STATISTICS IN MEDICINE
*Statist. Med.* **17**, 2661–2674 (1998)

# STATISTICAL REVIEWING FOR MEDICAL JOURNALS

DOUGLAS G. ALTMAN*

*ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford OX3 7LF, U.K.*

## SUMMARY

This paper reviews the difficulties associated with being a statistical reviewer for a medical journal. As background, I consider first the use of statistical reviewers by medical journals, medical journals' policies on statistical peer review, and the limited evidence of its effectiveness. The assessment of a manuscript is considered under the headings of design, methods of analysis, presentation and interpretation, with many illustrative examples of the difficulties to be overcome. I emphasize the judgemental nature of many aspects. I suggest how to present and structure the reviewer's report to the editor. Finally, I consider wider issues, including the various other ways in which statisticians can interact with medical journals. © 1998 John Wiley & Sons, Ltd.

## STATISTICS IN MEDICAL PAPERS

Over the last 40 years there has been a great increase in the use of statistics in medical research, and thus in papers published in medical journals. For example, between 1952 and 1982 the proportion of papers in *Pediatrics* using statistical methods beyond descriptive statistics quadrupled, and there was a vast increase in the use of more advanced methods.[1] By 1982 only half of the research papers could be understood by somebody familiar with only simple statistical methods (dispersion, $t$, $\chi^2$ tests or correlation). A comparison of the *New England Journal of Medicine* in 1978–1979 and 1990 also revealed dramatic changes in the use of statistics.[2] In particular there was now much greater use of complex methods such as logistic regression and proportional hazards regression for survival data. The trend towards greater complexity (sophistication?) has continued.

The number of medical journals continues to increase. If all of this research were sound, even if 'worthy but dull', perhaps not too much harm would be done. The reality, though, is that there is a wealth of evidence that much published research is methodologically unsound.[2,3] Over the last 30 years or so there have been many published reviews of the quality of statistics in medical journals. Although there were some earlier studies, the first influential review was probably that of Schor and Karten,[4] in which statistical problems were found in the majority of papers. There have been many further studies of this kind. Table I gives a brief summary of the findings of some

* Correspondence to: Douglas G. Altman, ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford OX3 7LF, U.K. E-mail: altman@icrf.icnet.uk

CCC 0277–6715/98/232661–14$17.50

Table I. Summary of some reviews of the quality of statistics in medical journals, showing the percentage of 'acceptable' papers (of those using statistics)

| Year published | First author | Number of papers | Number of Journals | % papers acceptable |
|---|---|---|---|---|
| 1966 | Schor[4] | 295 | 10 | 28 |
| 1977 | Gore[5] | 77 | 1 | 48 |
| 1979 | White[6] | 139 | 1 | 55 |
| 1980 | Glantz[7] | 79 | 2 | 39 |
| 1982 | Felson[8] | 74 | 1 | 34 |
| 1982 | MacArthur[9] | 114 | 1 | 28 |
| 1983 | Tyson[10] | 86 | 4 | 10 |
| 1985 | Avram[11] | 243 | 2 | 15 |
| 1985 | Thorn[12] | 120 | 4 | < 40 |
| 1988 | Murray[13] | 28 | 1 | 61 |
| 1988 | Morris[14] | 103 | 1 | 34 |
| 1995 | McGuigan[15] | 164 | 1 | 60 |
| 1996 | Welch[16] | 145 | 1 | 30 |

of those reviews which quantified the proportion of papers with important statistical errors. There was considerable variation in the scope of these studies and in what the reviewers termed an error, so that we should probably not attempt a numerical summary. Nevertheless, however we look at these reviews they clearly point to a major problem in statistical analysis and reporting. Few of the reviews considered statistical design except for aspects of clinical trials. This may be that it is easier to assess analysis than design, or that (as discussed below) quality of design may be more subjective, and thus harder to assess, than quality of analysis.

Table I shows little (if any) evidence of rising standards but, as noted, the nature of statistics in medical journals has changed. It is likely that the general understanding of basic statistical methods ($t$ and $\chi^2$ tests, for example) has improved, even though errors still occur in the use of these simple methods. The increased use of more complex methods, such as survival analysis and multiple regression, has led to new problems, many of which cannot be detected in published papers. Both the existence and content of Andersen's book *Methodological Errors in Medical Research*[17] serve to highlight further the need for higher standards in research. In addition to this direct evidence of poor research methods, there have been several studies of the statistical knowledge of doctors, such as that by Wulff *et al.*,[18] which have consistently shown that few doctors have a good grasp of even basic statistical ideas.

