

Statistics and ethics in medical research

Misuse of statistics is unethical

DOUGLAS G ALTMAN

"Some people hate the very name of statistics but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man."

FRANCIS GALTON¹

In 1949 a divorce case was heard in which the sole evidence of adultery was that a baby was born almost 50 weeks after the husband had gone abroad on military service. To quote Barnett²: "The appeal judges agreed that the limit of credibility had to be drawn somewhere, but on medical evidence 349 (days), whilst improbable, was scientifically possible." So the appeal failed.

If we look at the distribution of length of gestation³ (fig 1), which the judges apparently did not do, I think that most people would feel that the husband was hard done by. Even if we take reports of extremely long pregnancies as accurate, it is clear that, although "scientifically possible," a pregnancy lasting 349 days is an extremely unlikely occurrence. For those who believe as I do that a pregnancy of 51 weeks* exceeds the bounds of credibility, suppose it had been only 48 weeks, or 45?

*Using the standard convention of counting in completed weeks from the first day of the last menstrual period and assuming conception to have occurred 14 days later.

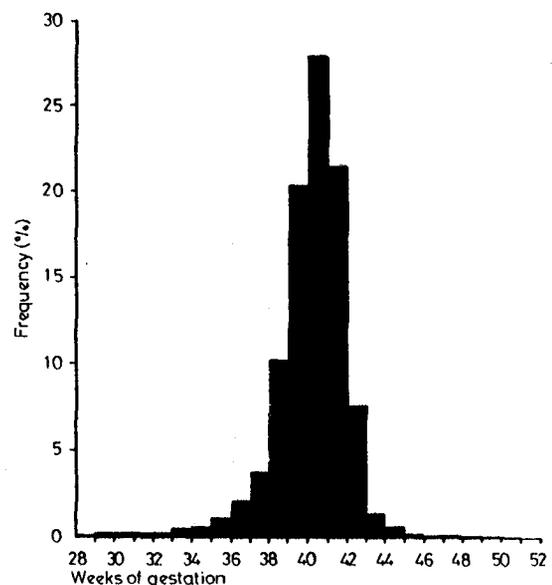


FIG 1—Frequency distribution of length of gestation.

If this case were heard now, where would *you* draw the line on the basis of fig 1?

This case illustrates a failure to use statistical methods when they ought to have been used, a fairly common occurrence. Saying that an event is possible is quite different from saying that it has a probability of, say, one in 100 000. Although not an example from medical research, this case concerned essentially the same difficulty as in many more frequently encountered problems, such as defining hypertension or obesity. Everything varies; it is in trying to draw lines between good

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSc, medical statistician (member of scientific staff)

and bad, high and low, likely and unlikely, and so on, that many problems arise. Although statistics cannot answer a given question, they can often shed considerable light on the problem.

Statistics and medical ethics

So what is the relation between statistics and medical ethics? It is well appreciated that ethical considerations may affect the design of an experiment. Perhaps the most obvious examples are clinical trials—we cannot, for example, carry out controlled trials of cigarette smoking. The purpose of this series of articles is to discuss in some detail a different and much neglected aspect of the relation—how the statistical aspects affect the ethics.

Stated simply, it is unethical to carry out bad scientific experiments.⁴ Statistical methods are one aspect of this. However praiseworthy a study may be from other points of view, if the statistical aspects are substandard then the research will be unethical. There are two principal reasons for this.

Firstly, the most obvious way in which a study may be deemed unethical, whether on statistical or other grounds, is the misuse of patients (or animals) and other resources. As May⁵ has said: "... one of the most serious ethical problems in clinical research is that of placing subjects at risk of injury, discomfort, or inconvenience in experiments where there are too few subjects for valid results, too many subjects for the point to be established, or an improperly designed random or double-blind procedure."

Secondly, however, statistics affects the ethics in a much more specific way: it is unethical to publish results that are incorrect or misleading. Errors in the use of statistics may occur at all stages of an investigation, and one error can be sufficient to render the whole exercise useless. A study may have been perfectly conceived and executed, but if it is analysed incorrectly then the consequences may be as serious as for a study that was fundamentally unsound throughout.

There are many ways in which the statistical content of research may be deficient. In a fascinating and somewhat frightening recent paper, Sackett⁶ identified 56 possible biases that may arise in "analytic research," over two-thirds of which related to aspects of study design and execution. Figure 2 shows

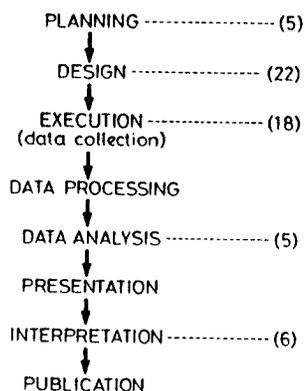


FIG 2—Structure of a research exercise. Explanation of numbers is given in text.

how these possible biases are distributed over the stages of a research exercise. In general this distribution also reflects very well the relative seriousness of statistical errors at each stage, and indicates where there is greatest need for statistical expertise. Errors in the analysis or interpretation of results can usually be rectified if detected in time—that is, before publication—but deficiencies in the design are nearly always irremediable. The end point of the process is usually publication. Problems may

well arise when this is considered to be the most important aspect of the whole exercise, a not uncommon occurrence.

Publication

Once published, a piece of research achieves both respectability and credibility so that it is important for journals to make strenuous efforts to detect substandard research. In recent years there have been several good studies of the quality of statistics in papers in medical journals to support the idea that there is much room for improvement. For example, Schor and Karten⁷ reported that, of 149 papers reporting analytical studies in several journals, only 28% were judged acceptable, 67% were deemed deficient but could be improved, and 5% were totally unsalvageable.

The editor of the journal wrote as follows:

"The study is an indirect argument for greater knowledge and appreciation of statistics by the medical author, for a reiteration on his part that the biostatistician is not a worrisome censor, but a valuable ally, and that biostatistics, far from being an unrelated mathematical science, is a discipline essential to modern medicine—a pillar in its edifice."⁸

More recent studies⁹⁻¹¹ have shown that there are still far too many papers being published in which the statistical analyses are incorrect. Conflicting results from similar studies can often be attributed to varying degrees of statistical competence.¹²⁻¹⁴

The ethical implications of publishing research containing incorrect or unfounded results or conclusions are little affected by the nature of the errors made, and are indeed much the same as the consequences of publishing spurious results. The cost in time and energy in trying to reproduce such results can be enormous.¹⁵ Alternatively, the results may rest unchallenged for many years. Suppose a randomised controlled trial is carried out in which a conclusion is reached that the new treatment is significantly better than the previous standard treatment. The publication of such a finding may well affect patient care, and it may then be considered to be unethical to carry out further trials as one group would be denied the new treatment that was "known" to be better. Clearly, both of these consequences of publication will hold whether or not the conclusions were justified unless any deficiencies are very obvious (and many that Sackett⁶ lists would not be) or if there is considerable protest. A solitary critical letter, perhaps from a statistician, hidden away on the correspondence page is unlikely to be sufficient. Similar consequences apply in the opposite case where a treatment is incorrectly found to be ineffective.

Summary

The ethical implications of statistically substandard research may be summarised as follows:

- (1) the misuse of patients by exposing them to unjustified risk and inconvenience;
- (2) the misuse of resources, including the researchers' time, which could be better employed on more valuable activities; and
- (3) the consequences of publishing misleading results, which may include the carrying out of unnecessary further work.

These are specific and highly undesirable outcomes. Failure to guard against these is surely as unethical as using experimental methods that offend against moral principles, such as failing to obtain fully informed consent from subjects. Surprisingly, this aspect seems to have been totally ignored by books on medical ethics.

All stages of research shown in fig 2 are vulnerable to statistical mismanagement. As an example consider one aspect of planning a study: "reading up published reports." If published papers are accepted uncritically you might be trying to verify someone else's spurious results. Remember too that authors will tend to

refer to other published work that supports their arguments and ignore papers that do not.

The next few articles will illustrate some ways in which errors at different stages of a study can compromise the ethical status of the research, and discuss some ways in which they may be avoided. These will serve only as examples, since it is impossible to be comprehensive. In the final article I will consider the role of the medical journals in this context.

References

- ¹ Galton F. *Natural inheritance*. London: Macmillan, 1889.
- ² Barnett V. The study of outliers: purpose and model. *Applied Statistics* 1978;**27**:242-50.
- ³ Chamberlain R. Birth weight and length of gestation. In: *British births 1970*. Vol 1. *The first week of life*. National Birthday Trust Fund and Royal College of Obstetricians and Gynaecologists. London: Heinemann, 1975.
- ⁴ Denham MJ, Foster A, Tyrrell DAJ. Work of a district ethical committee. *Br Med J* 1979;iii:1042-5.
- ⁵ May WW. The composition and function of ethical committees. *J Med Ethics* 1975;**1**:23-9.
- ⁶ Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;**32**:51-63.
- ⁷ Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;**195**:1123-8.
- ⁸ Anonymous. A pillar of medicine. *JAMA* 1966;**195**:1145.
- ⁹ Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J* 1977;ii:85-7.
- ¹⁰ Feinstein AR. A survey of the statistical procedures in general medical journals. *Clin Pharmacol Ther* 1974;**15**:97-107.
- ¹¹ Ambroz A, Chalmers TC, Smith H, Schroeder B, Freiman JA, Shareck EP. Deficiencies of randomized control trials. *Clin Res* 1978;**26**:280A.
- ¹² Gifford RH, Feinstein AR. A critique of methodology in studies of anticoagulant therapy for acute myocardial infarction. *N Engl J Med* 1969;**280**:351-7.
- ¹³ Peto R. Clinical trial methodology. *Biomedicine* (special issue) 1978;**28**:24-36.
- ¹⁴ Horwitz RJ, Feinstein AR. Methodologic standards and contradictory results in case-control research. *Am J Med* 1979;**66**:556-64.
- ¹⁵ Muller M. Why scientists don't cheat. *New Scientist* 1977;**74**:522-3.

*This is the first in a series of eight articles.
No reprints will be available from the author.*

Statistics and ethics in medical research

Study design

DOUGLAS G ALTMAN

The term "design" encompasses all the structural aspects of a study, notably the definition of the study sample, size of sample, method of treatment allocation, type of statistical design (randomised, cross-over, sequential, etc), and choice of outcome measures. The importance of this stage cannot be over-emphasised since no amount of clever analysis later will be able to compensate for major design flaws. In this article I will consider the relation between design and ethics in observational studies and clinical trials, but I will defer the problem of sample size until the next article.

Observational studies

In observational studies data from a sample of individuals are used, either implicitly or explicitly, to make inferences about the population of interest, such as men aged 20-65, hypertensives, or pregnant women. For this extrapolation to be valid, it is essential that the data obtained are as representative of the population as possible.

This usually entails some type of random sampling of subjects, for which a ready-made list of the whole population of interest (a sampling frame) is needed. Such lists, however, may be out of date (electoral registers) or inaccurate (doctors' lists of patients), in which case their use can lead to misleading results. Furthermore, it is often desirable to improve the representativeness of the sample by sampling separately from different subgroups—for example, by age and sex—but this additional information may not be available.

For many populations, such as the three examples above, no sampling frame exists, so that it may be impossible to obtain a representative sample. Consider, for example, trying to select a random sample of all the preschool children in an area to estimate the prevalence of vision or hearing defects. Yet for studies such as this, which set out to estimate the prevalence or incidence of some condition, the need for a truly representative sample is particularly great—otherwise the results are of uncertain value.

Even with a good selection procedure the study may be ruined by a poor response rate. Although deemed non-invasive, such studies may entail visiting people at home, expecting them to complete and return a questionnaire, or to attend a clinic, and thus may be liable to considerable non-cooperation.

Unfortunately, those who do not participate often tend to be somewhat different from those who do, both in respect of their medical condition (if this is relevant) and their social and demographic characteristics. This problem should be anticipated at the design stage, and plans made to "chase up" non-responders. It is generally advisable to keep questionnaires and other procedures short and simple to help reduce non-response. In the end, though, the response rate may largely depend on the subjects' perception of the importance of the study.

It is much less common in case-control studies to find researchers concerned about defining the subjects who will be eligible for a study, although Sackett¹ has described 22 biases that may arise at this stage. One of the most interesting is Berkson's bias, which Mainland² recently drew to the attention of readers of this journal. Case-control studies of hospital patients are often set up to study the relation between a specific disease and exposure to a suspected causal factor. If the hospital admission rates for exposed and unexposed cases and controls differ appreciably, then the observed association between the factor and the disease may be seriously biased (in either direction).³ Indeed, the choice of control group may affect the observed association between a disease and a suspected cause. A consequence of this is that such studies may need to be supported by prospective studies.

Another of Sackett's catalogue¹ is the membership (or "self-selection") bias. He cites the example of an apparent association between lack of exercise after myocardial infarction and the increased risk of recurrent attacks. This result was found in two observational studies where exercise was taken voluntarily, but was not substantiated by a prospective randomised study.

So the major problem of all observational studies is the selection of subjects for study. This aspect must be given considerable attention at the design stage, because if the sample is not representative of the population then the results will be unreliable and of dubious worth.

Clinical trials

Whatever one's view on the best type of design, clinical trials of some sort are clearly important for new treatments. As May⁴ says: "The ethical justification for such experimentation, which is outside the pure physician-patient relationship, is based on a judgment that in certain circumstances it is legitimate to put a subject at risk, with his or her consent, because of the overriding need of society for progress in combating certain diseases."

