STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:3275–3289

Statistics in medical journals: some recent trends

Douglas G. Altman*,†

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

SUMMARY

I review some areas of medical statistics that have gained prominence over the last 5–10 years: meta-analysis, evidence-based medicine, and cluster randomized trials. I then consider several issues relating to data analysis and interpretation, many relating to the use and misuse of hypothesis testing, drawing on recent reviews of the use of statistics in medical journals. I also consider developments in the reporting of research in medical journals. Copyright © 2000 John Wiley & Sons, Ltd.

INTRODUCTION

Medical statistics continues to evolve both in terms of methodological development and also in the use of statistics in medical research. In this paper I will consider first a few issues that have been prominent in medical statistics in the last 5–10 years: meta-analysis, evidence-based medicine, and cluster randomized trials. I then consider several issues relating to data analysis and interpretation, including recent reviews of the use of statistics in medical journals. Finally, I consider developments in the reporting of research in medical journals. I make no pretence of being comprehensive.

THE CONTINUING RISE OF META-ANALYSIS

Glass coined the term 'meta-analysis' in the 1970s to describe the process of gathering and combining information from many studies of the same type [1]. Among the earliest medical meta-analyses on Medline was one of 42 studies of treatment for stuttering published in 1980 [2]. Meta-analysis was beginning to make an impact in medicine towards the end of the 1980s [3], and such studies have become very familiar during the 1990s. With hindsight, it is hard to understand how the basic idea of reviewing all the relevant literature to answer a question – a self-evidently sensible approach – is such a modern development (with a few notable exceptions). One reason for the

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

Table I. Distinct stages of a systematic review (slightly adapted from Deeks [4]).

Specification of the question to be answered
Formulation of eligibility criteria that can be used to objectively identify studies with data appropriate to the chosen question
Production of a protocol which states the study selection criteria and all of the methods that will be used
Rigorous searching for all relevant trials
Evaluation of the results of the searches for relevant studies which fit the eligibility criteria.
Evaluation of studies for their quality and susceptibility to bias
Extraction of relevant data from the study reports, both summarizing pertinent information about the trial design and the results
Statistical combination of the data (meta-analysis) if appropriate and consideration of between-study differences
Investigation of the robustness of the results through plots and sensitivity analyses
Interpretation of the results

increase in meta-analyses is the new phenomenon of evidence-based medicine, of which more below, and the consequent heightened awareness of the desirability not just for evidence but reliable evidence obtained from all relevant studies. Meta-analysis is now a standard Medline publication type; Medline contains 589 such papers published in 1997 and 713 in 1999 (as of September 2000). These figures suggest a continuing increase in the level of activity.

While the term meta-analysis is still widely used to encompass the whole process of reviewing a body of similar studies, from study identification through to data analysis and interpretation, the term is somewhat ambiguous. In recent years the term 'systematic review' has been introduced. It is a useful term for the whole review process, as outlined in Table I, with the term meta-analysis reserved for the actual numerical combination of data from many studies. Unfortunately, we now have the situation where some people use meta-analysis in this restricted way while others use it to describe the whole process – synonymous with 'systematic review'.

As shown in Table I, a systematic review is a structured process with several key steps [4]. Most of the steps require judgement, so that a systematic review is not as objective as many would wish it to be. Meta-analysis, the statistical combination of the results of separate studies, is not an essential element of a systematic review. In particular cases meta-analysis may not be advisable because of important differences between the studies, poor quality of the studies, or perhaps because the results of the studies are very heterogeneous. Nevertheless, the other elements of a systematic review can and should be adhered to. It seems, though, that some investigators are reluctant to stop short of formal meta-analysis. For example, Poole and Greenland have noted that 'Meta-analysts rarely entertain the possibility of concluding that study-specific results are too heterogeneous to aggregate' [5].

The rise in the number of systematic reviews and meta-analyses has been accompanied by a considerable amount of methodological work, both theoretical and empirical. At the end of 1999 the database of methodological studies on the Cochrane Library included 1124 references (including conference abstracts) [6], yet this is surely a considerable underestimate of the amount of published

material. Statistical methodology for systematic reviews has been reviewed at length by Sutton *et al*. [7] and many practical issues are considered in Egger *et al.* [8].

Despite the huge uptake of the approach there remain many fundamental practical difficulties, and we should expect a continuation of the large literature of methodological and empirical research related to systematic reviews and meta-analysis. Some questions that continue to attract attention are considered in the next sections.

Meta-analyses of trials with binary endpoints

For meta-analyses of randomized trials with binary outcomes there is a continuing debate about which measure of effect to use. The main contenders are the odds ratio, the relative risk (risk ratio) and the risk difference. Although the risk difference (also called the absolute risk reduction) is routinely used in power/sample size calculations for individual trials, it is generally felt that this measure is often inappropriate for meta-analysis. Empirical evidence suggests that treatment effect is more likely to be similar across trials when using a relative rather than an absolute measure of treatment benefit [9].

Choosing between alternative relative measures is not so straightforward. The odds ratio is used to approximate the risk ratio in case-control studies, where the risk ratio cannot be assessed directly and where the event of interest is usually rare. It does not necessarily follow that the odds ratio is the right measure of effect in randomized trials. For these the risk ratio can be analysed directly and events are often not rare, so why use the odds ratio for trials and meta-analyses of trials?

