MEDICAL PRACTICE

*Occasional Review***Is the statistical assessment of papers submitted to the "British Medical Journal" effective?**

MARTIN J GARDNER, DOUGLAS G ALTMAN, D R JONES, D MACHIN

Abstract

A study was performed to judge the effectiveness of the statistical assessment scheme for papers submitted to the "British Medical Journal." Statistical criticism of the content of 200 submitted papers which had already been seen by a subject referee led directly or indirectly to 73 (37%) being rejected for publication. In most cases (53 out of 73) serious problems requiring more than minor revision were identified. A comparison of reports on subsequently unpublished and published papers showed that adverse statistical assessments—suggesting major problems—were more common in the papers that were not accepted for publication. Moderate, or less, revision was recommended for 63% of published papers but 39% of the remainder. A checklist of relevant questions was used in making a detailed comparison of 12 published papers, six of which had been statistically assessed and six of which had not. This comparison yielded little evidence that the papers that had been assessed were statistically more acceptable than those that had not been assessed but re-emphasised the subjectivity of refereeing and assessment.

MRC Environmental Epidemiology Unit, Southampton General Hospital, Southampton SO9 4XY

MARTIN J GARDNER, BSC, PHD, reader in medical statistics

MRC Clinical Research Centre, Harrow, Middlesex HA1 3UJ

DOUGLAS G ALTMAN, BSC, medical statistician

Westminster Medical School, London SW1P 2AR

D R JONES, MSC, PHD, senior lecturer in medical statistics

Department of Community Medicine, Southampton General Hospital, Southampton SO9 4XY

D MACHIN, MSC, senior lecturer in medical statistics

Correspondence to: Dr M J Gardner.

The use of a checklist similar to the one designed for this study might help to direct statistical assessors' comments, not only on the statistical competence of a paper but also on its acceptability for publication. But these findings also raise further questions—about the way editors select papers for statistical assessments and about the way authors respond to them.

Introduction

In February 1979 the *British Medical Journal* introduced a system for statistical review of submitted papers. The aim was to obtain an assessment of the statistical acceptability, or otherwise, of papers already passed as acceptable by subject referees. Since then whenever members of the editorial review group within the *BMJ*'s office have expressed doubts about a paper's statistics the paper has been sent to the statistical assessor(s). Their doubts have concerned the design of the study or the conduct, analysis, interpretation, or presentation of the findings.

Of the 5000 or so papers submitted to the journal each year,¹ about 100 are sent for statistical assessment after the first round of refereeing. This number is only a fraction (about 1 in 8) of those which are considered to be possibly acceptable for publication after standard refereeing. The small proportion reflects the fact that some medical papers are not of a statistical nature; that when they are the quality is often satisfactory; and that until recently a single statistician has done most of the assessments. As the *British Medical Journal* publishes 12-15 articles each week¹ and the statistical assessment has been done on a part time basis, only a limited number of papers can be reviewed critically in this way.

The statistician makes his assessment without knowing what doubts the editorial group has about the paper. His comments are then used by the group to help it decide whether to accept the paper, accept it subject to revision, or reject it. In all cases the final decision remains with the editorial group. About two

thirds of the assessed papers are subsequently published, often after revision, and the remaining one third rejected.

This statistical assessment scheme was introduced to help the refereeing and editorial processes to prevent the publication of inferior scientific articles. It was envisaged that when revision was recommended the statistical quality of the published papers would be improved. We describe here a study designed to try to gain insight into the statistical refereeing procedure. We took two approaches: firstly, we considered the content of the statistical reports on papers as submitted in relation to whether or not they were published subsequently and, secondly, we compared statistically assessed and unassessed papers that were published in the journal. Because we undertook the second part of the study to develop ideas, as well as for other reasons mentioned later, the number of papers included was small.

Methods

The study was based on the first 200 papers seen by the statistical assessor (MJG) since the scheme started. Of these 127 were published, and the statistical reports on the remaining 73 formed the material for investigating the statistical content of rejected papers. The reports on rejected papers were studied to identify whether the statistical criticism could have contributed to their rejection by the journal. The remarks made about each paper were classified into five categories (by MJG). The first category contained papers with problems statistically so serious that the assessor considered them irredeemable; the remaining categories contained papers judged to require major, moderate, minor, or no revision to be of an acceptable statistical standard for publication.