A revealing example concerns the epidemic of retrolental fibroplasia in the 1950s.⁵ The treatment of infants with early eye changes with adrenocorticotrophic hormone was thought

Division of Computing and Statistics, Clinical Research Centre,
Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSc, medical statistician (member of scientific staff)

to be a success as there was a cure rate of 75%. A clinical trial, however, would have shown that adrenocorticotrophic hormone was ineffective since 75% of such infants return to normal without treatment. The widespread use of this treatment meant that hundreds of infants were exposed to unnecessary risk, and that discovery of the cause of the epidemic (an oxygen-rich environment) was delayed.

The debate about the ethics of clinical trials is still very active. Some authors have suggested that it is unethical *not* to carry out a clinical trial on a new treatment, whereas others believe that such trials are unethical, at least in the way they are usually conducted.

IS IT ETHICAL TO RANDOMISE?

In most clinical trials subjects are allocated to the new treatment at random, others receiving either a standard treatment or a placebo. The main ethical problem is the balancing of the welfare of the individuals in the trial against the potential benefit to future patients.

It is the random allocation of subjects that comes in for most criticism. It is argued that even if at the beginning of a trial one may not know if a treatment is effective, as the study progresses it is unethical to continue to randomise ignoring the results so far.⁷ As Meier⁸ has observed, however, this attitude is based on the questionable premise "that it is unethical to deny an individual any expected benefit of treatment A over treatment B, regardless of how small that benefit may be or how uncertain."

Because of the difficulty in interpreting interim results of randomised studies, two types of non-randomised study have recently found some favour and deserve a closer look.

HISTORICAL CONTROLS

Is it really necessary to have a concurrent control group when carrying out a clinical trial? Cranberg⁹ has recently argued that instead one can use retrospective or "historical" controls—that is, previously collected data on patients who had received what would be the control treatment. Although widely practised, and perhaps of value in some circumstances,¹⁰ this can be extremely risky.

The main problem of studies using historical controls is their insensitivity to secular changes, most importantly in selection criteria.¹¹ The worst historical data to use are other people's published results, perhaps partly because of the publication bias towards positive results. Pocock¹² gives as an example 20 studies of fluorouracil for advanced cancer of the large bowel with reported success rates ranging from 8% to 85%. But data from a previous study in the same institution may also be unreliable. Pocock reports that in 19 instances where the same treatment was used in two consecutive trials of cancer chemotherapy in one organisation the changes in death rates from one trial to the next ranged from -46% to +24%, four of the differences being significant at the 2% level.

The use of historical controls is often advocated as being more ethical than using a concurrent randomised control group. The results of studies using historical controls are extremely unreliable, however, so that unless there is sound justification for their use such designs should themselves be rejected as unethical.

ADAPTIVE DESIGNS

Designs where the proportion of subjects allocated to each treatment depends on the accumulated results so far may appear preferable to randomised trials.⁷ It must be realised, however, that with such designs some subjects are still allocated to the treatment that is less successful so far, not so many as with randomised studies but still essentially at random. Further-

more, because of the unequal sample sizes for the two treatments, the study may require more subjects than an equal allocation study.^{8, 13}

Such designs require that the result for each individual is known quickly, which is often not the case. It is implicitly assumed that there is a single outcome of interest, whereas there may be several possible methods of assessment, as well as aspects such as side effects to be considered. They are also insensitive to any secular changes during the course of the study. For these reasons, although appealing in principle, adaptive designs have rarely, if ever, been used.⁸

SEQUENTIAL DESIGNS

Sequential designs¹⁴ may seem the best compromise in that they combine the many advantages of a randomised study with the desirable feature of taking account of the results so far in determining the length of the trial.

The main advantage over an ordinary randomised study is that the required sample size will be smaller if the treatment "effect" is larger. So the bigger the difference between treatments, the fewer subjects receive the less successful treatment.

Their main disadvantages are the same as for adaptive designs, especially the need for the results for each subject to be available quickly. Sequential designs are clearly of no value in long-term studies, where all the subjects will be recruited before any results are obtained. Nevertheless, in the right circumstances they can be useful and should probably be used more often.

CONSENT

Another problem of clinical trials is the need to obtain the "informed consent" of the subjects. In some cases this may be impossible because of the age or condition of the subjects, or because of the difficulty of explaining the scientific issues. Zelen¹⁵ has recently proposed a new design for comparing a new treatment with a standard one that neatly avoids the problem. He proposed that, of the subjects entering a trial, half are randomly assigned to receive the standard treatment (group 1). These subjects are treated as if they were not in the trial apart from the needs of standardised assessment and record keeping. The other half (group 2) are given a choice: they are offered the new treatment B, which is under investigation, but they may have the standard treatment A if they wish. The important point is that the subjects *choose*—this is quite different from agreeing to be randomised—so that the problems associated with informed consent do not arise.

If most of the second group elect to have the new treatment B, as is quite likely, then this design will probably be more efficient overall. It is of course essential to compare group 1 with group 2, not all those undergoing treatment A with those undergoing B. In this way two randomly selected groups will be compared. There will be some loss of efficiency because group 2 is "contaminated" by a minority undergoing treatment A, but this effect is likely to be outweighed by the advantage of having virtually no refusers.

This design, which seems perfectly ethical (Zelen¹⁵ discusses many of the issues), has two advantages over ordinary randomised trials—the ability to include all eligible subjects and the avoidance of the tricky problem of informed consent.

PLACEBOS

Too many studies compare a new treatment with a placebo rather than an existing treatment, and thus yield results that are of no practical importance. It is sometimes necessary to include placebos, but whenever possible they should be used

only when there is no appropriate treatment for comparison. Invasive placebo treatment is unlikely ever to be justified.

CONCLUSIONS

There is no one best design for all clinical trials. The choice for a specific trial must depend on the seriousness of the condition being treated, the nature of the treatments, the response time, the measures of outcome, and so on. The main ethical problem is balancing the interests of the individuals in the study with those of the much larger number who may benefit in the long term. But it is also vital that the research should provide useful results, and this may often be achieved best by a randomised study (double-blind if possible). If it is thought likely that highly favourable early results or a high incidence of side effects would argue in favour of premature termination of the study, then these considerations may be built in, using a sequential design.

The ethical difficulties associated with the widespread use of a new treatment without a trial are far greater than those associated with the trial itself. The importance of good design, however, is reflected in the many examples of conflicting results that may be found in series of case-control studies of the same topic.¹⁶ As a notable example, after 32 studies over 25 years there is still no consensus on the efficacy of anticoagulants following myocardial infarction.¹¹

References

- ¹ Sackett DL. Bias in analytic research. *J Chron Dis* 1979;**32**:51-63.
- ² Mainland D. Berkson's fallacy in case-control studies. *Br Med J* 1980;**280**:330.
- ³ Roberts RS, Spitzer WO, Delmore T, Sackett DL. An empirical demonstration of Berkson's bias. *J Chron Dis* 1978;**31**:119-28.
- ⁴ May WW. The composition and function of ethical committees. *J Med Ethics* 1975;**1**:23-9.
- ⁵ Silverman WA. The lesson of retrolental fibroplasia. *Sci Am* 1977;**236**(6):100-7.
- ⁶ Herbert V. Acquiring new information while retaining old ethics. *Science* 1977;**198**:690-3.
- ⁷ Weinstein MC. Allocation of subjects in medical experiments. *N Engl J Med* 1974;**291**:1278-85.
- ⁸ Meier P. Terminating a trial—the ethical problem. *Clin Pharmacol Ther* 1979;**25**:633-640.
- ⁹ Cranberg L. Do retrospective controls make clinical trials inherently fallacious? *Br Med J* 1979;**iii**:1265-6.
- ¹⁰ Gehan EA, Freireich EJ. Non-randomized controls in cancer clinical trials. *N Engl J Med* 1974;**290**:198-204.
- ¹¹ Doll R, Peto R. Randomised controlled trials and retrospective controls. *Br Med J* 1980;**280**:44.
- ¹² Pocock SJ. Allocation of patients to treatment in clinical trials. *Biometrics* 1979;**35**:183-97.
- ¹³ Byar DP, Simon RM, Friedewald WT, *et al.* Randomized clinical trials. Perspectives on some recent ideas. *N Engl J Med* 1976;**295**:74-80.
- ¹⁴ Armitage P. *Sequential medical trials*. 2nd ed. Oxford: Blackwell, 1975.
- ¹⁵ Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979;**300**:1242-5.
- ¹⁶ Horwitz BI, Feinstein AR. Methodologic standards and contradictory results in case-control research. *Am J Med* 1979;**66**:556-64.

*This is the second in a series of eight articles.
No reprints will be available from the author.*

Statistics and ethics in medical research

III How large a sample?

DOUGLAS G ALTMAN

Whatever type of statistical design is used for a study, the problem of sample size must be faced. This aspect, which causes considerable difficulty for researchers, is perhaps the most common reason for consulting a statistician. There are also, however, many who give little thought to sample size, choosing the most convenient number (20, 50, 100, etc) or time period (one month, one year, etc) for their study. They, and those who approve such studies, should realise that there are important statistical and ethical implications in the choice of sample size for a study.

A study with an overlarge sample may be deemed unethical through the unnecessary involvement of extra subjects and the correspondingly increased costs. Such studies are probably rare. On the other hand, a study with a sample that is too small will be unable to detect clinically important effects. Such a study may thus be scientifically useless, and hence unethical in its use of subjects and other resources. Studies that are too small are extremely common, to judge by surveys of published research.^{1 2} The ethical implications, however, have only rarely been recognised.^{3 4}

The approach to the calculation of sample size will depend on the complexity of the study design. I will discuss it here in the context of trying to ascertain whether a new treatment is better than an existing one, since it will help if the ideas are illustrated by one of the most common types of research.

Significant tests and power

Despite their widespread use in medical research significance tests are often imperfectly understood. In particular, few medical researchers know what the power of a test is. This is perhaps because most simple books and courses on medical statistics do not discuss it in any detail, even though it is a concept fundamental to understanding significance tests. Some of the general implications, however, are well appreciated, such as the awareness that the more subjects there are, the greater the likelihood of statistical significance.

Formally, the power of a significance test is a measure of how likely that test is to produce a statistically significant result for a population difference of any given magnitude. Practically, it indicates the ability to detect a true difference of clinical importance. The power may be calculated retrospectively to see how much chance a completed study had of detecting (as

significant) a clinically relevant difference. More importantly, it may be used prospectively to calculate a suitable sample size. If the smallest difference of clinical relevance can be specified we can calculate the sample size necessary to have a high probability of obtaining a statistically significant result—that is, high power—if that is the true difference. For a continuous variable, such as weight or blood pressure, it is also necessary to have a measure of the usual amount of variability. A simple example will, I hope, illustrate the relation between the sample size and the power of a test.

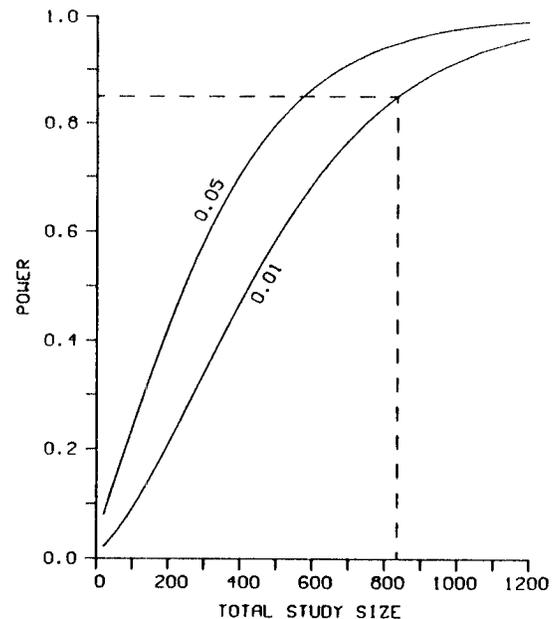


FIG 1—Relation between sample size and power to detect as significant ($p < 0.05$ or $p < 0.01$) a difference of 0.5 cm when standard deviation is 2 cm.

AN EXAMPLE

Suppose we wish to carry out a milk-feeding trial on 5-year-old children when a random half of the children are given extra milk every day for a year. We know that at this age children's height gain in 12 months has a mean of about 6 cm and a standard deviation of 2 cm. We consider that an extra increase in height in the milk group of 0.5 cm on average will be an important difference, and we want a high probability of detecting a true difference at least that large.