A suggestion that we should in general prefer the relative risk to the odds ratio [10] was heavily criticized [11]. These proponents of the odds ratio gave several arguments for the superiority of the odds ratio, but no direct evidence. There is, in fact, empirical evidence that across many hundreds of meta-analyses the odds ratio and risk ratio are about equally likely to be the better measure, in the sense of showing less heterogeneity of treatment effect across trials [9].

The odds ratio is of course an entirely valid measure, and is often the measure of choice. However, its use can mislead when it is interpreted as a risk ratio (for single studies or meta-analyses) [10, 12].

Importance of quality of primary studies

An important element of a systematic review should be the assessment of the methodological quality of the primary studies. Two difficult issues here are how to do it and what to do with the information. There are numerous published scales for assessing quality of randomized trials [13, 14]. Nurmohamed *et al.* [15] found a significant reduction in the risk of deep vein thrombosis with low molecular heparin. Jüni *et al.* [14] re-analysed this meta-analysis using the 25 quality scales identified by Moher *et al*. [13]. They compared the results of 'high' and 'low' quality trials using each scale in turn, and found that the results varied considerably depending on which scale was used. Using some scales, 'high quality' trials indicated that low molecular weight heparin was not superior to standard heparin, whereas 'low quality' trials showed better protection with the low molecular type. With other scales the opposite was found. Clearly, some pairs of these scales of quality are negatively correlated. Much of the problem here stems from the inclusion in many scales of aspects of reporting and methodology which would not be expected to be related to the study findings. Although some researchers have used quality scores as weights in meta-analysis [16], the more common view is that it is preferable to examine directly the influence of specific methodological features on the results [14, 17, 18].

Meta-analysis beyond randomized trials

While most meta-analyses have been of randomized trials, the basic principles have continued to spread to consideration of other types of study. Meta-analyses are quite common in epidemiology, although several authors have questioned their appropriateness [19, 20]. There has been increasing interest in carrying out systematic reviews and meta-analyses of diagnostic studies, for which new methods of analysis have been developed [21–23]. Meta-analyses are being performed of prognostic and economic studies, again with new methodological problems [8]. Indeed we may now expect to see systematic reviews of any type of study. In all of these cases, the methodological difficulties (including assessment of study quality) exceed those for systematic reviews and meta-analyses of randomized trials.

Analysing meta-analyses

As the number of published meta-analyses has increased a new type of study has arisen in which data from several meta-analyses are reanalysed in a single analysis. Some have termed this 'meta–meta-analysis'. Notably, this approach has been used to provide empirical evidence about some aspect of the methodology of the primary studies.

For example, a few studies have produced empirical evidence that aspects of design are related to research findings. Schulz *et al.* [24] reanalysed 250 randomized trials from 33 meta-analyses, and found that the treatment effect was 30 per cent to 41 per cent larger in trials without adequate concealment of treatment allocation. Also there was a 17 per cent bias in trials that were not double blind. Moher *et al.* [16] used a similar approach to investigate 127 randomized trials from 11 meta-analyses. They found that the treatment effect was 34 per cent larger in low quality trials than in high quality trials and 37 per cent larger in trials which did not report adequate concealment of treatment allocation.

The same general approach can be applied to other types of study. Lijmer *et al.* [25] investigated 218 evaluations of diagnostic tests from 11 published systematic reviews. They found empirical evidence that use of a case-control design led on average to overestimate of diagnostic performance compared with cohort studies (diagnostic odds ratio 3.0). They also found bias associated with the use of different reference tests for positive and negative results of the test under study.

The same idea has been used to answer clinical questions. For example, Katerndahl and Lawler [26] examined 23 meta-analyses of the efficacy of cholesterol reduction. They found that odds ratios depended on inclusion criteria and investigator variables and observed that methodologically better meta-analyses tended to report more beneficial odds ratios.

As these studies illustrate, this type of research can be very powerful.

EVIDENCE-BASED MEDICINE (EBM)

Anyone working in medicine cannot be unaware of the recent phenomenon of evidence-based medicine (EBM). What is meant by EBM, and what is its relevance to those working in other areas?

Evidence-based medicine has been defined as 'The conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients', and further elucidated as '...integrating individual clinical expertise with the best available external clinical evidence from systematic research' [27]. As Davidoff has observed, 'the rise of evidence-based

medicine does not represent the sudden appearance of a strange new element in clinical decision-making, but rather a shift in the balance among the existing elements, with greater weight now being put on "the evidence" than had been done in previous years' [28].

Many of the statistical issues relevant to EBM relate to systematic reviews and meta-analyses, already addressed above. Others, such as interpretation of single studies, are standard. Perhaps the one additional issue of statistical interest is the 'number needed to treat' (NNT), a valuable concept introduced about ten years ago [29]. Its use has increased in recent years, especially in systematic reviews and in journals of secondary publication such as *ACP Journal Club* and *Evidence-Based Medicine* and several recent specialty-specific titles.

The idea of the NNT is very simple. The NNT for a treatment is the number of patients who would need to receive the treatment, rather than standard care or placebo, for one additional patient to benefit (equivalently, to prevent one additional adverse outcome). Note the importance of the word 'additional' in this definition. The NNT is obtained very simply as the reciprocal of the absolute risk reduction (also called the risk difference).