To obtain comparative information on the content of the statistical reports of published and unpublished papers, we performed a further investigation. The second 100 of the 200 reports (papers 101-200) were used for this purpose. One assessor (this time DM) classified them into the same five categories as above without knowing whether or not the paper had been published subsequently—and without knowing the titles and names of authors of the papers. After the classification had been completed for the 100 reports, the results were analysed in relation to whether or not they had subsequently been published in the journal.

The design of the second part of the study was more complex. We decided that the most suitable way of obtaining the required data was to compare a number of statistically assessed published papers with an equal number of other "matched" published papers. Because of the large amount of time entailed in assessing thoroughly any one paper (normally two to three hours, but occasionally longer, not including "overnight cogitation"), we decided to start with six assessed papers and six "controls." The assessed papers were selected by taking every 30th of the 200 available papers on which the statistical assessor (MJG) had reported in sequential order of the date of the written report. The sequence was started randomly with the number 10, and if any thus selected paper had not been published then the first subsequent published paper was included instead. Papers were included without reference to their content, to the statistical report, or to whether any revisions had been suggested or incorporated.

The "matched" control papers were chosen from contemporary issues of the *British Medical Journal* by a statistician (DGA) completely unconcerned in the assessment. His guidelines were to select one paper on a related topic which was similar in the amount of statistical content and length to each of the six assessed papers but which had appeared in a different issue of the journal, so that the pairing was not obvious. This proved to be a difficult task.

The collection of the 12 papers was then arranged (by MJG) into ascending alphabetical order of the surname of the first author and numbered from 1 to 12. Two further medical statisticians (DRJ, DM), who were independent of the assessment process and without knowledge of the design of this pilot study, were then asked to carry out a detailed statistical review of the 12 papers and make their evaluation of quality on a specially prepared form (see figure). The forms identified each paper by its allotted number only. The main features looked at were the study design (question 2), the selection of participants (questions 3 and 4), the statistical methods used (questions 5 and 6), the conclusion reached (question 7), and the statistical acceptability of the paper (question 9). A form of this type has been used previously in a study on clinical trials (unpublished findings), but in the present study the questions were more general.

The use of two statisticians was designed to measure the agreement or disagreement between statistical reviewers of the same articles. The two reviews were carried out independently, and the reviewers had no means of identifying which or how many of the 12 papers had been statistically assessed before publication. After both statisticians had completed their reviews the answers were tabulated by one of the other participants (DGA), who did not know the identity of the numbered papers. Only after this preliminary analysis was the identification and pairing of the 12 papers revealed. The blindness built into the study design was to minimise potential biases.

Evaluation of statistical assessment—pilot study

Paper number If any answers are "No"
please make comments in
Section 11 below.

Name of assessor

(1) Is the objective (purpose) of the study sufficiently described ?	Yes/No
(2) Is an appropriate study design used to achieve the objective ?	Yes/No
(3) Is the selection of study members adequate in:	
(a) its description ?	Yes/No
(b) the source of patients ?	Yes/No
(c) sample chosen ?	Yes/No
(d) response rate ?	Yes/No
(4) (a) Is the reason for the sample size used discussed ?	Yes/No
(b) Does the sample size seem adequate ?	Yes/No
(5) (a) Is there a statement describing or referencing all statistical tests used ?	Yes/No
(b) Are the tests used appropriate ?	Yes/No
(6) Is the presentation of statistical material clear in the:	
(a) text ?	Yes/No
(b) tables ?	Yes/No
(c) figures ?	Yes/No
(7) Is an appropriate/correct conclusion drawn from the statistical analysis ?	Yes/No
(8) Was a statistician involved as:	
(a) author ?	Yes/No
(b) acknowledged help ?	Yes/No
(9) From the statistical viewpoint is the paper of acceptable standard to have been published ?	Yes/No
(10) Other remarks	
(11) Comments on questions 1-9 where answer was "No"	

Checklist used by the statisticians in reviewing 12 published papers, six of which had been assessed statistically and six of which had not.

Results

In the initial part of the study 18 of the 73 reports on papers not published subsequently were categorised as containing problems which could not be rectified. As such they were considered to be statistically unacceptable for publication. Another 35 of the reports indicated that major or moderate revisions of the statistical aspects would be necessary to improve the quality of the associated papers. These included studies, for example, where the data were of dubious quality, the analysis was inappropriate or incorrect, or the conclusions were unjustified from the statistical results. The reports on another 16 papers suggested the need for minor statistical revisions, often in presentation—for example indicating what followed a $\pm$ sign or better labelling of figures.