Figure 1 shows the power of the test for a true difference of

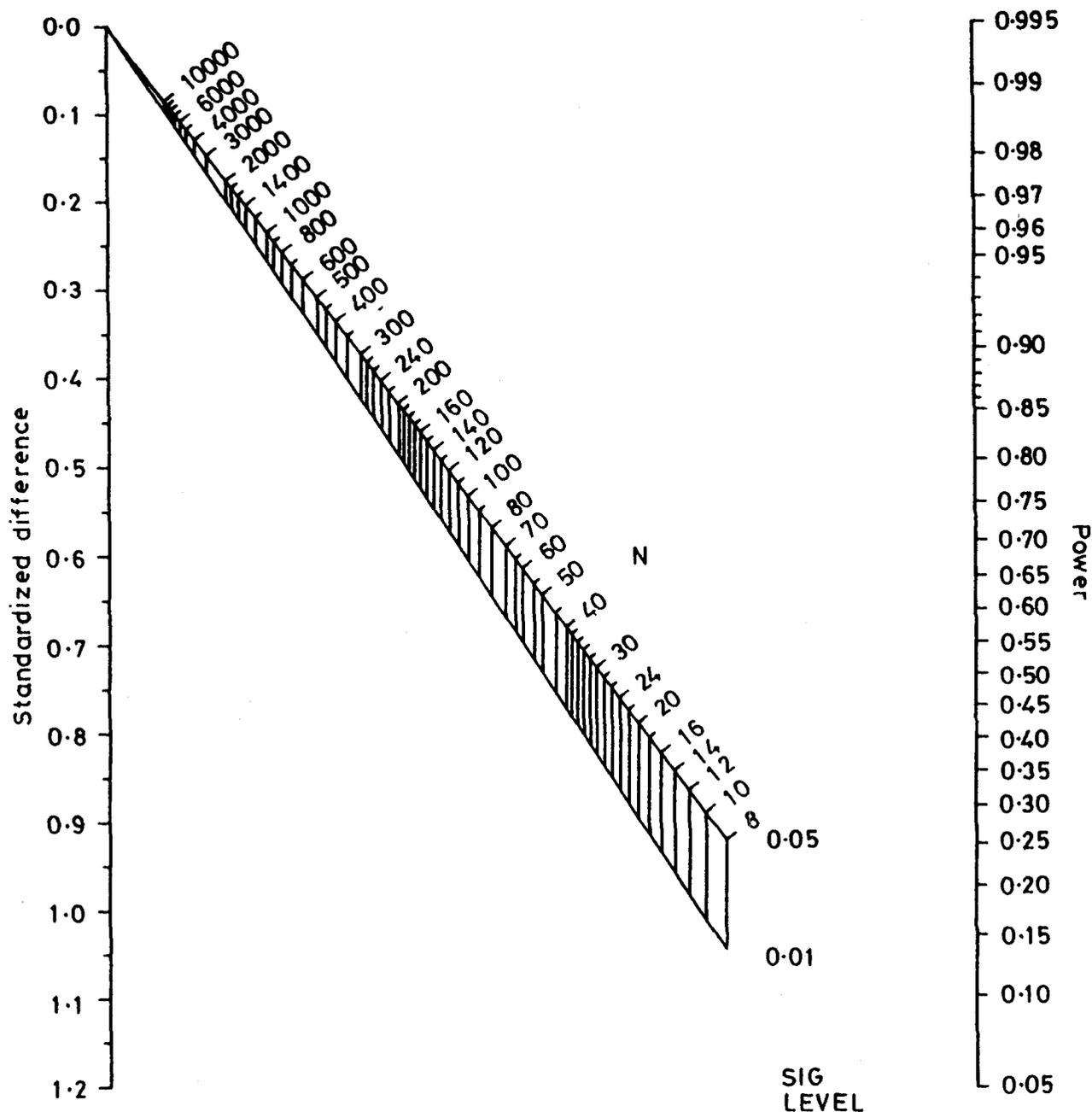


FIG 2—Nomogram for a two-sample comparison of a continuous variable, relating power, total study size, the standardised difference, and significance level.

0.5 cm. The increase in power with increasing sample size is clearly seen, as is the relation with the significance level. For any given sample size the probability of obtaining a result significant at either the 5% or 1% level, given a true difference in growth of 0.5 cm, can be read off. Power of 80-90% is recommended; fig 1 shows that to achieve an 85% chance of detecting the specified difference of 0.5 cm significant at the 1% level, we would need a total of about 840 children.

If we are told that we can have at most 500 children in all, what will the power be now? Figure 1 shows that the power drops from 85% to 60%. We are now more than twice as likely to miss a true difference of 0.5 cm at the 1% level, although the power is still about 80% for a test at the 5% level of significance. Alternatively, and not shown by fig 1, this size of study achieves the same power as the larger one for a difference of 0.65 cm instead of 0.5 cm. Whether or not this is thought sufficient will depend on how far one is prepared to alter one's criteria of acceptability for the sake of expediency. Although

they are to some extent arbitrary, it is generally advisable to stick closely to the pre-stated criteria.

A NEW SIMPLE METHOD

The formula on which these calculations are based is not particularly simple. Graphs are preferable, but because so many variables are concerned, a large set of graphs like fig 1 would be necessary to calculate sample size for any problem. Greater flexibility, however, is achieved by the nomogram shown in fig 2. This makes use of the standardised difference, which is equal to the postulated true difference (usually the smallest medically relevant difference) divided by the estimated standard deviation. So in the previous example the standardised difference of interest was $0.5/2.0=0.25$. The nomogram is appropriate for calculating power for a two-sample comparison of a continuous measurement with the same number of subjects in each

group. The only restriction is the common requirement that the variable that is being measured is roughly Normally distributed.

The nomogram gives the relation between the standardised difference, the total study size, the power, and the level of significance. Given the significance level (5% or 1%),* by joining with a straight line the specific values for two of the variables the required value for the other variable can easily be read off the third scale. By using this nomogram, it is both simple and quick to assess the effect on the power of varying the sample size, the effect on the required sample size of changing the difference of importance, and so on. It is easy to confirm the earlier calculations for the milk-feeding trial.

An estimate of the standard deviation should usually be available, either from previous studies or from a pilot study. Note that the nomogram is not strictly appropriate for retrospective calculations. Although it will be reasonably close for samples larger than 100, for smaller samples it will tend to overestimate the power.

QUALITATIVE DATA

For many studies the outcome measure is not continuous but qualitative—for example, where one is looking for the presence or absence of some condition or comparing survival rates. Peto *et al*⁵ have discussed calculating sample size for such studies, and they emphasise the problem of getting enough subjects when either the condition is rare or the expected improvement is not large. For example, about 1600 subjects would be needed to have a power of 90% of detecting (at $p < 0.05$) a reduction in mortality from 15% to 10%. Although the sample size will in general need to be much larger for studies including qualitative outcome measures, the logic behind the calculations is exactly the same as with continuous data, except that a prior estimate of the standard deviation is not needed. Several authors have published graphs for general use.⁶⁻⁸

OTHER TYPES OF STUDY

Sequential designs are similarly amenable to the incorporation of considerations of power at the design stage. Indeed, it is probably much more common here than for ordinary randomised studies. For these, and for more complicated designs, it may be particularly helpful to enlist the aid of a statistician when thinking about sample size.

Conclusions

The idea behind using the concept of power to calculate sample size is to maximise, so far as practicable, the chances of finding a real and important effect if it is there, and to enable us to be reasonably sure that a negative finding is strong grounds for believing that there is no important difference. The effect of the approach outlined above is to make clinical importance and statistical significance coincide, thus avoiding a common problem of interpretation.

Before embarking on a study the appropriate sample size should be calculated. If not enough subjects are available then the study should not be carried out or some additional source of subjects should be found.⁵ (It should also be borne in mind that expected accession rates tend to be over-optimistic.) The calculations affecting sample size and power should be reported when publishing results. A study² of 172 randomised controlled trials published in the *New England Journal of Medicine* and the *Lancet* from 1973 to 1976 found that none mentioned a prior estimate of the required sample size, and none specified a clinically relevant difference that might allow calculation of the

power of their study. Obviously in most of these studies such calculations were not done.

It is surprising and worrying that in such an ethically sensitive area as clinical trials so little attention has been given to an aspect that can have major ethical consequences. If the sample size is too small there is an increased risk of a false-negative finding. A recent survey¹ of 71 supposedly negative trials found that two-thirds of them had at least a 10% risk of missing a true improvement of 50%. In only one of the 71 studies was power mentioned as having been considered before carrying out the study. It is surely ethically indefensible to carry out a study with only a small chance of detecting a treatment effect unless it is a massive one, and with a consequently high probability of failure to detect an important therapeutic effect.

This is the third in a series of eight articles.

No reprints will be available from the authors.

References

- Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;299:690-4.
- Ambroz A, Chalmers TC, Smith H, Schroeder B, Freiman JA, Shareck EP. Deficiencies of randomized control trials. *Clinical Research* 1978; 26:280A.
- Newell DJ. Type II errors and ethics. *Br Med J* 1978;iv:1789.
- Anonymous. Controlled trials: planned deception? *Lancet* 1979;i:534-5.
- Peto R, Pike MC, Armitage P, *et al*. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I Introduction and design. *Br J Cancer* 1976;34:585-612.
- Aleong J, Bartlett DE. Improved graphs for calculating sample sizes when comparing two independent binomial distributions. *Biometrics* 1979;35:875-81.
- Boag JW, Haybittle JL, Fowler JF, Emery EW. The number of patients required in a clinical trial. *Br J Radiol* 1971;44:122-5.
- Mould RF. Clinical trial design in cancer. *Clin Radiol* 1979;30:371-81.

*As in the example these are two-tailed significance levels.

Statistics and ethics in medical research

Collecting and screening data

DOUGLAS G ALTMAN

Even with an impeccable design there are many ways in which a study can go wrong when the data are being collected. In general, the more complicated the design the more chance there is of the study not being carried out properly. As an example, consider this historic study. The story was related by "Student" (he of *t*-test fame):

"In the Spring of 1930 a nutritional experiment on a very large scale was carried out in the schools of Lanarkshire. For four months 10 000 schoolchildren received three-quarters of a pint of milk per day; 5000 of these got raw milk and 5000 pasteurised milk; another 10 000 children were selected as controls, and the whole 20 000 children were weighed and their height was measured at the beginning and end of the experiment."¹

There was no power problem here. The study found that children getting extra milk gained more weight in the period than did the controls. But did the extra milk cause the extra gain? The figure is a simplified chart showing the weight changes for girls during the study. Since the two milk groups are very similar, only one is shown here. There are two striking features of this graph. The first is that the controls were in all cases heavier than those getting extra milk (they were taller too). This can be easily explained by the discovery that some of the teachers who

allocated children to groups had juggled the randomisation to enable the poorer children to get the extra milk.

The second curious feature is that the observed growth rate in each group was much less than would be expected by looking at the next age group. The explanation for this is also very simple. The study began in February and ended in June, and the children were weighed on both occasions with their clothes on. The short-fall in weight increase is thus largely due to a different amount of clothing, and the smaller effect in the milk feeding group can be explained by the poorer children wearing relatively fewer clothes in winter.

It may be thought that errors such as these are really obvious, and nobody would make such mistakes nowadays. Two points may be made about the altruistic adjustment of the randomisation. Firstly, this procedure is not unknown in more recent times. Carleton *et al*² reported that strongly motivated doctors may upset trials by transilluminating envelopes containing the names of drugs in order to find the desired treatment. However well-intentioned, such underhand activities are by their nature likely to go undetected and can invalidate a whole study. Doctors should not agree to participate in a randomised controlled trial if they have a prior preference for one treatment. Equally, the study sample should not include subjects for which one treatment is clearly medically preferable. A trial where either of these conditions was broken would be unethical.³

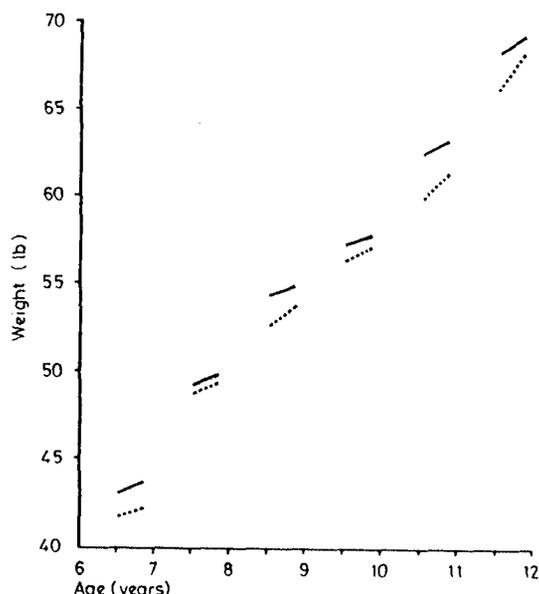
The second point relating to the allocation of subjects to treatments is that a major reason for random allocation is to eliminate the effect of both deliberate and unconscious biases. If the groups are not selected randomly it will be impossible to know whether any observed treatment effect is genuine, as in the

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSc, medical statistician (member of scientific staff)

Lanarkshire milk trial. So what reliability can we place on the results of a study in which patients were allocated to treatments "nearly at random"?

The other error in the Lanarkshire study, that of weighing children with full clothing at different times of the year, would be unlikely to be made in that form now. Errors of this sort, however, are very easy to make, and usually occur when a source of variation is overlooked. For example, in studies looking for small differences it may be important to allow for the fact that height and blood pressure are less in the evening than in the morning, or that lung function is better in summer than in winter. Failure



Lanarkshire milk experiment: comparison of control group (—) and milk feeding group (---) showing mean weight at beginning and end of study for each yearly age group.

to allow for such things can lead to two effects being "confounded" or inseparable. So, in the milk study we cannot say how much of the difference between the groups was due to the milk, how much to the non-random allocation, and how much to the changes in clothing.

