BMJ

RESEARCH

Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial

OPEN ACCESS

E Cobo *senior statistics editor and senior statistical lecturer*[1 2], J Cortés *statistical researcher*[2], J M Ribera *general secretary and chief of clinical haematology department*[1 3 4 5], F Cardellach *general secretary and professor of internal medicine*[1 6], A Selva-O'Callaghan *editorial committee member and senior lecturer in internal medicine*[1 3 7], B Kostov *statistical researcher* [8], L García *statistical researcher*[2], L Cirugeda *statistical researcher*[9], D G Altman *professor of statistics in medicine*[10], J A González *senior statistical lecturer*[2], J A Sànchez *senior statistical lecturer*[2], F Miras *statistical researcher*[2], A Urrutia *editorial committee member and senior lecturer in internal medicine*[1 3 4], V Fonollosa *editorial committee member and professor of internal medicine*[1 3 7], C Rey-Joly *current editor and professor of internal medicine*[1 3 4], M Vilardell *editor in chief and professor of internal medicine*[1 3 7]

[1]Medicina Clínica, Elsevier-Barcelona, Barcelona 08021, Spain; [2]Universitat Politècnica Catalunya, Barcelona; [3]Universitat Autònoma de Barcelona, Barcelona; [4]Hospital Germans Trias I Pujol, Badalona, Spain; [5]José Carreras Leukaemia Research Institute, Catalan Institute of Oncology, ICO, Badalona; [6]Universitat de Barcelona and Hospital Clínic, Barcelona; [7]Vall D'Hebron Hospital, Barcelona; [8]Primary Health Care Center Les Corts, GESCLINIC, Barcelona; [9]Centre for Research in Environmental Epidemiology, Barcelona; [10]Centre for Statistics in Medicine, University of Oxford, Oxford, UK

Abstract

Objective To investigate the effect of an additional review based on reporting guidelines such as STROBE and CONSORT on quality of manuscripts.

Design Masked randomised trial.

Population Original research manuscripts submitted to the *Medicina Clínica* journal from May 2008 to April 2009 and considered suitable for publication.

Intervention Control group: conventional peer reviews alone. Intervention group: conventional review plus an additional review looking for missing items from reporting guidelines.

Outcomes Manuscript quality, assessed with a 5 point Likert scale (primary: overall quality; secondary: average quality of specific items in paper). Main analysis compared groups as allocated, after adjustment for baseline factors (analysis of covariance); sensitivity analysis compared groups as reviewed. Adherence to reviewer suggestions assessed with Likert scale.

Results Of 126 consecutive papers receiving conventional review, 34 were not suitable for publication. The remaining 92 papers were allocated to receive conventional reviews alone (n=41) or additional reviews (n=51). Four papers assigned to the conventional review group deviated from protocol; they received an additional review based on reporting guidelines. We saw an improvement in manuscript quality in favour of the additional review group (comparison as allocated, 0.25, 95% confidence interval −0.05 to 0.54; as reviewed, 0.33, 0.03 to 0.63). More papers with additional reviews than with conventional reviews alone improved from baseline (22 (43%) *v* eight (20%), difference 23.6% (3.2% to 44.0%), number needed to treat 4.2 (from 2.3 to 31.2), relative risk 2.21 (1.10 to 4.44)). Authors in the additional review group adhered more to suggestions from conventional reviews than to those from additional reviews (average increase 0.43 Likert points (0.19 to 0.67)).

Correspondence to: E Cobo erik.cobo@upc.edu

Extra material supplied by the author (see http://www.bmj.com/content/343/bmj.d6783?tab=related#webextra)

Web appendix: Protocol trial design
Web table 1: Manuscript quality assessment instrument
Web table 2: Adherence recommendation assessment questionnaire

Conclusions Additional reviews based on reporting guidelines improve manuscript quality, although the observed effect was smaller than hypothesised and not definitively demonstrated. Authors adhere more to suggestions from conventional reviews than to those from additional reviews, showing difficulties in adhering to high methodological standards at the latest research phases. To boost paper quality and impact, authors should be aware of future requirements of reporting guidelines at the very beginning of their study.

Trial registration and protocol Although registries do not include trials of peer review, the protocol design was submitted to sponsored research projects (Instituto de Salud Carlos III, PI081903).

Introduction

The scientific value of biomedical journals relies on the peer review process and on editorial decisions, but the quality of these processes is far from guaranteed.[1-4] Processes, be they in patient care or in peer review, can be improved through interventions, which are then evaluated by trials. In 1992, Drummond Rennie[5] called for scientific proof of the value of the peer review system. A Cochrane review updated in 2007[6] concluded that little evidence supports the effectiveness of scientific peer review, since most studies tested the specific effects of masking either authors or reviewers. Investigators have also attempted to improve the quality of peer review,[7] reduce reviewer burden,[8] lessen reviewer bias,[9] and improve the detection of fraud.[10] However, only two of these studies were randomised controlled trials, and all of them focused on surrogate variables related to the review process and not on the true outcome: manuscript quality.