The results of the comparison of reports on subsequently unpublished and published papers are shown in table I. Of the 54 statistical assessments on papers which were not published, 15 (28%)

recommended that the paper was statistically unacceptable, compared with three (7%) for the 46 published papers. At the other extreme, all five papers where the reports indicated acceptability without revision were published. The relative likelihood of a paper not being published by the journal increased as the severity of the statistical comment increased, as indicated by the ratios in the final column of table I.

TABLE I—Classification of statistical content of reports on 100 papers submitted to the *British Medical Journal* according to whether they were subsequently published

Statistical content of paper	Subsequently published		Ratio of unpublished: published
	No	Yes	
Unacceptable	15	3	5.0
Acceptable after:			
major revision	18	14	1.3
moderate revision	13	11	1.2
minor revision	8	13	0.6
no revision	0	5	0
Total	54	46	1.2

In the second part of the study questions 7 and 9 in the figure were clearly most important in terms of the overall statistical content of the 12 papers. The detailed answers to these two questions are set out in table II. For nine of the 22 given responses the appropriate or correct statistical conclusions (question 7) were said to have been reached, entailing an assessed paper on six of the nine occasions. Only for one paper (assessed No 10) did the reviewers concur that the conclusion was appropriate, whereas for four papers (assessed No 3 and unassessed Nos 5, 6, and 11) they agreed on their inappropriateness. Where answers were given by both reviewers they disagreed on five papers, with one reviewer (DRJ) saying "yes" on each occasion. Among the paired responses, eight were the same—that is, either YY or NN—for both the assessed and unassessed papers, with a positive answer being given for the assessed paper in the three pairs where there were differential replies.

TABLE II—Answers to question 7 about appropriateness or correctness of the conclusion drawn from the statistical analysis and to question 9 about acceptability for publication (Y=yes, N=no, —=no answer) by reviewer and by pairings of previously assessed/unassessed papers

Question No	Reviewer	Code numbers of paired papers (statistically assessed paper first)					
		2/4	3/6	7/1	8/9	10/5	12/11
7	DRJ	YY	NN	YY	YY	YN	YN
	DM	NN*	NN	—†	NN	YN	NN
9	DRJ	YY	NN	NN	NN	YN	YN
	DM	NY*	NN	NN	NN	YN	NN

* The apparent conflict between these two answers occurs because DM answered question 9 as "yes, revision required however," which perhaps should be more appropriately interpreted as "no".

† For papers 7 and 1, DM answered "not applicable" because no statistical tests had been carried out; DRJ based his "yes" answers on the numerical results given in the papers.

Only two papers were said to be acceptable for publication (question 9) by both reviewers, of which one had been statistically assessed previously and one had not. A further eight of the papers were unacceptable to both reviewers; three of these papers had been through the statistical assessment scheme. The overall level of acceptability was low, with only six out of 24 affirmative answers to question 9. For this question eight of the 12 within pair answers were the same, and in seven of these both the assessed and unassessed papers were said to be unacceptable. On the four occasions where there was a differential response between the papers in the pair three were in favour of the assessed paper.

The numbers of "yes" and "no" responses given to each of the remaining questions are detailed in table III. Again there was no indication of any major differences between the assessed and unassessed papers. The individual reviewers and the pairing of papers are not shown since analyses taking these into account showed nothing extra. Apart from question 7, which has already been analysed, the question showing the largest contrast was that on the adequacy of the

TABLE III—Number of responses to each question as shown in the Figure according to whether or not the paper had initially been assessed by MJG

Question No	Assessed			Not assessed		
	Yes	No	No answer*	Yes	No	No answer*
1	10	2		10	2	
2	7	5		6	6	
3a	6	6		7	5	
3b	8	4		8	4	
3c	4	7	1	7	3	2
3d	3	5	4	2	6	4
4a	1	10	1	3	9	
4b	4	7	1	4	8	
5a	6	6		5	4	3
5b	6	5	1	4	4	4
6a	5	7		5	7	
6b	7	3	2	7	5	
6c	2	4	6	1	7	4
7	6	5	1	3	8	1
8a	1	10	1		12	
8b	2	9	1		12	
9	3	9		2	10	

* In nearly every case where a "yes" or "no" response was not given the assessor had written "not applicable."

chosen study sample (question 3c), with the unassessed papers performing better. The reviewers reported favourably on the statement of objectives (question 1) and rather less favourably on the source of patients for the study (question 3b) and the presentation of tabular information (question 6b). The other items were regarded as less satisfactory—for example, the explanation and size of the sample used (questions 4a and 4b) and the clarity of the figures (question 6c).

