Index

• If Nothing goes wrng, is everthing alright?
Probability of adverse events that have not yet occurred: a statistical reminder

BMJ 1995;311:619-620 (2 September)

Ernst Eypasch, ,a Rolf Lefering, ,a C K Kum, ,a Hans Troidl, director a
a II Department of Surgery, University of Cologne, Kliniken der Stadt Koln, Ostmerheimer Str 200, D-51109 Koln, Germany

Correspondence to: Dr Eypasch.

The probability of adverse and undesirable events during and after operations that have not yet occurred in a finite number of patients (n) can be estimated with Hanley's simple formula, which gives the upper limit of the 95% confidence interval of the probability of such an event: upper limit of 95% confidence interval=maximum risk=3/n (for n>30). Doctors and surgeons should keep this simple rule in mind when complication rates of zero are reported in the literature and when they have not (yet) experienced a disastrous complication in a procedure.

Just as aeroplanes should not crash, common bile ducts should not be cut and iliac vessels not be punctured during laparoscopic procedures. In reality, however, these things do happen.1 With the boom in endoscopic surgery, surgeons are claiming to have zero mortality or even zero morbidity in their series of operations. A little reminder, not only for surgeons, may be necessary. If a certain adverse event or complication does not occur in a series, it does not mean that it will never happen. Experience and Murphy's law teach us that catastrophes do happen, and their probability can in fact be calculated by a simple rule of thumb.

In 1983 Hanley, a Canadian statistician, published the paper If nothing goes wrong is everything alright?2 This paper deserves explanation and needs to be highlighted to surgeons in particular. The paper describes in detail the statistical implications if an event of interest fails to occur in a finite number of operations or subjects. Instead of assuming that a technique is safe because of zero numerators, we should look at confidence intervals between zero and a certain upper limit. Hanley gives a simple rule, which should be known by every practising surgeon, to calculate the upper limit of a 95% confidence interval.

Methods

THE FORMULA

Hanley wrote: "This rule of three states that if none of n patients showed the event about which we are concerned, we can be 95% confident that the chance of this event is at most 3 in n (i.e. 3/n). In other words, the upper 95% confidence limit of a 0/n rate is approximately 3/n."2 The calculations are based on the following consideration. Given the risk of a certain event, the probability of this event not occurring is (1-risk). The probability of this event not occurring in n independent observations (patients or operations) is then (1-risk)n. The higher the risk, the lower the chance of not finding at least one occurrence of the event. One can therefore determine the maximum risk of an event, with a 5% error, that is compatible with n observations of non-occurrence: (1-maximum risk)n=0.05, equal to 1-maximum risk=(n root 0.05), equal to 1-maximum risk=(0.05)1/n. For n>30 this can be approximated by 1-maximum risk=1-(3/n), equal to maximum risk=3/n.

Upper limits of 95% confidence intervals for occurrence of immediate intraoperative death from vascular injury in series of laparoscopic appendicectomies and cholecystectomies

Study	No of procedures	No of deaths due to injury	Upper limit of 95% confidence interval (rule of three)
Laparoscopic appendicectomy			
Hebebrand et al5	25	0	12/100
Attwood et al3	27	0	11/100
McAnena et al8	29	0	10/100
Frazee et al6	38	0	8/100
Kum et al4	57	0	5/100
Tate et al7	70	0	4/100
Pier et al9	653	0	4/1000
Total	842	0	1/1000
Laparoscopic cholecystectomy			
Peters et al10	100	0	3/100
Troidl et al11	400	0	8/1000
Cuschieri et al13	1236	0	2/1000
Southern Surgeons Club15	1518	0	2/1000
Larson et al14	1983	0	1/1000
Collet et al12	2955	0	1/1000
Total	8192	0	3/10000

This formula closely fits the upper limit of the 95% confidence interval.2 Even when n=20 the number based on the rule of three does not differ substantially from the exact value (15% v 14%2).

EXAMPLE

The event that most worries endoscopic surgeons is intraoperative vascular injury that leads to loss of a limb or death. We selected well known international reports of series of laparoscopic appendicectomies and cholecystectomies from the literature.3 4 5 6 7 8 9 10 11 12 13 14 15 None of them reported a major vascular injury with subsequent loss of a limb or death. We applied Hanley's rule of three to the data in the papers to calculate the upper limit of a 95% confidence interval for such an adverse event. The table shows the results of these calculations.