In recent years, the need to establish common, minimum standards of quality resulted in the development of reporting guidelines. These guidelines are defined as: "statements that provide advice on how to report research methods and findings . . . they specify a minimum set of items required for a clear and transparent account of what was done and what was found in a research study, reflecting in particular issues that might introduce bias into the research."[11] Specific guidelines have been developed for different kinds of medical investigation, such as those estimating intervention effects (CONSORT (consolidated standards of reporting trials),[12] TREND (transparent reporting of evaluations with non-randomised designs)[13]), assessing causes and prognosis (STROBE (strengthening the reporting of observational studies in epidemiology)[14]), quantifying accuracy of diagnosis and prognosis tools (STARD (standards for the reporting of diagnostic accuracy studies),[15] REMARK (reporting recommendations for tumour marker prognostic studies)[16]), testing genetic associations (STREGA (strengthening the reporting of genetic associations)[17]), and aggregating evidence (PRISMA (preferred reporting items for systematic reviews and meta-analyses)[18]). Some reporting guidelines have been shown to improve the quality of reports, such as CONSORT.[19 20]

Our team at the *Medicina Clínica* journal previously undertook two randomised trials to determine whether adding a statistical reviewer had any effect on final manuscript quality; the first trial[21] suggested a positive benefit that was confirmed in the second.[22] This second trial also investigated the effect of suggesting the use of reporting guidelines to reviewers, but did not observe any benefit.

The present study focuses on the merged effects of statistical reviews and reviewing guidelines. The intervention consisted of an additional review from a senior statistician asking authors to provide information about incomplete or missing items from reporting guidelines. This additional review was possible because the launch of the STROBE guideline in 2007 allowed us to systematically apply a reporting guideline to almost any paper submitted to *Medicina Clínica*. Thus, we aimed to quantify the effects of an additional review based on reporting guidelines on the quality of the final manuscript in a weekly medical journal with no specific requirements to follow reporting guidelines. After analysis of the overall results, we considered another hypothesis: when updating their manuscript, would authors adhere more to suggestions based on conventional reviews, rather than to those based on reporting guidelines?

Methods

Study design and population

The study was a randomised trial and the web appendix provides full details of the trial protocol. *Medicina Clínica* is a weekly journal based in Barcelona, Spain, with an impact factor of 1.4 that receives more than 300 original papers each year and publishes about a third of submissions. The journal did not ask authors to adhere to reporting guidelines. The current editors (MV and CRJ) assumed their roles on the journal in January 2000, and have since recruited AS, AU, VF, and EC. The two general secretaries (JMR and FC) have more than 15 years' experience.

During the selection process, the first editorial decision chose which papers were sent to reviewers (fig 1⇓). The second decision selected which papers were returned to authors for improvement. We defined the study population as all original research manuscripts received by *Medicina Clínica* from 1 May 2008 to 30 April 2009 that successfully passed through the second editorial decision after conventional peer review.

Intervention

All papers were reviewed by the usual referee team (usually two clinicians, or one clinician with either one statistician or one epidemiologist). Manuscripts in the intervention group also received an additional review based on reporting guidelines. A senior statistician (EC) did the additional reviews, and persistently provided suggestions on how to follow reporting guideline checklists. The manuscript study type determined the guideline used in the review: STROBE, CONSORT 2001,[23] TREND, STARD, STREGA, and REMARK. The box shows an example of a review based on reporting guidelines.

Although additional reviews were only sent to authors in the intervention group, these additional reviews were done for every paper accepted conditionally in the first editorial decision (fig 1). This step had three purposes: to maintain manuscript flow throughout the editorial process; to keep the main investigator masked; and to obtain a score for the initial quality, which was needed for subsequent random allocation.

The 92 papers eligible for randomisation received a mean of 12.8 (standard deviation 9.6) reviewer suggestions per paper. Each paper receiving an additional review had 13.5 (3.9) suggestions, was 446 (146) words long, took 28.4 (11.2) minutes to read, and needed 40.8 (15.6) minutes to draft a review. The reporting guidelines used were: STROBE (85 papers, 92%), CONSORT (17, 18%), TREND (14, 15%), STARD (nine, 10%), STREGA (two, 2%), and REMARK (one, 1%). For some papers, suggestions related to more than one reporting guideline: CONSORT, STROBE, and TREND (13 papers, 14%); STROBE and STARD (seven, 8%); STROBE and STREGA (two, 2%); and CONSORT and TREND (one, 1%).

Most suggestions followed the guideline wording very closely ("STROBE 14: Please provide baseline characteristics of study

Box: Example of review based on reporting guidelines

In accordance with the STROBE statement applied to cross sectional studies and with the aim of increasing transparency and clarity (www.strobe-statement.org/Checklist.html), authors should consider the following modifications:

Strobe 3: Please specify secondary objectives. Of the large number of relationships you have studied indicate, where appropriate, which of them had previous hypotheses. Otherwise, comment on this in the discussion—for example, whether or not these results should be considered as exploratory.

Strobe 5: Please indicate data collection dates.

Strobe 10: Please specify the reasons for collecting a specific sample size. If it was previously stated, please provide the rationale as well as either power or accuracy.

Strobe 11: Please explain the rationale for quantitative outcome cut points and whether or not they were previously specified.

Strobe 13: It is implicitly understood that all consecutive patients were selected and all of them agreed to participate. It is also understood that there are missing data for four deaths only. Please make these points explicit and detail any deviations.

Strobe 16: Please provide 95% confidence intervals for the estimated proportion of the primary objective. If the normal, large sample approximation cannot be applied, consider exact binomial methods: confidence intervals are especially relevant for small samples.