Despite its apparent simplicity, use of the NNT is neither without difficulty nor controversy. The difficulty comes primarily from the problem of confidence intervals, which encompass an NNT of infinity when the confidence interval for the ARR includes zero (that is, when the treatment effect is not statistically significant) [30]. The inaccurate term 'number needed to harm' (NNH) has been developed for adverse effects of an intervention. It would be preferable to replace NNT and NNH by NNTB and NNTH, where B and H refer to benefit and harm, respectively [30]. Similar measures have been developed for other types of study, such as the number needed to screen [31], to diagnose, or to follow (for prognostic studies) but none has achieved wide use as yet.

The controversy comes from the quoting of the NNT as a single measure of effect from a trial or meta-analysis. As noted above, treatment effects are more often consistent in a relative rather than absolute scale. It follows that the ARR and thus the NNT must vary across groups with different event rates. To quote a single NNT is thus misleading for heterogeneous populations [32]. This issue is especially relevant to meta-analyses, although it applies also to many individual controlled trials.

The place of non-randomized studies

Randomized trials are sometimes impossible, and even when possible may simply not have been done [33]. There is increasing interest in the value of non-randomized studies to address questions of effectiveness of health interventions [34]. Attention has focused on empirical evidence of the difference between the results of randomized and non-randomized studies, with mixed findings. It is clear that sometimes non-randomized studies do give very similar answers to randomized trials [35–37], but they may also differ markedly [20].

In hierarchies of levels of evidence, non-randomized studies provide weaker evidence than randomized studies [38, 39]. As noted already, however, not all randomized studies are done well, whereas some non-randomized studies are carried out to a high standard. A better understanding of when we can rely on non-randomized evidence would be very valuable. Instruments for assessing the quality of non-randomized studies are being developed [40] but are likely to have similar problems to quality scores for randomized trials without support from further empirical studies.

CLUSTER RANDOMIZED TRIALS

One of the main recent developments in clinical trials has been the increased uptake and understanding of the cluster randomized design [41–44]. Typically, this design may be adopted either because the intervention is targeted at health professionals but the outcome is assessed in their patients, or because individual randomization is not feasible. My (admittedly crude) Medline search suggests that there has been an increasing use of such designs in recent years, perhaps partly as a result of a greater interest in performing trials in the context of routine patient care.

Here, too, as the methods are applied more widely a greater number of problems arises, so there has been an increase in methodological studies. Some areas of continuing concern include ethical issues (notably the question of informed consent) [45], the effect of drop-outs within clusters, interim analysis and meta-analysis [41]. Papers in a forthcoming special issue of *Statistics in Medicine* address many practical issues in the conduct of such trials [46].

The sample size for a cluster randomized trial is greater than that for a conventional trial using individual randomization by a factor of $1 + (m - 1)\rho$ (known as the design effect), where m is the average cluster size and ρ is the intracluster correlation coefficient [41]. It follows that for a given total sample size many small clusters will yield more statistical power than fewer large clusters, although in practice it is often not possible to make such a choice. Because clusters are often large, even very small values of the intracluster correlation coefficient can lead to a large design effect.

The magnitude of the design effect can be surprising. For example, if $\rho = 0.05$ the design effect is 11 for a trial with clusters of size 200. The size of ρ and hence the design effect will differ according to nature of intervention. A real practical problem in designing a cluster randomized trial is that often it is hard to know what value of ρ might be plausible. Two surveys of published papers found that fewer than 20 per cent of cluster randomized studies properly accounted for clustering in sample size calculations [47, 48].

Failure to take account of the design can have a profound effect on the analysis, and can lead to seriously distorted results. There are two related consequences of ignoring the fact that the data include multiple observations on the same individuals. First, this procedure violates the widespread assumption of statistical analyses that the separate data values should be independent. Secondly, the sample size is inflated, sometimes dramatically so, which may lead to spurious statistical significance. Using the wrong units in the analysis is also a common error. Two reviews of published cluster randomized trials found this mistake in 50 per cent [47] and 57 per cent [48] of papers. (A similar problem arises in individually randomized trials with multiple measures per patient. In a review of 196 randomized trials of non-steroidal anti-inflammatory agents, Gøtzsche [49] found that 63 per cent of reports used the wrong units of analysis.)

DEVELOPMENTS IN STATISTICAL METHODOLOGY

New statistical methods may take some while to find their way into medical research, although there is some evidence to suggest that the speed of transfer has picked up in recent decades. Altman and Goodman [50] suggested that the following methods were likely to be seen more often in coming years:

(i) bootstrap (and other computer-intensive methods);
(ii) Gibbs sampler (and other Bayesian methods);

(iii) generalized additive models;
(iv) classification and regression trees (CART);
(v) models for longitudinal data (general estimating equations);
(vi) models for hierarchical data;
(vii) neural networks.

While these methods are all sometimes used in medical research, none seems yet to appear at all widely. Some, certainly, have a rather limited range of applicability compared with, say, the proportional hazards model. Also, it can take many years before the importance of new methodology becomes clear.

As far as I know, of the above methods only the use of neural networks has been subjected to a formal review. Schwartzer *et al.* [51] reviewed the oncology literature for 1991–1995 and found 173 papers using neural networks or artificial intelligence. They carried out a detailed review of the 43 papers which reported the use of feed-forward neural nets for classification. Of these 23 related to diagnosis, 6 to automatic tumour grading, 10 to prognosis, and there were 5 other applications.