Discussion

Several conclusions can be drawn from the table. It is obvious that a small series of any procedure can say hardly anything about the safety of the technique. Even though a major vascular injury with subsequent loss of a limb or death never occurred, the statistical analysis shows that, depending on the study selected, there was the threat that it might occur in four out of every 1000 procedures or even 12 out of every 100. This makes statements like "laparoscopic appendectomy is the method of choice"3 premature or even irresponsible if they are based on single studies.

The non-occurrence of an adverse event in a surgical series does not mean that it cannot happen. It can, and the true rate of occurrence can be estimated from its 95% confidence interval. It is a good estimate of the worst case that is compatible with the observed data. The smaller the sample, the wider the confidence interval. This means that the upper limit of a confidence interval from a small sample is greater than that from a large sample, but this does not mean that the true probability of an adverse event occurring is larger in a small series.

Doctors and surgeons should keep this simple rule of three in mind when complication rates of zero are reported in the literature and when they have not (yet) experienced a disastrous complication in a procedure.

Troidl H, Backer B, Langer B, Winkler-Wilfurth. Fehleranalyse-- Evaluierung und Verhutung von Komplikationen; ihre juristische Implikation. Langenbecks Archive fur Chirurgie Supplement Kongressbericht. Heidelberg: Springer Verlag, 1993:59-72.

Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything alright? JAMA 1983;259:1743-5.

Attwood SEA, Hill ADK, Murphy PG, Thornton J, Stephens RB. A prospective randomised trial of laparoscopic versus open appendectomy. Surgery 1992;112:497-501. [Medline]

Kum CK, Ngoi SS, Goh PMY, Tekant Y, Isaac JR. Randomized controlled trial comparing laparoscopic appendectomy to open appendectomy. Br J Surg 1993;80:1599-600. [Medline]

Hebebrand D, Troidl H, Spangenberger W, Neugebauer E, Schwalm T, Gunther MW. Laparoscopic or conventional appendectomy? A prospective randomised trial. Der Chirurg 1994;65:112-20.

Frazee RC, Roberts JW, Symmonds RE, Snyder SK, Hendricks JC, Smith RW, et al. A prospective randomised trial comparing open versus laparoscopic appendectomy. Ann Surg 1994;219:725-31. [Medline]

Tate JJT, Dawson J, Chung SCS, Lau WY, Li AKC. Laparoscopic versus open appendectomy: prospective randomised trial. Lancet 1993;342:633-7. [Medline]

McAnena OJ, Austin O, Hederman WP, Gorey TF, Fitzpatrick J, O'Connell PR. Laparoscopic versus open appendicectomy. Lancet 1991;338:693.

Pier A, Gotz F, Bacher C. Laparoscopic appendectomy in 625 cases: from innovation to routine. Surg Laparosc Endosc 1991;1:8-13. [Medline]

Peters JH, Ellison EC, Innes JT, Liss JL, Nichols KE, Lomano JM, et al. Safety and efficacy of laparoscopic cholecystectomy. Ann Surg 1991;213: 3-12.

Troidl H, Spangenberger W, Langen R, Al-Jaziri A, Eypasch E, Neugebauer E, et al. Laparoscopic cholecystectomy. Technical performance, safety, and patient benefits. Endoscopy 1992;24:252-61. [Medline]

Collet D, Edye M, Perissat J. Conversions and complications of laparoscopic cholecystectomy. Results of a survey conducted by the French Society of Endoscopic Surgery and Interventional Radiology. Surg Endosc 1993;7: 334-8.

Cuschieri A, Dubois F, Mouiel J, Mouret P, Becker H, Buess G, et al. The European experience with laparoscopic cholecystectomy. Am J Surg 1991; 161:385-7.

Larson GM, Vitale GC, Casey J, Evans JS, Gilliam G, Heuser L, et al. Multipractice analysis of laparoscopic cholecystectomy in 1983 patients. Am J Surg 1992;163:221-6. [Medline]

The Southern Surgeons Club. A prospective analysis of 1518 laparoscopic cholecystectomies. N Engl J Med 1991;324:1073-8. [Abstract]

(Accepted 15 June 1995)

Intention-To-Treat Analysis

Gerard E. Dallal, Ph.D.