Strobe 22: If applicable, specify all the sponsors and their control over the publication of results. Clarify similarities and differences with previous studies.

participants”), but some were more elaborate (“STROBE 12 and 16: Numerical covariates employed for adjusting were categorised: please check whether you get the same results if you choose a different cut point, or whether you treat them as numerical”). The table⇓ classifies the number of suggestions based on reporting guidelines for every paper section.

Allocation

Before randomisation, all manuscripts were given an ad hoc assessment by the senior statistician (EC) using a score ranging from 1 to 9, in order to give a global measure of report quality at baseline. With these scores, we were able to use a random minimisation algorithm to balance mean differences in the ad hoc score as well as differences in study type counts (that is, intervention, longitudinal, cross sectional, and other type), but not to equilibrate the overall number of manuscripts in both groups. The algorithm gave probabilities from 0.5 (in the case of indifferent allocation to one or another group) to 0.8 (if both minimisation factors indicated allocation to the same group).

Allocation concealment

The second editorial decision (after peer review and before randomisation) took place without committee members knowing which papers were allocated to receive the additional review (fig 1). At later editorial decisions, committee members saw the additional reviews of papers in the intervention group.

Outcome assessment

We obtained the baseline and final scores by using a manuscript quality assessment instrument, designed by Goodman and colleagues[24] and used in our previous trial (web table 1).[20] The instrument uses a 5 point Likert scale from 1 (low) to 5 (high), and comprises 37 items that assess the quality of the research report—not the quality of the research itself. The first item refers to the overall quality of the manuscript (our primary outcome). Of the remaining 36 specific items, 28 (78%) refer directly to key items in reporting guidelines and eight (22%) refer to paper format and style.

The secondary outcome was the average of all pertinent items—that is, after excluding specific items that did not apply to the current study. The evaluators were three junior statisticians (JC, BK, and LG) with experience in teaching scientific critical reading to health professionals. The evaluators first rated each paper individually but, because they were expected to raise different methodological concerns, they were allowed to know each other’s opinions before reaching a consensus. If a consensus was not met, the final score was the average of the individual scores.

Statistical analyses

Main hypotheses

The main statistical analysis specified in the protocol compared papers according to their initial allocation (“as allocated” comparison), adjusted for baseline quality and study type using an analysis of covariance. We did a secondary “as reviewed” comparison (that is, comparing papers according to the reviews they actually received, irrespective of initial allocation) to assess the effect of protocol deviations on the conclusion. We also compared the proportion of papers that improved from the baseline. The reliability of individual ratings was assessed with the intraclass correlation coefficient.

Post hoc hypothesis

A statistical researcher (LC) classified, pooled, and masked reviewer suggestions. Two junior statisticians (BK, LG) then rated each suggestion’s relevance and the authors’ adherence in the final manuscript version to each suggestion using two Likert scales (web table 2). Because manuscripts had a different number of suggestions, the author’s adherence to suggestions was averaged within the paper and compared between groups with a t test weighted by the root of the number of suggestions. We also did an equivalent weighted paired t test to compare the adherence to reviewer suggestions between conventional reviews and additional reviews in the intervention group. We fitted a mixed model with random effects accounting for both reviewer and author variability to analyse sensitivity to the statistical analysis.

Sample size calculation indicated that 50 papers per group allowed 80% power for a 55% standardised difference between groups, in relation to the mean change in scores from the initial version to the final version of the paper. In the previous year (2007), the first editorial decision rejected 186 (57%) of 328 received manuscripts and the second decision rejected 24 (17%) of the remaining 142; therefore, 118 papers were sent to authors with reviewer suggestions. Consequently, we defined an entire year as the recruitment period. For this study, both authors and referees were informed that their material could be used to assess the quality of the editorial process.

Results

Flow

From May 2008 to April 2009, 126 consecutive original papers were included in the study (fig 1). Of these papers, 34 (27%) were rejected on the basis of the conventional review, resulting in 92 randomised papers. Study types of the included manuscripts were: 16 (17%) intervention studies, mainly before-after studies with only five randomised trials; 38 (41%) longitudinal studies; 26 (28%) cross sectional studies; and 12 (13%) studies of other types (mainly diagnostic studies). We saw protocol deviations in four papers in the conventional review group, which underwent an additional review based on reporting guidelines before the scheduled date (that is, cross-in manuscripts).

Outcome reliability and validity assessment

Individual ratings before the consensus discussion shared 0.46 of common information (using the intraclass correlation coefficient). While rating the updated manuscripts, masked evaluators guessed the allocated group in 62% (56/90) of papers (95% confidence interval 51% to 72%; individual percentages of success 56% (50/90), 57% (51/90), and 68% (61/90). Overall, when looking at the author's adherence to any reviewer suggestion, the evaluators were also able to guess the suggestion's origin (that is, from a conventional review or an additional review) in 56% (n=855, 95% confidence interval 54% to 59%) of the 1521 suggestions. Evaluators also guessed whether the type of review was from a clinician or statistician in 63% (n=961, 61% to 66%) of the suggestions.

Selected papers

According to the ad hoc 1 to 9 scale for baseline quality, the 34 papers rejected after the second editorial decision (fig 1) had lower mean score than the 92 accepted papers (3.68 (standard deviation 2.24) v 4.75 (2.20), difference 1.07 (95% confidence interval 0.19 to 1.95)).

