What errors do peer reviewers detect, and does training improve their ability to detect them?

Sara Schroter[1] • Nick Black[2] • Stephen Evans[2] • Fiona Godlee[1] • Lyda Osorio[2] • Richard Smith[1]

[1] BMJ, BMA House, Tavistock Square, London WC1H 9JR, UK

[2] London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

Correspondence to: Dr Sara Schroter. E-mail: sschroter@bmj.com

DECLARATIONS

Competing interests
FG is the editor of the BMJ, SS is a senior researcher for the BMJ, RS is the former editor of the BMJ and NB, SE, and SS review for the BMJ

Funding
This study was funded by the NHS London Regional Office Research & Development Directorate. The views and opinions expressed in this paper do not necessarily reflect those of NHSE (LRO) or the Department of Health

Ethical approval
The ethics committee of the London School of Hygiene and Tropical Medicine approved the study

Guarantor
SS

Abstract

Objective To analyse data from a trial and report the frequencies with which major and minor errors are detected at a general medical journal, the types of errors missed and the impact of training on error detection.

Design 607 peer reviewers at the BMJ were randomized to two intervention groups receiving different types of training (face-to-face training or a self-taught package) and a control group. Each reviewer was sent the same three test papers over the study period, each of which had nine major and five minor methodological errors inserted.

Setting BMJ peer reviewers.

Main outcome measures The quality of review, assessed using a validated instrument, and the number and type of errors detected before and after training.

Results The number of major errors detected varied over the three papers. The interventions had small effects. At baseline (Paper 1) reviewers found an average of 2.58 of the nine major errors, with no notable difference between the groups. The mean number of errors reported was similar for the second and third papers, 2.71 and 3.0, respectively. Biased randomization was the error detected most frequently in all three papers, with over 60% of reviewers rejecting the papers identifying this error. Reviewers who did not reject the papers found fewer errors and the proportion finding biased randomization was less than 40% for each paper.

Conclusions Editors should not assume that reviewers will detect most major errors, particularly those concerned with the context of study. Short training packages have only a slight impact on improving error detection.

Introduction

Peer reviewers are responsible for improving the quality of manuscripts to be published and 'should weed out serious methodological errors'.[1] Despite the use of peer review, errors, inconsistencies and methodological weaknesses are commonly found in published medical research[2–4] and peer review has been criticized as being an ineffective and expensive procedure.[5,6]

Three studies have reported on the rate of detection of errors by reviewers. The first used two fictitious reports submitted to all reviewers of a general medical journal and found that reviewers missed many of the deliberate errors in the manuscripts.[7] A second study introduced 10 major and 13 minor errors in a manuscript and distributed it to 262 reviewers of the *Annals of Emergency Medicine*.[8] Reviewers failed to identify two thirds of the major errors and about 7% recommended

Contributorship
NB and RS initiated the study; SS, NB, RS, and FG designed it; NB created the test papers; SS conducted the study; SS and SE did the data analysis; and SS, NB and SE, interpreted the results. All authors assisted in writing the paper

Acknowledgements
We thank all the reviewers who participated, the editors who assisted with the face-to-face training and the Critical Appraisal Skills Programme (CASP) team, the authors of the original manuscripts for allowing us to use them, and Joe Kim for his help with the graphics

acceptance. The third study reported that, on average, reviewers detected only two out of eight areas of weakness in a modified paper.[9]

We conducted a single blind randomized controlled trial (RCT) on the effect of training on the performance of peer reviewers of a general medical journal, the BMJ.[10] Reviewers were randomized to one of three groups (control, face-to face training and self-taught) and invited to review three manuscripts during the study period. The training package focused on what editors want from reviewers and how to critically appraise RCTs. For all groups, we inserted nine major and five minor methodological errors into each manuscript before sending the papers out for review. The authors of the original manuscripts gave their consent for the insertion of errors and their use in the trial. The quality of review, assessed using a validated instrument,[11] was the primary outcome measure and the number of major errors detected was secondary. The objective of this paper is to report the frequency with which the nine major and five minor errors were detected and the impact that training had on each of the 14 errors studied. As the methods of the trial and primary results have previously been reported,[10] they are described only briefly in this paper. In this paper the data from the RCT is used as observational data.