The main reason for the plethora of statistical errors is that the majority of statistical analyses are performed by people with an inadequate understanding of statistical methods. They are then peer reviewed by people who are generally no more knowledgeable. Sadly, much research may benefit researchers rather more than patients, especially when is carried out primarily as a ridiculous career necessity.[19]

## MEDICAL JOURNALS' POLICIES REGARDING STATISTICAL REVIEW

Peer review of manuscripts submitted to scientific journals has become standard practice in recent times. In parallel with the increased use of statistics, medical journals have increasingly

 

brought statisticians into the review process. There is, however, very little evidence about how much use is made of statisticians or what the effect of this effort is, nor is there much published guidance about what is expected from the statistician who serves in this capacity. I will address these issues in turn.

First, a semantic point. I prefer the term reviewer to referee, as their role is to assess (review) the quality of the manuscript, and by such means to help the editor to decide whether or not to accept a paper, not to make this decision themselves. The process is widely known as peer review, not peer refereeing. Also, for a statistician, the majority of the effort is put into improving the quality of papers that are published, not determining which should be rejected. However, I will use the two terms interchangeably.

There have been two surveys of the use of statistical reviewers by medical journals. George[20] surveyed 98 journals with a high citation impact factor, with an 85 per cent response rate. The general picture was that most papers were not seen by a statistician. He made several recommendations for journals:

(i) they should require that all papers have statistical review;
(ii) they should recruit qualified statisticians as reviewers;
(iii) reviewers should see the revised manuscript (or be offered the option);
(iv) they should publish their policy on statistical refereeing;
(v) they should adopt written guidelines for statistical reporting.

At that time few of these practices were common, and he noted that it might not be easy to achieve them. A second survey about ten years later (1993–1995) found some marked changes.[21] This survey of 171 journals (67 per cent response rate) was of a different sample of journals, although it had similar inclusion criteria and it used similar and often identical questions.

The most notable change was an increase from 15 per cent to 37 per cent in the journals which had a policy that meant that all published papers had received statistical review. In the later survey 29 per cent of the journals reported that over half of their published research papers had been statistically reviewed. None the less, it is clear that a high proportion of papers, even in these high impact journals, are still published without such a review. About two-thirds of the editors in the later survey thought that statistical review led to important changes in over half of the papers reviewed. Apart from the evidence from these surveys, some journals, in particular the general medical journals, publish their policy on statistical review. A few journals indicate that all papers will undergo statistical review before acceptance for publication.

Some journals also have policies regarding specific statistical issues. Some of these will reflect the recommendations in the 'Vancouver' guidelines.[22] Perhaps the most common is the requirement that authors should provide confidence intervals with their main results. The *British Medical Journal* (*BMJ*) will not accept non-randomized trials when randomization was feasible.[23]

It is likely that many more journals would like to make use of statisticians in their review process. However, there seems to be a shortage of statisticians who are both available and willing to do this work.

## THE EFFECT OF STATISTICAL REFEREEING

It seems self-evident that statistical refereeing must be beneficial, but there have been surprisingly few studies of the effect of statistical refereeing on the quality of papers published in medical journals. Schor and Karten[4] reviewed 149 papers in 10 journals. They considered that 28 per cent

 

of them were statistically acceptable, 68 per cent needed revision, and 5 per cent were 'unsalvageable'. After the institution of a statistical refereeing programme at the *Journal of the American Medical Association* they repeated the exercise (on a smaller sample) and found that 74 per cent of papers were statistically acceptable. They reported that a further before–after study at another journal showed an improvement from 35 per cent to 70 per cent acceptable.

Gardner and Bond[24] reported a pilot study of 45 papers assessed both as originally submitted to the *BMJ* and as actually published. They found that 11 per cent were acceptable as submitted and 84 per cent acceptable as published. In 4 of the 7 papers unacceptable as published, criticisms by the statistical referee had not been adequately dealt with.

Surprisingly, these two studies – one 30 years old and the other rather small – provide the only direct evidence I am aware of about the effect of statistical input into peer review.

The preceding comments might suggest that statistical review is the solution to the ills of the medical literature. This is not so, for several reasons. Most obviously, many aspects of manuscript review are subjective, an issue I consider in more detail below. In addition, reviewing papers is not easy. Statisticians rarely receive explicit training in this role, although they will probably take part in various activities relating to critical appraisal. There have been occasional papers which give a personal view on the issues involved in being a referee;[13,25,26] here I will offer another, rather more detailed, view.