Baseline quality based on 1 to 5 Likert scale

The groups receiving conventional reviews alone and additional reviews had similar mean Goodman scores overall at baseline (3.00 *v* 2.84, fig 2⇓). Specific items in reporting guidelines with high initial scores were oversight (4.54), analysis of multiple measures (4.05), and organisation (4.04). The worst scoring items were masking (1.45), dropout analysis (1.86), and dropout description (1.92). Standard deviations of the specific items varied from 0.69 (oversight) to 1.72 (confidence intervals). Pooled standard deviations of the overall and average quality of papers were 1.01 and 0.50, respectively.

Intervention effect

Overall quality (primary outcome) was higher in papers receiving additional reviews than in those receiving conventional reviews alone (0.55 (standard deviation 0.83) *v* 0.27 (0.59); adjusted improvement 0.25 (95% confidence interval –0.05 to 0.54); fig 3⇓); this difference in quality was significant in the "as reviewed" population (0.33, 0.03 to 0.63). We obtained almost identical results for the average quality (secondary outcome) of all valid items (as allocated comparison, 0.11 (–0.01 to 0.22); as reviewed comparison, 0.15 (0.04 to 0.27)). A post hoc interaction test showed that additional reviews had an increased effect on quality of the 16 intervention studies (0.87, 0.01 to 1.74; P=0.04; fig 3).

More papers improved from baseline in overall quality in the additional review group than in the conventional review group (22 (43%) *v* eight (20%); relative risk 2.21, 95% confidence interval 1.10 to 4.44; difference 23.6%, 3.2% to 44.0%; number needed to treat 4.2, 2.3 to 31.2; fig 4⇓). This effect increased if we incorporated the four manuscripts with protocol deviations into the "as reviewed" analysis (relative risk 3.36, 95% confidence interval 1.42 to 7.99). The estimated effect was fairly similar in papers that did or did not include a statistician in the conventional reviews (data not shown).

Adherence to reviewer suggestions

Since most papers were located above the diagonal line in fig 5⇓, the graph showed that, in the intervention group, authors adhered more to suggestions from conventional reviews than to those from additional reviews based on reporting guidelines (3.08 (0.90) *v* 2.70 (0.80). The difference was significant in a weighted paired comparison of means (0.43, 95% confidence interval 0.19 to 0.67). The weighted correlation between both adherences was 0.46 (P=0.01), showing the consistency among authors to consider and to include both kinds of reviewer suggestions. The estimated weighted mean difference in the conventional group (3.37) was higher than that in the additional review group (3.14), although the between group difference was not significant (0.23; –0.17 to 0.63). The mixed model sensitivity analysis showed similar results.

Discussion

Summary of findings

Our data indicated that specific reviewer recommendations to authors in order to fulfil reporting guidelines boosted the number of improved papers from 20% to 43% during the peer review process. But a high proportion of papers (45 (88%), fig 4) did not reach the maximum quality of reporting, based on Goodman score assessments by junior statisticians. Furthermore, although the postulated effect on the primary outcome corresponded to an average improvement of 0.40 on the Likert scale (that is, a 1 point improvement in two of five manuscripts), the observed improvement was only 0.25 (that is, a 1 point improvement in one of four manuscripts). Therefore, although we had secondary evidence of effect, its size was too small to consider that our intervention reached its objectives.

The finding that authors adhered more to suggestions from conventional reviews than to those from additional reviews invites several interpretations. Firstly, authors might not be aware of reporting guidelines in the previous phases of the study, which might make it difficult to incorporate their recommendations in the report. Secondly, authors might prefer to concentrate their efforts on more conventional suggestions. Thirdly, authors in the additional review group might have to cope with a much higher number of suggestions than authors in the conventional review group, and could automatically adhere less to any recommendation, which was supported by the observed trend towards lower adherence to conventional suggestions in the additional review group than that in the conventional review group.

Strengths and limitations of the study

We included any manuscript conditionally accepted after peer review within the scheduled time window. By following published guidelines, we also facilitated the definition of our intervention and its future replication. However, our study had several limitations. Firstly, the estimated effect could be interpreted as mainly a STROBE effect, since this reporting guideline was used in the vast majority of assessments. However, the observed effect in this type of study was significantly reduced (fig 3). The increased effect in the 16 intervention studies (fig 3) could be attributed to the longer tradition of intervention study guidelines.

Secondly, since our trial was conducted in one journal and the intervention relied on only one statistician, external validity was limited. Thirdly, we designed a masked assessment method, but observed that evaluators were perhaps not completely blinded. If they subconsciously rated manuscripts higher in the additional review group, the true intervention effect would have been slightly smaller. Fourthly, we had four protocol deviations; this number is fairly low if we consider the complexity of the process, but high enough to complicate the interpretation of the results.

Furthermore, the second editorial decision meeting had a higher rejection rate (27%) than the previous year (17%), resulting in only 92 randomised papers and a slightly lower power than that designed for 100 papers. Finally, although the Goodman score was especially developed to measure quality improvement during the peer review process, it has not been updated since the development of reporting guidelines, and its validity has not been formally assessed—probably owing to the absence of a gold standard. As Friedberg recently highlighted, "developing useful instruments to measure manuscript quality remains a huge challenge."[25] Ultimately, paper quality is a surrogate for the true purpose of research: to have clarity, transparency, reproducibility, and impact, both on healthcare and on scientific research.