Methods

The trial was approved by the London School of Hygiene & Tropical Medicine ethics committee. On invitation to take part in the study, participants were asked to give written consent to review three papers for the study and to agree to attend a full day of training if selected to do so.

Participants

We performed a power calculation (reported in the previous paper[10]) based on our primary outcome measure, review quality, and estimated that 190 reviewers were needed in each group. All *BMJ* reviewers (n=1256) resident in the UK who had reviewed at least one paper between January 1999 and February 2001 were invited to take part. No exclusion criteria were applied, other than non-residence in the UK.

Consenting reviewers were randomized into three groups – two intervention groups and a control group – using a stratified permuted blocks randomization method. Previous studies identified several factors that affect review quality and so the stratification was based on age, current investigators in medical research projects, postgraduate training in epidemiology, postgraduate training in statistics, and membership of an editorial board of a scientific or medical journal.[12,13]

Assessments and procedures

Three previously published papers each describing an RCT on a general medical subject were selected for use in this study, and the authors and journal editors were contacted for permission to use them. The papers selected described studies evaluating the effects of discharge summaries, personalized computer-generated health records, and patients holding their own records (i.e. they were general articles). Papers describing RCTs were specifically chosen as they usually provide more structure for review than other research designs. The names of the original authors were removed and the titles of the manuscripts and references to study locations were changed.

Deliberate errors were introduced into the first test paper. To determine the level of difficulty of the errors inserted, we piloted the paper on a sample of three editors and two epidemiology postgraduate students. The paper was subsequently modified to exclude the errors not detected by any of the sample reviewers, and the remaining errors were classified individually as major (nine) or minor (five) by members of the research team (NB, SS, RS, FG). Where two people indicated major and two minor, the difficulty of the error was discussed as a group until a consensus was reached. Similar errors, in terms of type and level of difficulty, were then inserted in the other two test papers.

The nine major errors focused on methodological weaknesses, inaccurate reporting of data and unjustified conclusions, while the five minor errors focused on omissions and inaccurate reporting of data (Table 1). As a result of severe editing of the manuscripts to insert the deliberate errors, there were some additional unintended inconsistencies in the papers. Whilst these were reported by many reviewers, we considered only the identification of the 14 deliberate errors. One major error (unknown reliability and validity of outcome measure) was introduced into each paper to act as a control – that is, no training was provided and we did not expect to see improvement in the detection of this error.

All consenting reviewers were asked to review the first paper. After this baseline assessment, one intervention group received a full day of face-to-face training and the other intervention

Table 1
Descriptions of 14 deliberate errors

Major errors	
Poor justification for conducting the study	A statement that the authors failed to put their study into context by providing relevant information from previous studies and justifying why there was a need for a new study
Biased randomization procedure	A statement that the randomization method (e.g. randomization by family name or day of the week) was inadequate and could result in systematic bias
No sample size calculation reported	A statement that a sample size calculation had not been reported
Unknown reliability and validity of the outcome measures	A statement that there was insufficient information about the measurement properties (i.e. reliability and validity) of the outcome measures used e.g. quality of life questionnaires A simple statement that the instruments were not referenced was not counted as this did not comment on the unknown measurement properties of the instruments
Failure to analyse the data on an intention-to-treat basis	A statement that the authors were incorrect in their assumptions that intention-to-treat analysis was inappropriate
Poor response rate	A statement that the response rate was low
Unjustified conclusions	A statement that the authors made inappropriate conclusions beyond their findings by inappropriately generalizing their results to other areas of care which were not under study
Discrepancy between data reported in the abstract and results	A statement that there were discrepancies between the data reported in the main text of the paper and that reported in the abstract
Inconsistent denominator	A statement that the number of patients / cases reported in the papers are inconsistent and difficult to follow
Minor errors	
No ethics committee approval	A statement that there was no indication that the study had been approved by an ethics committee
No explanations for ineligible or non-randomized cases	A statement that the flow of participants through each stage of the study was not clear and that the authors failed to provide explanations for ineligible or non-randomized cases
Inconsistency between data reported in main text and tables	A statement that there were discrepancies between figures reported in the main text and those reported in the tables
Failure to spot word reversal in text leading to wrong interpretation of results	A statement that the words in the text describing the findings reported in a table were inverted leading to the wrong interpretation
Hawthorne effect	A statement that the authors failed to report a possible Hawthorne effect i.e. that participants were aware that they were in a study and may have behaved differently from usual.

group was mailed a self-taught training package. Details of the training are described in a previous publication.[10] Reviewers who completed the first review were sent the second paper to review two to three months after the intervention; the third paper was sent approximately six months later if they completed the second review.