They found that most papers made important errors in the application of the new technology. They observed that 'Application of artificial neural networks ... is often accompanied by grossly overstated claims, praising neural networks as the ultimate solution to the problem of diagnosis and prognosis' [51]. Despite comments by enthusiastic users of neural networks, such as '... formidable tools in the fight against cancer' and '... may be beneficial to medicine and urology in the twenty-first century as molecular biology has been in the twentieth', Schwartzer *et al.* [51] observed that 'There is no evidence that artificial neural networks have provided real progress in the field of diagnosis and prognosis in oncology'. It is interesting to note that the types of error found in this review are similar to those observed many years ago in a review of the then new technique of discriminant analysis [52].

Exhortations to adopt a Bayesian approach continue to appear as do predictions of the demise of frequentist methods. The availability online of the whole text of the *British Medical Journal* allows searching for particular methods. In the period 1996–1999 (November) there were 34 'hits' for Bayes, but only one of these was a research paper [53].

SOME ISSUES IN DATA ANALYSIS AND INTERPRETATION

Many quite basic aspects of analysis and interpretation continue to cause much confusion and, occasionally, controversy. Here I highlight a few of the most prominent, each of which is related to some degree to the still wide practice of hypothesis testing.

Inference based on comparison of P values

Comparisons between subgroups are common in medical papers. For example, in a controlled trial comparing a new treatment with a standard treatment authors may examine whether the observed benefit was the same for different subgroups of patients. A common approach to answering this question is to analyse the data separately in each subgroup and compare the two P-values. Quite often one is significant ($P < 0.05$) and the other is not, leading to the inference that the treatment is beneficial in one group but not the other. A similar approach can be seen in other types of study. For example, in a meta-analysis of trials of clozapine two subsets of trials were examined, according to whether the trial was or was not sponsored by a drug company [54]. The authors

reported that the odds of relapsing on clozapine versus comparators was 'clearly less in sponsored trials' – here the pooled odds ratio was 0.5 (95 per cent CI 0.3 to 0.7). By contrast, they reported 'equivocal findings in the non-sponsored studies', for which the odds ratio was 0.4 (95 per cent CI 0.1 to 1.4). Note that the two sets of trials had very similar results – the difference is just in the width of the confidence interval, which is determined by the amount of data available. The authors' line of reasoning is false and very likely to lead to misinterpretation, as here. Differences in *P*-values can arise because of differences in effect sizes or differences in standard errors or a combination of the two [55].

It is questionable whether it is ever justified to compare *P*-values like this. The correct approach is to consider whether there is direct (rather than indirect) evidence that the two groups differ. In other words we look for a possible interaction, here between treatment effect and subgroup [56], constructing the difference or ratio of the estimates with a confidence interval, as appropriate. For the example the ratio of odds ratios is 1.2 and the 95 per cent confidence interval is approximately 0.3 to 5.0. Clearly there is no evidence to support the idea that the two subsets of trials showed different effects [57].

Dichotomizing continuous variables

In medical research, values of continuous variables are often grouped into two or more categories. While there may be clinical value in such classifications, there is no statistical reason why all continuous variables should be treated this way. Grouping for descriptive purposes may allow a simpler presentation and is not especially problematic. In data analysis more serious problems may arise.

Grouping into categories effectively introduces an extreme form of measurement error with an inevitable loss of power, especially when data are collapsed into only two groups. A binary split leads to a comparison of groups of individuals with high or low values of the measurement and a simple estimate of the difference between the groups (with its confidence interval). However, this simplicity is gained at some expense. Cohen [58] observed that dichotomizing is equivalent to throwing a third of the data away, with a consequent major loss of power. Using such binary variables in regression models may lead to biased estimates [59]. Even reducing ordinal variables to two categories may be inadvisable [60].

It is highly desirable for the choice of cutpoint not to be influenced by the actual data values. A common approach is to take the sample median, to give two equal groups. However, when different studies each use the sample median their results cannot easily be compared or combined in a meta-analysis. Note that moving the cutpoint to a higher values leads to higher mean values in both groups.

The arbitrariness of the choice of cutpoint may lead to the idea of trying more than one value and choosing that which, in some sense, gives the most satisfactory result. The temptation should be strongly resisted. Taken to extremes, this approach leads to trying every possible cutpoint and choosing the value that minimizes the *P*-value. Because of the multiple testing the overall false positive rate will be very high, being around 40 per cent rather than the nominal 5 per cent [61]. This unacceptable strategy has become common in the cancer literature, especially in breast cancer [61, 62], and a similar proposal has appeared in the epidemiological literature [63]. Fortunately it has not spread throughout medicine.

Further research is desirable to explore the dangers of residual confounding and the use of methods of analysis that avoid grouping [64, 65].

P-Values versus confidence intervals

While there is now wide use of confidence intervals in clinical medicine, the uptake of confidence intervals has not been equal throughout medicine. Cozens [66] found confidence intervals in just 16 per cent of 100 papers in two radiology journals in 1993 compared with 52 per cent of 50 concurrent papers in the *British Medical Journal*. Curran-Everitt *et al.* [67] found confidence intervals in only one out of 370 papers published in the *American Journal of Physiology in* 1996!

The increased use of confidence intervals has been not so much instead of *P*-values but as a supplement to them [68]. Rothman [69] was an early advocate of confidence intervals in medical papers, and as editor of *Epidemiology* he has gone much further:

> 'When writing for *Epidemiology*, you can also enhance your prospects if you omit tests of statistical significance. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals discourages them, and every worthwhile journal will accept papers that omit them entirely. In *Epidemiology*, we do not publish them at all.
> ... we prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is "significant", as if neither chance nor bias could then account for the findings' [70].