Conclusions and policy implications

As mentioned previously, very few randomised trials have assessed interventions to improve manuscript quality after peer review. Health research is responsible for improvements in healthcare. But before implementing its results, the last and essential step is communication and dissemination, which relies on the peer review process. If we want to further improve our health system, we should develop and select efficient interventions and accurate prognosis and diagnosis tools. To avoid poor health research,[2 3] we need competent communication and editorial processes, with transparent publications and a low false discovery rate.[26-29] Although reporting guidelines are a profoundly reasonable procedure for boosting paper quality, reasonability does not imply effect.[20]

Authors in a mid-level medical journal have more difficulties in following suggestions based on reporting guidelines than those from conventional reviews alone. If authors do not consider key methodological features at the design and execution phases of their study, they will have difficulties in improving the paper at the later scientific phases. To boost paper quality and impact, authors should be aware of future requirements of reporting guidelines at the very beginning of their study, and peer reviewers should be made aware of the importance of transparent reporting and receive training if needed.

We thank David Moher, Umair Mallick, and Julie Morris for their helpful suggestions to the manuscript; Matthew Elmore for English editing; and Mercedes Belmonte for managing the manuscript follow-up.

Contributors: EC (guarantor) and JMR had the original idea, and with FC, AS, DGA, AU, VF, CRJ, and MV, designed the study. JMR, FC, and AS managed and monitor the manuscript flow. EC did the additional review based on reporting guidelines. BK, LG, and JC rated the manuscript quality. BK and LG assessed the suggestion relevance and adherence. LC managed the database, classified, merged, and masked the reviewer suggestions. JC and JAS programmed and did the analysis. FM and JAG designed, programmed, and maintained the allocation algorithm. EC, JAG, and JC designed the figures. JAG and JC audited the database, statistical programming, and editing process. All authors contributed to and approved the final manuscript.

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf and declare that: JC, LG, LC, BK, and FM are supported by the "Bioestadística para no estadísticos" learning programme; they have no relationships with companies that might have an interest in the submitted work in the previous 3 years; their spouses, partners, or children have no financial relationships that may be relevant to the submitted work; and they have no non-financial interests that may be relevant to the submitted work.

Ethics approval: both authors and referees were informed that their material might be used to assess the quality of the editorial process.

Data sharing: Technical appendix, statistical code, and dataset available from the corresponding author at erik.cobo@upc.edu.

1 Rennie D. Editorial peer review: its development and rationale. In: Godlee F, Jefferson T, eds. Peer review in health sciences. BMJ Books, 2003:1-13.
2 Altman DG. The scandal of poor medical research. *BMJ* 1994;308:283-4.
3 Von Elm E, Egger M. The scandal of poor epidemiological research. *BMJ* 2004;329:868-9.
4 García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. *BMC Med Res Methodol* 2004;4:13.
5 Rennie D. Editorial peer review: let us put it on trial. *Controlled Clinical Trials* 1992;13:443–5.
6 Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Systematic Rev* 2007;2:MR000016. doi: 10.1002/14651858.MR000016.pub3.
7 Schroter S, Black N, Evans S, Smith R, Carpenter J, Godlee F. Effects of training on quality of peer review: a randomised controlled trial. *BMJ* 2004;328:673-5.
8 Johnston SC, Lowenstein DH, Ferriero DM, Messing RO, Oksenberg JR, Hauser SL. Early editorial manuscript screening versus obligate peer review: a randomized trial. *Ann Neurol* 2007;61:A10-12.
9 Ross JS, Gross CP, Desai MM, Hong Y, Grant AO, Daniels SR, et al. Effect of blinded peer review on abstract acceptance. *JAMA* 2006;295:1675-80.
10 Peer review and fraud. *Nature* 2006;444:971-2.
11 EQUATOR Network. Library for health research reporting. 2011. www.equator-network.org/resource-centre/library-of-health-research-reporting.
12 Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
13 Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health* 2004;94:361-6.
14 Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al; STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007;4:1628-54.
15 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41-4.
16 McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br J Cancer* 2005;93:387-91.
17 Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al. Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. *Eur J Epidemiol* 2009;24:37-55.
18 Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
19 Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust* 2006;185:263-7.
20 Moher D, Turner L, Shamseer L, Plint A, Weeks L, Peters J, et al. The influence on CONSORT on the quality of RCTs: an updated review. Society for Clinical Trials 32nd Annual Meeting, 2011 May 16-18, Vancouver, Canada 2011;A79.
21 Arnau C, Cobo E, Ribera JM, Cardellach F, Selva A, Urrutia A. Effect of statistical review on manuscript quality in MEDICINA CLÍNICA (Barcelona): a randomized study. *Med Clin (Barc)* 2003;121:690-4.
22 Cobo E, Selva-O'Callagham A, Ribera JM, Cardellach F, Dominguez R, Vilardell M. Statistical reviewers improve reporting in biomedical articles: a randomized trial. *PLoS ONE* 2007;2(3):e332.
23 Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al; the CONSORT Group. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94.
24 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.
25 Friedberg EC. Peer review of scientific papers—a never-ending conundrum. *DNA repair (Amst)* 2010;9:476-7.
26 Meng XL. Desired and feared—what do we do now and over the next 50 years? *The American Statistician* 2009;63:202-10.
27 Young NS, Ioannidis JPA, Al-Ubaydli O. Why current publication practices may distort science. *PLoS Med* 2008;10:e201.