Reviewers were sent the manuscripts in a style similar to the standard *BMJ* review process, but were told these papers were part of the study and were not paid for the reviews. Reviewers were asked to review the papers within three weeks and were sent the standard *BMJ* guidance for reviewers (see bmj.com for details) and a prepaid return envelope. Reminders were sent to increase response rates.

Outcome measure: number of deliberate errors detected

The number of major and minor errors reported in each review was assessed independently by two researchers (SS and LO) blind to the identity and study group of the reviewer. A strict marking scheme was used; an identification of error was only counted if there was a clear statement describing the error and explaining the problem, so that the review would be of practical use to the authors and the

Table 2
Characteristics of reviewers completing each review

Characteristic	*Paper1 (n=522)*	*Paper 2 (n=440)*	*Paper 3 (n=418)*
Loss from paper 1 (%)		16	20
Mean (SD) age (years)	49.5 (8.3)	49.4 (8.4)	49.3 (8.4)
Age range (years)	27–80	30–80	30–80
Number (%) male	379 (73)	320 (73)	304 (73)
Current investigator in medical research project	432 (83)	364 (83)	346 (83)
Postgraduate training in epidemiology	193 (37)	166 (38)	156 (37)
Postgraduate training in statistics	280 (54)	241 (55)	231 (55)
Editorial board member	253 (49)	208 (47)	196 (47)

editor. One point was allocated for each error if the reviewer clearly identified the error and half a point was given if there was some evidence that the error had been identified. If the reviewer returned the manuscript, it was also checked for comments indicating the identification of an error. A point was only awarded if the error had been clearly identified – the underlining of text on the manuscript alone was not considered sufficient.

Statistical analysis

The intra-class correlation coefficient was used to assess the level of agreement between raters for each error and for the total error score. Generally, values >0.70 are considered acceptable for the intra-class correlation coefficient.[14]

To calculate the percentage of reviewers reporting each error, half points were rounded to full points for each rater and an average of the two raters' scores was calculated. Scores were then rounded again (0=0, 0.5=1, 1=1) so that if at least one rater indicated that the reviewer had identified the error, the reviewer was given a mark.

Results

Reviewer characteristics

Five hundred and twenty two (86%) of the 607 reviewers randomized completed a review of the first paper, 440 of 522 (84%) completed the second, and 418 of 440 (95%) completed the third. The self-reported characteristics of the reviewers in terms of age, sex, postgraduate experience in statistics and/or epidemiology, current research investigator and member of a journal editorial board are shown in Table 2. Characteristics were similar for reviewers completing each of the papers.

Reliability of ratings

A good level of agreement was reached between the two independent raters for the assessment of the reporting of individual errors (Table 3). The intra-class correlation coefficients were >0.70 for each error when averaged across the three papers. An intra-class correlation coefficient >0.90 for the nine-item total major error score reflects excellent agreement.

Detection of errors

For all groups combined (control, self-taught, face-to-face) the average number of the nine major errors reported was 2.58 (standard deviation [SD] 1.9) in Paper 1, 2.71 (SD 1.6) in Paper 2 and 3.05 (SD 1.8) in Paper 3. The average number of the five minor errors reported was 0.91 (SD 0.8) in Paper 1, 0.85 (SD 0.8) in Paper 2 and 1.09 (SD 0.8) in Paper 3. Table 4 shows the data for the combined group and for each study group.