It is now commonplace to see requests for proposals specify that study data be subjected to an intention-to-treat analysis (ITT) with "followup and case ascertainment continued regardless of whether participants continued in the trial". *Regardless* means regardless of compliance, changing regimens, reason for outcome [accidental death is death]... A popular phrase used to describe ITT analyses is **"Analyze as randomized!"** Once subjects are randomized, their data **must be** used for the ITT analysis! This sounds...well, the polite word is *counter-intuitive*. *Bizarre* is closer to the mark.

When Richard Peto first introduced the idea of ITT, the cause was taken up by many prominent statisticians, including Paul Meier, then of the University of Chicago and now Columbia University, whom I have heard speak eloquently in its defense. Others thought that Peto's suggestion was a sophisticated joke and awaited the followup article, which never came, to reveal the prank. I sympathize with this latter camp.

There are three major lines of justification for intention-to-treat analysis.

1. Intention-to-treat simplifies the task of dealing with suspicious outcomes, that is, it guards against conscious or unconscious attempts to influence the results of the study by excluding odd outcomes.
2. Intention-to-treat preserves the baseline comparability between treatment groups achieved by randomization.
3. Intention-to-treat reflects the way treatments will perform in the population by ignoring compliance when the data are analyzed.

Dealing with questionable outcomes and guarding against conscious or unconscious introductions of bias

One of Meier's examples involves a subject in a heart study where there is a question of whether his death should be counted against his treatment or set aside. He subject died from falling off his boat after having been observed carrying a few six-packs of beer on board for his solo sail. Meier argues that most researchers would set this event aside as probably unrelated to the treatment, while intention-to-treat would require the death be counted against the treatment. But suppose, Meier continues, that the beer is eventually recovered and every can is unopened. Intention-to-treat does the right thing in any case. By treating all treatments the same way, deaths unrelated to treatment should be equally likely to occur in all groups and the worst that can happen is that the treatment effects will be watered down by the occasional, randomly occurring outcome unrelated to treatment. If we pick and choose which events should count, we risk introducing bias into our estimates of treatment effects.

Preserving baseline comparability between treatment groups achieved by randomization.

When the drug clofibrate was studied, there was no treatment effect but subjects who were more compliant, whether on clofibrate or placebo, tended to have better outcomes. When there is no control group, the effect of compliance will be mistakenly attributed to treatment under the assumption that because the better outcomes were observed for those who followed the treatment more closely, the treatment must be effective. In many studies, potentially noncompliant subjects may be more likely to quit a particular treatment. For example, a noncompliant subject might be more likely to quit when assigned to strenuous exercise than to stretching exercises. In an on treatment analysis, the balance in compliance achieved at baseline will be lost and the resulting bias might make one of two equivalent treatments appear to be better than it truly is simply because one group of subject, on the whole, are more compliant.

As a more extreme case of Paul Meier's example, consider a study in which severely ill subjects are randomly assigned to surgery or drug therapy. There will be early deaths in both groups. It would be tempting to exclude the early deaths of those in the surgery group who died before getting the surgery on the grounds that *they never got the surgery*. However, this has the effect of making the drug therapy group much less healthy on average at baseline.

Reflecting performance in the population

Intention-to-treat analysis is said to be more realistic because it reflects what might be observed in actual clinical practice. In practice, patients may not comply, they may change treatments, they may accidentally die. ITT factors this into its analysis. It answers the public health question of what happens when a recommendation is made to the general public and the public decides how to implement it. The results of an intention-to-treat analysis can be quite different from the treatment effect observed when compliance is perfect.

My own views

What troubles me most about intention-to-treat analyses is that the phrase *intention-to-treat* is sometimes used as an incantation to avoid thinking about research issues. Its use often seems to be divorced from any

research question. It is easy to imagine circumstances where researchers might argue that the actual research question demands an intention-to-treat analysis to evaluate the results--for example, "For these reasons, we should be following everyone who enters the study regardless of compliance". What worries me is hearing ITT recommended for its own sake without any reference to the specific questions it might answer.