What is already known on this topic

Manuscript quality in biomedical journals is far from guaranteed, despite the continued use of peer review after submission

Reporting guidelines have been developed to improve manuscript quality and transparency

What this study adds

Additional reviews based on reporting guidelines resulted in a moderate improvement in manuscript quality

Authors have difficulties in adhering to high standards of reporting during the writing phase; awareness of guidelines should be guaranteed during the design and execution of the study

28 Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ. False-positive results in cancer epidemiology: a plea for epistemological modesty. *J Natl Cancer Inst* 2008;100:988-95.
29 Blair A, Saracci R, Vineis P, Cocco P, Forastiere F, Grandjean P, et al. Epidemiology, public health, and the rhetoric of false positives. *Environ Health Perspect* 2009;117:1809-13.

Accepted: 23 September 2011

Cite this as: *BMJ* 2011;343:d6783

This is an open-access article distributed under the terms of the Creative Commons Attribution Non-commercial License, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited, the use is non commercial and is otherwise in compliance with the license. See: http://creativecommons.org/licenses/by-nc/2.0/ and http://creativecommons.org/licenses/by-nc/2.0/legalcode.

Table

Table 1| **Total number of suggestions based on reporting guidelines**

Common and specific items	STROBE (N=85)		CONSORT plus extensions (N=17)*		TREND (N=14)		STARD (N=9)		Total (N=92)†
	Item	n (%)	Item	n (%)	Item	n (%)	Item	n (%)	n (%)
Title and abstract									
	1	53 (62)	1	9 (53)	1	11 (79)	1	1 (11)	76 (72)
	1a	0							
	1b	2 (2)							
Introduction, background									
	2	13 (15)	2	1 (6)	2	0	–	–	14 (14)
Methods									
Participants and recruitment	5	47 (53)	3	11 (65)	3	16 (100)	3	5 (56)	138 (75)
	6	49 (54)					4	2 (22)	
	6a	4 (5)							
	6b	0							
Objectives	3	49 (58)	5	7 (41)	5	7 (50)	2	3 (33)	66 (63)
Variables, measurements									
Interventions	–	–	4	12 (71)	4	8 (57)	–	–	20 (13)
Standard, outcomes	7	36 (40)	6	11 (59)	6	6 (36)	7	2 (22)	96 (67)
	8	31 (35)			10	0	8	0	
							9	1 (11)	
							10	9 (100)	
Sample size	10	81 (95)	7	14 (82)	7	12 (86)	–	–	108 (92)
Bias, randomisation, study design	4	26 (29)	8	3 (18)	8	2 (14)	5	4 (33)	91 (64)
	9	44 (51)	9	5 (29)			6	2 (22)	
			10	4 (24)					
Masking	–	–	11	10 (35)	9	3 (14)	11	8 (89)	21 (15)
Statistical methods	11	32 (38)	12	29 (94)	11	18 (93)	12	11 (78)	197 (86)
	12	77 (60)					13	9 (100)	
	12a	7 (8)							
	12b	4 (4)							
	12e	6 (6)							
Missing data	12c	25 (30)	16	9 (47)	–	–	22	1 (11)	53 (38)
	12d	12 (14)							
	14b	4 (5)							
Funding	22	81 (95)	–	–	–	–	–	–	81 (88)
Results									
Participant flow	13	81 (77)	13	13 (59)	12	6 (43)	16	2 (22)	103 (78)
Recruitment	14c	18 (20)	14	5 (29)	13	1 (7)	14	2 (22)	30 (26)
							17	3 (33)	
Baseline data	14	23 (26)	15	4 (24)	14	0	15	1 (11)	34 (32)
	14a	3 (4)			15	1 (7)			
Numbers analysed	15	20 (24)	16	9 (47)	16	1 (7)	18	1 (11)	31 (30)

No commercial reuse: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe

(continued)

Common and specific items	STROBE (N=85)		CONSORT plus extensions (N=17)*		TREND (N=14)		STARD (N=9)		Total (N=92)†
	Item	**n (%)**	**Item**	**n (%)**	**Item**	**n (%)**	**Item**	**n (%)**	**n (%)**
Outcomes and estimation	16	123 (89)	17	25 (88)	17	19 (79)	19	2 (22)	188 (94)
	16a	1 (1)					21	4 (44)	
	16b	0					23	0	
	16c	1 (1)					24	9 (100)	
Ancillary analyses	17	8 (8)	18	5 (29)	18	1 (7)	–	–	14 (12)
Adverse events	–	–	19	14 (82)	19	10 (71)	20	7 (78)	31 (23)
Discussion									
Interpretation	18	65 (58)	20	20 (88)	20	4 (21)	–	–	170 (80)
	19	80 (71)							
Generalisability	21	47 (46)	21	2 (12)	21	0	25	3 (33)	52 (48)
Overall evidence	20	83 (68)	22	3 (18)	22	0	–	–	86 (66)
Total (100%)	**–**	**1236**	**–**	**225**	**–**	**126**	**–**	**92**	**1700**

N=total number of manuscripts; n=number of times each reporting guideline item was used (since each item might have more than one suggestion, n can be greater than N); %=manuscripts with at least one suggestion divided by total number of manuscripts (N).

*Includes CONSORT 2001 and CONSORT for non-pharmacological treatment interventions.

†Two further reporting guidelines were used sporadically: STREGA in two manuscripts with suggestions about participants (three), statistical methods (two), baseline data (two) and outcomes (two); and REMARK in one manuscript with suggestions about participants (one), sample size (one), study design (one), statistical methods (four), participant flow (one), recruitment (one), outcomes (two), and interpretation (one).