Table 5 shows the proportion of reviewers reporting each error for each paper by study group. The detection of errors was relatively consistent across papers. Overall, the errors most frequently reported were biased randomization procedure and no explanations for ineligible or non-randomized cases. The least often reported errors were word reversal, no mention of a Hawthorne effect (a temporary change in behaviour – typically an improved response – in response to altered environmental conditions), and inconsistency between text and tables. There was consistency between the three groups in the errors detected (i.e. the interventions had little effect on the detection of errors).

Figures 1a and 1b show the proportion of reviewers reporting each error labelled with a number from 1 to 14 (the order based on frequency of reporting shown separately for major and minor errors) for those who recommended rejection of Paper 1 and those who did not recommend rejection, respectively. Figures 1c and 1d show these proportions for Paper 2, and Figures 1e and 1f for Paper 3. The proportion of reviewers

Table 3
Agreement for assessment of error reporting in all three papers

Error	*Average intra-class correlation coefficient for the three reviews*
Major	
Poor justification for study	0.74
Biased randomization procedure	0.86
No sample size calculation	0.98
Unknown reliability & validity of outcome measure	0.86
Failure to analyse the data on an Intention-to-treat basis	0.90
Poor response rate	0.85
Unjustified conclusions	0.73
Discrepancy between abstract & results	0.81
Inconsistent denominator	0.84
Total Major error score	*0.91*
Minor	
No ethics approval	0.98
No explanations for ineligible or non-randomized cases	0.82
Inconsistency between text & tables	0.92
Word reversal	0.81
No mention of Hawthorne effect	0.75

reporting each error in each paper was higher for reviewers recommending rejection than for those who did not. For each of the three papers, over 60% of the reviewers who recommended rejection reported that the randomization procedure was biased [error 1]. Other errors frequently reported by those rejecting the papers included inadequate reporting of ineligible or non-randomized cases [error 10] (58% averaged across the three papers), a poor response rate [error 4] (48%) and unjustified conclusions [error 3] (46%). Whilst these same errors were those most frequently reported by reviewers not recommending rejection, the proportions were considerably lower (34, 32, 29 and 35%, respectively).

Table 4
Mean (SD) errors identified by group for each paper

	Major errors	*Minor errors*
Paper 1		
Control group (n=173)	2.38 (2.0)	0.99 (0.9)
Self-taught group (n= 166)	2.68 (1.7)	0.79 (0.8)
Face-to-face group (n= 183)	2.68 (1.8)	0.94 (0.8)
All groups combined (n= 522)	2.58 (1.9)	0.91 (0.8)
Paper 2		
Control group (n= 162)	2.13 (1.6)	0.71 (0.8)
Self-taught group (n= 120)	3.14 (1.4)	1.05 (0.9)
Face-to-face group (n= 158)	2.96 (1.7)	0.84 (0.8)
All groups combined (n= 440)	2.71 (1.6)	0.85 (0.8)
Paper 3		
Control group (n= 156)	2.71 (1.8)	0.96 (0.9)
Self-taught group (n= 111)	3.37 (1.7)	1.21 (0.8)
Face-to-face group (n= 151)	3.18 (1.8)	1.12 (0.8)
All groups combined (n= 418)	3.05 (1.8)	1.09 (0.8)

Discussion

Principal findings

On average, reviewers reported only three out of nine major errors in their reviews, with almost a quarter of the reviewers reporting one or less. This is similar to two previously reported studies. Baxt *et al*.[8] found reviewers failed to identify two thirds of the major errors in a manuscript and Godlee *et al*.[9] found the mean number of weaknesses in design, analysis or interpretation commented on was only two out of eight, with only 10% of reviewers identifying four or more areas of weakness and 16% failing to identify any.

The poor detection rate we observed was not due to over-demanding expectations of reviewers. For example, we classified 'inconsistencies between text and tables' as only a minor error, despite this being an important issue that needs to be picked up somewhere in the review process as it has an impact on the readability of the manuscript and the understanding of the results.