Intention-to-treat analysis answers a certain kind of research question. On treatment analysis answers a different kind of research question. My own view is to ignore labels, understand the research question, and perform the proper analysis *whatever it's called*. In some cases it may even be ITT! Usually, I perform both an intention-to-treat analysis and an on treatment analysis, using the results from the different analyses to answer different research questions.

If the purpose of a study is to answer "the public health question", then an ITT analysis should be performed. An ITT analysis should not be performed simply to perform an ITT analysis. An ITT analysis should be performed because the researchers are interested in answering the public health question **and they have determined that an ITT analysis will answer it**.

There are two components to how a treatment will behave in the population at large: efficacy and compliance. However, these are separate issues that should not be routinely combined in a single intention-to-treat analysis. A treatment's efficacy is often of great scientific importance (all exaggeration aside) regardless of compliance issues. Compliance during a trial might be quite different from compliance once a treatment has been proven efficacious. One can imagine, for example, what compliance might be like during a vitamin E trial and then what they would be like if Vitamin E were shown to prevent most forms of cancer! Should a treatment be found to be highly effective but unpalatable, future research might focus on ways to make it more palatable while other research, exploiting the active components of the treatment, might come up with new, more effective treatments. There may be cases, such as the treatment of mental disease, where an intention-to-treat analysis will truly reflect the way the treatments will behave in practice. In the fields in which I work, these situations tend to be exceptions rather than the rule.

Meier's example does not strike me as a compelling reason for ITT. The subject is on treatment. What is unclear is the way the outcome should be classified. This can be an issue even for ITT analyses. In the example, we don't know whether the subject suffered a heart attack. The beer might change the likelihood of various possibilities but the cause of death is still a guess whether the bottles were opened or unopened. In cases like this it makes sense to perform the analysis in many ways--for those outcomes where the cause of death is certain and then for all outcomes.

ITT does preserve the comparability at baseline achieved by randomization, but it is not the only way to do so. There might be a run-in period before subjects are randomized in order to identify noncompliant subjects and exclude them before they are assigned to treatment. As in the clofibrate study, compliance can be used as a covariate so that it is not confounded with treatment. In cases such as the surgery/drug therapy example, all deaths within a certain number of days of assignment might be excluded regardless of treatment.

David Salsburg once asked what to do about an intention-to-treat analysis if at the end of a trial it was learned that everyone assigned treatment A was given treatment B and vice-versa. I am living his joke. In a placebo-controlled vitamin E study, the packager delivered the pills just as the trial was scheduled to start. Treatments were given to the first few dozen subjects. As part of the protocol, random samples of the packaged pills were analyzed to insure the vitamin E did not lose potency during packaging. We discovered the pills were mislabeled--E as placebo and placebo as E. Since this was discovered a few weeks into the trial, no one had received refills, which might have been different from what was originally dispensed. We relabeled existing stores properly and I switched the assignment codes for those who had already been given pills to reflect what they actually received. How shall I handle the intention-to-treat analysis?

This slip-up aside, an intention-to-treat analysis is appropriate here. The primary research question asks what will happen if vitamin E is recommended for all residents. However, because the study pill is administered along with a subject's drugs, it's hard to imagine how compliance might change, even if the results of the trial were overwhelmingly positive. This makes the ITT the analysis of choice in this instance.

On the other hand, one could argue that because of the way treatments are administered, the ITT and on treatment analyses will be identical **except** for those who cannot tolerate the pill or whose physician decides, after enrollment, that they should not be in a study in which they might receive a vitamin E supplement. If a recommendation for supplements were made, such subjects would not be able to follow it, so perhaps it is inappropriate to include their data in the analyses.

In summary, I will recant a bit of my opening paragraph. ITT is not bizarre. In some circumstances, it may be the right thing to do. A slavish devotion to ITT is worse than bizarre. It could be harmful. The proper approach is to ignore labels, understand the research question, and perform the proper analysis *whatever it's called*!

[back to The Little Handbook of Statistical Practice]

• The odds ratio

J Martin Bland, *professor of medical statistics* a, **Douglas G Altman**, *professor of statistics in medicine* b.

a Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE, b ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Correspondence to: Professor Bland

In recent years odds ratios have become widely used in medical reportsalmost certainly some will appear in today's *BMJ*. There are three reasons for this. Firstly, they provide an estimate (with confidence interval) for the relationship between two binary ("yes or no") variables. Secondly, they enable us to examine the effects of other variables on that relationship, using logistic regression. Thirdly, they have a special and very convenient interpretation in case-control studies (dealt with in a future note).

