742 journals, the authors were selected by paired sets of random numbers. Only examined indexed articles published in and originating from the US and no data were gathered on rejected manuscripts that were never published or published in a non-indexed journal, or were published in a journal indexed by a service other than MEDLINE. Additionally, authors' opinions were sought only after they had published successfully. This may have biased views in favour of peer review.

In summary, the quality of randomised studies was superior to the design of the non-randomised studies. The most interesting methodological issue we identified in the design of the studies was the consistent difficulty in ensuring robust double blinding procedures. This was probably a reflection of the relatively small world of editorial peer review, especially when applied to specialist areas of knowledge. No study commented on potential adverse effects of peer review.

Effect of methods

Editorial peer review, as performed by certain journals, appears to make papers more readable and improve the general quality of reporting, but the generalisability of evidence for this conclusion is limited. We found limited evidence that peer review may improve validity of published research but evidence of its effects on readability was equivocal.

The results of our descriptive review show that the relatively well-researched practice of author concealment, which is laborious and expensive, appears to have some effect on the outcome of the quality assessment process. There is no reliable evidence that reviewer concealment has a benefit. There is some evidence that referees produce more courteous reports when their name is to be revealed. However, it is not clear whether this translates to an effect on the quality of the final publication.

Checklists, while convenient for editors, have limited reliable evidence to support their use.

We found limited evidence that training engendered changes in behavior but these effects are limited and seem to be short-lived.

Different methods of communication with reviewers and means of dissemination do not appear to have an effect on quality.

On the basis of two small studies, little can be said about the ability of the peer-review process to detect bias against unconventional interventions.

DISCUSSION

The available research has not clearly addressed the more important outcomes identified in our conceptual model. One of us (FD) took the view that the outcomes listed as ideal in Table 1 were unrealistic, and that the aims of editorial peer review are much more limited.

There seems to have been great interest in the area of blinding or concealment of reviewers and/or authors and little or no effort in other, conceptually far more important, areas of peer-review research. For example, virtually no attention has been paid to the study of manuscripts rejected by journals, and the effects of peer review on those manuscripts. Since many articles rejected by one journal are eventually published in another, the effects of editorial peer review may not become apparent unless they are studied across entire groups of journals. In addition, a wide variety of ad hoc instruments have been developed and used. As none of these are the same, aggregate evaluation is practically impossible.

The studies that we identified address marginal aspects and evaluate low-rank surrogate indicators of effect and are likely to be a reflection both of the relative youth of the research effort on editorial peer review and of the preponderance of journal editors. Editors understandably are more likely to engage in research which is directly related to the day-to-day management of their journals,

rather than in longer-term large projects to test the validity of a process they apply daily. However, the evaluation of editorial peer review should be based on its aims rather than practitioners' expediency or specific interests. Although not specifically an aim of the review, we were disappointed by the lack of identification and evidence for possible alternatives to the current model of peer review (such as open peer review, no peer review and other forms of manuscript improvement).

We tested the robustness of our findings by applying less sensitive inclusion criteria to the studies which we had identified in our initial search. These included studies such as Cho 1998 (a survey looking for associations between success of reviewer masking, journals' policy on masking and other characteristics). Different inclusion criteria made no difference to our findings.

One particular problem is that of generalisability of study results. Most studies were carried out in small numbers of atypical journals run by pioneer editors. This problem could be overcome by a large multicentre effort to assess the effects of editorial peer review.

Given the widespread use of peer review and its importance, it is surprising that so little is known of its effects. However, the research needed to address these questions would require a well-funded and coordinated effort involving several sectors of the scientific community and the cooperation of large numbers of authors and editors. Such research could, perhaps, best be coordinated by a central body, such as a scientific council organised with the specific aim of scientifically assessing the quality of peer review. Until such research is undertaken, peer review should be regarded as a long-standing, potentially expensive, untested process with uncertain outcomes. At the same time, the current absence of evidence on the efficacy and effectiveness of editorial peer review is very likely a reflection of the conceptual and methodologic difficulties of studying the scientific process itself, rather than evidence of the absence of efficacy and effectiveness. Phillips, in his letter about testing peer review for grant applications commented that "if we're going to spend billions of the taxpayers' dollars on science, the least we can do is spend it scientifically" (Phillips 2000). Although the money involved may be less, we believe that the system of ensuring the availability of reliable evidence to decision-makers lies at the heart of improving healthcare. Information quality filters are essential and may even help "shape the future of medicine and public health" (Horton 2000), however, given the current lack of evidence about the effects of editorial peer review, we believe it would be premature to make either positive or negative judgments about the ultimate value of this long-standing element in the scientific process.

AUTHORS' CONCLUSIONS

Implication for systematic reviews and evaluations of healthcare

We conclude that at present there is little systematic, empirical evidence to support the use of editorial peer review as a mechanism to ensure quality of reports of biomedical research in biomedical journals. Practitioners of editorial peer review should recognise the lack of convincing empirical evidence of its effects and bear this in mind when making editorial decisions. At the same time, editors and reviewers should be aware of the conceptual and methodological difficulties involved in studying an activity as complex as the scientific process, of which editorial peer review is an integral part.

Implication for methodological research

A large, well-funded programme of research on the effects of editorial peer review should be urgently launched. The aim of the programme should be to identify the procedures that guarantee better quality of scientific output of both individual journals and larger groups of journals in the aggregate.

During our searches we identified the existence of evidence of the effects of peer review in other areas of science, such as industrial economics. The value of examining and learning from this literature is unclear.

ACKNOWLEDGEMENTS

For the 2005 update: Marit Johansen carried out the searches. Professor Lisa Bero and Dr Sarah Schroter provided additional information.

For the 2003 review: Ms Philippa Middleton carried out the searches, Drs Iain Chalmers, Drummond Rennie, John Overbeke, Mike Clarke commented on the protocol and the review, and Ms Sarah Moore provided administrative support.

REFERENCES

References to studies included in this review

Arnau 2003 *{published data only}*

Arnau C, Cobo E, Ribera JM, Cardellach F, Selva A, Urrutia A. [[Efecto de la revision estadistìca en la calidad de los manuscritos publicados en Medicina Clìnica: estudio aleatorizado]. *Med Clin (Barc)* 2003;**121**(18):690–4].

Bingham 1998 *{published data only}*

* Bingham CM, Higgins G, Coleman R, Van der Weyden MB. The Medical Journal of Australia internet peer-review study. *The Lancet* 1998;**352**:441–5.

Callaham 1998 *{published data only}*

* Callaham ML, Wears RL, Waeckerle JF. Effect of attendance at a training session on peer reviewer quality and performance. *Annals of Emergency Medicine* 1998;**32**(3):318–22.

Callaham 2002 *{published data only}*

Callaham ML, Schriger DL. Effect of structured workshop training on subsequent performance of journal peer reviews. *Annals of emergency medicine* 2002;**40**(3):323–28.

Callaham 2002a *{published data only}*

Callaham ML, Knopp RK, Gallagher EJ. Effect of written feedback by editors on quality of reviews: Two randomized trials. *Journal of the American Medical Association* 2002;**287**(21):2781–2783.

Das Sinha 1999 *{published data only}*

* Das Sinha S, Sahni P, Nundy S. Does exchanging comments of Indian and non-Indian reviewers improve the quality of manuscript reviews?. *The National Medical Journal of India* 1999;**12**(5):210–3.

Day 2002 *{published data only}*

Day FC, Schriger DL, Todd C, Wears RL. The use of dedicated methodology and statistical reviewers for peer review: a content analysis of comments to authors made by methodology and regular reviewers. *Annals of Emergency Medicine* 2002;**40**(3):329–33.

Elvik 1998 *{published data only}*

* Elvik R. Are road safety evaluation studies published in peer reviewed journals more valid than similar studies not published in peer reviewed journals?. *Accid Anal and Prev* 1998;**30**(1):101–18.