The detection rate varied between the nine major errors. Those most likely to be detected (>50%) related to the sampling and randomization techniques. In contrast, those least likely to be detected (<30%) related to the analysis of data and inconsistencies in the reporting of results. Baxt *et al*.[8] reported similar findings: 68% of the reviewers in

Table 5
Proportion of reviewers identifying each error by group for the three papers

	Paper 1			*Paper 2*			*Paper 3*		
	Control	*Self-taught*	*Face-to-face*	*Control*	*Self-taught*	*Face-to-face*	*Control*	*Self-taught*	*Face-to-face*
Major									
Poor justification for study	31	36	36	22	36	35	30	37	30
Biased randomization procedure	49	58	53	46	72	65	48	65	64
No sample size calculation	21	24	21	22	31	34	25	36	34
Unknown reliability & validity of outcome measure	13	19	21	15	28	22	45	60	47
Failure to analyse the data on an intention-to-treat basis	22	18	22	9	13	15	34	45	38
Poor response rate	34	36	37	43	53	51	46	43	47
Unjustified conclusions	43	40	41	36	48	45	44	45	47
Discrepancy between abstract & results	23	25	28	20	30	30	33	39	37
Inconsistent denominator	38	45	53	40	58	56	17	18	21
Minor									
No ethics approval	18	14	14	25	43	30	33	41	32
No explanations for ineligible or non-randomized cases	50	48	58	30	46	33	56	75	73
Inconsistency between text & tables	5	2	2	12	12	12	4	5	5
Word reversal	13	9	10	9	15	16	7	9	11
No mention of Hawthorne effect	21	12	19	2	2	2	3	1	2

their study did not realize that the conclusions of the work were not supported by the results. We found that whilst many reviewers acknowledged that the conclusions went beyond the results, about 40% failed to report that the authors had extrapolated their results to areas of care not studied.

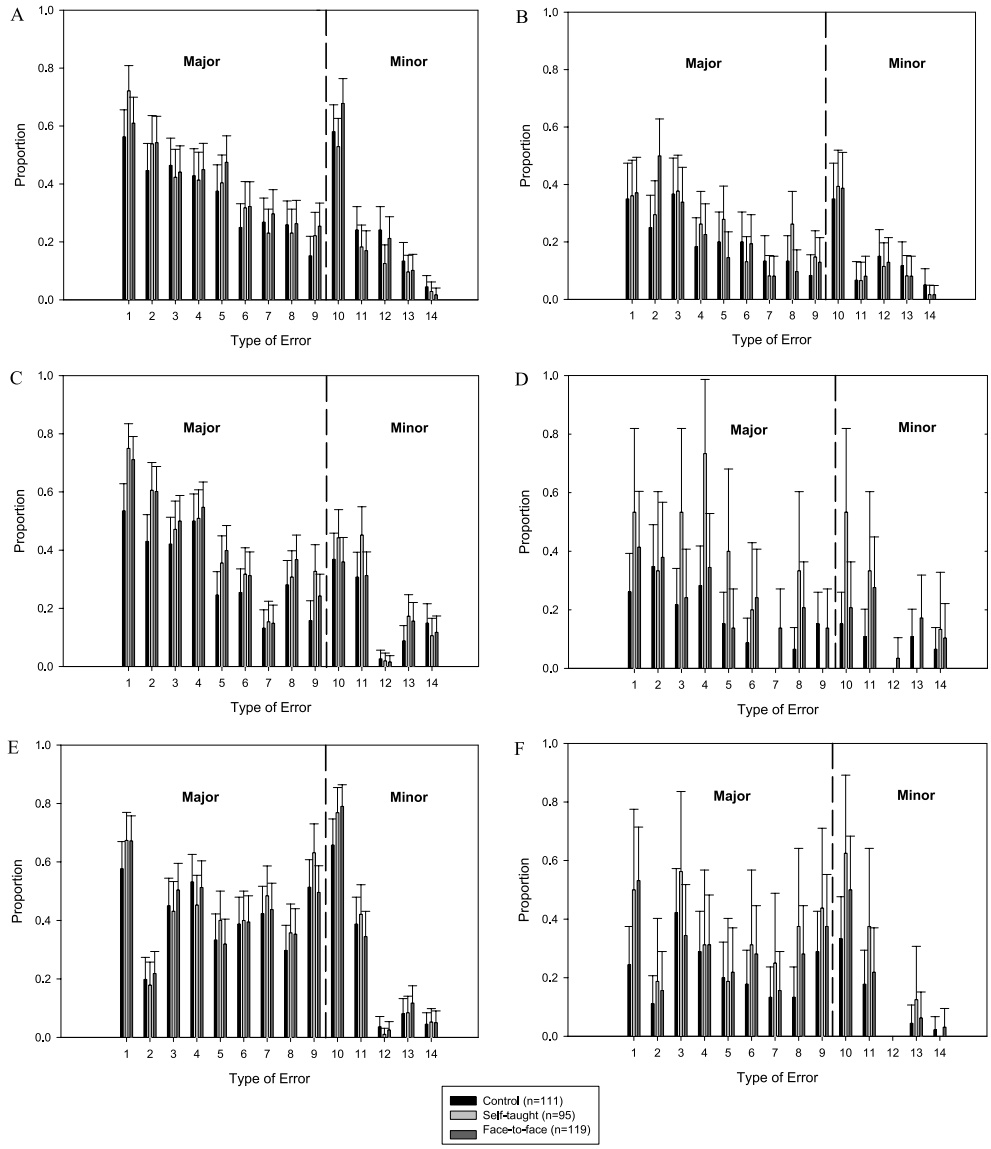
Training led to some improvements in error detection. Broadly speaking, the errors that were detected more frequently after training were those to do with technical aspects such as the response rate, randomization procedure and sample size calculation. Areas in which little or no improvement occurred were to do with putting the study in context, both in terms of the pre-existing literature and in terms of the implications of the findings for policy or practice. The dramatic improvement in detection of the control error is partly explained by the low level of detection in the pre-intervention paper (18%). However, it also suggests that the Hawthorne effect may have contributed to improvements in detection of several errors.