The odds are a way of representing probability, especially familiar for betting. For example, the odds that a single throw of a die will produce a six are 1 to 5, or 1/5. The odds is the ratio of the probability that the event of interest occurs to the probability that it does not. This is often estimated by the ratio of the number of times that the event of interest occurs to the number of times that it does not. The table shows data from a cross sectional study showing the prevalence of hay fever and eczema in 11 year old children.1 The probability that a child with eczema will also have hay fever is estimated by the proportion 141/561 (25.1%). The odds is estimated by 141/420. Similarly, for children without eczema the probability of having hay fever is estimated by 928/14 453 (6.4%) and the odds is 928/13 525. We can compare the groups in several ways: by the difference between the proportions, 141/561928/14 453=0.187 (or 18.7 percentage points); the ratio of the proportions, (141/561)/(928/14 453)=3.91 (also called the relative risk); or the odds ratio, (141/420)/(928/13 525)=4.89.

Association between hay fever and eczema in 11 year old children1

Now, suppose we look at the table the other way round, and ask what is the probability that a child with hay fever will also have eczema? The proportion is 141/1069 (13.2%) and the odds is 141/928. For a child without hay fever, the proportion with eczema is 420/13 945 (3.0%) and the odds is 420/13 525. Comparing the proportions this way, the difference is 141/1069420/13 945=0.102 (or 10.2 percentage points); the ratio (relative risk) is (141/1069)/(420/13 945)=4.38; and the odds ratio is (141/928)/(420/13 525)=4.89. The odds ratio is the same whichever way round we look at the table, but the difference and ratio of proportions are not. It is easy to see why this is. The two odds ratios are
which can both be rearranged to give
If we switch the order of the categories in the rows and the columns, we get the same odds ratio. If we switch the order for the rows only or for the columns only, we get the reciprocal of the odds ratio, 1/4.89=0.204. These properties make the odds ratio a useful indicator of the strength of the relationship.

The sample odds ratio is limited at the lower end, since it cannot be negative, but not at the upper end, and so has a skew distribution. The log odds ratio,2 however, can take any value and has an approximately Normal distribution. It also has the useful property that if we reverse the order of the categories for one of the variables, we simply reverse the sign of the log odds ratio: log(4.89)=1.59, log(0.204)=1.59.

We can calculate a standard error for the log odds ratio and hence a confidence interval. The standard error of the log odds ratio is estimated simply by the square root of the sum of the reciprocals of the four frequencies. For the example,

A 95% confidence interval for the log odds ratio is obtained as 1.96 standard errors on either side of the estimate. For the example, the log odds ratio is loge(4.89)=1.588 and the confidence interval is 1.588±1.96Å~0.103, which gives 1.386 to 1.790. We can antilog these limits to give a 95% confidence interval for the odds ratio itself,2 as exp(1.386)=4.00 to exp(1.790)=5.99. The observed odds ratio, 4.89, is not in the centre of the confidence interval because of the asymmetrical nature of the odds ratio scale. For this reason, in graphs odds ratios are often plotted using a logarithmic scale. The odds ratio is 1 when there is no relationship. We can test the null hypothesis that the odds ratio is 1 by the usual 2 test for a two by two table.

Despite their usefulness, odds ratios can cause difficulties in interpretation.3 We shall review this debate and also discuss odds ratios in logistic regression and case-control studies in future Statistics Notes.

• Survival probabilities (the Kaplan-Meier method)

J Martin Bland, *professor of medical statistics*, a **Douglas G Altman**, *head*. b

a Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE, b ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Correspondence to: Professor Bland

As we have observed,1 analysis of survival data requires special techniques because some observations are censored as the event of interest has not occurred for all patients.For example, when patients are recruited over two years one recruited at the end of the study may be alive at one year follow up, whereas one recruited at the start may have died after two years. The patient who died has a longer observed survival than the one who still survives and whose ultimate survival time is unknown.