Discussion

This review of statistical reports on submitted papers suggests that the assessment scheme has made an important contribution to excluding articles of poor quality from appearing in the *BMJ*. All the papers that were sent to the statistical assessor had been previously refereed by the standard process and found to be suitable from this viewpoint, but the editorial group felt that a statistical opinion would nevertheless be desirable. Over one third of the first 200 papers which were assessed in this way were not subsequently published. In 53 of these 73 unpublished papers serious statistical problems were raised which would have required at least major revision before becoming acceptable for publication. To some extent on the basis of the assessments, along with editorial judgment on other aspects of the papers, many statistically poor papers have been appropriately rejected.

A comparison of the statistical assessments on 100 consecutive papers showed that those which were later published were reported to contain fewer problems. This is as would be expected if the comments in the reports were incorporated into the decision about publication. None the less, the reports on subsequently published papers were not all favourable, with only 18 out of 46 said to require minimal revision. In most cases changes were made to the submitted manuscripts before publication, to deal with the statistical points raised.

One noticeable outcome of the second part of this study was the level of adverse comment made by the two independent statistical reviewers, both of papers previously assessed for statistical content and of those unassessed. For questions directly related to statistical content, there were a similar number of "yes" and "no" responses. The two reviewers, both senior medical statisticians, concurred in this finding, although one of the reviewers was more favourably disposed to the papers than the other.

There are several possible reasons for these findings. To take the differences between reviewers first, it must be recognised that statistical refereeing is necessarily somewhat subjective. There are often various plausible ways to design, carry out, or analyse a study to answer a particular research question, with no firm rules to indicate the preferred approach, and so some disagreement between reviewers is not surprising. The questions in the figure were produced as a checklist of the main factors

that might be looked for by a statistical referee—but the use of the checklist was also not entirely objective, particularly where words such as “adequate” and “acceptable” were concerned.

On the apparent discrepancies between the initial statistical assessment (MJG) and the later reviews there are two practical points that can be made. Firstly, there was a difference in approach since the original assessment was unstructured, and this may have resulted in a degree of lenience because no systematic attempt to answer a predetermined list of questions was made. Secondly, the two reviewers of the published, rather than submitted, papers were taking part in an experiment—rather than doing routine refereeing—and for one of the reviewers at least this probably led to stricter judgments.

We have looked in detail at the six papers which were assessed both by the *BMJ*'s statistical assessor and, during the study, by the two outside reviewers. When we compared the reviews and contrasted the submitted and published papers for evidence of any revisions resulting from the initial statistical assessment a number of interesting points emerged. One paper was concerned with the effectiveness of different treatments, and each of the three statistical reviewers raised doubts because a non-randomised study had been carried out with dubiously selected controls. In a second paper, reporting a comparative study, the inappropriateness of the control group was also mentioned by all reviewers. A third paper contained an analysis of treatments used for a rare cancer over a long period of time; both the later reviewers raised the potential problems of non-randomised treatment allocation, although one regarded the study as the best possible in the circumstances. In another two papers many points relating to the presentation and interpretation of the results were raised by all three reviewers—a consequence of incomplete, but also incorrect, revision. In the sixth paper, the only one found by both subsequent reviewers to draw an appropriate or correct conclusion and be acceptable for publication, a major error had been detected in the original assessment and suitably dealt with.

Many other matters of detail were, of course, included in the reviews, but the main ones mentioned raise some important issues, such as how the statistical reports should be written and used. The emphasis of the assessments has been on statistical competence and not so much on the acceptability of the articles for publication. The difficulties of design in the first two papers—which no revision could have overcome—were perhaps not pointed out strongly enough to the editors, or the editors may have chosen to overlook them for reasons possibly associated with the medical content of the paper. Where revisions were made not all the points raised in the statistical assessment were covered and further errors were sometimes introduced. The revised papers were not seen again by the statistical assessor before publication. Such a reassessment is probably desirable in practice, although it would clearly increase the workload.