Perhaps to try to insure against this sort of problem, it is quite common for a study to collect information on anything that might possibly be of some value or interest. This seems particularly common in surveys, where one is not always investigating a specific issue but looking at a general situation. If information is being collected by questionnaire, however, then increasing the number of questions may lower the response rate, with the results being less reliable as a consequence. Further, excessive amounts of information may reduce the care given to data collection.

Data screening

Before proceeding to the analysis, some degree of data screening should be carried out. By screening is meant checking so far as is possible that the recorded values are plausible, since one can not usually know if the data are correct. Simple data sets obviously need minimal checking in comparison with studies concerning a large amount of information for each subject.

Screening the data (sometimes called cleaning or validation) entails checking that for each variable all the observations are within reasonable limits. Where feasible, each variable should also be cross-checked against other relevant information. This may show inconsistencies such as an 18-year-old woman with six children. It may also show that values that appeared odd are quite compatible with other data.

Much can be learnt from an initial close examination of the data, taking variables both one and two at a time, using histograms and scatter diagrams.⁵ As well as identifying outliers, such screening of the data should disclose whether it will be necessary to transform any of the variables before analysis. It will also help to discover if any observations are missing. All of these aspects merit examination.

WHAT CAN WE DO ABOUT OUTLIERS?

Outliers are observations that are not compatible with the rest of the data. Typically there may be one or two such values in a set of data, but they can have an unduly large influence on the results of an analysis.

The first thing to do with suspicious values is to make sure they have not been incorrectly transcribed. Any impossible value should be treated as missing data, but defining what is impossible may be very difficult. For example, how large would a value for length of gestation or maternal age be before it was considered impossible?

If an outlying observation appears correct in that the value is possible (although unlikely) and there is no evidence to suggest that it is wrongly recorded, then it should not be excluded from the analyses. It is particularly bad to remove such values purely on the grounds that they are the smallest or largest.

In small samples outlying values may have a very large influence on the results—for example, a regression line will be "pulled towards" outlying values. Ranking methods can be used, but they are generally only useful for testing hypotheses, not for the estimation of means, standard deviations, regression slopes, and so on.

WHY TRANSFORM DATA?

When analysing continuous variables (height, blood pressure, serum cholesterol, etc) it is usual to make use of a "family" of statistical analyses, including *t* tests, regression, and the analysis of variance, that make important assumptions about the data. Such analyses are not valid if these criteria are not met.

The best known example of this is when data display skewness instead of the required symmetric Normal (Gaussian) distribution. All of the above methods have some sort of Normality assumption. In such cases it is often possible to find a mathematical transformation for the data that will make the analysis valid.⁶ By far the most common transformation used in medical research is the logarithmic transformation, needed, for example, for various biochemical measurements.⁷ It is worth noting that an appropriate transformation may also have the effect of making previously suspicious values become quite reasonable.

Although it is obvious that the more nearly the underlying assumptions are met the more reliable will be the results, it is unfortunately not possible to say how far the raw data can deviate from the ideal before the results become invalid. Because of the subjective nature of this problem expert help can be particularly helpful here.

WHAT CAN WE DO ABOUT MISSING DATA?

An important distinction must be made between data that are missing through random misfortune (if some forms are mislaid, for instance) or for a reason directly or indirectly related to the study itself. Most studies have a few accidentally missing observations. These cases can usually be omitted without greatly affecting the results. It may be thought preferable to include a subject for any analyses for which data exist, only excluding him when the relevant observation is missing. This procedure can cause complications in interpretation, however, as each analysis will be based on different subjects, and is better avoided if possible.

It is also common to have data missing through a subject's refusal to supply information or to participate in a study. The problem here is that refusers are often an atypical subgroup. In a survey it may be possible to study what is known about the refusers to see if and how they do differ from participants, and to try to estimate the effect on the results. Clearly a high refusal rate will mean that little sensible extrapolation from the sample to the population is possible.

In a randomised trial it is essential that refusers (or withdrawals) are considered as part of the group to which they were allocated.³ A good example is given by a study⁸ of the sudden infant death syndrome. High-risk infants were randomly allocated to observed and control groups, where observation consisted of increased health visitor surveillance. In the control group, where active participation did not need to be sought, there were nine unexpected deaths out of 922 infants, a rate of 9.8 per thousand. In those allocated to the "observed" group, there were two unexpected deaths out of 627 who agreed to participate (3.2 per thousand), and three out of 210 among those who refused (14.3 per thousand). This is a good example of the commonly found poor prognosis among refusers.

The purpose of a randomised trial is to be able to make comparisons between randomly allocated groups. Some trials have "observed controls" where one randomly chosen group is offered treatment while the other group is just observed. Any refusing treatment must still be considered with the treated group; otherwise the two groups will no longer be comparable (the control group do not have a chance to refuse), and it will not be possible to draw valid conclusions. Such trials are thus comparisons of different treatment policies. Alternatively trials can have "placebo controls," when only those subjects who give their informed consent to participate are randomised. Such studies give a direct comparison of treatments, although on a less representative group of subjects, but they are not always practical. The two approaches are discussed and illustrated in Meier's fascinating and very readable account of the Salk vaccine trial.⁹

The health visitor surveillance study had observed controls, so that all of those allocated to the observation group should be considered together. This gives five unexpected deaths out of 837, which is a rate of 6.0 per thousand, and is not nearly significantly different from the control group. The authors excluded the refusers from their analysis, giving a much larger apparent effect of observation (although still not statistically significant). In contrast, a recent study¹⁰ comparing treatments for suspected myocardial infarction included withdrawals from the trial when analysing the data.

Another class of missing data is censored data—that is, values that cannot be measured. One common source is in the measurement of substances present in such low concentrations that some of the samples are below the sensitivity of the equipment being used. Another is where records are kept of the length of time for some event to happen (survival data) or the length of duration of some phenomenon, and the experiment is terminated before an answer can be obtained for all subjects. Censored data are clearly very different from missing observations, and must not be excluded from analysis; this would severely affect the results as these are the most extreme observations. Such data sets can be analysed by non-parametric (ranking) methods if only a few observations are censored at the same point. If censoring is at different values (as in survival studies) more rigorous statistical methods are necessary.

Conclusions

Problems with data collection are often the result of the failure at the design stage to anticipate unusual circumstances. This is one reason why large studies ought to have a pilot phase to try to spot any major deficiencies. It is because we cannot foresee everything that may be relevant that randomisation is so important, but it must be strictly adhered to.

The wide availability of computers and calculators has made

it much easier to carry out statistical analyses. Unfortunately, they have also made it easy to produce results without ever really studying the raw data. Before embarking on analysis there is much that can be learnt from simple inspection of variables both singly and in pairs. Such screening of the data, especially graphically, as well as greatly helping to prepare the data for analysis, can also provide considerable insight into the relationships between variables.

The issues of data screening discussed in this article generally receive scant attention. Yet they concern strategic decisions that can have major implications for the ensuing results, as the criticism¹¹ of the Anturane study¹² has shown. They directly affect the validity and thus the ethics of research.

This is the fourth in a series of eight articles. No reprints will be available from the author.

References

- 1 "Student." The Lanarkshire milk experiment. *Biometrika* 1931;23:398-406.
- 2 Carleton RA, Sanders CA, Burack WR. Heparin administration after acute myocardial infarction. *N Engl J Med* 1960;263:1002-5.
- 3 Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I Introduction and design. *Br J Cancer* 1976;34:585-612.
- 4 Clarke BF, Campbell IW. Long-term comparative trial of glibenclamide and chlorpropamide in diet-failed, maturity-onset diabetics. *Lancet* 1975;i:245-7.
- 5 Healy MJR. The disciplining of medical data. *Br Med Bull* 1968;24:210-4.
- 6 Armitage P. *Statistical methods in medical research*. Oxford; Blackwell, 1971:350-9.
- 7 Flynn FV, Piper KAJ, Garcia-Webb P, McPherson K, Healy MJR. The frequency distributions of commonly determined blood constituents in healthy blood donors. *Clin Chim Acta* 1974;52:163-71.
- 8 Carpenter RG, Emery JL. Final results of study of infants at risk of sudden death. *Nature* 1977;268:724-5.
- 9 Meier P. The biggest health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine. In: Tanur JM, Mosteller F, Kruskal WH, et al, eds. *Statistics: a guide to the study of the biological and health sciences*. San Francisco; Holden-Day, 1977:88-100.
- 10 Wilcox RG, Roland JM, Banks DC, Hampton JR, Mitchell JRA. Randomised trial comparing propranolol with atenolol in immediate treatment of suspected myocardial infarction. *Br Med J* 1980;280:885-8.
- 11 Kolata GB. FDA says no to Anturane. *Science* 1980;208:1130-2.
- 12 The Anturane Reinfarction Trial Research Group. Sulfipyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med* 1980;302:250-6.

Statistics and ethics in medical research

V—Analysing data

DOUGLAS G ALTMAN

The incorrect analysis of data is probably the best known misuse of statistical methods, largely due to a series of reviews¹⁻³ that have shown how common such errors are in published papers. Nevertheless, these mistakes, which tend to be in the use of the simpler techniques, continue to proliferate. The mishandling of statistical analysis is as bad as the misuse of any laboratory technique. Both can lead to incorrect answers and conclusions and are thus unethical because they render research valueless.

In this article I will look briefly at problems associated with simple significance tests and will consider in more depth some less well-appreciated difficulties associated with two other common techniques—correlation and regression. I will then look at two specific medical problems that often result in incorrect analyses.

Errors in common statistical analyses

Nowadays some types of statistical analyses are seen so often in medical publications that their use is taken for granted. Everyone knows them, but the evidence suggests that many people do not know how to use them properly, or when *not* to use them. For example, Gore *et al*⁴ found at least one such error in about half of the papers containing statistical analyses that they reviewed.

t TESTS AND χ^2 TESTS

The *t* tests to compare two groups of measurements are used extremely widely, but often incorrectly.²⁻⁴ The problems usually relate to the data not complying with the underlying statistical assumption that the two sets of data come from populations that are Normal and have the same variance. Another serious error is to ignore the fact that the two sets of measurements relate to the same (or matched) individuals, in which case the paired *t* test is needed. These problems are fairly familiar and have been well illustrated by White⁵ so I will not consider them further here.

Although generally posing fewer problems, χ^2 tests for comparing proportions also suffer some abuse, notably where there are too few observations. The sample size constraint also

applies to the form of χ^2 test which simply entails comparing observed and expected frequencies. This method was used to compare observed numbers of deaths from five types of leukaemia (0, 1, 2, 4, 0) with their respective "expected" numbers (2, 1, 1, 3, 0),⁵ but seven deaths is far too few for such an analysis to be valid.

CORRELATION

Perhaps one harmful side effect of the vast increase in availability of computing power is that the distinct statistical analyses of correlation and regression have become greatly confused. This is probably because of the close similarity between the mathematical calculations rather than for any logical reason, for it is relatively rare that one is truly interested in both analyses.

The correlation coefficient is a measure of the degree of linear (or "straight line") association between two continuous variables. If the relationship between the two variables is curved the correlation may be an artificially low measure of association. Alternatively, the correlation may be artificially high if a few observations are very different from the rest. For these reasons it is unwise to place any importance on the magnitude of the correlation without looking at a scatter plot of the data.

Misleading correlations can also be obtained if the data relate to different groups of subjects having different characteristics. Adam⁶ looked at the relationship between body weight and the proportion of sleep that was rapid eye movement sleep in 16 adults, and found a rank correlation of 0.78. The original high correlation, however, was partly due to the men having higher values of both variables, for the correlations for men and women separately were 0.61 and 0.37 respectively. A further incorrect procedure is to use data comprising more than one observation per individual.

The main problem is that the test of significance of a correlation coefficient, which is a test of the null hypothesis of no association (zero correlation), is based on the assumption of joint Normality of the two variables. This is characterised by the data points having a roughly elliptical shape in the scatter diagram. If this is not so the correlation will be misleading and the test of significance invalid. The distributional assumption may be overcome either by transformation of the data, or by the calculation of "rank" correlation, which makes no important assumptions.

In medical research correlations are greatly overused, perhaps because they are easy to calculate and are measured on a scale that is independent of the data. Correlation ought really to be considered to be mainly an investigative analysis, suggesting

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSC, medical statistician (member of scientific staff)

areas for further research; for forming hypotheses rather than for testing them.

REGRESSION

The rationale for regression analysis is very different. In regression we are interested in describing mathematically the dependence of one variable on one or more other variables. In the simple linear case we are calculating the equation of the "best" straight line relating to the so-called "dependent" variable (Y) to the "independent" (or explanatory) variable (X).^{*} For example, we might be interested in the dependence of lung function on height or of blood pressure on age. The appropriateness of a linear relationship can again best be verified by means of a scatter plot.

The most important underlying assumption in regression is that the Y variable is Normally distributed with the same variance for each value of X, and major departures from this condition can usually be detected by eye. There are no restrictions on X, so that it is perfectly valid, for example, to choose a wide range of X values to get a better estimate of the regression line. This would, however, artificially inflate the correlation coefficient, although correlations are often calculated from such data.

Regression is used to estimate a dependence relationship. The resulting equation can be used to predict Y (say, lung function) from X (height) for an individual. The difference between an individual's actual and predicted lung functions can be used as a measure of lung function standardised for height.