Several recent publications show that the multiple comparisons debate is alive and well. I will not consider this topic in detail (see Goodman [71] for this), but observe that it is hard to see views such as the following being reconciled:

> 'No adjustments are needed for multiple comparisons' [72].
> 'Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference' [73].
> '... Type I error accumulates with each executed hypothesis test and must be controlled by the investigators' [74].
> 'Methods to determine and correct type 1 errors should be reported in epidemiologic and public health research investigations that include multiple statistical tests' [75].

The fact that *P*-values are widely misunderstood is sometimes used as an argument in favour of the Bayesian approach. While there are excellent discussions of the value of Bayesian methods in medical research [76–78], some proponents gloss over the not inconsiderable difficulties in applying Bayesian methods widely. This is surely one of the main reasons for their continuing relative rarity in the medical journals. Further, in the light of experience with other statistical methodology, we should expect that if Bayesian methods were widely adopted they would also frequently be misunderstood and misused by researchers without a proper training or understanding.

Spiegelhalter *et al.* [77] have made the following excellent recommendations:

(i) the preparation of an extensive set of case studies showing practical aspects of the Bayesian approach, in particular for prediction and handling multiple substudies, in which mathematical details are minimized;
(ii) the development of standards for the performance and reporting of Bayesian analyses;
(iii) the development and dissemination of software for Bayesian analysis, preferably as part of existing programs.

Table II. Assessment of the quality of methodology of 364 randomized controlled trials published in 10 leading surgical journals (1988–1994) [88].

Criterion	Acceptable (%)
Clear description of intervention	94
Adequate control group	93
Inclusion criteria	75
Randomization technique	27
Sample size calculation	19
Definition of endpoint	65
Unbiased outcome assessment	48
Adverse events documented	77

Insufficient evidence

As I have already noted, there is often a temptation to place too much emphasis on results based on limited information. A particular problem is the interpretation of a non-significant result as providing direct evidence that there is no real difference. This common false interpretation is potentially serious. (A short note on this theme [79] unwittingly had the same title of a much earlier paper [80].)

Nor should researchers or clinicians believe that a statistically significant result from a small study gives much information. Indeed, it is quite possible that having a small amount of data – for example from a single small randomized trial – may be worse than having no evidence at all, as it can easily lead to belief that we have reliable information.

REPORTING RESEARCH

The value of research papers is greatly diminished by poor standards of reporting. The widely-adopted Vancouver guidelines include the following reasonable requirement:

> 'Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results' [81].

Recent examples showing that journals do not fully enforce this requirement include reviews of phase II clinical trials in cancer [82] and the reporting of logistic regression analyses [83].

Reporting of randomized controlled trials (RCTs) has been subjected to particular scrutiny. Inadequate reporting has been highlighted by many reviews of methodology [84–87] and as part of systematic reviews of particular medical areas. Table II shows for illustration the findings of a review of 364 RCTs published in surgery journals [88]. Many papers failed to report essential features of study methodology. Particular areas of concern include reporting of the method of treatment allocation and the question of whether all patients were included in the analysis. Hotopf *et al.* [89] reviewed 122 RCTs of selective serotonin uptake inhibitors in patients with depression. Only one trial report included details of the method of randomization.

Hollis and Campbell [90] reviewed RCTs published in 1997 in leading general medical journals. About half of the trials (119) mentioned an intention-to-treat analysis, but only five reports explicitly stated that there were no deviations from random allocation. Some 89 trials (75 per cent) had

some missing data on the primary outcome variable. It is clear that the intention-to-treat approach is another aspect of randomized trials that authors need to explain in the context of their trial.

The CONSORT statement [91] presented a list of 21 items which should be reported in a paper describing an RCT, based on empirical evidence where possible. Also they proposed that reports should include a flow chart describing patient progress through the trial. Most leading general medical journals and many specialist journals (>70) have adopted the CONSORT recommendations [92]. In principle, authors should no longer be able to hide study inadequacies by omission of important information.

Reporting of other types of study has also been found to be deficient, leading to thoughts of CONSORT-type initiatives for other types of study. A CONSORT style exercise has led to the development of the QUOROM statement, providing guidance for reporting systematic reviews and meta-analyses [93]. Draft guidelines for reporting studies of diagnostic accuracy have appeared [94].

There are bound to be difficulties in ensuring adherence to such guidelines. Journals cannot devote major time to checking that authors really do provide every bit of information required. Nevertheless, the heightened awareness of authors, referees and editors (and readers) of the major elements of good reporting does lead to improvements. As noted above, this has been observed for the use of confidence intervals, and there is emerging evidence of the benefits of CONSORT (Moher *et al.*, submitted for publication).

General problems associated with the inadequate reporting of statistical methods are unlikely to be fully overcome by such initiatives. As Williams *et al.* noted, in a review of papers in the *American Journal of Physiology*. 'In contrast to lengthy descriptions of other experimental methods and protocols, descriptions of statistical analyses that were used in a study sometimes only consist of two or three sentences. In this case, a reader can effectively evaluate the experiment but not necessarily the experimental design, analysis, and conclusions' [95].