Strengths and weaknesses of study

There are several limitations to our study to consider when drawing any lessons or implications:

- First, some reviewers felt the papers they were assessing were so methodologically flawed that they recommended rejection and did not complete a full review; had they continued they may have reported more of the errors.
- Second, reviewers knew they were taking part in a study and this may have affected their behaviour. In passing, it is interesting to note whilst many reported that the experience felt artificial and that they were behaving differently than usual, very few reported the fact that the papers they reviewed had failed to address this same issue (i.e. the Hawthorne effect).
- Third, some reviewers may not have been very familiar with RCT methodology. Whilst many such people may have declined to participate in the study (in the knowledge we were focusing on critical appraisal), some may have participated with the deliberate intention of gaining knowledge and experience in this area.
- Fourth, we confined our study to the review of RCTs, so the findings may not be generalizable. We cannot extrapolate our findings to the performance of reviewers tackling other types of study design. Some might argue

Figure 1
Proportion of reviewers identifying each error for those who did and did not recommend rejection of each paper



A: Reviewers rejecting Paper 1 (n=335); B: Reviewers not rejecting Paper 1 (n=156); C: Reviewers rejecting Paper 2 (n=346); D: Reviewers not rejecting Paper 2 (n=71); E: Reviewers rejecting Paper 3 (n=325); F: Reviewers not rejecting Paper 3 (n=74)
Errors: 1, Biased randomization procedure; 2, Inconsistent denominator; 3, Unjustified conclusions; 4, Poor response rate; 5, Poor justification; 6, Discrepancy between abstract & results; 7, ITT would be appropriate; 8, No sample size calculation; 9, Unknown reliability & validity; 10, No explanation of drop outs; 11, No ethics approval; 12, Hawthorne effect; 13, Word reversal; 14, Inconsistency between text & tables

that it is easier to detect flaws in RCTs than other study designs due to their structure and the availability of CONSORT criteria.[17]

- Fifth, we restricted the study to UK reviewers and so we should be cautious in extrapolating the results to reviewers outside of the UK.
- Sixth, whilst we tried to standardize the errors inserted into the three papers in terms of the level of difficulty, some of the errors may have been harder or easier to detect in the different papers. While this can be taken into account in comparing the impact of training

interventions, it means that direct comparisons of error detection rates between the three papers are difficult to interpret. It is also possible that the errors we introduced to the papers may not be the most important errors for reviewers to detect, but the *BMJ* does expect its reviewers to be able to detect them.