Examples of improper practices are the use of the regression equation to predict the Y variable for values of the X variable outside the range of the original data set (called extrapolation); the fitting of a straight line where the data show curvature; the use of a Y on X regression equation to predict X from Y (except in certain circumstances); and the use of simple regression where there are heterogeneous subgroups (the correct technique being analysis of covariance). Unless there is a plot of the data most of these procedures may be undetectable in a published paper.

Method comparison studies

Some of the practical problems in analysing data, notably the choice of the correct analysis to match the relevant hypothesis, are well illustrated by the problems of method comparison studies.

In medical research it is quite common to carry out a study to compare two different methods of measuring something. This may be to compare measurements made with some new piece of equipment with the "true" measurements, but it is more often to compare two different measuring devices where neither can be said to give the truth. (A similar problem arises when comparing the same measurement on different occasions.)

The obvious first step in the analysis is to plot the values obtained by each method as a scatter diagram. To judge from publications, the apparently obvious second step is to calculate the correlation between the two measurements. This is, however, a completely misguided approach, stemming from the common failure to appreciate what information the correlation coefficient gives.

An example of the false reasoning that is very common in published work is given by a study⁷ comparing two methods of assessing the gestational age of newborn babies; one was the much-used Dubowitz method based on neurological and physiological signs and the other the Robinson method, which is based on neurological signs only. The scatter diagram showed only moderate agreement. The correlation between the two methods, however, was 0.85, and the authors argued directly

^{*}These terms simply denote which variable is considered to be dependent on the other.

from this that the two methods agreed well and that it would be reasonable to use the simpler method.

To test an observed correlation coefficient for statistical significance is to test how likely the observed result would be under the "null hypothesis" that the two variables were not associated at all. This is patently ludicrous when the two variables are obviously associated by their very nature; we would be astonished to find that two methods of measurement were uncorrelated. In fact, it can be shown that in these circumstances the magnitude of the correlation largely reflects the spread of the measurements. As such, its use is completely erroneous in this context.

What we really want to know in these studies is how well the two measures agree. The simplest approach is to calculate the difference between two measurements for each subject. The mean of these differences will then be a measure of accuracy (or bias) and the standard deviation a measure of precision. Both bias and precision are necessary in order to assess agreement. The between-method differences may tend to increase as the measurements increase, in which case it may be necessary to transform the data before analysis. With more than two methods, or if repeat observations are made (which is desirable), the more general analysis of variance must be used.

Hunyor *et al*⁸ did calculate the mean and standard deviation of paired differences when comparing various sphygmomanometric methods with intra-arterial blood pressures, but then based their statements about relative accuracy on the high correlations they found. They studied hypertensives only; had they studied some normotensives as well they would undoubtedly have observed higher correlations, but these would not have implied any better agreement between methods.

One last point about method comparison studies is that they are often carried out on such small numbers of subjects that the two methods will not be found significantly different unless there is an enormous difference between them. There is considerable potential here for incorrectly finding a new method acceptable, and for such methods to be recommended for widespread use without justification.

Reference ranges

Another area where simple statistical methods are often applied blindly is in the construction of reference (or normal) ranges against which to judge future observations. For example, some people believe that since a range is required, all that is needed is to obtain results from some "normal" subjects and quote the range of values. Apparent differences in reference ranges for the same index can often be attributed to one or more of them having been calculated incorrectly. Also, the sample size taken is often too small to get reliable answers. I have seen a reference range calculated from seven subjects, incorrectly at that, whereas at least 100 observations are needed to get a reliable range.

The usual calculation of a 95% reference range as the mean ± 2 standard deviations is yet again based on the assumption that the data follow a Gaussian or Normal distribution. Often this condition is not fulfilled and we see statements like "The mean ^{99m}Tc uptake in this group was 1.8% \pm SD 1.1%, making the upper limit of normal (mean ± 2 SD) 4.0%."⁹ The unstated lower limit is negative, however, which is nonsense. This type of calculation of a normal range on skew data results in considerably more than the nominal 5% of subjects being classified as "abnormal." The consequence of such a classification may be to perform further tests, so that there is a clear ethical aspect to the construction and interpretation of normal ranges. Even where the range is calculated sensibly there is a strong case for quoting the standard error of the limits, to emphasise the considerable uncertainty involved.

Whether or not the use of such ranges is sensible is beyond the scope of this article; the issues have been clearly discussed by Oldham¹⁰ and Healy.¹¹

Selecting which data to analyse

A rather more subtle problem that can occur in any study is the selection of which data to analyse. Errors may occur when analyses are carried out as a direct result of having seen the data. In a comparison of several groups of subjects it is not valid to select those groups with the highest and lowest values and apply the usual significance test to the means purely on that basis, because the null hypothesis of no difference is inappropriate when the largest difference is being examined. More generally, selection of comparisons to test because they "look interesting" will in the long run result in more than the nominal (say 5%) proportion of falsely positive results.

A second form of selection is to analyse only a subset of the subjects on the basis of their results. In a recent study 30 patients with idiopathic hypercalciuria were given a dietary supplement of unprocessed bran.¹² Only 22 patients "achieved a reduction in urinary calcium," and only these 22 patients were analysed. No data were provided on the other eight subjects, so we can not tell whether they really were a different group or just one end of a distribution of differing responses to the bran, which seems more likely. This procedure is completely unacceptable without justification—anyone can show significant results by analysing only those subjects with the greatest response.

The basic principle is to analyse according to the original hypothesis and experimental design. Other results that look interesting are pointers for further research.

Summary

It is of no value collecting good data if the analysis is inadequate or invalid. The results obtained may then be worthless,

or at best they will fail to realise the true potential of the data. Either way, the value of the whole experiment is diminished to a point where the ethics of the investigation must be called into question.

References

- ¹ Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;195:1123-8.
- ² Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J* 1977;ii:85-7.
- ³ White SJ. Statistical errors in papers in the *British Journal of Psychiatry*. *Br J Psychiatry* 1979;135:336-42.
- ⁴ Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1980;81:1-7.
- ⁵ Tabershaw IR, Lamm SH. Benzene and leukaemia. *Lancet* 1977;ii:867-8.
- ⁶ Adam K. Bodyweight correlates with REM sleep. *Br Med J* 1977;ii:813-4.
- ⁷ Serfontein GL, Jaroszewicz AM. Estimation of gestational age at birth. *Arch Dis Child* 1978;53:509-11.
- ⁸ Hunyor SN, Flynn JM, Cochineas C. Comparison of performance of various sphygmomanometers with intra-arterial blood pressure readings. *Br Med J* 1978;iii:159-62.
- ⁹ Van 'T Hoff W, Pover GG, Eiser NM. Technetium-99m in the diagnosis of thyrotoxicosis. *Br Med J* 1972;iv:203-6.
- ¹⁰ Oldham PD. The uselessness of normal values. In: Arcangeli P, Cotes JE, Cournand A, eds. *Introduction to the definition of normal values for respiratory function in man*. Turin: Panminerva Medica, 1969:49-56.
- ¹¹ Healy MJR. Normal values from a statistical viewpoint. *Bulletin de l'Académie Royale de Médecine de Belgique* 1969;9:703-18.
- ¹² Shah PJR, Green NA, Williams G. Unprocessed bran and its effect on urinary calcium excretion in idiopathic hypercalciuria. *Br Med J* 1980;281:426.

This is the fifth in a series of eight articles. No reprints will be available from the author.

Statistics and ethics in medical research

VI—Presentation of results

DOUGLAS G ALTMAN

A very important aspect of statistical method is the clear numerical and graphical presentation of results. Although many statistical textbooks and courses discuss simple visual methods such as histograms, bar charts, pie charts, and so on, they are usually introduced as descriptive or investigative techniques. It is uncommon to find discussion of how best to present the results of statistical analyses. This is surprising, since the interpretation of the results, both by the researcher and by later readers of the paper, may be critically dependent on the methods used to present the results.

Little need be said here about the simple visual methods already mentioned—they are well covered by Huff.¹ The problems associated with graphs, however, are rather more important.

Graphical presentation

In 1976 a Government publication² gave examples of some past successes in preventive medicine. One of these examples concerned the introduction in the 1930s of mass immunisation against diphtheria. Figure 1(a) shows their presentation of childhood mortality from diphtheria from 1871 to 1971. This appears to show that the introduction of immunisation resulted in a rapid decline in mortality. In their figure, however, mortality is plotted on a logarithmic scale and shows proportional changes. When the data are plotted on a linear scale,³ as in fig 1(b), the visual effect is quite different, as is the interpretation. From this figure we can see that over the period in question mortality from diphtheria had been dropping very quickly, and this specific preventive measure was adopted relatively late in the day. This is not to say that the introduction of immunisation was not effective, but that the degree of its effectiveness that one accepts depends considerably on which way the data are presented.

For experimental data it is unlikely to be appropriate to transform the scale of one or both axes unless it has been necessary to carry out the analysis on transformed data. For example, if analysis has been carried out on log data, it is probably better to show a scatter diagram with a log scale to demonstrate that the transformed data comply with the appropriate assumptions.

Scatter diagrams and regression

For simple data sets scatter diagrams are tremendously helpful. By showing all the data it is much easier for the reader to evaluate the analyses that were carried out. It is essential, however, that coincident points are indicated in some way. If there are different subgroups within the data set (different sexes perhaps) these may be indicated by means of different symbols. This will provide extra information at no expense, and will help to show the appropriateness (or otherwise) of analysing the data as one set, or for each subgroup separately.

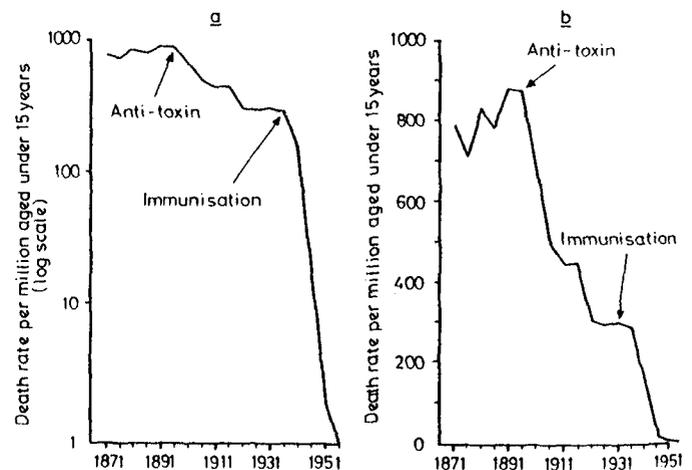


FIG 1—Childhood mortality from diphtheria (a) on a log scale² (b) on a linear scale.³

Unfortunately, to many people scatter diagrams automatically suggest the calculation of correlations and the fitting of regression lines, even though one or both of these methods may be invalid or of no interest. One often sees scatter diagrams where a straight line has been drawn through the data but no reference is made to it, either in the figure or in the text. Perhaps the intention is to show that the data have been “properly analysed,” but presentations like this demonstrate the reverse.

How should results of regression analyses be presented? This will depend partly on the context. For example, if the analysis shows that the relationship between two variables is too weak to be of practical value, then there may be little point in quoting the equation of the line of best fit. If the equation is given then the standard error of the slope (and of the intercept if this is of practical importance) and the number of observations are important information. One other quantity is necessary, how-

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ
DOUGLAS G ALTMAN, BSc, medical statistician (member of scientific staff)

ever, before one can make full use of a regression equation. The equation can be used to estimate the variable Y for any new value of the variable X. Such an estimate is, however, of limited value without some measure of its uncertainty, for which it is additionally necessary to have the residual standard deviation.⁴ This is a useful quantity in its own right, as it is a measure of the variability of the discrepancies (residuals) between the observations and the values predicted by the equation and is thus a measure of the "goodness of fit" of the regression line to the data. The residual standard deviation is rarely supplied in papers, so that it is impossible to know what uncertainty is attached to the use of the regression line for estimating Y from X.

Whatever information is presented, it is vital that it is unambiguous. The following equation may be meant to give much of the information but the meaning of the last term is unclear:

$$\text{TBN(g)} = (28.8 \cdot \text{FFM(kg)} + 288) \pm 8.5\%$$

The paper⁵ from which this example comes also includes an example of a type of incorrect visual presentation of a regression equation—namely, the extension of the line well beyond the range of the data. This practice is extremely unreliable and potentially misleading, and can rarely be justified.

Variability

Despite its obvious importance and its almost universal presence in scientific papers, the presentation of variability in medical journals is a shambles. It is quite clear that some practices are now considered obligatory purely because they are widely used and accepted, not because they are particularly informative.

Much of the confusion may arise from imperfect appreciation of the difference between the standard deviation and the standard error. In simple terms the standard deviation is a measure of the variability of a set of observations, whereas the standard error is a measure of the precision of an estimate (mean, mean difference, regression slope, etc) in relation to its unknown true value. Despite this clear distinction in meaning, many people seem to have an innate preference for one or the other; some time ago I looked at all the issues of the *BMJ*, *Lancet*, and *New England Journal of Medicine* for October 1977 and found only three papers that used both, although 50 used either one or the other. Similar results were found in a much larger study.⁶ It has been suggested that perhaps the standard error of the mean is more popular because it is always much smaller,^{6, 7} and this may well be so.