## ASSESSING THE STATISTICAL QUALITY OF A PAPER

Research papers in medical journals almost all follow the same structure: introduction, methods, results, and discussion. From the statistical perspective the only information of direct relevance in the introduction is likely to be the aim(s) of the study. The methods section should describe the study's objectives and all aspects of the design of the study, including which statistical methods of analysis were used. The results section naturally will contain the results, and the presentation of these needs to be considered. The interpretation of the results will usually appear in the discussion. Lastly, the abstract of the paper (which comes first) should provide a brief but honest summary of the methods, results and conclusions. That is the theory; in practice, almost any information can be found in any section of the paper, and it is not uncommon to find information given in the abstract that does not appear elsewhere or that disagrees with the main paper.

The main areas for the statistical reviewer to consider are design, methods of analysis, presentation of results and interpretation. In an attempt to help discussion of these I have tried to categorize criticisms as relating to 'definite errors', matters of judgement, and poor reporting (minor points, but not necessarily trivial).

Having made these categories, I am not completely comfortable about using them. In particular, the 'definite' errors are mostly not as definite as the name implies. Some require judgement too, for example in deciding what is 'inappropriate' or 'inadequate'. Also, definite errors are not necessarily important ones. I use the category of reporting errors to indicate either minor technical matters or aspects of completeness of reporting which could be detected by a suitably trained sub-editor.

Tables II to V show some examples of each category for study design, analysis, presentation, and interpretation. In each case I will comment on just a few of these items. My focus will be on problems with what was done or how it was reported. However, errors of omission are also common and may be just as serious.[27]

 

Table II. Some examples of errors in design

| |
|---|
| *Definite errors* |
| Failure to use randomization in a controlled trial |
| Use of an inappropriate control group |
| Use of a crossover design for a study of a condition that can be cured, such as infertility |
| Failure to anticipate regression to the mean |
| *Matters of judgement* |
| Is the sample size large enough? |
| Is the response rate adequate? |
| *Poor reporting* |
| Study aims not stated |
| Justification of sample size not given |
| In a controlled trial, method of randomization not stated |

Table III. Some examples of errors in analysis

| |
|---|
| *Definite errors* |
| Unpaired method for paired data |
| Using a *t*-test for comparing survival times (some of which are censored) |
| Use of correlation to relate change to initial value |
| Comparison of *P*-values |
| Failure to take account of ordering of several groups |
| Wrong units of analysis |
| *Matters of judgement* |
| Whether to suggest that the authors adjust the analysis for potential confounding variables? |
| Is the rationale for categorization of continuous variables clear? |
| Are categories collapsed without adequate justification? |
| Is use of parametric methods acceptable for data that are non-Normal (for example, skewed or ordinal)? |
| *Poor reporting* |
| Failure to specify all methods used |
| Wrong names for statistical methods, such as variance analysis, multivariate analysis (for multiple regression) |
| Misuse of technical terms, such as quartile |
| Citing non-existent methods such as 'arc sinus transformation' and 'impaired *t*-test' (seen in published papers) |
| Referring to unusual/obscure methods without explanation or reference |

## Design

This category well illustrates the softness of some of these issues. For example, the failure to state study aims is common. However, in some cases this could be a very serious error while in other cases the aim would be quite obvious.

Consider a case-control study in which the minimum age of controls was the mean age of the cases, and cases and controls were from different geographical areas.[28] This seems a clear case

 

Table IV. Some examples of errors in presentation

| |
|---|
| *Definite errors* |
| Giving SE instead of SD to describe data |
| Pie charts to show the distribution of a continuous variable |
| Results given only as *P*-values |
| CI given for each group rather than for the contrast |
| Use of scale changes or breaks in histograms |
| Failure to show all points in scatter diagrams |
| *Matters of judgement* |
| Would the data be better in a table or a figure? |
| Should we expect authors to have considered (and commented on) goodness-of-fit? |
| *Poor reporting* |
| Numerical results given to too many or, occasionally, too few decimal places |
| $r$ or $\chi^2$ values to too many decimal places |
| $P = \text{NS}$, $P = 0{\cdot}0000$ etc. |
| Reference to 'non-parametric data' |
| Tables that do not add up, or which do not agree with each other |

Table V. Some examples of errors in interpretation

| |
|---|
| *Definite errors* |
| Failure to consider confidence interval when interpreting non-significant difference, especially in a small study |
| Drawing conclusions about causation from an observed association without supporting evidence |
| Interpreting a poor study as if it was a good one (for example, a small study as a large one, a non-randomized study as a randomized one) |
| *Matters of judgement* |
| Have the authors taken adequate account of possible sources of bias? |
| How should multiplicity be handled (for example, multiple time points or multiple groups)? |
| Is there over-reliance on *P*-values? |
| *Poor reporting* |
| Discussion of analyses not included in the paper |
| Drawing conclusions not supported by the study data |

of an inappropriate control group, but how different do the groups have to be before such a judgement is made? So this may be a definite error but it may also be a matter of judgement.