Ernst 1999 *{published data only}*

* Ernst E, Resch K-L. Reviewer bias against the unconventional?A randomized double-blind study of peer review. *Complementary Therapies in Medicine* 1999;**7**:19–23.

Fisher 1994 *{published data only}*

* Fisher M, Friedman SB, Strauss B. The effects of blinding on acceptance of research papers by peer review. *JAMA* 1994;**272**: 143–6.

Gardner 1990 *{published data only}*

* Gardner MJ, Bond J. An exploratory study of statistical assessment of papers published in the British Medical Journal. *JAMA* 1990;**263**(10):1355–7.

Godlee 1998 *{published data only}*

* Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding peer reviewers and asking them to sign their reports. *JAMA* 1998;**280**(3):237–40.

Goodman 1994 *{published data only}*

Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at Annals of Internal Medicine.. *Ann Intern.Med.* 1994;**121**:11–21..

Jadad 1996 *{published data only}*

* Jadad AR, Moore A, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary?. *Controlled Clinical Trials* 1996;**17**:1–12.

Jefferson 1998 *{published data only}*

* Jefferson T, Smith R, Yee Y, Drummond M, Prat M, Gale R. Evaluating the BMJ guidelines for economic submissions: prospective audit of economic submissions to BMJ and The Lancet. *JAMA* 1998;**280**(3):275–7.

Justice 1998 *{published data only}*

* Justice AC, Cho MK, Winker MA, Berlin JA. Does masking author identity improve peer review quality?. *JAMA* 1998;**280**(3): 240–2.

McNutt 1990 *{published data only}*

* McNutt RA, Evans AT, Fletcher RH, Fletcher SW. The effects of blinding on the quality of peer review. *JAMA* 1990;**263**(10): 1371–6.

Neuhauser 1989 *{published data only}*

* Neuhauser D, Koran CJ. Calling Medical Care reviewers first: a randomized trial. *Medical Care* 1989;**27**(6):664–6.

Pierie 1996 *{published data only}*

Pierie J, Walvoort HC, Overbeke AJPM. Readers' evaluation of effect of peer review and editing on quality of articles in the Nederlands Tijdschrift voor Geneeskunde. *Lancet* 1996;**348**: 1480–3.

Pitkin 2002 *{published data only}*

Pitkin RM, Burmeister LF. Identifying manuscript reviewers: randomized comparison of asking first or just sending. *Journal of the American Medical Association* 2002;**287**(21):2795–2796.

Resch 2000 *{published data only}*

Resch KI, Ernst E, Garrow J. A randomized controlled study of reviewer bias against an unconventional therapy. *Journal of the Royal Society of Medicine* 2000;**93**(4.):164–167.

Rochon 2002 *{published data only}*

Rochon PA, Bero LA, Bay AM, Gold JL, Dergal JM, Binns MA, Streiner DL, Gurwitz JH. Comparison of review articles published in peer-reviewed and throwaway journals. *Journal of the American Medical Association* 2002;**287**(21):2853–2856.

Schroter 2004 *{published data only}*

Schroter S, Black N, Evans S, Carpenter J, Godlee F, Smith R. Effects of training on quality of peer review: Randomised controlled trial. *British Medical Journal.* 2004;**328**(7441):673–675.

Strayhorn 1993 *{published data only}*

* Strayhorn J, McDermott JF, 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**(6):947–52.

van Rooyen 1998 *{published data only}*

Van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of blinding and unmasking on the quality of peer review. *JGIM* 1999; **14**(10):622–4.

* van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of blinding and unmasking on the quality of peer review. *JAMA* 1998; **280**(3):234–7.

van Rooyen 1999 *{published data only}*

van Rooyen S, Black N, Goodlee F. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *J Clin Epidemiol* 1999;**52**(7):625–629.

* van Rooyen S, Godlee F, Evans S, Black N, Smith R. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ* 1999;**318**:23–7.

Walsh 2000 *{published data only}*

Walsh E, Rooney M, Appleby L, Wilkinson G. Open peer review: a randomised controlled trial. *British Journal of Psychiatry* 2000;**176**: 47–51.

Weller 1996 *{published data only}*

Weller AC. Editorial peer review: A comparison of authors publishing in two groups of U.S. medical journals. *Bulletin of the Medical Library Association* 1996;**84**(3):359–366.

References to studies excluded from this review

Abby 1994 *{published data only}*

Abby M, Massey MD, Galandiuk S, Polk HC. Peer review is an effective screening process to evaluate medical manuscripts. *JAMA* 1994;**272**(2):105–6.

Bacchetti 2002 *{published data only}*

Bacchetti P. Peer review of statistics in medical research: The other problem. *British Medical Journal* 2002;**324**(7348):1271–1273.

Baxt 1998 *{published data only}*

Baxt WG, Waeckerle JF, Berlin JA, Callaham ML. Who reviewers the reviewers?Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Annals of Emergency Medicine* 1998;**32**: 310–17.

Blank 1991 *{published data only}*

* Blank RM. The effects of double-blind versus single-blind reviewing; experimental evidence from the American Economic Review. *American Economic Review* 1991;**81**(5):1041–67.

Cho 1998 *{published data only}*

* Cho MK, Justice AC, Winker MA, Berlin JA, Waeckerle, Callaham ML, Rennie D. Masking author identity in peer review: what factors influence masking success?. *JAMA* 1998;**280**(3): 243–5.

Cicchetti 1992 *{published data only}*

Cicchetti DV, Rourke BP, Wass P. Peer review for manuscript and grant submissions: Relevance for research in clinical neuropsychology. *Journal of Clinical and Experimental Neuropsychology* 1992;**14**(6):976–80.

Cicchetti 1998 *{published data only}*

Cicchetti DV. Good science and good peer reviewing: Are they related?. *Journal of Clinical and Experimental Neuropsychology* 1998; **20**(3):428–31.

Cleary 1988 *{published data only}*

Cleary J, Alexander B. Blind versus non blind review: a reevaluation of selected medical journals. *DICP, The Annals of Pharmacotherapy* 1990;**24**:1117–8.

Cleary JD, Alexander B. Blind versus nonblind review: Survey of selected medical journals. *Drug Intelligance and Clinical Pharmacy* 1988;**22**:601–2.

Coronel 1999 *{published data only}*

Coronel R, Opthof T. The role of the reviewer in editorial decision-making. *Cardiovascular Research* 1999;**43**:261–4.

Cox 1993 *{published data only}*

Cox D, Gleser L, Perlman M, Reid N, Roeder K. Report of the ad hoc committee on double-blind refereeing. *Statistical Science* 1993; **8**(3):310–30.

Cullen 1992 *{published data only}*

Cullen DJ, Macaulay A. Consistency between peer reviewers for a clinical specialty journal. *Academic Medicine* 1992;**67**(12):856–9.

Dixon 1983 *{published data only}*

Dixon GF, Schonfeld SA, Altman M, Whitcomb ME. The peer review and editorial process: A limited evaluation. *The American Journal of Medicine* 1983;**74**:494–5.

Feurer 1994 *{published data only}*

Feurer ID, Becker GJ, Picus D, Ramirez E, Darcy MD, Hicks ME. Evaluating peer reviews: Pilot testing of a grading instrument. *JAMA* 1994;**272**(2):98–100.

Goldbeck-Wood 1999 *{published data only}*

Goldbeck-Wood S. Secrecy and openness in peer review - time for a change of culture?. *Italian Journal of Gastroenterology and Hepatology* 1999;**31**(8):659–62.

Hatch 1998 *{published data only}*

Hatch CL, Goodman SN. Perceived value of providing peer reviewers with abstracts and preprints of related published and unpublished papers. *JAMA* 1998;**280**(3):273–4.