STANDARD DEVIATION

The standard deviation, which describes the variability of raw data, is often presented by attaching it to the corresponding mean using a \pm sign: "The mean... was 30 mg (SD ± 4.6 mg)," or something similar. This presentation suggests that the standard deviation is ± 4.6 mg, but the standard deviation is always a positive number.⁸ More importantly, it also suggests that the range from mean - SD to mean + SD (25.4 to 34.6 mg) is meaningful, but this is not so unless one is genuinely interested in the range encompassing about 68% of the observations. In general, the most useful range is probably the mean ± 2 SD, within which about 95% of the observations lie. This range is 20.8 to 39.2, which is twice as wide as that implied by " ± 4.6 mg." Such ranges apply only if the observations are approximately Normally distributed. Otherwise, although the standard deviation can be calculated, it may not convey much information about the spread of the data. In such cases the median and two centiles (say the 10th and 90th or the 5th and 95th for larger samples) will provide better information.^{9, 10} The range of values may also be of interest, but it is highly dependent on the number of observations and is very sensitive to extreme or outlying

observations. Alternatively, the omission of the \pm sign leads to an unambiguous although much less informative presentation: "The mean was 30 mg (SD 4.6 mg)."

STANDARD ERRORS

Similar comments apply to the presentation of standard errors. Here the most often quoted range of \pm SE around an estimate is that within which we can be about 68% sure that the true value lies, whereas the 95% range is twice as wide. (For practical purposes these "confidence intervals" apply even when the data are not Normally distributed.) The presentation most usually used (mean \pm SE) is thus misleading in giving the impression of greater precision than has been achieved. Quoting the range mean ± 2 SE is much better, but this is rarely seen. Much confusion would be eliminated if the sign \pm was used only when referring to a range.

ERROR BARS

Error bars are a popular way of displaying means and standard errors. They are usually a visual representation of the range mean \pm SE such as in fig 2. In this example the error bars for A and B do not overlap: does this tell us anything about the difference between the groups?

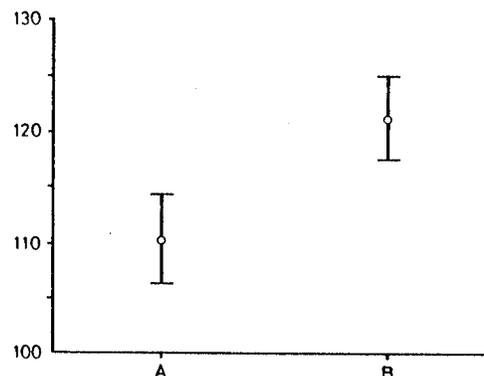


FIG 2—Mean (\pm SE) diastolic blood pressure from two sets of observations.

Suppose A and B represent two different types of sphygmomanometer, and we measure the diastolic pressure of 15 people using each machine. Figure 3(a) shows the results of such an experiment where the agreement is clearly good, but machine B tends to give slightly higher readings. Figure 3(b) shows some data where agreement is generally very poor. Yet both of these

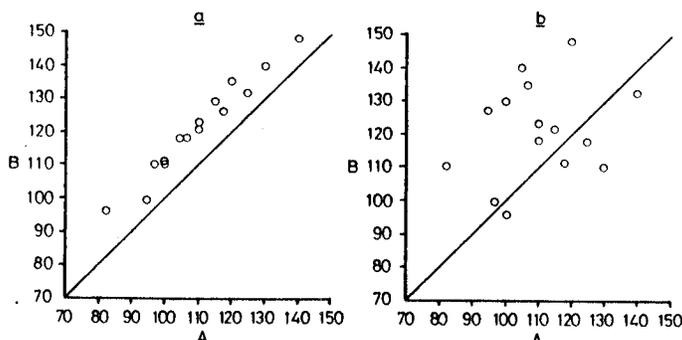


FIG 3—Comparison of diastolic blood pressures measured by two sphygmomanometers on 15 subjects (a) with good agreement but some bias (b) with very poor agreement.

sets of data can be described exactly by the means and SEs in fig 2. This is because fig 2 tells us nothing about differences between machines for each subject. Error bars are thus useless in the case of paired observations.

Now suppose that we wish to compare the diastolic blood pressures of two distinct groups of people, say doctors (group A) and bus-drivers (group B). Figures 4(a) and 4(b) show two possible outcomes. In which case, if either, are the two groups significantly different? It is not easy to tell from the raw data shown that the groups are significantly different in fig 4(a) ($p < 0.05$) but not in fig 4(b) ($p > 0.1$). What would an "error-bar" plot show? Well, again both examples would yield fig 2, showing that the visual impression of non-overlapping bars does not by itself give any information about statistical significance. If the error bars do overlap, however, then the difference between the means is not statistically significant.¹¹

For error bars to be useful they ought to convey useful information about either the precision of individual means or the differences between means. In their usual form they do neither, although my impression is that many people believe that they do both. The use of confidence intervals (mean \pm 2 SE) instead of error bars does at least give useful information about individual means. Although it is sometimes possible to make the visual presentation give an indication of statistical significance, it is probably better to give confidence intervals and, if desired, report on the significance separately.

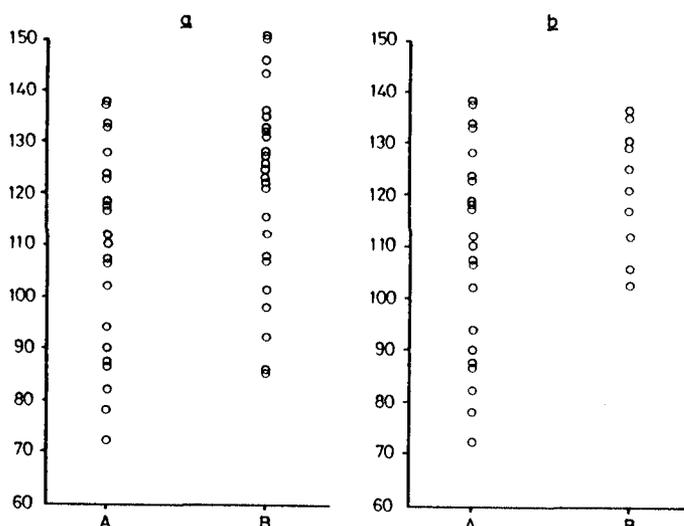


FIG 4 (a) and (b)—Comparisons of diastolic blood pressure in two different groups of subjects.

Numerical precision

One other aspect of presentation that deserves some comment is numerical precision. It is rarely necessary to quote results—means, standard deviations, and so on—to more than three significant figures (that is, excluding leading or trailing zeros). For tabular presentation it may be a positive advantage to reduce the precision of each entry to make any patterns or trends more obvious.¹²

Spurious precision should also be avoided. Examples are the quoting of t or χ^2 values to four decimal places, and a regression slope with seven significant figures (12.97642). My favourite is the summary¹³ of a test of significance as $p < 10^{-54}$, although I must concede that there is only one significant figure here!

Some suggestions

More thought should be given to numerical and visual presentation, rather than automatically following precedent.

Some ways of supplying more information without using more space are:

(1) In a plot information about the spread of data (by ± 2 SD ranges or centiles) can be given as well as means and confidence intervals.

(2) A figure and a table may be combined by using the X axis labels as table column headings. For example, in fig 2 I could have given the mean, SD, range, and sample size for the two groups under the figure using little extra space.

(3) When scatter plots have the same variable on each axis as in fig 3(a) and 3(b), a small histogram of the within-person differences can be added in an otherwise empty corner.

Summary

Whatever results are presented it is vital that the methods are identified. In one survey of over 1000 papers¹⁴ as many as 20% of the procedures were unidentified, and in another it was not clear whether the SD or SE was given in 11% of 608 papers.⁵ It is impossible to appraise a paper in the presence of such ambiguities.

Visual display is a particularly effective way of presenting results. Given alternatives, however, many people might opt for the method of display that fits in better with their beliefs. If decisions are taken as a result of such presentations then there is scope for manipulating events by choice of presentation. This practice is well recognised in the way statistics are sometimes presented in the mass media and advertisements; we should not rule out this phenomenon in the medical world.

This is the sixth in a series of eight articles. No reprints will be available from the authors.

References

- Huff D. *How to lie with statistics*. Harmondsworth: Penguin, 1973.
- Department of Health and Social Security. *Prevention and health: everybody's business*. London: HMSO, 1976.
- Radical Statistics Health Group. *Whose priorities?* London: Radical Statistics, 1976.
- Armitage P. *Statistics in medical research*. Oxford: Blackwell, 1971:150-6.
- Hill GL, Bradley JA, Collins JP, McCarthy I, Oxby CB, Burkinshaw L. Fat-free body mass from skinfold thickness: a close relationship with total body nitrogen. *Br J Nutr* 1978;39:403-5.
- Bunce H, Hokanson JA, Weiss GB. Avoiding ambiguity when reporting variability in biomedical data. *Am J Med* 1980;69:8-9.
- Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1980;61:1-7.
- Gardner MJ. Understanding and presenting variation. *Lancet* 1975;ii:230-1.
- Mainland D. SI units and acidity. *Br Med J* 1977;iii:1219-20.
- Feinstein AR. Clinical biostatistics. XXXVII Demeaned errors, confidence games, nonplussed minuses, inefficient coefficients, and other statistical disruptions of scientific communication. *Clin Pharmacol Ther* 1976;20:617-31.
- Browne RH. On visual assessment of the significance of a mean difference. *Biometrics* 1979;35:657-65.
- Ehrenberg ASC. Rudiments of numeracy. *Journal of the Royal Statistical Society Series A* 1977;140:277-97.
- Vaughan Williams EM, Tasgal J, Raine AEG. Morphometric changes in rabbit ventricular myocardium produced by long-term beta-adrenoceptor blockade. *Lancet* 1977;ii:850-2.
- Feinstein AR. Clinical biostatistics. XXV A survey of the statistical procedures in general medical journals. *Clin Pharmacol Ther* 1974;15:97-107.

Statistics and ethics in medical research

VII—Interpreting results

DOUGLAS G ALTMAN

"... it is a function of statistical method to emphasise that precise conclusions cannot be drawn from inadequate data."

E S PEARSON AND H O HARTLEY¹

The problems of interpretation have already appeared several times in the preceding articles. Obviously the sorts of error already discussed, most likely in design or analysis, may lead to incorrect results and thus erroneous conclusions. But some errors are specific to the interpretation of results, and these I will consider in this article. Most emphasis will be given to tests of significance, since these quite clearly cause great difficulty.

Significance tests

Before tackling some of the trickier issues it is worth making the general point that the sensible interpretation of statistical analysis cannot be independent of the knowledge of what the data are (and how they were obtained).

Table I, for example, shows the results of the comparison of two groups of subjects given different treatments with the outcome for each subject recorded as positive or negative. A

TABLE I—Comparison of outcomes for two treatment groups

Treatment	1	Outcome		Total
		A	B	
	2	8	24	32
Total		12	28	40

χ^2 test $p < 0.05$.

χ^2 test on these data shows a significant association between the grouping and the outcome, but in the absence of further information we are unable to interpret these results. Knowing that the subjects were all pregnant and the outcomes were male and female babies is likely to aid interpretation and increase interest, but the further knowledge that the subjects were all cows will probably lessen interest again, unless you are a farmer. Yet you may be curious to know what the "treatments" were—perhaps there is some relevance for people. Well, all the cows were artificially inseminated; those in group 1 were facing north at the time and those in group 2 were facing south.²

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSc, medical statistician (member of scientific staff)

Given all the information, most people would probably dismiss this as a chance finding, rather than accept it as evidence of an association between the direction the cows were facing and the sex of their calves. This is quite reasonable behaviour if we consider the meaning of statistical significance.

INTERPRETING SIGNIFICANT RESULTS

Like several statistical terms, "significant" is perhaps an ill-chosen one. It should be realised that the level of significance is just an indication of the degree of plausibility of the "null hypothesis," which in the above example was that the outcomes of the two groups were really the same. If the null hypothesis is deemed too implausible we reject it and accept the "alternative hypothesis" that the treatments differ in their effect.

It is ridiculous to lay down rigid rules for something so subjective, especially as interpretation will be greatly influenced by other evidence—few studies are carried out in isolation. As Box *et al*³ have said: "If the alternative hypothesis were plausible a priori, the experimenter would feel much more confident of a result significant at the 0.05 level than if it seemed to contradict all previous experience." Indeed, in the long run one in 20 comparisons of equally effective treatments will be significant at the 5% level (by definition), so to accept all significant results as real⁴ is extremely unwise, as the above data illustrate.

Conventional significance levels (5%, 1%, 0.1%) are useful, but only as guides to interpretation, not as strict rules. To describe a result of $p = 0.05$ as "probably significant"⁵ implies that the interpretation depends on which side of 0.05 p really is. On the contrary, values of p of, say, 0.06 and 0.04 should not lead to opposite conclusions, but to closely similar ones.

One prevalent misconception relates to the precise meaning of p , the significance level; p is the probability of obtaining a result at least as unlikely as the observed one, if the null hypothesis of no effect is true. The last part of this definition is essential; to omit it leads to the common error of believing that p is also the probability so that we make a mistake by accepting the significant result as a real finding. This is just not so, and it is sad to see this view in a paper trying to explain the meaning of significance.⁶ All we can say is that p is the probability of such a result arising if the null hypothesis is true. We obviously do not actually know whether the null hypothesis is true, so the probability of rejecting it in error is also unknown, although this clearly reduces as p gets smaller.