Figures

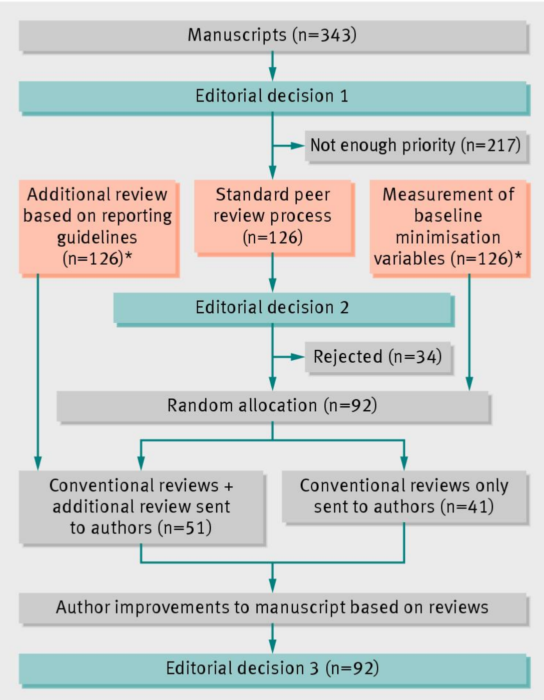

Fig 1 Study design and manuscript flow. *Additional reviews and measurement of minimisation variables were undertaken during the standard peer review process, but this information was concealed until the later editorial stages

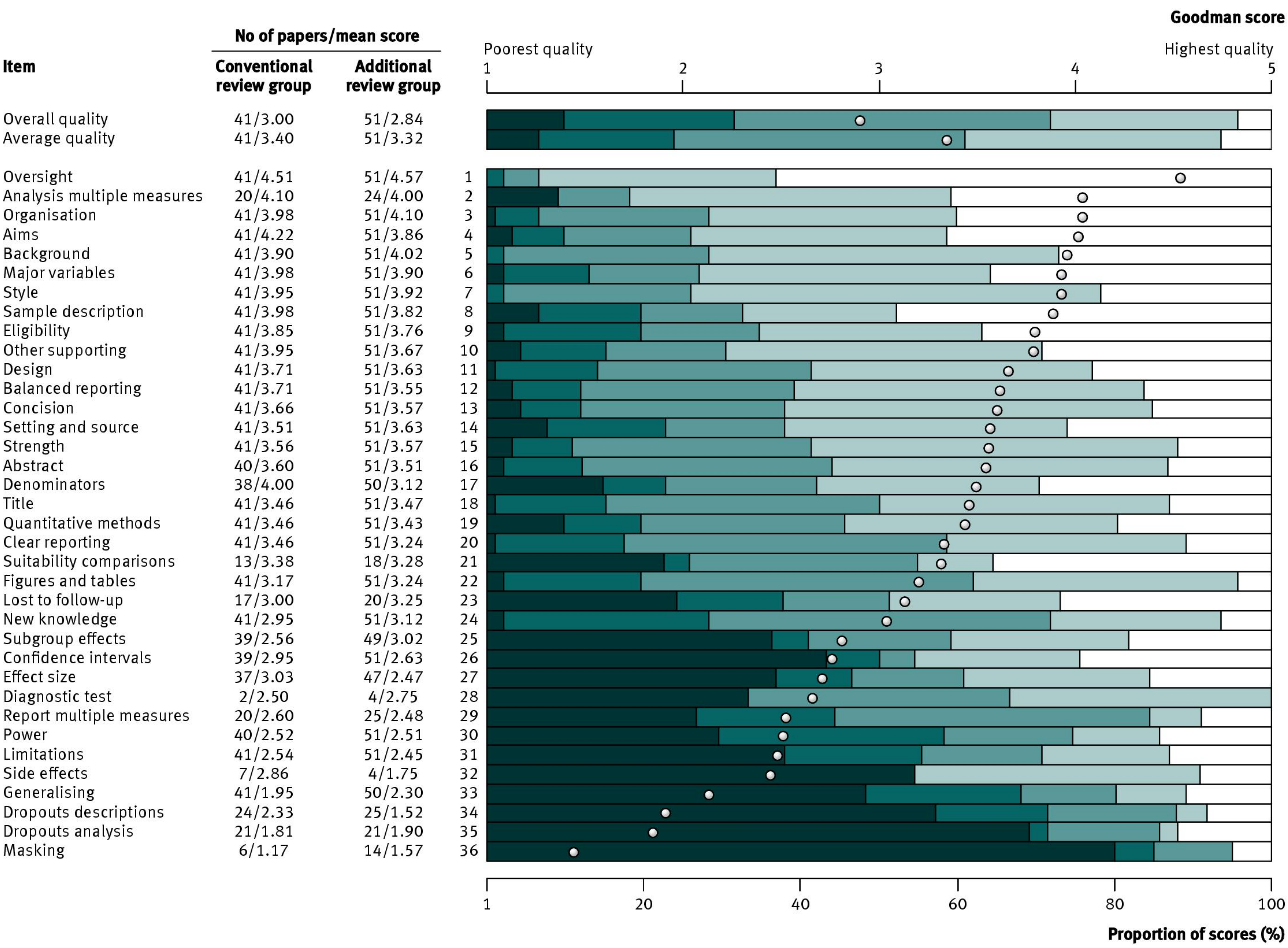

Fig 2 Goodman quality scores at baseline. Bars contain proportion of scores from 1 (dark shade) to 5 (light shade), with cumulative percentages shown in the bottom scale. Gradation colour for the average quality was consequently adapted (break points are 2.5, 3.0, 3.5, and 4.0). Dots represent total mean for each specific item