Hemlin 1999 *{published data only}*

Hemlin S. (Dis) Agreement in peer review. In: Juslin P Montgomery H editor(s). *Judgment and decision making: Neo-Brunsurikian and process-tracing approaches*. Nahwah, New Jersey: Lawrance Erlbaum Associates, 1999:275–301.

Jadad 1998 *{published data only}*

Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, Moher D. Methodology and reports of systematic reviews and meta-analyses: A comparison of Cochrane reviews with articles published in paper-based journals. *JAMA* 1998;**280**(3):278–80.

Katz 2002 *{published data only}*

Katz DS, Proto AV, Olmsted WW. Incidence and nature of unblinding by authors: Our experience at two radiology journals with double-blinded peer review policies. *American Journal of Roentgenology* 2002;**179**(6):1415–1417.

Kumar 1999 *{published data only}*

Kumar PD. How do peer reviewers of journal articles perform? Evaluating the reviewers with a sham paper. *J Assoc Physicians India* 1999;**47**:198–200.

Laband 1994 *{published data only}*
* Laband DN, Piette MJ. Does the 'blindness' of peer review influence manuscript selection efficiency?. *Southern Economic Journal* 1994;**60**:896–906.

Lee 2002 *{published data only}*
Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. *Journal of the American Medical Association* 2002;**287**(21): 2805–2808.

Mahoney 1977 *{published data only}*
Mahoney MJ. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* 1977;**1**(2):161–175.

Morrow 1992 *{published data only}*
Morrow J, Bray M, Fulton J, Thomas J. Interrater reliability of 1987-1991 Research Quarterly for Exercise and Sport reviews. *Research Quarterly for Exercise and Sport* 1992;**63**:200–4.

Nylenna 1994 *{published data only}*
Nylenna M, Riis P, Karlsson Y. Multiple blinded reviews of the same two manuscripts: Effects of referee characteristics and publication language. *JAMA* 1994;**272**(2):149–51.

Opthof 1999 *{published data only}*
Opthof T. Submission, acceptance rate, rapid review system and impact factor. *Cardiovascular Research* 1999;**41**:1–4.

Oxman 1991 *{published data only}*
Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchinson BG, Milner RA, Streiner DL. Agreement among reviewers of review articles. *Journal of Clincal Epidemiology* 1991;**44**(1):91–8.

Peters 1982 *{published data only}*
Peters DP, Ceci SJ. Peer-review practices of psychological journals: The fate of published articles, submitted again. *The Behavioural and Brain Sciences* 1982;**5**:187–255.

Pitkin 2002a *{published data only}*
Pitkin RM, Burmeister LF. Prodding tardy reviewers. A randomised comparison of telephone, fax, and e-mail. *JAMA* 2002;**287**(21): 2794–5.

Presser 1980 *{published data only}*
Presser S. Collaboration and the quality of research. *Social Studies of Science (SAGE, London & Beverly Hills)* 1980;**10**:95–101.

Purcell 1998 *{published data only}*
Purcell GP, Donovan SL, Davidoff F. Changes to manuscripts during the editorial process: Characterizing the evolution of a clinical paper. *JAMA* 1998;**280**(3):227–8.

Rosenblatt 1980 *{published data only}*
Rosenblatt A, Kirk SA. Recognition of authors in blind review of manuscripts. *Journal of Social Service Research* 1980;**3**(4):383–94.

Schriger 2002 *{published data only}*
Schriger DL, Cooper RJ, Wears RL, Waeckerle JF. The effect of dedicated methodology and statistical review on published manuscript quality. *Annals of Emergency Medicine* Schriger DL, Cooper RJ, Wears RL2002;**40**(3):334–7.

van Rooyen 1999b *{published data only}*
van Rooyen, Black N, Godlee F. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology* 1999;**52**:625–9.

References to ongoing studies

Delamothe/BMJ *{published data only}*
Ongoing study Starting date of trial not provided. Contact author for more information.

Lee/Bero 2004 *{published and unpublished data}*
A Qualitative Study of Editorial Decision Making. Ongoing study Starting date of trial not provided. Contact author for more information.

Schroter/BMJ 2004 *{published data only}*
Ongoing study Starting date of trial not provided. Contact author for more information.

Additional references

Antman 1992
Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized controlled trials and recommendations of clinical experts. *JAMA* 1992;**268**: 240–8.

Godlee 2003
Godlee F, Jefferson T. Introduction. In: Godlee F, Jefferson T editor(s). *Peer review in health sciences*. Second Edition. London: BMJ Books, 2003:xiii–xv.

Hagstrom 1965
Hagstrom WO. *The Scientific Community*. Southern Illinois University Press, 1965.

Horton 2000
Horton R. The refiguration of medical thought. *The Lancet* 2000; **356**:2–4.

Jefferson 1999
Jefferson T. What are the costs and benefits of editorials and non-systematic reviews?. *BMJ* 1999;**318**:135.

Jefferson 2002
Jefferson TO, Alderson P, Davidoff F, Wager E. Effects of editorial peer review: a systematic review. *JAMA* 2002;**287**:2784–2786.

Jefferson 2002a
Jefferson TO, Wager E, Davidoff F. Measuring the quality of editorial peer review. *JAMA* 2002;**287**:2786–90.

Knoll 1990
Knoll E. The communities of scientists and journal peer-review. *JAMA* 1990;**263**:1330–2.

Kronick 1990
Kronick DA. Peer-review in 18th-century scientific journalism. *JAMA* 1990;**263**:1321–2.

Overbeke 2003
Overbeke J, Wager E. The state of the evidence: what we know and what we don't know about journal peer review. In: Godlee F, Jefferson T editor(s). *Peer review in health sciences*. Second Edition. London: BMJ Books, 2003:45–61.

Phillips 2000
Phillips M. Peer review. *The Lancet* 2000;**355**:660.

Rennie 1990
Rennie D. Editorial peer-review in biomedical publication. The first international congress. *JAMA* 1990;**263**:1317.

Rennie 2003
Rennie D. Editorial peer review:its development and rationale. In: Godlee F, Jefferson T editor(s). *Peer review in health sciences*. Second Edition. BMJ Books, 2003:1–13.

Sackett 1997
Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine*. Edinburgh: Churchill Livingstone, 1997:8–9.

Saint 2000
Saint S, Christakis DA, Saha S, Elmore JG, Welsh DE, Baker P, et al.Journal reading habits of internists. *J Gen Intern Med* 2000;**15**: 881–4.

Sievert 1996
Sievert M, McKinin EJ, Johnson ED, Reid JC, Mitchell JA. Beyond relevance - characteristics of key papers for clinicians: an exploratory study in an academic setting. *Bull Med Libr Assoc* 1996; **84**:351–8.

Vandenbroucke 1998
Vandenbroucke JP. Medical journals and the shaping of medical knowledge. *Lancet* 1998;**352**:2001–6.