INTERPRETING NON-SIGNIFICANT RESULTS

Every significance test measures the credibility of a null hypothesis—for example, that two treatments are equally

effective. A non-significant result just means that the results were not strong enough to reject the null hypothesis; "not significant" does not imply either "not important" or "non-existent." To consider all non-significant results as indicating no effect of importance is clearly wrong. Conversely, to believe that an observed difference is a real one with an insufficient degree of certainty is to run a large risk of chasing shadows. Thus when reporting "negative" results, it is especially important to give a confidence interval around the observed effect⁷—for example, around the difference between two means.

In the third article I discussed at length the idea of the power of a significance test. It is appropriate to return to the topic of power here. Studies with low power (as a result of inadequate sample size) will often yield results showing effects which, if real, would be of clinical importance, but which are not statistically significant. In general it is safest to consider such non-significant results as being inconclusive (or "not proven"), preferably backed up with a recommendation that further data be collected. When this is not feasible and there are ethical implications, as in the following example, the problem of interpretation is particularly great.

Carpenter and Emery⁹ investigated the possible effect on the incidence of sudden unexpected infant death of an increase in the number of visits by the health visitor to high-risk babies. They found fewer unexplained deaths in the "treatment" group (five out of 837) than in the control group (nine out of 922), but the difference is not nearly statistically significant ($p > 0.5$). From a statistical point of view, the results are inconclusive. Because such deaths are rare, the power of the study was very low; it would have needed a much larger sample to get a clear answer.¹⁰ The authors asked: "Can we reasonably withhold increased surveillance from all high-risk infants?"¹¹ More dispassionately we might ask whether the evidence is really strong enough to justify a change in policy that would presumably necessitate withdrawing health visitors from other activities.

MULTIPLE TESTS OF SIGNIFICANCE

A further difficulty arises when several tests of significance are carried out on one set of data. This may, for example, take the form of looking to see which pairs of a number of groups are significantly different from each other, or which of a number of different factors are related to a variable of interest. Unfortunately, the greater the number of tests carried out, the higher the overall risk of a "false-positive" result. As Meier has pointed out,¹² it is not reasonable to restrict the number of aspects of the data that are investigated purely to relieve the statistician's problems of interpretation. He suggested a good compromise, which is to treat a small number of tests as being of primary importance, "and to regard other findings as tentative, subject to confirmation in future experiments." The level of significance will have some bearing here, since we will be more ready to accept a highly significant finding (say, $p < 0.001$) even in the context of numerous tests.

Association and causation

It is widely believed that "you can prove *anything* with statistics," but it is much more realistic to say that you can establish *nothing* by statistics alone. This is especially true when considering the interpretation of observed associations between two variables. It is easy and often tempting to assume that the underlying relationship is a causal one, even in the absence of any supporting evidence, but many associations are not causal. In particular, misleading associations appear when each of the variables is correlated with a third "hidden" variable. A simple example of this phenomenon is when two variables that change with time display an association in the complete absence

of any causal relationship—for example, the divorce rate and the price of petrol.

The deduction of a causal relationship from an observed association can rarely be justified from the data alone. Support is needed from prior knowledge, including other experimental or observational data. Sometimes, however, such information is not obtainable, and the causal hypothesis can be supported only by allowing for the most likely hidden variables. There are several examples of epidemiological studies producing associations that are not unanimously believed to be causal, such as that between water hardness and cardiovascular mortality. A few people do not even accept that the association between smoking and lung cancer is causal despite the great volume of collateral evidence.

A recent paper¹³ concerning the failure to show a relationship between diet and serum cholesterol concentration gave a salutary reminder that variables may falsely appear to be unrelated. Although a strong relationship between dietary cholesterol and serum cholesterol has been shown in closely controlled dietary studies, the authors showed that a straightforward population study would be likely to miss such an association because of several sources of variability in both variables.

Another difficulty that can beset the interpretation of observed associations is where two possibly causal factors are inseparable. A simple example is where two alternative methods of measurement are compared with only one experimenter using each method.¹⁴ Any observed differences may be due either to differences between the methods or between the experimenters, or both. The two effects are *confounded*. A much more complex version of the same problem arises when trying to explain different mortality rates for the same disease in different countries.

Prediction

The use of observed relationships to make predictions about individuals is another area with many pitfalls. Just as it is dangerous to generalise from the particular, we must be very careful about particularising from the general.

For continuous variables, relationships are usually described by regression equations. It must be remembered that such fitted equations are approximate, both because they are calculated from a sample of data, and also because the imposition of an exact relationship (straight-line or curved) may be more convenient than realistic. The degree of scatter of the observations around the fitted line indicates the closeness of the relationship between the variables, and thus the uncertainty associated with predicting one from the other for specific cases. For example, a regression of height on weight for adult men would show a clear positive relationship with a large amount of scatter.

Regression equations should be used for prediction only within their limitations, so the regression line described above would be inappropriate for either boys or women. Such extrapolation is completely invalid. Also the prediction of height would be more certain for someone of average weight than for a very light or very heavy man, and this is borne out by the correct 95% confidence intervals for prediction which become wider further from the mean. It is very common to see a single figure quoted for the precision of any possible estimate; this is quite wrong.

Prediction also poses problems where the data are categorical. Table II shows the relationship between two diagnostic tests and the presence or absence of two diseases. Data such as these are usually described by the sensitivity and specificity, confusing terms for the proportions of correctly diagnosed positives and negatives. In both cases the sensitivity and specificity are high at 0.9 (maximum 1.0). These do not, however, measure the value of such tests for predictive purposes; in fact they become more misleading the lower the prevalence of the disease. The best approach is to consider what proportions of the diagnosed

positives and negatives were true positives and negatives respectively. In table IIa, where the prevalence is 50%, these figures are also both 0.9, indicating high predictive ability. In table IIb, the prevalence is 2%. Although virtually all of those with a negative test were truly negative (4410/4420), only 16% (90/580) of those diagnosed as positive were true positives. So the value of the test is low, even though 90% of the true positives give positive results. The usefulness of such a test depends on the cost of a false-positive finding. This is the problem when deciding whether or not screening for rare conditions (such as breast cancer) is worth while. For such purposes, the sensitivity is of no use at all—a high sensitivity is a necessary but not sufficient condition for a good predictive test.

Exactly the same considerations apply to the interpretation of a value exceeding a reference (or normal) range as automatically indicating abnormality without consideration of the prevalence of abnormality. Indeed, this is equivalent to looking at only the top row in table IIb. Such a procedure can lead to ludicrous interpretations of data—for example, that it is safer to drive very fast as few accidents are caused by cars travelling at more than 100 miles an hour.

TABLE II—Relation between diagnostic test and disease state with prevalence of disease (a) 50% and (b) 2%

(a)				(b)			
Disease	+	Test		Disease	+	Test	
		+	-			+	-
+	180	20	200	+	90	10	100
-	20	180	200	-	490	4410	4900
	200	200	400		580	4420	5000

Conclusions

The enormous amount of published research makes it inevitable that papers will often be judged, in the first instance

at least, by the authors' own conclusions or summary. It is thus vitally important that these contain valid interpretations of the results of the study, since the publication of misleading conclusions may both nullify the research in question and falsely influence medical practice and further research.

This is the seventh in a series of eight articles. No reprints will be available from the author.

References

- Pearson ES, Hartley HO. *Biometrika tables for statisticians*. Vol 1. 3rd ed. Cambridge: University Press, 1970: 83.
- Wood PDP. On the importance of correct orientation to sex in cattle. *Statistician* 1977;26:304-6.
- Box GEP, Hunter WG, Hunter JS. *Statistics for experimenters*. New York: Wiley, 1978:109.
- Dudley H. When is significant not significant? *Br Med J* 1977;ii:47.
- Newton J, Illingworth R, Elias J, McEwan J. Continuous intrauterine copper contraception for three years: comparison of replacement at two years with continuation of use. *Br Med J* 1977;ii:197-9.
- Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1980;61:1-7.
- Rose G. Beta-blockers in immediate treatment of myocardial infarction. *Br Med J* 1980;280:1088.
- Chalmers TC, Matta RJ, Smith H, Kunzler A-M. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977;297:1091-6.
- Carpenter RG, Emery JL. Final results of study of infants at risk of sudden death. *Nature* 1977;268:724-5.
- Bland JM. Assessment of risk of sudden death in infants. *Nature* 1978;273:74.
- Carpenter RG, Emery JL. Reply to J M Bland. *Nature* 1978;273:74-5.
- Meier P. Statistics and medical experimentation. *Biometrics* 1975;31:511-29.
- Jacobs DR, Anderson JT, Blackburn H. Diet and serum cholesterol: do zero correlations negate the relationship? *Am J Epidemiol* 1979;110:77-87.
- Serfontein GL, Jaroszewicz AM. Estimation of gestational age at birth. *Arch Dis Child* 1978;53:509-11.

Statistics and ethics in medical research

VIII—Improving the quality of statistics in medical journals

DOUGLAS G ALTMAN

Publication of a paper implies that the work is both sound and worth while. As I pointed out in my first article, it bestows both respectability and credibility on the work—a “seal of approval.” Once a paper has been published the results may influence both medical practice and further research by other scientists, and if the subject is of general interest the “mass media” may report the findings.

The ultimate responsibility for the general standard of published research rests with the medical journals. Perhaps unwillingly, the journals have the role of guardians of quality. This is particularly important with regard to statistical methods, which the majority of readers of medical papers are not able to judge for themselves and so must take on trust. The system of appraisal by independent referees is not ideal, but it is probably

the most practical method of quality control. Referees are usually selected, however, for their expertise in the relevant medical topic; their ability to assess the statistical aspects is left somewhat to chance. The result is that the statistical methods used in many research papers do not receive adequate scrutiny, with the consequences described in the previous articles.

The poor quality of statistics in published papers has been a cause of concern for many years, and is not confined to medical research. In 1964 Yates and Healy¹ wrote: “It is depressing to find how much good biological work is in danger of being wasted through incompetent and misleading analysis of numerical results.” Concern should be particularly great in the medical field because of the ethical implications, but the medical journals have generally been slow to appreciate that the statistical aspects can be fundamental to the validity of research.

Division of Computing and Statistics, Clinical Research Centre,
Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSc, medical statistician (member of scientific staff)

Statistics in medical papers

Probably as a reflection of widespread unease, there have been several reviews of the quality of statistics in published papers

over the past 15 years.²⁻⁶ These views are not strictly comparable because they looked at different statistical aspects in different journals at different times. Nevertheless, they all found many statistical errors or important errors of omission—in 72%, 49%, 52%, 45%, and 44% of papers studied, respectively. Further, a review of papers in five general medical journals found that 20% of the statistical procedures used were unidentified.⁷

It is impossible to assess the seriousness of many of the errors found. For example, an invalid analysis may give the same answer as an appropriate one, omission of information about randomisation does not necessarily mean that subjects were not allocated to treatments at random, and so on. It is, though, a measure of the disturbingly high prevalence of bad statistics that the reviewers of 62 papers in the *BMJ*¹ thought that it was "some comfort that only five papers drew a false conclusion."

Reviews of statistical procedures have sometimes been accompanied by editorials^{8,9} reinforcing the suggestions made in most of the papers that the standards of teaching should be improved and that there should be greater participation by statisticians in medical research. Such articles, however, stop short of the obvious suggestion that many of the papers should not have been published, at least as they stood, since any errors detected after publication could equally well have been detected at the refereeing stage.

Not all journals are equally culpable. The number of journals that use statisticians as referees, and sometimes also as members of editorial boards, has gradually increased, and several journals have publicly recognised the need to improve their statistical reviewing.¹⁰⁻¹² As Rennie¹⁰ says: "Our goal is the publication of data that are correctly observed and properly analysed." Such sentiments should be endorsed by all medical journals.

Raising statistical standards

Later I shall examine in some detail what the journals can do to improve standards. It is, however, important to realise that there are other aspects to the problem, which can broadly be summarised by the question: "Why is the standard of statistics so low in papers submitted for publication?"

TEACHING OF STATISTICS

The recent widespread move to include statistics in the syllabus for medical students and other science undergraduates is a welcome development. Such teaching is likely to be most beneficial when it gets away from a rigid method-orientated approach and concentrates more on general concepts. For medical students it may be more successful when not taught as an isolated subject, but closely related to another course such as epidemiology.¹³

Statistics is not an easy subject, however. A short introductory course is not sufficient to equip qualified doctors or scientists to carry out their own statistical analyses adequately, both because of the necessarily limited scope of such courses and also because several years may elapse before they need to use the knowledge. Thus although there is room for improvement in undergraduate teaching, it is unlikely to have much effect on the quality of statistics in medical research.

Of greater value in this respect would be postgraduate courses in statistics for those who had previously had an introductory course, and aimed particularly at those intending to do research. Such courses should try to give a greater understanding of statistical concepts: to help researchers to understand properly the simpler statistical methods (including when not to use them), to appreciate the principles of more advanced methods, and to know when to seek expert help. If such courses exist they are rare.

Similar comments apply to textbooks, where there is a wide

gap between the elementary¹⁴ and the comprehensive.¹⁵ Simple textbooks are usually much too strongly method-orientated to give a good grasp of the underlying principles behind much of statistics.