Likewise, failure to specify the method of randomization may no longer be seen as a minor matter. There is empirical evidence showing that the quality of the randomization procedure relates to trial findings, with trials which do not report concealed allocation obtaining larger treatment effects.[29]

 

## Methods of analysis

First, it is essential that authors specify which statistical methods they used. This is a basic requirement of the widely adopted uniform guidelines, often called the 'Vancouver guidelines', which includes the statement: 'Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results'.[22]

Using an unpaired method (such as the two-sample *t*-test) for comparing two sets of paired observations may seem a definite error, but there are situations when the pairing is more cosmetic rather than actual – for example, in a case-control study comparing cases who are new-born babies with control babies who happen to be the next baby born in that hospital. Also, in some situations, paired analysis is either very difficult or simply impossible.

Examples of the comparison of *P*-values include subgroup analyses, within-group analyses of change over time (for example, from baseline in RCT), and a serial (repeated) measurements analysed independently at multiple time points. In each case comparisons are explicitly or implicitly made between *P*-values.[30] I have even encountered sets of (non-independent) *P*-values compared by a further test.[31] While basing inferences on the comparison of *P*-values is common, in my view it is (almost?) never justified. As a referee I would not let it pass.

In some conditions it is possible to take several measurements on the same patient, but the focus of interest usually remains the patient. Failure to recognize this fact results in multiple counting of individual patients and can lead to seriously distorted results. Analysis ignoring the multiplicity violates the widespread assumption of statistical analyses that the separate data values should be independent. Also, the sample size is inflated, sometimes dramatically so, which may lead to spurious statistical significance. There can be problems of interpretation too. Commenting on one trial, Andersen wryly observed that ' … this trial resulted in the apparent conclusion that after 1 year 22 per cent of the patients, but only 16 per cent of the legs, have expired'[17]. The issue of units of analysis is of course also an issue of design as well as analysis, showing that these categories too are not so clear-cut.

Other areas which cause difficulty, among many I could mention, include judgements about the use of parametric or non-parametric methods, and the use (or non-use) of Bonferroni-type corrections.

## Presentation

As examples of spurious precision, I have seen a mean gestational age at birth given as 40 weeks, 5·68 days. (Note that 0·01 day is about 15 minutes.) Likewise the regression equation birthweight $= -3{\cdot}0983527 + 0{\cdot}142088$ chest circumference $+0{\cdot}158039$ midarm circumference purports to predict birthweight to 1/1000000 g. Note here the common error of giving the constant to greatest precision.

Poor presentation may provide clues that there may be serious errors elsewhere. This might include evidence that results have been taken straight from computer printout (for example, spurious precision, *P*-values of 0·0000, use of * in regression equations), the presence of a large number of reporting errors, many numerical errors, and even outright stupidity. To illustrate this last category, Andersen[17] refers to a paper which summarized patient characteristics before and two years after jejuno-ileal bypass operation. The authors reported a highly significant reduction in weight, a non-significant change in height, and a highly significant increase in age of about two years!

 

## Interpretation

The failure to draw appropriate inferences from a non-significant result can be illustrated by the study of Sung *et al.*[32] They randomized patients to octreotide infusion or emergency sclerotherapy for acute variceal haemorrhage. They reported that they would have needed 900 patients per group to have reasonable power to detect an improvement in response rate from 85 per cent to 90 per cent. As this was way beyond what they could achieve, they 'arbitrarily set a target of 100 patients and accepted a chance of a type II error'. The observed rates of controlled bleeding were 84 per cent in the octreotide group and 90 per cent in the sclerotherapy group, giving $P = 0{\cdot}55$. They quoted a confidence interval (CI) for the treatment difference as 0 to 19 per cent – it should have been $-7$ per cent to 19 per cent. More seriously, they drew the unjustified conclusion that 'octreotide infusion and sclerotherapy are equally effective in controlling variceal haemorrhage'.

Another common difficulty is the interpretation of data-derived analyses – analyses not prespecified and generally suggested by the results obtained. For example, Newnham *et al.* carried out a large randomized trial comparing routine and intensive ultrasound during pregnancy.[33] They found significantly more low birthweight babies in the frequent ultrasound group (although the difference in mean birthweight was negligible). This was not one of the main outcomes and indeed was the only one of more than 20 variables they looked at to show a significant difference. Most of the paper's discussion was about birthweight. Incidentally, one of the authors' arguments in favour of this being a genuine association was plausibility. This is an almost worthless argument – doctors can find a plausible explanation for any finding.[34] Analyses suggested by the data should be acknowledged as exploratory; for generating hypotheses rather than testing them.

In addition, there are some problem areas that cut across the categories just discussed. For example, many difficulties arise through multiplicity involving issues such as multiple time points, multiple comparisons and subgroup analyses. These can be seen as issues of analysis or interpretation, but may stem from poor study design.