Wager 2007
Wager E, Middleton P. Technical editing of research reports in biomedical journals. *Cochrane Database of Systematic Reviews* 2007, Issue 2. [DOI: 10.1002/14651858.MR000002.pub2]

* *Indicates the major publication for the study*

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Arnau 2003

Methods	Double-blind randomised trial of consecutive manuscripts of original research received by Medicina Clinica (Barcelona), a Spanish journal of internal medicine, published weekly. The study aimed to assess the effect of joint clinician-statistician (arm 1) review compared to clinician only review (standard review, arm 2). The study took place between May 2000 and February 2001. Randomisation took place in blocks of four and was centralised in the editorial office
Data	Eighty-two manuscripts submitted between 1 May 2000 and 31 May 2001 judged by the editorial board to fit journal content. All manuscripts fitting these criteria but not requiring statistical review (presumably editorials)were excluded from the study. During the study period the journal received 270 submissions of which 82 fit the inclusion criteria. Of the 16 that were rejected while undergoing review (9 in Arm 1 and 7 in Arm 2), 8 were lost to follow up because of administrative errors (3 and 5), 7 were still undergoing review after the end of the study period (4 and 3) and 8 (all in arm 1) were lost because of failure to respond at various stages. Four manuscripts needing statistical review were assigned from arm 2 to arm 1, so the final totals were 21 and 22 manuscripts respectively
Comparisons	Quality of final manuscript compared to its original version was assessed by two statisticians (blinded to authors' and reviewers' identity). The two also underwent preliminary assessment of published manuscripts. Their inter-rater agreement was 0.32
Outcomes	Quality was assessed using a modified version of the 9-point Goodman scale, assessing quality of manuscript by sections scoring from 0 to 10. The modification was necessary because of poor inter-rater agreement and of cultural factors such as statisticians' understanding of some of the scale questions. Analysis of referees' reports revealed four deviations involving 'co-opting' of statisticians into arm 2 or presence of clinical reviewers who were also statistically qualified. The effect (median score) of additional statistical review (ITT analysis) was 1.35 (95% CI -0.45 to 3.16). However when the arm 2 'co-opted' statistical input is taken into account, the difference is significant (1.96, 95% CI 0.25 to 3.67)
Notes	The authors conclude that adding a statistical reviewer improves the process of peer review. A well-written study, but the tiny denominator and setting-specific issues may limit the generalisability of its conclusions

Bingham 1998

Methods	Before and after study
Data	56 journal papers with accompanying reviewers' commentaries
Comparisons	Publishing the paper with reviewers' comments on the world wide web and asking for comments
Outcomes	1. Number of readers' comments 2. Manuscript revisions 3. Quality of reviewers' reports compared to before intervention

Bingham 1998 (*Continued*)

Notes	No change in quality of commissioned reviewers was noted before or during the study. Electronic comments from readers were few.

Callaham 1998

Methods	Prospective controlled study with two control groups; one group matched and one group unmatched
Data	Referees for one journal
Comparisons	4 hour training workshop for reviewers
Outcomes	1. Quality of reviewers' reports, assessed by journal editors 2. Recommendation for acceptance 3. Congruence with editors' decision
Notes	There was no significant change in any performance measure after the workshop.

Callaham 2002

Methods	Two randomised controlled trials designed to asses if peer reviewers from Annals of Emergency Medicine (AEM) attending a formal interactive training session produced better reviews
Data	In Study 1, 25 AEM reviewers randomised by standard invitation who volunteered attendances were matched to 25 control reviewers who were invited but did not attend, using the same inclusion criteria - Average review quality score (median < 4). Study 2 (intense recruitment)was conducted to reduce self-selection bias (limitation of study 1, as attendees and controls were randomised by invitation and not attendance), the same average review quality score. However, non-responders were aggressively followed up for a response by e-mail, etc until a confirmed response of the invitees was obtained. In study 2, 29 reviewers accepted the invitation and 12 actually attended
Comparisons	Attendance at a formal 4-hour highly interactive workshop on peer review. The format of the workshop included small group format with questions, discussion and debate. Before workshop: -Sent a fictitious manuscript to review in writing -Attendee's review of manuscript returned before workshop Workshop: -Brief written test on peer review (study 2 only) -Introduction to the process and goals of peer review -Presentation of the specific expectations of the journal for a quality peer review -Discussion of how to critically appraise a research manuscript (synopsis of evidence-based critical appraisal techniques) -Detailed review of the fictitious manuscript and discussions of its strengths and weaknesses -Discussion of how to write a review and communicate these strengths and weaknesses to the journal -Discussion of several dozen actual peer reviews, illustrating desirable and undesirable examples and features and their alternatives -Discussion of actual reviews of the fictitious manuscript, illustrating desirable and undesirable examples and features and their alternatives

Callaham 2002 (*Continued*)

	-Survey of attendees' opinion of the course -Brief written test on peer review (study 2 only)
Outcomes	Outcomes for study 1 and 2 are the same: The number and mean rating of reviews in preceding 2 years were compared with the number and mean rating of reviews in the 2 years after the workshop The difference was expressed mean rating change (95% confidence interval) Study 1 - The mean change in rating after the workshop was 0.11 (95% confidence interval (CI) - 0.25 to 0.48) for control reviewers and 0.10 (95% CI - 0.20 to 0.39) for attended. Study 2 - of 75 reviewers intensively recruited, only 12 (41%) of those who said they would attend did. All of the participants thought the workshop would improve their performance ratings. Test scores at the end of the workshop improved in 73% of participants compared with scores on pretest. The control reviewers' average rating changed by - 0.10 (95% CI - 0.49 to 0.29) versus (95% CI - 0.34 to 0.23) for attendees
Notes	The authors conclude that among invited peer reviewers, voluntary attendance at a highly structured and interactive workshop was low and did not improve the quality of subsequent reviews, contrary to the predictions of attendees. Efforts to aggressively recruit average reviewers to a second workshop (study 2) were time consuming, had low success rates, and showed a similar lack of effect on ratings, despite improvement in scores on a test instrument. Workshop teaching formats, although traditional are of unproven efficacy. In view of the many biases present in both studies it is difficult to reach any conclusion on the basis of the evidence presented. The text is reported in a confusing manner with terms such as 'matched controls' engendering confusion

Callaham 2002a

Methods	Report of two randomised controlled trials assessing whether written feedback to reviewers of AEM improved subsequent reviews. Authors, editors and reviewers were blinded to each other's identity and reviewers to editor's rating of their reviews
Data	Reviewers were selected from an eligible pool. Study 1 included 35 reviewers with a median quality score of 3 or lower (low-volume, low-quality reviewers)and study 2 included 95 reviewers with median score of 4 or lower (low-volume, average quality reviewers) For study 1, 51 reviewers were eligible and randomized and 35 (182 reviews) had sufficient data for analysis. For study 2, 127 (324 reviews) reviewers were eligible and randomized, and 95 had sufficient data
Comparisons	Reviewers in both studies were randomized to intervention or control (standard procedure). The intervention in study 1 consisted of feed back with other reviewer's assessment of the same submission and a copy of the editor's letter to the authors. Study 2 had the same intervention as Study 1 but with the addition of the editor's ratings of reviewers' performance and a sample high grade review of another submission to the journal. Follow-up was two further reviews for each study
Outcomes	Editor's routine quality rating (on a scale of 1 to 5) of all reviews (blinded to study enrollment). For study 1, the mean individual reviewer rating change was 0.16 (95% confidence interval [Cl], -0.26 to 0.58) for control and -0.13 (-0.49 to 0.23) for intervention. For study 2, controls had a mean individual rating change of 0.12 (95% Cl, -0.20 to 0.26) and intervention reviewers, 0.06 (-0.19 to 0.31)
Notes	The authors conclude that in study 1, minimal feedback from editors on review quality had no effect on subsequent performance of poor-quality reviewers, and the authors thought there might have been a negative effect. In study 2,

Callaham 2002a (*Continued*)

	feedback to average reviewers, despite being more extensive and supportive, produced no improvement in reviewer performance. In summary, simple written feedback to reviewers seems to be an ineffective educational tool. The description of the studies is very brief. Randomisation and the possible effects of selection bias are not described. It may be that as the authors state the average level of reviewers was so low that improvement was difficult. One wonders how generalisable the results may be.

Das Sinha 1999

Methods	Randomised controlled trial, allocation concealed by sealed envelopes
Data	Journal submissions
Comparisons	Each submission was sent to Indian and non-Indian reviewers. In one group the reviewers' were told that their comments would be sent to the other reviewer, in the other group no exchange of comments occurred
Outcomes	Quality of reviews assessed by journal editors, blinded to name of reviewer
Notes	The authors concluded that their study showed that non-Indian peer reviewers produced better quality reviews than Indians. In addition, exchanging reviews among reviewers was not found to make a difference to their quality.