Relationship to other studies

Two other studies looking at the effects of peer review have found limited improvements in manuscript quality.[15,16] A recent study comparing the quantity and quality of data tables and figures in reports of RCTs submitted to the *BMJ* and subsequently published in peer-reviewed journals found peer review to be limited in improving the presentation of data.[15] *BMJ* external peer reviewers seldom commented on the tables or figures and the numbers of tables and figures did not change markedly between submission and publication. Goodman *et al*.[16] found manuscript quality improved after peer review and editing at *Annals of Internal Medicine*, but that improvement was modest and there was still substantial room for improvement. Aspects that showed the most improvement were discussion of study limitations, acknowledgement and justifications of generalizations, appropriateness of the strength and tone of the conclusions, use of confidence intervals and description of the setting. However, due to the study design it was not possible to distinguish the effects of external peer review from internal editing.

Implications

The principal implication of our findings, when taken together with the previous studies cited above, is that journal editors should not assume that their reviewers will detect most major flaws in manuscripts. The study paints a rather bleak picture of the effectiveness of peer review. Improvements after training were minor despite using the types of papers easiest to review for errors, our reviewers being better trained and qualified than those at many smaller journals, and despite focusing on technical errors that are easier to detect than more fundamental errors involving flawed assumptions and theoretical models. Clearly, using more than one reviewer may increase the total numbers of errors detected, though some errors are likely to remain undetected. This may be of no immediate consequence if the major errors which have been detected lead to a decision to reject the manuscript. However, as a principal aim of peer review is to improve the quality of published papers, it seems that it is only partially successful. The improvements after training were trivial and were largest in the technical aspects of review, which could be identified by well-trained journal staff and probably don't require expert professional external reviewers. The shortcomings of peer reviewing that have been revealed in this and other studies cannot be easily resolved by short training interventions: the small effects observed were not worth the resources and time required for training. The question remains as to how best to improve the peer review process.

References

1 Altman DG. Poor-quality medical research. What can journals do? *JAMA* 2002;**287**:2765–7
2 Altman DG. Statistics in medical journals. *Stat Med* 1982;**1**:59–71
3 Andersen B. *Methodological errors in medical research.* Oxford: Blackwell; 1990
4 Altman DG. The scandal of poor medical research. *BMJ* 1994;**308**:283–4
5 Smith R. Peer review: Reform or revolution? *BMJ* 1997;**315**:759–60
6 Jefferson T, Alderson P, Wager E, Davidoff F. Effects of peer review: A systematic review. *JAMA* 2002;**287**:2784–6
7 Nylenna M, Riis P, Karlsson Y. Multiple blinded reviews of the same two manuscripts. *JAMA* 1994;**272**:149–51
8 Baxt WG, Waeckerle JF, Berlin JA, Callaham ML. Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Ann Emerg Med* 1998;**32**:310–7
9 Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomised controlled trial. *JAMA* 1998;**280**:237–40
10 Schroter S, Black N, Evans S, *et al.* Effects of training on the quality of peer review: A randomised controlled trial. *BMJ* 2004;**328**:657–8
11 van Rooyen S, Black N, Godlee F. Development of the Review Quality Instrument (RQI) for assessing peer reviews of manuscripts. *J Clin Epidemiol* 1999;**52**:625–9
12 Evans AT, McNutt RA, Fletcher SW, Fletcher RH. The characteristics of peer reviewers who produce good-quality reviews. *J Gen Intern Med* 1993;**8**:422–8
13 Black N, van Rooyen S, Godlee F, Smith R, Evans S. What makes a good reviewer and a good review in a general medical journal. *JAMA* 1998;**280**:231–3
14 Scientific Advisory Committee for the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: Attributes and review criteria. *Qual Life Res* 2002;**11**:193–205
15 Schriger DL, Sinha R, Schroter S, Liu PY, Altman DG. From submission to publication: a retrospective review of the tables and figures in a cohort of randomised controlled trials submitted to the British Medical Journal. *Ann Emerg Med* 2006;**48**:750–6
16 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
17 Moher D, Schulz KF, Altman DGfor the CONSORT Group. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Ann Intern Med* 2001;**134**:657–62