## Issues specific to certain types of study

As indicated in some of the above examples, specific considerations apply to specific study types, such as surveys, case-control studies, controlled trials and systematic reviews (meta-analyses). It helps if the reviewer is familiar with the type of study, although those who review frequently for a medical journal, especially a general medical journal, must expect to receive papers on all types of study. They should, however, always send back a paper if they feel that it goes into areas in which they are not especially confident or competent. Likewise, it is desirable if the reviewer is familiar with the medical subject matter. Specialist journals usually recruit statistical referees who work in the relevant medical specialty, and it is certainly helpful to be familiar with the medical issues being addressed. Again, this will not apply to general journals, where the regular reviewer may expect to receive papers on almost any medical area.

Check-lists tailored to a particular type of study can help the reviewer, if only by acting as a memory jogger. It is in general harder to spot things missing from a paper than those which are included but incorrect.

 

**Referee's report**

The referee's comments will need to be put together as a written report, primarily for the editor but also for the authors. It is helpful to structure the report, for example by grouping comments under headings (methods, results, discussion, abstract). It is also helpful to indicate for each comment the relevant page(s) of the manuscript. I find it useful to have a large number of separated comments rather than long paragraphs. If the referee numbers the comments, assessment of the revised version is greatly aided if authors are asked, by the editors, to respond in a covering letter to each of the referee's points in turn.

The referee should be constructive. For example, it is better to indicate how the analysis could be improved than simply observing that the present analysis is incorrect. The referee should use language that the authors will understand; technical terms such as interaction and heteroscedasticity should be avoided.

It is valuable to indicate which are the most important criticisms. While the relative importance may be obvious to a statistician, it is unlikely to be so for either editors or authors, who in general will be equally ignorant of statistical niceties. Also, further to my preceding classification, the reviewer should try to distinguish cases where they feel that there is a definite error from those where it may be felt preferable to do something different.

A common difficulty is that key information is missing from a manuscript. The reviewer should point out the deficiencies, and request that the authors rectify the omissions. Quite often this can reveal new problems, occasionally fatal ones, which is one of the main reasons for the reviewer to see the revised manuscript. At the *BMJ* statistical referees are asked to say if they wish to see a revision, if there is one.

Should the reviewer express an opinion about whether the editor should accept or reject the paper? Some journals expressly ask reviewers, including statisticians, to indicate whether they think that the paper should be accepted. As I have indicated, I do not think that this is in general the role of the reviewer. However, there are occasions when I feel strongly that a paper should be rejected. In such cases I draw attention to this opinion and the reasons for it in a covering letter to the editor. The editors of one journal noted that 'Biomedical statisticians ... probably come nearest to having the veto on the publication of a paper ... '.[35] While rejection may be suggested because of fatal methodological flaws, I have occasionally encountered dishonesty, such as failing to mention that most or all of the results have been published already, or changing some important aspect of the stated design in comparison with an earlier paper. Such behaviour is unacceptable. Very occasionally the reviewer may encounter results that suggest outright fraud. Such suspicions should naturally be discussed with the editor.

The *BMJ* published two check-lists for statistical referees to use.[36] They should also be useful for authors and editors. Table VI shows a summary of 100 check-lists I completed for papers refereed for the *BMJ* during 1991–1993. There was clearly considerable room for improvement, but I felt that only one paper was unsalvageable, to use the term of Schor and Karten.[4]

I have discussed many issues, but there are many troublesome questions about refereeing for which there are no simple answers. These include:

(i) How much of a purist should the referee be (especially if it is unlikely that the 'correct' analysis will alter the findings)?
(ii) How much detail should the referee require of highly complex analyses that would not be understood by most readers?

 

Table VI. Summary of check-lists for 100 consecutive papers (excluding controlled trials) reviewed for the *British Medical Journal* ('not relevant' or missing answers are excluded)

| | Yes | Unclear | No |
|---|---|---|---|
| Objective clear? | 83 | 6 | 11 |
| Appropriate study design? | 72 | 25 | 3 |
| Source of subjects? | 83 | 6 | 10 |
| Sample size calculation? | 0 | 0 | 63 |
| Satisfactory response rate? | 49 | 23 | 2 |
| Methods described adequately? | 47 | — | 53 |
| Statistical analyses appropriate? | 41 | 37 | 22 |
| Statistical presentation satisfactory? | 14 | — | 86 |
| Confidence intervals given? | 51 (+8*) | — | 41 |
| Conclusions justified? | 40 | 49 | 11 |
| Paper statistically acceptable? | 4 | — | 96 |
| If not could it become acceptable? | 89 | 6 | 1 |

* Confidence intervals given inappropriately

(iii) Should the referee take account of the expectation that the authors have no access to expert statistical help? If so, how?
(iv) How should the referee deal with papers using methods that are (widely) used by other statisticians but which he/she does not approve of?
(v) When is a definite error a fatal flaw?