Day 2002

Methods	Study assessing the use of dedicated methodology and statistical reviewers for peer review. Journal staff randomly selected reviews from 1998 and 1999 written by methodology reviewers (two reviewers)and blinded their identity
Data	Thirty studies (15 in each arm) reviewed by one of two methodology and statistical reviewers versus non specialist reviewers randomly selected from all original research articles sent to Annals of Emergency Medicine (AEM) in 1997
Comparisons	The authors first developed a checklist using existing taxonomies (99 items in eight categories for classification of comments made in manuscript reviews). The checklist had a bias towards methodology (taxonomy is detailed in the study appendix)
Outcomes	To present the results the authors used summary scales. In table 1 they compare methodology reviewers A versus B versus regular peer reviewers. The outcomes were: 1. Number of reviews (number of comments 2. Mean number of comments per review (range) Table 2 - distribution of reviewers comments , by category from the 99 item taxonomy (checklist)as follows: 1. Hypothesis/purpose/theoretic model 2. Study design/power 3. Research and analytic methods 4. Statistical methods 5. Presentation of methods 6. Presentation of results 7. Interpretation of results/limitations 8. Other comments (non methodology/statistical)

Day 2002 (*Continued*)

Notes	The authors conclude that "this small study provides evidence that two dedicated methodology and statistical reviewers provided reviews that were generally consistent and emphasised methodology issues that were distinct from those raised by regular reviewers. Although these findings are insufficient to establish the value of dedicated methodology review, they do highlight the potential of such reviews to improve the methodological quality of manuscripts". Random extraction is biased as it was done by journal staff who selected reviews from 1998 and 1999 written by Methodology reviewers A+B. How the randomisation took place is not described. In addition, they are assessing specialised versus regular reviewers and we are not told how many.

Elvik 1998

Methods	Comparative study comparing quality of papers published in journals with and without peer review
Data	Published evaluation studies in the field of road safety
Comparisons	Journal policy: with or without peer review
Outcomes	Validity against 7 criteria: sampling techniques, total and mean sample size, specification of accident or injury severity, study design, number of confounding factors controlled and number of moderator variables specified
Notes	Studies of authors whose affiliation to universities was published in peer-reviewed journals scored higher on validity.

Ernst 1999

Methods	Randomised controlled trial, allocation 'by computer' Designed to test for reviewer bias against unconventional interventions
Data	Medical practitioners invited to act as reviewers
Comparisons	Reviewers allocated to receive fictitious papers identical except for the intervention: in one it was metoprolol, in the other beef spleen extract
Outcomes	Quality of paper, and quality of study, rated by reviewer
Notes	No evidence of reviewer bias towards the 'unconventional' drug but showed poor inter-rater reliability.

Fisher 1994

Methods	Randomised controlled trial - allocation of reviewers to blind or non-blind group used a random numbers table
Data	114 reviewers for the Journal of Developmental and Behavioural Pediatrics reviewing 57 manuscripts
Comparisons	Reviewers allocated to review the manuscript either with or without the cover sheet, headers and footers, which identify the authors.
Outcomes	Five-point scale rating the quality of the manuscript and its suitability for publication.

Fisher 1994 (*Continued*)

Notes	Blinded peer review is effective in producing less biased reviews.

Gardner 1990

Methods	Before and after study
Data	45 papers published in the BMJ that had been assessed statistically on submission
Comparisons	Review of submitted papers by a statistical referee using a checklist, followed by revision and publication
Outcomes	Proportion of papers considered statistically acceptable on submission and publication
Notes	Statistical assessment with the use of checklists is beneficial to increasing the statistical quality of published studies.

Godlee 1998

Methods	Randomised controlled trial - allocation by random numbers assigned by a statistician
Data	420 reviewers from the BMJ's bank, reviewing one paper with 8 weaknesses inserted
Comparisons	1) authors' details removed, reviewer asked to sign report 2) authors' details removed, reviewer not asked to sign 3) authors' details not removed, reviewer asked to sign 4) authors' details not removed, reviewers not asked to sign 5) As for 4, but reviewers unaware they were in a study
Outcomes	Mean number of weaknesses detected by reviewers
Notes	There was no evident effect of concealment on the quality of reviews.

Goodman 1994

Methods	Before and after study
Data	111 consecutive manuscripts accepted for publication in the Annals of Internal Medicine, reviewed by 44 reviewers
Comparisons	Revision of original manuscript by authors, incorporating comments from outside reviewers, editors and production editor
Outcomes	Quality of manuscript on 34 item scale. Rated by 'expert' assessors
Notes	Peer review and editing improve the quality of medical research reporting, however, the authors were cautious about the generalisability of their findings.

Jadad 1996

Methods	Randomised controlled trial, using random number table
Data	14 reviewers from 3 categories (researchers, clinicians, others) assessing 36 reports of pain research
Comparisons	1) 7 reviewers blinded to the authors, affiliations, journal, date of publication, financial support and acknowledgements 2) 7 reviewers not blinded to above details
Outcomes	Quality measured with a series of related scales with 11, 6 and 3 items
Notes	Blinded assessment produced lower quality scores and more consistent results than open assessment.

Jefferson 1998

Methods	Before and after study carried out in two settings, one with the intervention and one without
Data	105 (before) and 87 (after) reports of economic studies submitted to the BMJ and Lancet
Comparisons	1) Active introduction of guidelines for review and revision of economic submissions at the BMJ 2) No active introduction of guidelines at the Lancet, though the guidelines were in the public domain
Outcomes	Quality of economic submissions using 26-item checklist Acceptance rate
Notes	Guidelines had no apparent impact on the quality of submissions but helped editors in managing submissions.

Justice 1998

Methods	Randomised controlled trial using random number table
Data	118 submissions to 5 journals
Comparisons	1) 1 of 2 reviewers blinded to manuscript authors and institutions (n = 92 manuscripts) 2) Journal usual practice (no blinding in 4 journals, 1 blinding authors' identity) (n = 26)
Outcomes	Quality of reviewers' reports as judged by journal editor and author, using 4-item scales
Notes	Analysis based on 77 (84%) manuscripts from intervention arm where data available for editors' rating, and 40 manuscripts from intervention arm for authors' rating

McNutt 1990

Methods	Randomised controlled trial, for each manuscript one reviewer randomized to intervention and one to control
Data	131 original research manuscripts submitted to the Journal of General Internal Medicine

McNutt 1990 (*Continued*)

Comparisons	1) Reviewer blinded to authors and institutions 2) Reviewer not blinded
Outcomes	Quality rated by editor (from editor and author perspective), and by author
Notes	No association between signing and review quality was detected, but blinding the reviewer resulted in a statistically significantly higher quality review.

Neuhauser 1989

Methods	Randomised controlled trial
Data	Reviewers of 95 manuscripts submitted to Medical Care
Comparisons	1) Reviewer telephoned to warn that manuscript was being sent 2) Reviewer not warned
Outcomes	Time for various steps in the editorial process
Notes	Total turnaround time was significantly longer in the group with warned reviewers and calling reviewers increases costs.