DISCUSSION

Over several decades there has been considerable evidence of the wide misuse of statistics [3, 96]. Despite widespread attempts to improve statistical understanding it is clear from reviews of research in particular medical specialties that errors in statistics continue much as before. Kanter and Taylor [97] found that 75 per cent of 59 articles in *Transfusion* used an inappropriate statistical test or contained an error in calculation or interpretation, and 22 per cent of papers reported a conclusion not supported by the data. McGuigan [98] found serious errors in 40 per cent of 164 papers published in the *British Journal of Psychiatry*. Likewise, Welch and Gabbe [99] reviewed 145 articles published in the *American Journal of Obstetrics and Gynecology*. They found only 30 per cent where the methods were clearly appropriate and found serious statistical errors in 19 per cent. Two of these studies were accompanied by editorials suggesting or promising greater editorial vigilance in future [100, 101] and the other by a commentary [102] which observed that there had been little improvement since an earlier review of the same journal 15 years earlier [103].

There seems to be a general feeling that the level of understanding of statistics has improved. Even if true, such reviews illustrate that there remains much scope for further improvement. The biggest change over the last two decades or so has been the increased availability of sophisticated statistical software on each researcher's desk. While there are obvious benefits associated with this massive change, it may provide the user with only the illusion of greater statistical sophistication.

Most software does nothing to assist in the understanding of statistical principles and cannot help researchers to design sensible studies.

Medical practitioners and researchers clearly find learning statistics difficult, not least in the exercise of judgement. Not surprisingly, perhaps, it is quite common to see requests or attempts to standardize approaches and remove the element of judgement. For example, in a discussion of different methods for analysing method comparison studies, Rankin and Stokes [104] observed: 'A consensus needs to be reached to establish which tests ... are the most relevant ones to be adopted universally'. This questionable aim indicates a 'best buy' approach to statistics. In general there may be several valid ways of analysing a data set. While it would make life easier if, say, we abandoned all non-parametric (or, indeed, parametric) analyses, this would certainly not be a 'Good Thing'.

A similar view seems to lie behind the development of simple rules to help assess the 'quality' (that is, reliability) of research findings. For example, it is well recognized that incomplete follow-up of patients in randomized trials compromises the reliability of the findings, but can we say what is acceptable loss? The figure of 80 per cent follow-up is being used as the minimum acceptable. For example, the journal *Evidence-Based Medicine* will not include reports of trials with worse follow-up than this (see also 'levels of evidence' [39]). Assessing whether a trial was a good one should take account of circumstances, including what is achievable. In trials of lifestyle interventions, for example, such as dieting or smoking cessation, such drop-out rates are rarely achieved, unless using an unrealistically short follow-up period.

While there is potential value in guidelines, these should not in general be interpreted as rules, and we should not disguise the fact that exercising judgement is a major element of statistics. As Vandenbroucke observed, 'we should not be afraid of teaching [physicians] uncertainty' [105].

I wish to end with a few broad observations which, I concede, are not drawn directly from the above material but from a rather wider perspective [106]:

(i) The misuse of statistics is very important.
(ii) A general climate of sloppiness is bad for science.
(iii) Statistics is much more subjective (and difficult) than is usually acknowledged (this is why statisticians have not been replaced by computers).
(iv) Major improvements in the quality of research published in medical journals are unlikely in present research climate.
(v) Too much research is done primarily to benefit the careers of researchers.
(vi) It need not be like this!

REFERENCES

1. Glass GV. Primary, secondary, and meta-analysis of research. *Educational Research* 1976; **5**:3–8.
2. Andrews G, Guitar B, Howie P. Meta-analysis of the effects of stuttering treatment. *Journal of Speech and Hearing Disorders* 1980; **45**:287–307.
3. Altman DG. Statistics in medical journals: developments in the 1980s. *Statistics in Medicine* 1991; **10**:1897–1913.
4. Deeks JJ. Systematic reviews of published evidence: miracles or minefields? *Annals of Oncology* 1998; **9**:703–709.
5. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology* 1999; **150**:469–475.
6. The Cochrane Collaboration. *The Cochrane Library* [database on disk and CDROM]. Issue 4. Update Software: Oxford, 1999 (updated quarterly).
7. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health Technology Assessment* 1998; **2**(19):1–276.
8. Egger M, Davey Smith G, Altman DG (eds). *Systematic Reviews in Health Care*: *Meta-analysis in Context*. 2nd edn. BMJ Books: London, 2000.