The reviewer will need to address such questions as best they can when they arise.

## CONCLUDING REMARKS

Some years ago I encountered a very thought-provoking (unpublished) quotation attributed to Michael Healy:

> 'The difference between medical research and agricultural research is that medical research is done by doctors but agricultural research is not done by farmers.'

While we cannot assume that all agricultural research is impeccable, there seems little doubt that many of the ills of the medical literature are due to the fact that much medical research is carried out by clinicians with little training in research methods, primarily as a career necessity. There is clear evidence of the harmful effects of this situation, but no evidence of any initiatives which will make any impact.[19]

The absence of professional researchers in so much medical research points to a clear need for statisticians to be involved in research at some stage. As numerous statisticians have pointed out over the past 60 years at least, the earlier the involvement the better.[3,7]

Most statistical errors are probably relatively unimportant, but some can have a major bearing on the validity of a study, especially errors in design. Further exploration of the nature of statistical errors, their causes and possible remedies are considered elsewhere.[2,3,38]

 

To what extent can and should journals intervene to try to stop bad research getting published? The most obvious way is by using statistical referees to assess the quality of submitted manuscripts. The aims here are to avoid publishing studies which are unsound or unreliable and to improve as far as is practicable the quality of the papers which are published. As I have discussed, the assessment of quality is a highly subjective affair, and all authors will be familiar with the differences of opinion between reviewers of the same manuscript. Despite its evident shortcomings, I believe that statistical review is a valuable part of the publication system.

As Finney[26] has recently noted, there is very little published on the role of the statistical referee. Exceptions are papers by Vaisrub[25] and Murray,[13] although the latter, despite its title, is more a review of errors encountered than comments on the nature of reviewing a manuscript. As is probably clear from the preceding material, refereeing papers is not easy, yet statisticians can expect little or no training in this role. However, refereeing is a form of critical appraisal, and this is an especially important skill for a statistician. Reviewing manuscripts can thus be educational, and I would recommend people to try it if they get the opportunity. However, it can be very time-consuming and may not be especially rewarding unless one is actually interested in the papers (which may well not be the case).

Refereeing may be getting harder. Papers are likely to include much more statistical material than previously, and new techniques continue to be developed and absorbed into medical research. These can pose considerable difficulties for reviewers. Some more or less recent statistical techniques include: the bootstrap; Gibbs sampling; generalized additive models; classification and regression trees; generalized estimating equations; multi-level models; and neural networks.[39] Not only may a paper describe unfamiliar methods, these may be described in inadequate detail to judge if their use was appropriate.

Statistical reviewing seems like an area needing research. There has been very little research into the benefits of statistical review, and none that I know of relating to the manner in which statistical review is carried out. More importantly, perhaps, research is needed into how best to improve the quality of papers submitted to medical journals and indeed improve the quality of the research itself. To this end some journals try to influence the quality of submissions. Most obviously, many journals include some statistical issues in their instructions to authors. When I reviewed many of these some years ago I found that most such sections were very brief. They included some rather surprising statements, and some suggestions that were, to say the least, opaque. Two examples of the latter type are:

> 'When possible give the range, SD (standard deviation) or ME (mean error), and significance of differences between numerical values.'
> 'To aid in the review process, include the statistical worksheet (not for publication), if applicable.'

Many, though were quite sensible, including:

> 'Authors should beware of placing undue emphasis on secondary analyses, especially when they were suggested by an inspection of the data.'
> 'Error bars should not be used to represent SEM etc. – bars in graphs and histograms should represent 95 per cent confidence intervals.'
> 'The decision to publish is not based on the direction of the results.'

If authors read the instructions to authors and took heed of them, the task of the reviewer would be eased. However, personal experience suggests that many authors do neither of these things.

 

Occasional submissions come from authors who seem never to have seen a copy of the journal let alone its instructions for submission.

Statistical guidelines are another way in which journals can try to influence the quality of papers submitted. Some general guidelines have been published.[40–42] Many published guidelines have focused on specific areas, for example clinical trials (for example, Grant[43]). The most recent example is the CONSORT statement,[44] which is unique in being 'adopted' by over 70 journals by the end of 1997. Here adoption implies that journals require authors to comply with the recommendations for reporting trials. This status may at least partly reflect the fact that, unusually, journal editors were among the authors, but it probably also relates to the widespread recognition that the reporting of controlled trials is generally inadequate for those carrying out systematic reviews. Similar publications covering other study designs are likely in coming years. Some journals have adopted the earlier suggestion[45] that authors are required to complete a check-list when they submit their paper, indicating that the CONSORT requirements have been met and where each item can be found. This principle could usefully be extended to all types of study, with authors required to complete a check-list to show that they had dealt with certain aspects of statistics, such as specifying all the methods used, providing confidence intervals, and so on.