Pierie 1996

Methods	Retrospective comparative study
Data	50 articles in submitted, accepted and published forms, from Nederlands Tijdschrift voor Geneeskunde
Comparisons	1) Peer review (submitted to accepted versions) 2) Technical editing (accepted to published versions)
Outcomes	Quality of each report rated by one person from each of 4 categories (medical student, recent medical graduate, general practitioner, specialist)using 25 and 17-item questionnaires
Notes	There was a statistically significant improvement of published articles after both peer review and editing

Pitkin 2002

Methods	Randomised controlled trial comparing the effects of sending the same manuscript for review either with no prior warning ('justsend')or with a preliminary fax asking for reviewer availability to review ('askfirst'). Randomisation took place by random number generator
Data	Two hundred and eighty three consecutive manuscripts submitted to the Journal of Obstetrics & Gynaecology between September 1999 and May 2000 and 566 reviewers. Reviewers were chosen by the editor on the basis of availability of address and fax number Two hundred and forty-seven of the 'justsend' referees and 177 of the 'askfirst' referees produced a review. Twenty-

Pitkin 2002 (*Continued*)

	two of the 'justsend' declined at several stages to review and the manuscripts were sent on other referees (all 22 of first refusals were eventually refereed). One hundred and eighty-one of the 'askfirst' referees agreed to review, whereas 102 of the initial referees either did not respond (59) or declined (43). All 102 manuscripts were eventually refereed by 'substitute' referees
Comparisons	Manuscripts sent to referees versus preceded by a fax asking for availability to review
Outcomes	1. Proportion of 'justsend' referees who failed to optout 2. Proportion of 'askfirst' referees who agreed to review 3. Time (in days) from enrolment in study reception of review (time 1) 4. Time (in days) from mailing of manuscript to reception of review (time 2) 5. Blinded quality assessment of review on a 5-point scale
Notes	Although differences in frequency of specific declines were significant (8% versus 15%), the production rate of reviews by original 'justsend' referees who did not optout and 'askfirst' reviewers who agreed to referee were not significant (247/261 versus 177/181). Time 1 was not significantly different between arms (24.7 + 9.7 versus 25.9 + 10.5), whereas time 2 was significantly different (21.0 + 9.2 versus 25 + 10.1). Review quality in the subset of 151 referee reports agreed to review and returned reviews (147-22, as the allocator was blinded) was not significantly different. The authors conclude that advance warning of review did not affect quality, but elicited 36% of turndowns or actual reviews. However the extra work involved in finding 'substitute' reviewers was made up by the 'askfirst' speedier review turnaround. A reasonably reported study, with an unexplained and possibly unvalidated five-point quality scale

Resch 2000

Methods	Double-blind randomised trial to assess the hypothesis that peer review favours an orthodox form of treatment over conventional therapy. To test this hypothesis the authors produced a fictitious short report of a placebo controlled trial of appetite suppressants. Manuscript A was testing a orthodox compound, whereas manuscript B tested a homeopathic equivalent. Manuscripts originated from a fictitious institution and were identical except for the name of the drug. Reviewers were unaware that the short report was fictitious. The score sheet developed by the authors for peer-review assessment consisted of dichotomous questions, summarising questions on importance (1= trivial even if true, 5= major contribution to knowledge in the field), and a visual analogue scale recommending to reject or accept the paper
Data	Three hundred and sixty-nine reviewers selected by scanning MEDLINE, Jan 93 to June 96 for articles dealing with the treatment of obesity. 1137 articles were retrieved. Addresses of first authors were collected. They excluded authors who had previously reviewed manuscripts for the European Journal of Clinical Nutrition. If more than one paper from one institution was identified, whether from the same or different authors, only the latest was used. This left 396 addresses which were sorted by country, randomised into group blocks (randomisation with blocks of four)
Comparisons	Reviewers were randomised into group blocks of four. Groups A and B received the fictitious short report and the evaluation sheet. Group A's evaluation sheets title was underlined, where as group B's was not. The authors of the report did this to allow identification of group A or B. The response rate was 166 (41.7%), 141 which were used for valuation (35.4%)
Outcomes	The authors used a score sheet developed by themselves for peer-review assessment consisting of dichotomous questions, summarizing questions on importance (1= trivial even if true, 5= major contribution to knowledge in the field), and a visual analogue scale recommending to reject or accept the paper. After dichotomization of the rating scale, a significant difference in favour of the orthodox version with an odds ratio

Resch 2000 (*Continued*)

	of 3.01 (95% confidence interval, 1.03 to 8.25), was found. This observation mirrored that of the visual analogue scale for which the respective medians and interquartile ranges were 67% (51% to 78.5%) for version A and 57% (29.7% to 72.6%) for version B
Notes	The authors conclude that reviewers showed a wide range of responses to both versions of the paper, with a significant bias in favour of the orthodox version. Authors of technically good unconventional papers may therefore be at a disadvantage in the peer-review process. A clearly written report. However, the low response rate (35.4% of the total) was probably affected by the impossibility of sending out a reminder. This would have jeopardised confidentiality and increased their ethical dilemma. (they did not get the report authorised by the ethics board).

Rochon 2002

Methods	Cohort study comparing review articles in published peer-reviewed versus throwaway journals focusing on diagnosis or treatment of medical conditions. Scales were used to assess outcomes
Data	Three hundred and ninety-four review articles published in 1998 either in the top five leading peer-reviewed journals (Annals of Internal Medicine, BMJ, JAMA, The Lancet and New England Journal of Medicine) or in highest circulation throwaway journals (Consultant, Hospital Practice, Patient Care and Postgraduate Medicine). Articles published in peer-reviewed journals were classified as systematic and non-systematic reviews, whereas those in throwaway journals were classified all as non-systematic reviews
Comparisons	Quality scales were used to assess methodological quality and readability of articles. In addition, six recently qualified physicians (Review Article Study Group) rated clinical relevance of articles
Outcomes	Barnes and Bero tool (12 point questionnaire evaluating, for example, inclusion criteria used on a 3-point scale: 0-no; 1-partial; 2-yes) plus reference count of article Presentation (font size, use of colour, number of tables, photographs and figures) Readability: Flesch index score and Gunning FOG index score The Review Article Study Group answered two statements on a five point scale (0=strongly disagree; 1=strongly agree): 1. This article may provide useful information for my practice 2. I would consider reading this article and for the 30 cardiology articles and graded the general content question using the scale: -The article addressed an important issue -The topic is of interest to me -The topic is relevant to my practice -The article provides practice strategies for physicians such as myself -I will use the information to help care for patients Of the 394 articles in the sample, 16 (4.1 %) were peer-reviewed systematic reviews, 135 (34.3%) were peer-reviewed non-systematic reviews, and 243 (61.7%) were non-systematic reviews published in throwaway journals. The mean (SD) quality scores were highest for peer-reviewed articles (0.94 [0.09] for systematic reviews and 0.30 [0.19] for non-systematic reviews) compared with throwaway journal articles (0.23 [0.03], F2,391 = 280.8, P<.001). Throwaway journal articles used more tables (P=.02), figures (P=.01), photographs (P<.001), colour (P<.001), and larger font sizes (P<.001) compared with peer-reviewed articles. Readability scores were more often in the college or higher rang for peer-reviewed journals compared with the throwaway journal articles (104 [77.0%] versus 156 [64.2%]; P=.01). Peer-reviewed article titles were judged less relevant to clinical practice than throwaway journal article titles(P<.001)

Rochon 2002 (*Continued*)

Notes	The authors conclude that although lower in methodological and reporting quality, review articles published in throwaway journals have characteristic that appeal to physician readers. This is an interesting study. Possible biases include: 1. The scoring tables had a bias towards systematic reviews 2. The reviewers were not truly representative of the population as they were new graduates 3. There may be a heavy selection bias as the top five peer reviewed journals and four highest circulation throwaway journals have entirely different purposes. 4. Not comparing like with like and using an instrument (The Barnes and Bero scale) in a biased way, makes the results difficult to interpret