INVOLVEMENT OF STATISTICIANS

In general, the larger a project the more likely it is that a statistician will be directly concerned. Yet a survey¹⁶ of 211 cancer treatment studies in progress in 1978 showed that in only 47% was a statistician fully concerned (in design, data collection, and analysis). There was some involvement in a further 44%, but in 9% there was none. Unfortunately, not all medical researchers have direct access to a statistician, but large collaborative studies usually need considerable statistical advice,¹⁷ preferably with a statistician as an active participant. Even for small studies statistical advice before the research begins may be very valuable, especially in helping to match the design to the objectives of the study, and also to give the statistician a greater understanding of the research. Yet, despite common pleas for early involvement, most consultancy concerns the analysis of data that have already been collected. A bigger problem, though, is that many projects are carried out without the benefit of any statistical advice at all. Increased involvement of statisticians in medical research would clearly improve the overall standard of statistics, but this requires greater availability of medical statisticians than at present.

Successful consultancy relies on the ability of both researcher and statistician to understand each other's language, which is not always easy. Sprent¹⁸ has suggested that "Interdisciplinary communication is probably the most pressing problem in the pursuit of knowledge." The difficulties from the statistician's viewpoint have been discussed so often that a 1977 bibliography¹⁹ gave nearly 40 references. One aspect not often mentioned is that statisticians receive little or no preparation for consultancy work, either with respect to the sort of practical statistical problems that arise, or the role of consultant. This is a definite shortcoming in the education of statisticians, especially important because of their influence on the conduct of medical research.

ETHICAL COMMITTEES

Ethical committees have the opportunity to review many protocols for intended research on human subjects, and have the important sanction of withholding their approval. In view of the ease with which research can be rendered unethical by statistical mismanagement (as discussed in previous articles) it should be an automatic part of the review by ethical committees to look formally at the experimental design, and preferably also at the intended form of analysis. May²⁰ has written: "A poorly designed or poorly conceived experiment is unethical by definition and should not be permitted. Further it is the responsibility of the review committee to ensure that the conception and design meet the accepted canons of scientific method because we are dealing with experimentation which may not be for the individual subject's direct benefit." We can share his surprise that statisticians are not universally represented on ethical committees.

WHY PUBLISH?

One reason for the relentless production of low quality papers (not only with respect to the statistics) is the pressure on many individuals to publish as much as possible, with quantity being much more important than quality. At present it is known that other papers with poor statistics are being published, so a scientist may well think that there is no incentive (or need) to do better. But if journals were more careful about what they published we might advance to a state where fewer papers of a

higher standard were produced. This might also help to stem the counter-productive flow of new journals.

Role of the medical journals

There is general agreement among the medical journals in their attitude towards publishing the results of unethical research. Such research may have yielded valuable findings, but, as one editor wrote²¹: "publication in a reputable journal automatically implies that the editor and his reviewers condone the experimentation." In effect, papers describing unethical research are treated as "inadmissible evidence." For papers that may be deemed unethical because of their incorrect use of statistical methods, however, the attitudes of the journals vary enormously. Surely the same sort of argument as above should be extended, with publication similarly implying editorial approval of the data analysis and interpretation of results. It is illogical to refuse (quite rightly) to publish possibly useful findings of unethical research and yet be prepared to publish papers in which the results are invalidated by incorrect use of statistical methods.

One of the more obvious dangers of publishing questionable papers is that the conclusions may be quoted uncritically in the national press (since journalists are not usually qualified to criticise). Any ensuing critical letters will not receive similar publicity.

STATISTICAL REVIEW OF PAPERS

Since the reviews of published papers²⁻⁶ have found errors in about half of the papers examined, it is obvious that statistical review before publication ought to be highly effective. In 1964 the *Journal of the American Medical Association* raised the proportion of published papers considered statistically acceptable from one-third to three-quarters when it introduced a comprehensive statistical reviewing procedure.²²

Some of the following suggestions about ways in which journals can raise the quality of statistics in published papers have been made before,^{8 11 22} most notably in two recent papers.^{23 24} The most important recommendations are:

Statisticians should help referee

Journals should recruit statistically experienced people as referees, preferably with representation on editorial boards. Statistical review should be a formal procedure and not based on a casual inquiry to the nearest available statistician to "check that everything is all right." This is particularly important for specialist journals, where some depth of knowledge of the subject is often necessary.

All papers using any statistical procedure should be refereed by a statistician

Any paper in which inferences are drawn from the data presented should be seen by a statistician, whatever the level of statistical content. Indeed, the papers that cause the most trouble are usually those using only simple statistical methods "... where formal statistical review had seemed unwarranted,"¹⁰ rather than those with more complicated analyses. Short reports should not be exempt but should get higher priority. To reduce the work load the statistical assessment could be carried out only when a paper is likely to prove otherwise acceptable.

Revised papers should be returned to the same referee for reappraisal

A statistical refereeing system cannot work well without this condition. Failure to do this was the main reason why only 75% of published papers were completely acceptable even after the introduction of such a scheme.²²

Journals using a statistical refereeing system should state clearly what their policy is

This may help to discourage the submission of poor papers, and it would be valuable information for readers to know whether or not a journal uses such a system.

There should be statistical guidelines for contributors

All journals have instructions for contributors; very few mention statistics, and these rarely say much. It would obviously be undesirable for each journal to have different guidelines, but some agreement on this could be achieved in the same way as it has been on formats for references, perhaps in collaboration with the statistical societies. Some suggestions are given below.

All research papers should include a separate section on statistical methods

This should include information on relevant aspects of design, data collection, and analysis. Particularly important (if relevant) are the treatment allocation policy, response rate (and how non-responders were dealt with), and clear descriptions of analyses. Unusual methods of analysis should be given a specific reference (not a whole textbook!) with the reason for their use. This is a very important section of a paper, and should not be shortened at the expense of essential information.

Journals should give priority to well-executed and well-documented studies

Editorial boards should carefully consider the quality of study design, performance, analysis, and presentation of results when evaluating manuscripts. Standards should not be relaxed just because a paper is topical or interesting. Also, journals should not reject statistically valid papers purely because the findings were negative. (Obviously, this does not extend to those studies discussed in the third article, that are too small to detect important differences.) As Bradford Hill said 25 years ago: "A negative result may be dull but often it is no less important than the positive; and in view of that importance it must, surely, be established by adequate publication of the evidence."²⁵

Less important but still desirable additional features are:

Authors should be encouraged to supply additional information (especially on methodology) to help the referees but not for publication

One of the problems when assessing papers is lack of information necessary for proper statistical assessment; this is the main reason for the fifth recommendation above. The extra information could be a more detailed account of the design, a fuller description of the methods used and the results, and copies of other related papers.

Authors should be encouraged to include the raw data in their papers

Obviously this is only practicable for small studies, but could be eased by using "miniprint" tables.

Journals should employ editorial staff with some understanding of statistics

This is perhaps less important if a comprehensive statistical refereeing system is adopted but is still highly desirable especially in the event of disagreement between authors and referees.

For all journals to implement a comprehensive statistical refereeing system might well require many more medical statisticians than are currently available. It is much more likely, however, that there will be a continued steady increase in the use of statistical referees by journals, which should not cause major problems. Even the appointment by a journal of a single statistician can be enormously successful in raising the quality of statistics in published papers.

GUIDELINES FOR STATISTICAL REFEREES

Apart from checking on the validity of the statistical methods used, referees should ensure that there is adequate explanation and justification of what was done. It is also particularly important that the conclusions are reasonable, and that the summary is a fair reflection of the content.

The referee's report should be able to be understood by the authors, who may have only minimal statistical training.

GUIDELINES FOR CONTRIBUTORS

What sort of statistical guidelines should journals provide? Clearly these should not include advice on how to carry out research, although they might include discussion of the merits of different types of design. Such guidelines would not be a set of rules, but rather advice. The main emphasis should be on how best to describe clearly what procedures were used and what inferences were drawn.

Comprehensive guidelines would be of great benefit; these could perhaps be produced by a working party including representatives of medical journals and statistical societies. The following general suggestions relate to some of the more important aspects; they cannot be taken as comprehensive.

Design—This should be described clearly with, if relevant, information on treatment allocation, sample selection, if and how randomisation was used, whether or not the study was "blind" in any way, how sample size was determined (power), etc.

Data collection—Surveys should have response rates specified, and the representativeness of the sample and the possible effects of non-response should be discussed.

Analysis—The use of unusual forms of analysis should be justified, preferably with a reference, but all analyses should be very clearly described. It may be necessary to demonstrate the validity of the assumptions for some analyses (*t* tests, regression, etc).

Presentation of results—The results presented should be those most relevant to the question asked. Thus analysis of paired data should be accompanied by information—for instance, mean and standard deviation—about the within-person differences. Significance levels should not be given in place of quantitative results.

Interpretation of results—Special care should be taken to distinguish between statistical significance and clinical significance. Confidence intervals may greatly aid interpretation, especially where results are not statistically significant.

CONCLUSIONS

Reviews of published papers²⁻⁶ have all found unacceptably high proportions of papers with statistical errors. Some journals may feel that their policy of publishing letters criticising individual papers is an adequate safeguard. To take this attitude is to fail to appreciate the responsibility of the journals, both for ethical and scientific reasons, to avoid publishing sub-standard papers. In any case letters to journals usually produce a reply from the authors repeating their incorrect claims. Further, most papers are never read by anyone with the statistical knowledge to detect the flaws. If the credibility of published research is to

be raised it is essential that more journals introduce comprehensive statistical review procedures.

Summary

In these articles I have concentrated very much on one aspect of research. This is not meant to imply that statistics is of overriding importance, but rather that it is an area where much improvement is both highly desirable and possible.

By emphasising the ethical implications of carrying out research and publishing papers with incorrect statistics, I have argued that this is not just a matter for the individual researcher. There needs to be a wider appreciation of the importance of correct statistical thinking, and a great improvement in the standard of published research so that the sorts of errors discussed become very much the exception rather than commonplace. In the long term improved teaching and the greater involvement of statisticians will help; in the short term it is essential to have higher standards for published papers.

I am especially grateful to Martin Bland, Ted Coles, Stewart Mann, Charles Rossiter, and Patrick Royston for their perceptive criticism of earlier drafts of these articles. I must also thank Nicola Wilson Smith for the large amount of typing she has done.

This is the eighth in a series of eight articles. No reprints will be available from the author.

References

- Yates F, Healy MJR. How should we reform the teaching of statistics? *Journal of the Royal Statistical Society A* 1964;127:199-210.
- Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;195:1123-8.
- Lionel NDW, Herxheimer A. Assessing reports of therapeutic trials. *Br Med J* 1970;iii:637-40.
- Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J* 1977;ii:85-7.
- White SJ. Statistical errors in papers in the *British Journal of Psychiatry*. *Br J Psychiatry* 1979;135:336-42.
- Glantz SA. Biostatistics: how to detect, correct, and prevent errors in the medical literature. *Circulation* 1980;61:1-7.
- Feinstein AR. Clinical biostatistics. XXV A survey of the statistical procedures in general medical journals. *Clin Pharmacol Ther* 1974;15:97-107.
- Anonymous. A pillar of medicine. *JAMA* 1966;195:1145.
- Anonymous. Statistical errors. *Br Med J* 1977;ii:66.
- Rennie D. Vive la différence (p < 0.05). *N Engl J Med* 1978;299:828-9.
- Shuster JJ, Bimion J, Moxley J, et al. Statistical review process. Recommended procedures for biomedical research articles. *JAMA* 1976;235:534-5.
- Rosen MR, Hoffman BF. Statistics, biomedical scientists, and *Circulation Research*. *Circ Res* 1978;42:739.
- Clarke M, Clayton DG, Donaldson LJ. Teaching epidemiology and statistics to medical students—the Leicester experience. *Int J Epidemiol* 1980;9:179-85.
- Swinscow TDV. *Statistics at square one*. London: BMA, 1976.
- Armitage P. *Statistical methods in medical research*. Oxford: Blackwell, 1971.
- Tate HC, Rawlinson JB, Freedman LS. Randomised comparative studies in the treatment of cancer in the United Kingdom: room for improvement? *Lancet* 1979;ii:623-5.
- Breslow N. Perspectives on the statistician's role in cooperative clinical research. *Cancer* 1978;41:326-32.
- Sprent P. Some problems of statistical consultancy (with discussion). *Journal of the Royal Statistical Society A* 1970;133:139-64.
- Woodward WA, Schucany WR. Bibliography for statistical consulting. *Biometrics* 1977;33:564-5.
- May WW. The composition and function of ethical committees. *J Med Ethics* 1975;1:23-9.
- Woodford FP. Ethical experimentation and the editor. *N Engl J Med* 1972;286:892.
- Schor S. Statistical reviewing program for medical manuscripts. *American Statistician* 1967;21:28-31.
- O'Fallon JR, Dubey SD, Salsburg DS, Edmonson JH, Soffer A, Colton T. Should there be statistical guidelines for medical research papers? *Biometrics* 1978;34:687-95.
- Mosteller F, Gilbert JP, McPeck B. Reporting standards and research strategies for controlled trials. Agenda for the editor. *Controlled Clinical Trials* 1980;1:37-58.
- Hill AB. Contribution to the discussion of a paper by D J Finney. *Journal of the Royal Statistical Society A* 1956;119:19-20.