Reviewing manuscripts is only one of many ways in which statisticians can help medical journals. Other aspects include:

(i) helping to formulate journal policy;
(ii) auditing the quality of statistics in published papers (generally and in specific areas);
(iii) helping to produce statistical guidelines or check-lists for authors;
(iv) educating editors;
(v) providing explanatory statistical comments on published papers/letters;
(vi) writing expository articles about statistical matters.

Interaction with journals in this way can be highly rewarding, especially when working with one journal over many years.

As a final comment, I would summarize reviewing medical papers as difficult, time-consuming, sometimes interesting, sometimes boring, appreciated by journals, appreciated by authors (but perhaps not appreciated by employers), usually unpaid, occasionally frustrating, and educational. Many journals are desperate for expert statistical help. I recommend statisticians to try it.

ACKNOWLEDGEMENTS

I thank Mike Campbell and David Finney for helpful suggestions.

## REFERENCES

1. Hayden, G. F. 'Biostatistical trends in pediatrics: implications for the future', *Pediatrics*, **72**, 84–87 (1983).
2. Altman, D. G. 'Statistics in medical journals – developments in the 1980s', *Statistics in Medicine*, **10**, 1897–1913 (1991).
3. Altman, D. G. 'Statistics in medical journals', *Statistics in Medicine*, **1**, 59–71 (1982).
4. Schor, S. and Karten, I. 'Statistical evaluation of medical manuscripts', *Journal of the American Medical Association*, **195**, 1123–1128 (1966).

 

5. Gore, S., Jones, I. G. and Rytter, E. C. 'Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976', *British Medical Journal*, **1**, 85–87 (1977).
6. White, S. J. 'Statistical errors in papers in the *British Journal of Psychiatry*', *British Journal of Psychiatry*, **135**, 336–342 (1979).
7. Glantz, S. 'Biostatistics: how to detect, correct, and prevent errors in the medical literature', *Circulation*, **61**, 1–7 (1980).
8. Felson, D. T., Cupples, L. A. and Meenan, R. F. 'Misuse of statistical methods in *Arthritis and Rheumatism*. 1982 versus 1967–68', *Arthritis and Rheumatism*, **27**, 1018–1022 (1984).
9. MacArthur, R. D. and Jackson, G. G. 'An evaluation of the use of statistical methodology in the *Journal of Infectious Diseases*', *Journal of Infectious Diseases*, **149**, 349–354 (1984).
10. Tyson, J. E., Furzan, J. A., Reisch, J. S. and Mize, S. G. 'An evaluation of the quality of therapeutic studies in perinatal medicine', *Journal of Pediatrics*, **102**, 10–13 (1983).
11. Avram, M. J., Shanks, C. A., Dykes, M. H. M., Ronai, A. K. and Stiers, W. M. 'Statistical methods in anesthesia articles: An evaluation of two American journals during two six-month periods', *Anesthesia and Analgesia*, **64**, 607–611 (1985).
12. Thorn, M. D., Pulliam, C. C., Symons, M. J. and Eckel, F. M. 'Statistical and research quality of the medical and pharmacy literature', *American Journal of Hospital Pharmacy*, **42**, 1077–1082 (1985).
13. Murray, G. D. 'The task of a statistical referee', *British Journal of Surgery*, **75**, 664–667 (1988).
14. Morris, R. W. 'A statistical study of papers in the Journal of Bone and Joint Surgery (BR)', *Journal of Bone and Joint Surgery* (*BR*), **70-B**, 242–246 (1988).
15. McGuigan, S. M. 'The use of statistics in the *British Journal of Psychiatry*', *British Journal of Psychiatry*, **167**, 683–688 (1995).
16. Welch, G. E. and Gabbe, S. G. 'Review of statistics usage in the *American Journal of Obstetrics and Gynecology*', *American Journal of Obstetrics and Gynecology*, **175**, 1138–1141 (1996).
17. Andersen, B. *Methodological Errors in Medical Research*, Blackwell, Oxford, 1990.
18. Wulff, H. R., Andersen, B., Brandenhoff, P. and Guttler, F. 'What do doctors know about statistics?', *Statistics in Medicine*, **6**, 3–10 (1987).
19. Altman, D. G. 'The scandal of poor medical research', *British Medical Journal*, **308**, 283–284 (1994).
20. George, S. L. 'Statistics in medical journals: a survey of current policies and proposals for editors', *Medical and Pediatric Oncology*, **13**, 109–112 (1985).
21. Goodman, S. N., George, S. L. and Altman, D. G. 'Statistical reviewing policies of medical journals: caveat lector?', *Journal of General Internal Medicine*, in press.
22. International Committee of Medical Journal Editors. 'Uniform requirements for manuscripts submitted to biomedical journals', *Journal of the American Medical Association*, **277**, 927–934 (1997).
23. Altman, D. G. 'Randomisation', *British Medical Journal*, **302**, 1481–1482 (1991).
24. Gardner, M. J. and Bond, J. 'An exploratory study of statistical assessment of papers published in the *British Medical Journal*', *Journal of the American Medical Association*, **263**, 1355–1357 (1990).
25. Vaisrub, N. 'Manuscript review from a statistician's perspective', *Journal of the American Medical Association*, **253**, 3145–3147 (1985).
26. Finney, D. J. 'The responsible referee', *Biometrics*, **53**, 715–719 (1997).
27. Mosteller, F. 'Problems of omission in communications', *Clinical Pharmacology and Therapeutics*, **25**, 761–764 (1979).
28. Olsson, H. and Ingvar, C. 'Left handedness is uncommon in breast cancer patients', *European Journal of Cancer*, **27**, 1694–1695 (1991).
29. Schulz, K. F., Chalmers, I., Hayes, R. and Altman, D. G. 'Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials', *Journal of the American Medical Association*, **273**, 408–412 (1995).
30. Matthews, J. N. S. and Altman, D. G. 'Interaction 2: Compare effect sizes not P values', *British Medical Journal*, **313**, 808 (1996).
31. Jannink, I., Bennen, J. N., Blaauw, J., Vandiest, P. J. and Baak, J. P. A. 'At convenience and systematic random sampling – effects on the prognostic value of nuclear-area assessments in breast-cancer patients', *Breast Cancer Research and Treatment*, **36**, 55–60 (1995).
32. Sung, J. J. Y., Chung, S. C. S., Lai, C.-W., Chan, F. K. L., Leung, J. W. C., Yung, M. Y., Kassianides, C. and Li, A. K. C. 'Octreotide infusion or emergency sclerotherapy for variceal haemorrhage', *Lancet*, **342**, 637–641 (1993).