We think that the general use of a checklist such as that used in the study would be advantageous for statistical refereeing, and we would recommend its introduction, although it should not be regarded as a straitjacket. Answers to these questions would not necessarily produce perfect judgments about each paper, since many statistical points—for example, the degree of skewness in the data which invalidates *t* tests—are somewhat subjective, but the use of the questions would lead to a careful consideration of the statistical components and guide the reviewer to answer those questions that were particularly important to the editorial committee. Some revision of the form is needed—for example, question 5 should more suitably use the word “methods” rather than “tests,” and question 7 should be divided into two separate questions about appropriateness and correctness. Not all the formulated questions will be relevant to all studies, so a “not applicable” category is desirable, as shown in tables II and III. A modified version has, in fact, recently been used in an analysis of the statistical content of another journal (unpublished findings).

A more extensive evaluation of the statistical assessment scheme would need to consider other issues in addition to those

which have been looked at here. For example, are the appropriate papers being sent to the statistical assessor? Should the subject referee be asked to suggest whether there is a need for a statistical review? Does statistical assessment improve submitted papers that contain errors, but only to the level of unassessed published papers? Also, it would be important to know whether papers which were not published as a result of the statistical assessment contained serious errors that would be agreed on by other statisticians. Some “good” papers may have incorrectly been given a negative statistical assessment and unjustifiably prevented from publication.

In general the statistical content of papers appearing in various medical journals has become a subject of much concern and some inquiry in recent years.² * The resulting reports have indicated that in about half of all published papers there are statistical errors, and there have been several suggestions for ways of improving statistics in papers.⁹ ¹¹ These include improved communication between the statistical and medical professions in education and collaborative work, greater participation by statisticians in refereeing, the use of a catalogue of potential biases, and the availability of a comprehensive list of statistical guidelines (see accompanying paper, p 1489). Another suggestion is the addition of a statistician to the editorial board, which has recently been adopted by some journals. Overall, the whole issue of refereeing (not just statistical refereeing) is one for scrutiny, amendment, and improvement whenever possible.¹²

References

- 1 Smith R. Steaming up windows and refereeing medical papers. *Br Med J* 1982;**285**:1259-61.
- 2 Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;**195**:1123-8.
- 3 Lionel NDW, Herxheimer A. Assessing reports of therapeutic trials. *Br Med J* 1970;**iii**:637-40.
- 4 Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J* 1977;**i**:85-7.
- 5 Rao NSN, Seth AK, Murthy NS, Marwah SM. Some statistical deficiencies in medical research in India. *Indian J Med Res* 1977;**66**:696-703.
- 6 White SJ. Statistical errors in papers in the *British Journal of Psychiatry*. *Br J Psychiatry* 1979;**135**:336-42.
- 7 Glantz SA. Biostatistics: how to detect, correct, and prevent errors in the medical literature. *Circulation* 1980;**61**:1-7.
- 8 DerSimonian R, Charette J, McPeck B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med* 1982;**306**:1332-7.
- 9 Sackett DL. Bias in analytic research. *J Chron Dis* 1979;**32**:51-63.
- 10 Altman DG. Statistics and ethics in medical research: VIII—Improving the quality of statistics in medical journals. *Br Med J* 1981;**282**:44-7.
- 11 Altman DG. Statistics in medical journals. *Statistics in Medicine* 1982;**1**:59-71.
- 12 Lock S. Peer review weighed in the balance. *Br Med J* 1982;**285**:1224-6.

What effect does the herb yarrow (Achillea millefolium) have on the cardiac rhythm? Can it relieve troublesome extrasystoles in a patient with no cardiac abnormalities? What are the long term effects and symptoms of overdose?

Yarrow has been used as a medicinal herb since ancient times. Achilles is said to have treated his warriors' wounds with an ointment made from the leaves—hence its generic name. It is said to have antipyretic, diaphoretic, carminative, and haemostatic properties and has been used for a wide variety of conditions including fever, toothache, haemorrhage, and as a tonic. I have, however, been unable to find any reference to effects on cardiac rhythm. Yarrow contains an essential oil. Excessive doses of essential oils can produce nausea, vomiting and diarrhoea, headache, and dizziness. Severe poisoning produces either excitation of the central nervous system and convulsions or depression of the central nervous system and coma. Hepatic and renal toxicity have been described in people working in contact with essential oils for three to 12 years.—LINDA BEELEY, director, drug and therapeutic unit, Birmingham.

Martindale. *The extra pharmacopoeia*. 28th ed. London: Pharmaceutical Press, 1982:670.
I.e. Strange R. *A history of herbal plants*. Brighton: Angus and Robertson, 1977:4-6.