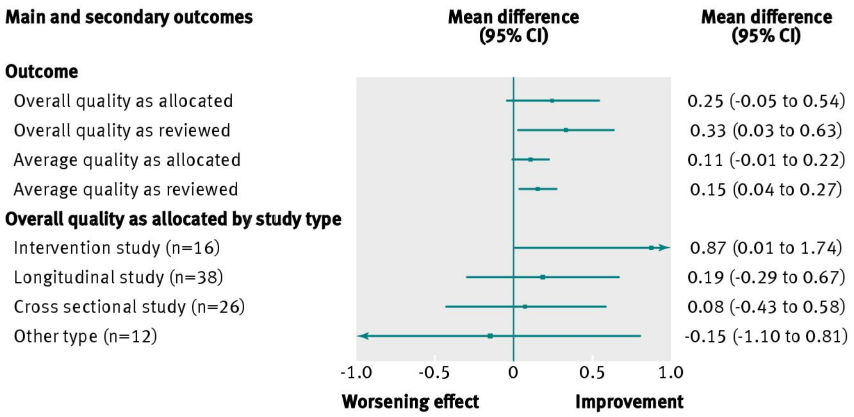

Fig 3 Effect of additional reviews on overall and average quality in "as allocated" and "as reviewed" populations, and primary analysis stratified by study type

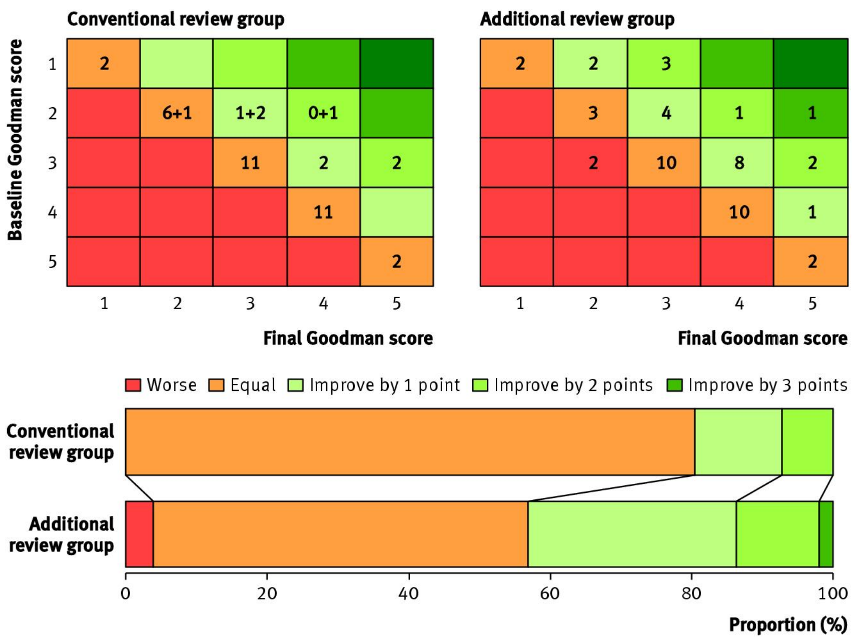

Fig 4 Baseline and final Goodman quality scores in allocated groups. Numbers after the plus signs indicate the four manuscripts with protocol deviations

No commercial reuse: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe

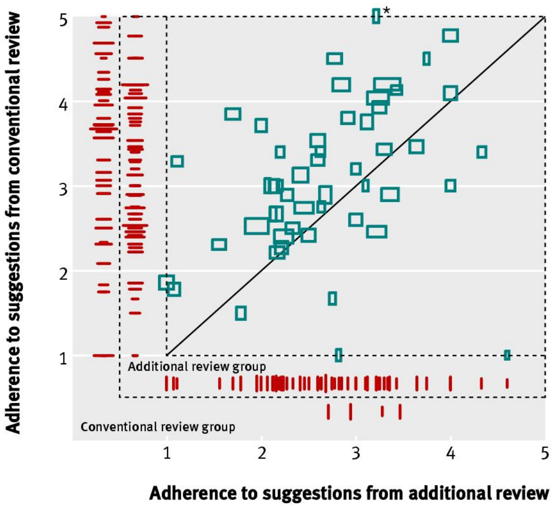

Fig 5 Average author adherence to repeated reviewer suggestions based on 5 point Likert scale (1=minimum, 5=maximum). Rectangles represent the 51 papers from the additional review group, with paired suggestions both from conventional reviews (shown as the horizontal lines on each rectangle) and additional reviews (shown as the vertical lines of each rectangle). The side length of rectangles represents the amount of information from any type of review (square root of the number of suggestions per manuscript) and the rectangle area represents each paper's overall information. A rectangle above the diagonal line indicates that a paper adhered more to the conventional review than to the additional review. For example, the asterisked rectangle corresponds to a manuscript receiving 14 suggestions (proportional to the square of the vertical sides) from the additional review with a 3.21 average level of adherence, and two suggestions (the square of the horizontal sides) from the conventional review with a mean adherence score of 5. Lines in the external margin represent papers from the conventional review group, and lines on the internal margin represent papers from the additional review group that received both conventional (lines along the vertical axis) and additional (lines along the horizontal axis) reviews; lines are repeated here to assist between group comparison.