 

33. Newnham, J. P., Evans, S. F., Michael, C. A., Stanley, F. J. and Landau, L. I. 'Effect of frequent ultrasound during pregnancy: a randomised controlled trial', *Lancet*, **342**, 887–891 (1993).
34. Sackett, D. L., Haynes, R. B., Guyatt, G. H. and Tugwell, P. *Clinical Epidemiology*: *A Basic Science for Clinical Medicine*, 2nd edn, Little Brown, Boston, 1991, p. 296.
35. Selby, P. and Twentyman, P. 'The role of the clinical editor', *British Journal of Cancer*, **63**, 1–2 (1991).
36. Gardner, M. J., Machin, D. and Campbell, M. J. 'Use of check lists in assessing the statistical content of medical studies', *British Medical Journal*, **292**, 810–812 (1986).
37. Anonymous. 'Mathematics and medicine', *Lancet*, **i**, 31 (1937).
38. Altman, D. G. and Bland, J. M. 'Improving doctors' understanding of statistics (with discussion)', *Journal of the Royal Statistical Society*, *Series A*, **154**, 223–267 (1991).
39. Altman, D. G. and Goodman, S. N. 'Transfer of technology. Transfer of technology from statistical journals to the biomedical literature: past trends and future predictions', *Journal of the American Medical Association*, **272**, 129–132 (1994).
40. Altman, D. G., Gore, S. M., Gardner, M. J. and Pocock, S. J. 'Statistical guidelines for contributors to medical journals', *British Medical Journal*, **286**, 1489–1493 (1983).
41. Bailar, J. C. and Mosteller, F. 'Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations', *Annals of Internal Medicine*, **108**, 66–73 (1988).
42. Murray, G. D. 'Statistical guidelines for the *British Journal of Surgery*', *British Journal of Surgery*, **78**, 782–784 (1991).
43. Grant, A. 'Reporting controlled trials', *British Journal of Obstetrics and Gynaecology*, **96**, 397–400 (1989).
44. Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Simel, D. and Stroup, D. F. 'Improving the quality of reporting of randomized controlled trials: the CONSORT Statement', *Journal of the American Medical Association*, **276**, 637–639 (1996).
45. Altman, D. G. and Doré, C. J. 'Randomisation and baseline comparisons in clinical trials', *Lancet*, **335**, 149–153 (1990).

 