9. Deeks JJ, Altman DG, Dooley G, Sackett DL. Choosing an appropriate dichotomous effect measure for meta-analysis: empirical evidence of the appropriateness of the odds ratio and relative risk. (Abstract). *Controlled Clinical Trials* 1997; **18**(3S):84S–85S.
10. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *Evidence-Based Medicine* 1996; **1**:164–166.
11. Senn S, Walter S, Olkin I. Odds ratios revisited. *Evidence-Based Medicine* 1998; **3**:71.
12. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* 1994; **47**:881–890.
13. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials* 1995; **16**:62–73.
14. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association* 1999; **282**:1054–1060.
15. Nurmohamed MT, Rosendaal FR, Bueller HR, Dekker E, Hommes DW, Vandenbroucke JP, Briet E. Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. *Lancet* 1992; **340**:152–156.
16. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352**:609–613.
17. Greenland S. Quality scores are useless and potentially misleading. *American Journal of Epidemiology* 1994; **140**: 300–301.
18. Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, Jonas WB. Impact of study quality on outcome of placebo-controlled trials of homeopathy. *Journal of Clinical Epidemiology* 1999; **52**:631–636.
19. Greenland S. Can meta-analysis be salvaged? *American Journal of Epidemiology* 1994; **140**:783–787.
20. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *British Medical Journal* 1998; **316**:140–144.
21. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* 1993; **12**:1293–1316.
22. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making* 1993; **13**:313–321.
23. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1995; **48**:119–130.
24. Schulz KF, Chalmers I, Hayes R, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* 1995; **273**:408–412.
25. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GK, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association* 1999; **282**:1061–1066.
26. Katerndahl DA, Lawler WR. Variability in meta-analytic results concerning the value of cholesterol reduction in coronary heart disease: a meta–meta-analysis. *American Journal of Epidemiology* 1999; **149**:429–441.
27. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine. How to Practice and Teach EBM*. Churchill-Livingstone: London, 1997; 2.
28. Davidoff F. In the teeth of the evidence: the curious case of evidence-based medicine. *Mount Sinai Journal of Medicine* 1999; **66**:75–83.
29. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* 1988; **318**:1728–1733.
30. Altman DG. Confidence intervals for the number needed to treat. *British Medical Journal* 1998; **317**:1309–1312.
31. Rembold CM. Number needed to screen: development of a statistic for disease screening. *British Medical Journal* 1998; **317**:307–312.
32. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading. *British Medical Journal* 1999; **318**:1548–1551.
33. Black N. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal* 1996; **312**:1215–1218.
34. Moses LE. Measuring effects without randomized trials? Options, problems, challenges. *Medical Care* 1995; **33**: AS8–AS14.
35. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal* 1998; **317**:1185–1190.
36. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment* 1998; **2**(13):1–124.
37. Reeves BC, Maclehose RR, Harvey IM, Sheldon TA, Russell IT, Black AMS. Comparisons of effect sizes derived from randomised and non-randomised studies. In *Health Services Research Methods: a Guide to Best Practice*, Black N, Brazier J, Fitzpatrick R, Reeves B (eds). London: BMJ Books, 1998; 73–85.
38. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Statistics in Medicine* 1984; **3**:361–373.
39. Centre for Evidence-Based Medicine. Levels of evidence and grades of recommendations. http://cebm.jr2.ox.ac.uk/docs/levels.html (accessed 10 December 1999).

40. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health* 1998; **52**:377–384.
41. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics* 1998; **47**:95–113.
42. Kerry SM, Bland JM. Trials which randomize practices I: how should they be analysed. *Family Practice* 1998; **15**:80–83.
43. Kerry SM, Bland JM. Trials which randomize practices II: sample size. *Family Practice* 1998; **15**:84–87.
44. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 1999; **28**:319–326.
45. Edwards SJL, Braunholtz DA, Lilford RJ, Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. *British Medical Journal* 1999; **318**:1407–1409.
46. Campbell MJ, Elbourne D, Donner A. (eds). Statistical issues in the design and analysis of cluster randomized trials. *Statistics in Medicine* 2000; in press.
47. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *International Journal of Epidemiology* 1990; **19**:795–800.
48. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American Journal of Public Health* 1995; **85**:1378–1383.
49. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials* 1989; **10**:31–56.
50. Altman DG, Goodman S. Transfer of technology from statistical journals to the biomedical literature: past trends and future predictions. *Journal of the American Medical Association* 1994; **272**:129–132.
51. Schwarzer G. Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Statistics in Medicine* 2000; **19**:541–561.
52. Lachenbruch PA. Some misuses of discriminant analysis. *Methods of Information in Medicine* 1977; **16**:255–258.
53. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal* 1998; **316**:1701–1705.
54. Wahlbeck K, Adams C. Sponsored drug trials show more-favourable outcomes. *British Medical Journal* 1999; **318**:465.
55. Matthews JNS, Altman DG. Interaction 2: Compare effect sizes not P values. *British Medical Journal* 1996; **313**:808.
56. Matthews JNS, Altman DG. Interaction 3: How to examine heterogeneity. *British Medical Journal* 1996; **313**:862.
57. Matthews JNS. Sponsored trials do not necessarily give more-favourable results. *British Medical Journal* 1999; **318**:1762.
58. Cohen J. The cost of dichotomization. *Applied Psychological Measurement* 1983; **7**:249–253.
59. Selvin S, Two issues concerning the analysis of grouped data. *European Journal of Epidemiology* 1987; **3**:284–287.
60. Sankey SS, Weissfeld LA. A study of the effect of dichotomizing ordinal data upon modelling. *Communications in Statistics – Simulation* 1998; **27**:871–887.
61. Altman DG, Lausen B, Sauerbrei W, Schumacher S. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994; **86**:829–835.
62. Altman DG. Suboptimal analysis using "optimal" cutpoints. *British Journal of Cancer* 1998; **78**:556–557.
63. Wartenberg D, Northridge M. Defining exposure in case-control studies: a new approach. *American Journal of Epidemiology* 1991; **133**;1058–1071.
64. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 1997; **8**:429–434.
65. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* 1995; **6**:356–365.
66. Cozens NJ. Should we have confidence intervals in radiology papers? *Clinical Radiology* 1994; **49**:199–201.
67. Curran-Everett D, Taylor S, Kafadar K. Fundamental concepts in statistics: elucidation and illustration. *Journal of Applied Physiology* 1998; **85**:775–786.
68. Savitz DA, Tolo KA, Poole C. Statistical significance testing in the *American Journal of Epidemiology*, 1970–1990. *American Journal of Epidemiology* 1994; **139**:1047–1052.
69. Rothman KJ. A show of confidence. *New England Journal of Medicine* 1978; **299**;1362–1363.
70. Rothman KJ. Writing for *Epidemiology. Epidemiology* 1998; **9**(3).
71. Goodman SN. Multiple comparisons, explained. *American Journal of Epidemiology* 1998; **147**:807–812.
72. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1**:43–46.
73. Perneger TV. What's wrong with Bonferroni adjustments. *British Medical Journal* 1998; **316**:1236–1238.
74. Moyé LA. P-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology* 1998; **8**:351–357.
75. Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *American Journal of Epidemiology* 1998; **147**:615–619.
76. Breslow N. Biostatistics and Bayes. *Statistical Science* 1990; **5**:269–284.
77. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to Bayesian methods in health technology assessment. *British Medical Journal* 1999; **319**:508–512.
78. Goodman SN. Towards evidence-based medical statistics. Part 2. The Bayes factor. *Annals of Internal Medicine* 1999; **130**:1005–1021.

79. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *British Medical Journal* 1995; **311**:485.
80. Hartung J, Cottrell JE, Giffen JP. Absence of evidence is not evidence of absence. *Anesthesiology* 1983; **58**:298–300.
81. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine* 1997; **126**:36–47 (see also http://www.acponline.org/journals/resource/unifreqr.htm dated May 2000—accessed 5 September 2000).
82. Kramar A, Potvin D, Hill C. Multistage designs for phase II clinical trials: statistical issues in cancer research. *British Journal of Cancer* 1996; **74**:1317–1320.
83. Bender R, Grouven U. Logistic regression models used in medical research are poorly presented. *British Medical Journal* 1996; **313**:628.
84. Koes BW, Bouter LM, van der Heijden GJMG. Methodological quality of randomized clinical trials on treatment efficacy in low back pain. *Spine* 1995; **20**:228–235.
85. Lent V, Langenbach A. A retrospective quality analysis of 102 randomized trials in four leading urological journals from 1984–1989. *Urological Research* 1996; **24**:119–122.
86. Skovlund E. A critical review of papers from clinical cancer research. *Acta Oncologica* 1998; **37**:339–346.
87. Thornley B, Adams C. Content and quality of 2000 controlled trials in schizophrenia over 50 years. *British Medical Journal* 1998; **317**:1181–1184.
88. Hall JC, Mills B, Nguyen H, Hall JL. Methodologic standards in surgical trials. *Surgery* 1996; **119**:466–472.
89. Hotopf M, Lewis G, Normand C. Putting trials on trial—the costs and consequences of small trials in depression: a systematic review of methodology. *Journal of Epidemiology and Community Health* 1997; **51**:354–358.
90. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal* 1999; **319**:670–674.
91. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz K, Simel D, Stroup D. Improving the quality of reporting of randomized controlled trials: the CONSORT Statement. *Journal of the American Medical Association* 1996; **276**:637–639.
92. Moher D. CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. *Journal of the American Medical Association* 1998; **279**:1489–1491.
93. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF for the QUOROM Group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999; **354**:1896–1900.
94. Bruns DE, Huth EJ, Magid E, Young DS. Toward a checklist for reporting of studies of diagnostic accuracy of medical tests. *Clinical Chemistry* 2000; **46**:893–895.
95. Williams JL, Hathaway CA, Kloster KL, Layne BH. Low power, type II errors, and other statistical problems in recent cardiovascular research. *American Journal of Physiology* 1997; **273**:H487–H493.
96. Altman DG. Statistics in medical journals. *Statistics in Medicine* 1982; **1**:59–71.
97. Kanter MH, Taylor JR. Accuracy of statistical methods in TRANSFUSION: a review of articles from July/August 1992 through June 1993. *Transfusion* 1994; **34**:697–701.
98. McGuigan SM. The use of statistics in the *British Journal of Psychiatry. British Journal of Psychiatry* 1995; **167**:683–688.
99. Welch GE, Gabbe SG. Review of statistics usage in the *American Journal of Obstetrics and Gynecology. American Journal of Obstetrics and Gynecology* 1996; **175**:1138–1141.
100. McCullough J, Aster R, Bove JR, Garratty G, Issitt P, Oberman H, Perkins H. Statistics in articles published in TRANSFUSION. *Transfusion* 1994; **34**:654–655.
101. Anonymous. Statistics usage in the *American Journal of Obstetrics and Gynecology*: toward a better understanding. *American Journal of Obstetrics and Gynecology* 1996; **175**:1137.
102. Hand D, Sham P. Improving the quality of statistics in psychiatric research. *British Journal of Psychiatry* 1995; **167**: 689–690.
103. White SJ. Statistical errors in papers in the *British Journal of Psychiatry. British Journal of Psychiatry* 1979; **135**:336–342.
104. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation* 1998; **12**:187–199.
105. Vandenbroucke JP. Observational reserarch and evidence-based medicine: what should we teach young physicians? *Journal of Clinical Epidemiology* 1998; **51**:467–472.
106. Altman DG. The scandal of poor medical research. *British Medical Journal* 1994; **308**:283–284.