Schroter 2004

Methods	Single-blind randomised trial assessing the effects of training on the quality of peer review. Three arms randomising consenting reviewers (no better defined)
Data	Six hundred and nine reviewers at BMJ. Arm one - control group. Arm two - self taught group who received a training package and CD Rom, based on the material used for group three, covering what the BMJ editor requires from reviewers and techniques of critical appraisal from randomised controlled trials. Group three - face to face group, received a full days training and the CD Rom. Drop out rate was, arm one; 46 (23%), arm two; 92 (45%) and arm three; 53 (26%). No reasons are given
Comparisons	Reviewers were asked to rate three previously published papers, each describing an RCT of alternative generic ways of organising and managing clinical work. Names of the authors were removed and titles of manuscript and any reference to study location were changed. Fourteen deliberate errors were introduced as follows: - Major errors 1. Poor justification for conducting the study 2. Biased randomisation procedure 3. No sample size calculation reported 4. Unknown reliability and validity of the outcome measures 5. Failure to analyse the data on an intention-to-treat basis 6. Poor response rate 7. Unjustified conclusions 8. Discrepancy between data reported in the abstract and results 9. Inconsistent denominator Minor errors 10. No ethics committee approval 11. No explanations for ineligible or non-randomized cases 12. Inconsistency between data reported in main text and tables 13. Failure to spot word reversal in text leading to wrong interpretation of results 14. Hawthorne effect The first paper was reviewed and used as a baseline. After this, the training was given to group three and the study material mailed to group two. Two or three months later, the second paper was sent, followed approximately six month later by the third paper
Outcomes	Review quality - eight item quality instrument taken from van Rooyen et al 1999. Grading from 1 (not at all) to 5 (discussed extensively)as follows: - 1. Did the reviewer discuss the importance of the research question? 2. Did the reviewer discuss the originality of the paper? 3. Did the reviewer clearly identify the strengths and weaknesses of the method (study design, data collection and

Schroter 2004 (*Continued*)

	data analysis)? 4. Did the reviewer make specific useful comments on the writing, organisation, tables and figures of the manuscript? 5. Were the reviewer's comments constructive? 6. Did the reviewer supply appropriate evidence using examples from the paper to substantiate their comments? 7. Did the reviewer comment on the author's interpretation of the results? 8. How would you rate the quality of this review overall? -Number of deliberate errors as described in Interventions/Exposure -Time taken and recommendations on publication
Notes	The authors conclude that short training packages have only a slight impact on the quality of peer review. The value of longer interventions needs to be assessed. Reviewers in the self taught group scored higher in review quality after training than did the control group (score 2.85 versus 2.56; difference 0.29, 95% confidence interval 0.14 to 0.44; P = 0.001), but the difference was not of editorial significance and was not maintained in the long term. Both intervention groups identified significantly more major errors after training than did the control group (3.14 and 2.96 versus 2.13; P < 0.001), and this remained significant after the reviewers' performance at baseline assessment was taken into account. The evidence for benefit of training was no longer apparent on further testing six months after the interventions. Training had no impact on the time taken to review the papers but was associated with an increased likelihood of recommending rejection (92% and 84% versus 76%; P = 0.002). There are a number of points which may have introduced bias into this trial. 1. The trial has a high drop out rate. Arm one (control group) 77%, arm two (self taught) 55% and arm three (training) 74%. We are not told why. However, withdrawals may be linked to the unpaid status of reviewers, and their awareness that they were part of a trial. 2. No description of randomisation. 3. All people involved were employed by BMJ, peer review and assessment was carried out externally, but by whom we do not know. 4. The trial used consenting reviewers.

Strayhorn 1993

Methods	Before and after study
Data	296 pairs of reviewers (before) and 272 pairs of reviewers (after) for the Journal of the American Academy of Child and Adolescent Psychiatry
Comparisons	Introduction of new rating scales for use by reviewers, along with training manuals in using the scales and research design
Outcomes	Degree of inter-rater agreement on quality of paper, acceptance recommendation
Notes	Inter-rater reliability increased after the new scales were introduced.

van Rooyen 1998

Methods	Randomised controlled trial, allocation by independent researcher
Data	527 manuscripts submitted to the BMJ, each reviewed by a pair of reviewers
Comparisons	1) Reviewers masked to co-reviewers' identities (n=149 manuscripts) 2) Identities of coreviewers not masked (n=150 manuscripts) 3) Reviewers not informed of study (n=158 manuscripts) 4) Preference arm if reviewers refused unmasking, i.e. remained masked (n=10 manuscripts)
Outcomes	Review quality on 7 item scale, scored by an editor Time to review manuscript
Notes	Knowledge of one or both reviewers' or authors' identity made no difference to the quality of reviews, recommendations or time taken in reviewing.

van Rooyen 1999

Methods	Randomised controlled trial, allocation by researcher. Paired design, i.e. randomisation within pair of reviewers for a given manuscript
Data	113 pairs of reviewers for manuscripts submitted to the BMJ
Comparisons	1) Reviewer asked that her/his identity be revealed to author 2) Reviewer's identity to remain concealed
Outcomes	1. Quality of review rated by editor and author, using 7-item scale 2. Time to review article 3. Recommendation to publish
Notes	Asking reviewers' consent to identification did not affect quality of reviews, time taken to review or recommendations to publish, but significant number of reviewers are likely to refuse to give their opinions.

Walsh 2000

Methods	Randomised controlled trial, allocation concealment unclear. Unbalanced groups. Outcome assessors blinded.
Data	408 reviews of consecutive submissions to the journal
Comparisons	1) Author's and reviewer's identities open. 2) Author's and reviewer's identity concealed.
Outcomes	Review quality using 7-item scale previously used elsewhere.
Notes	An open system is feasible and that signed reviews are possibly of better quality although they take longer to complete.

Weller 1996

Methods	Cohort study with two arms.
Data	Group one articles were taken from 17 'large prestigious medical journals'. Group two consisted of 742 'small speciality journals'. All of the journals were published in the USA and indexed on MEDLINE. The inclusion criteria for the two groups was different, why this is so is not explained. Group one criteria are is as follows: 1. In MEDLINE 2. Cited more than 5,000 per year 3. Circulation of more than 10,000 in the USA 4. Included on the following three recommended journals a. Brandon-Hill barded Index Medicus Cambrian College of Physicians, A Library for Internist Group two were published in the USA, indexed on MEDLINE, but did not meet any of the other criteria
Comparisons	Articles from both groups were randomly selected from MEDLINE in 1992. The random extraction (however) for both groups was different. Group one used a random number table. Group two had too many articles indexed from the 742 journals, therefore the authors were selected by paired sets of random numbers. The authors were then mailed a survey (questionnaire)
Outcomes	1. Reviewer understood manuscript 2. Review had constructive suggestions 3. Review improved content 4. Review improved organisation 5. Review clarified conclusions 6. Review changed conclusions 7. Review improved statistics 8. Reviewers had conflicting advise 9. Reviewer gave biased review Outcome analysis was stratified into 'all manuscripts' and 'previously rejected manuscripts'
Notes	The author concludes that many authors made very positive comments about the review process, stating, in effect, that editorial peer review, while imperfect, is the best process available. The study has many weaknesses. The inclusion criteria for the two groups was different, the report only examined indexed articles published in and originating from the US, no data was gathered on rejected manuscripts that were never published or published in a non-indexed journal, or were published in a journal indexed by a service other than MEDLINE and authors opinions were sought only after they had published successfully. This may have biased views in favour of peer review.

Characteristics of excluded studies *[ordered by study ID]*

Abby 1994	Non-comparative study
Bacchetti 2002	Non-comparative study
Baxt 1998	A study of association of reviewer characteristics and performance

(*Continued*)

Blank 1991	This study was of submisions to an economic journal
Cho 1998	The study was a descriptive survey looking for associations between success of reviewer masking, journals' policy on masking and other characteristics.
Cicchetti 1992	No original data
Cicchetti 1998	No original data
Cleary 1988	Non-comparative study
Coronel 1999	Non-comparative study
Cox 1993	No original data
Cullen 1992	This study compares characteristics of peer reviewers
Dixon 1983	Non-comparative study
Feurer 1994	This pilot tests the validity of an instrument for assessing quality.
Goldbeck-Wood 1999	No original data
Hatch 1998	Non-comparative study
Hemlin 1999	Narrative review and non-comparative study
Jadad 1998	Not about the effects of peer review
Katz 2002	Study testing ease of identifying blinded authors
Kumar 1999	Non-comparative study
Laband 1994	This study was of submissions to economic journals.
Lee 2002	Study assessing association between journal quality indicators and quality of published studies
Mahoney 1977	Not about the effects of peer review
Morrow 1992	Non-comparative study
Nylenna 1994	Not about the effects of peer review
Opthof 1999	Non-comparative study
Oxman 1991	Instrument testing

(*Continued*)

Peters 1982	Non-comparative study
Pitkin 2002a	Study looking at prompting reviewers to return reviews to journal
Presser 1980	Non-comparative study
Purcell 1998	Non comparative study
Rosenblatt 1980	Non-comparative study
Schriger 2002	Non-comparative study
van Rooyen 1999b	Development of outcome measure

Characteristics of ongoing studies *[ordered by study ID]*

Delamothe/BMJ

Trial name or title	
Methods	
Data	
Comparisons	Comparative study to see whether telling reviewers that their signed reviews might be posted on the World Wide Web affected the quality of their reviews
Outcomes	
Starting date	
Contact information	Dr A Delamothe BMJ Editorial BMA House Tavistock Square London WC1H 9JR Email: TDelamothe@bmj.com
Notes	

Lee/Bero 2004

Trial name or title	A Qualitative Study of Editorial Decision Making
Methods	
Data	
Comparisons	
Outcomes	
Starting date	
Contact information	Prof Lisa A. Bero. Grant # RO1NS044500 Funding Agency: National Institute of Neurological Disorders and Stroke (through Office of Research Integrity) Project period: 9/15/2002-8/31/2004
Notes	

Schroter/BMJ 2004

Trial name or title	
Methods	
Data	
Comparisons	Comparative study at 10 BMJ Journals to compare author and editor suggested reviewers. Comparing them in terms of the quality of review, timeliness and recommendation to the editor. Will not have any result until the summer
Outcomes	
Starting date	
Contact information	Dr Sara Schroter Senior Researcher BMJ Editorial BMA House Tavistock Square London WC1H 9JR Email: SSchroter@bmj.com
Notes	

DATA AND ANALYSES

This review has no analyses.

APPENDICES

Appendix 1. Databases searched

The following databases were searched for the 2003 review:

AUSTRALASIAN MEDICAL INDEX 1980 to 2000 (searched 11 February 2000)
Text words: peer-review and peer review
BEST EVIDENCE 4, 2000 (searched July 2000)
Text words: peer-review and peer review
BIOETHICSLINE 1973 to 2000 (searched 7 February 2000)
Text words: peer-review and peer review
CINAHL 1997 to 1999 (searched July 2000)
Text words: peer-review and peer review
COCHRANE LIBRARY Issue 2, 2000 - searched 11/4/00 (including Cochrane Controlled Trials Register, Cochrane Database of Systematic Reviews, Cochrane Methodology Register, Database of Abstracts of Reviews of Effectiveness) Text words: peer-review and peer review
CURRENT CONTENTS 1999 to 2000 (searched February 2000)
Text words: peer-review and peer review
DISSERTATION ABSTRACTS 1861 to 2000 (searched February 2000)
Text words: peer-review and peer review
EMBASE
Text words: peer-review and peer review (January 1989 to January 2000 - searched July 2000)
HEALTHSTAR 1975 to 1999 (searched December 1999)
Text words: peer-review and peer review
MEDLINE 1966 to February 2000
peer-review and peer review (MeSH exploded and text terms)
(search carried out by Anne Lusher, UKCC for the Cochrane Methodology Register, February 2000)
NATIONAL RESEARCH REGISTER (searched March 2000)
Text words: peer-review and peer review
PsycLIT 1887 to 2000 (searched February 2000)
Text words: peer-review and peer review
PUBMED (1966 to 2000)
Text words: peer-review and peer review
1998-2000 (searched 2nd February 2000)
1997 (searched 28th March 2000)
PUBSCIENCE - journal coverage varies but is generally 1998 to date (searched February 2000)
Text words: peer-review and peer review
SIGLE 1980 to 6/1999 (searched 3rd April 2000)
Text words: peer-review and peer review

The following databases were searched for the 2005 update:

CINAHL, Ovid, Searched June 11, 2004
Search: exp "Peer Review"/ or peer review$.tw.
Entry Week from January 1999 to June Week 1 2004
CMR (Cochrane Methodology Register)
Searched May 28, 2004
Search: CMR code like *peer review*
DISSERTATION ABSTRACTS, Ovid
Searched June 13, 2004
Search: peer review$.tw.
Entry Month from January 2000 to June 2004
EMBASE, Ovid
Searched June 4, 2004
Search: Peer Review/ or peer review$.tw.
Entry Week from January 1993 to January 1998, from January 2000 to June Week 22 2004
Evidence Based Medicine Reviews (*formerly Best Evidence*): ACP Journal Club (ACP), Ovid
Issue 1991 to March/April 2004
Searched June 18, 2004
Search: peer review$.mp.
MEDLINE, Ovid
Searched June 3, 2004
Search: exp Peer Review/ or peer review$.tw.
Entry Date from January 2000 to May Week 4 2004
PsycINFO (*CD-rom version is PsycLit*)
Searched June 4, 2004
Search: Peer Evaluation/or peer evaluat$.tw. or peer review$.tw.
From January 2000 to May Week 5 2004
PUBMED Searched June 2, 2004
Search: peer review
Entry Date from January 1, 2000 to June 1, 2004
Note: BIOETHICLINE is now part of MEDLINE and HEALTHSTAR is now part of PubMed.

Appendix 2. Handsearches

The following were handsearched for the 2003 review:

- The reference list of relevant articles, posted requests on the EASE and WAME web sites, searched the Locknet site (last searched September 2000).
- Godlee F, Jefferson T (eds). Peer review in the health sciences. London: BMJ Books, 1999
- Ethics and policy in scientific publication. Chicago,Ill: Council of Biology Editors, 1993
- Behavioural and Information Technology (1997 to 2000)
- Cognitive Psychology (1993 to 2000)
- EASE Bulletin (35 September 1988, 47 September 1992, 48 January 1993, 57 February 1996, 58 June 1996, 59 October 1996, 23(2) July 1997, 25(1) February 1999, 26(1) February 2000, 26(2) July 2000, 26(3) October 2000
- JAMA special issues on peer review (1990, 1994, 1998)
- Journal of Information Science (1985 to 2000)
- Perceptual and Motor Skills (1993 to 2000)

The following were handsearched for the 2005 update:

- Godlee and Jefferson (Godlee 2003)
- The 2002 JAMA special issue on peer review.

WHAT'S NEW

Last assessed as up-to-date: 19 February 2007.

27 December 2007	Amended	Converted to new review format.

HISTORY

Protocol first published: Issue 2, 2001

Review first published: Issue 1, 2003

20 February 2007	New citation required and conclusions have changed	Substantive amendment

CONTRIBUTIONS OF AUTHORS

For the 2005 update: Tom Jefferson, Melanie Rudin and Suzanne Brodney Folse selected studies, extracted data and wrote the updated draft of the review. All reviewers contributed to the editing of the draft review.

For the 2003 review: Elizabeth Wager helped obtain some of the studies, Tom Jefferson and Phil Alderson selected studies, extracted data and wrote the first draft of the review. All reviewers contributed to the editing of the draft protocols and review.

DECLARATIONS OF INTEREST

Tom Jefferson co-edited the book "Peer review in health sciences". All authors are active peer reviewers and have published articles in peer-reviewed journals.

SOURCES OF SUPPORT

Internal sources

- NHS Research and Development Programme, UK.
- Commonwealth Dept of Health and Aged Care, Australia.

External sources

- No sources of support supplied

INDEX TERMS

Medical Subject Headings (MeSH)

Biomedical Research [*standards]; Peer Review, Research [*standards]