The table shows data from a study of conception in subfertile women.2 The event is conception, and women "survived" until they conceived. One woman conceived after 16 months (menstrual cycles), whereas several were followed for shorter time periods during which they did not conceive; their time to conceive was thus censored.

Time (months) to conception or censoring in 38 sub-fertile women after laparoscopy and hydrotubation2
We wish to estimate the proportion surviving (not having conceived) by any given time, which is also the estimated probability of survival to that time for a member of the population from which the sample is drawn. Because of the censoring we use the Kaplan-Meier method. For each time interval we estimate the probability that those who have survived to the beginning will survive to the end. This is a conditional probability (the probability of being a survivor at the end of the interval on condition that the subject was a survivor at the beginning of the interval). Survival to any time point is calculated as the product of the conditional probabilities of surviving each time interval. These data are unusual in representing months (menstrual cycles); usually the conditional probabilities relate to days. The calculations are simplified by ignoring times at which there were no recorded survival times (whether events or censored times).

In the example, the probability of surviving for two months is the probability of surviving the first month times the probability of surviving the second month given that the first month was survived. Of 38 women, 32 survived the first month, or 0.842. Of the 32 women at the start of the second month ("at risk" of conception), 27 had not conceived by the end of the month. The conditional probability of surviving the second month is thus 27/32=0.844, and the overall probability of surviving (not conceiving) after two months is 0.842Å~0.844=0.711. We continue in this way to the end of the table, or until we reach the last event. Observations censored at a given time affect the number still at risk at the start of the next month. The estimated probability changes only in months when there is a conception. In practice, a computer is used to do these calculations. Standard errors and confidence intervals for the estimated survival probabilities can be found by Greenwood's method.3 Survival probabilities are usually presented as a survival curve (figure). The "curve" is a step function, with sudden changes in the estimated probability corresponding to times at which an event was observed. The times of the censored data are indicated by short vertical lines.

There are three assumptions in the above. Firstly, we assume that at any time patients who are censored have the same survival prospects as those who continue to be followed. This assumption is not easily testable. Censoring may be for various reasons. In the conception study some women had received hormone treatment to promote ovulation, and others had stopped trying to conceive. Thus they were no longer part of the population we wanted to study, and their survival times were censored. In most studies some subjects drop out for reasons unrelated to the condition under study (for example, emigration) If, however, for some patients in this study censoring was related to failure to conceive this would have biased the estimated survival probabilities downwards.

Secondly, we assume that the survival probabilities are the same for subjects recruited early and late in the study. In a long term observational study of patients with cancer, for example, the case mix may change over the period of recruitment, or there may be an innovation in ancillary treatment. This assumption may be tested, provided we have enough data to estimate survival curves for different subsets of the data.

Thirdly, we assume that the event happens at the time specified. This is not a problem for the conception data, but could be, for example, if the event were recurrence of a tumour which would be detected at a regular examination. All we would know is that the event happened between two examinations. This imprecision would bias the survival probabilities upwards. When the observations are at regular intervals this can be allowed for quite easily, using the actuarial method.3

Formal methods are needed for testing hypotheses about survival in two or more groups. We shall describe the logrank test for comparing curves and the more complex Cox regression model in future Notes.

• Confidence intervals for the number needed to treat

Douglas G Altman, *professor of statistics in medicine*.

Imperial Cancer Research Fund Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

The number needed to treat (NNT) is a useful way of reporting the results of randomised controlled trials.1 In a trial comparing a new treatment with a standard one, the number needed to treat is the estimated number

of patients who need to be treated with the new treatment rather than the standard treatment for one additional patient to benefit. It can be obtained for any trial that has reported a binary outcome.

Summary points

The number needed to treat is a useful way of reporting results of randomised clinical trials
---
When the difference between the two treatments is not statistically significant, the confidence interval for the number needed to treat is difficult to describe
---
Sensible confidence intervals can always be constructed for the number needed to treat
---
Confidence intervals should be quoted whenever a number needed to treat value is given

Trials with binary end points yield a proportion of patients in each group with the outcome of interest. When the outcome event is an adverse one, the difference between the proportions with the outcome in the new treatment (pN) and standard treatment (pS) groups is called the absolute risk reduction (ARR=pNpS). The number needed to treat is simply the reciprocal of the absolute risk difference, or 1/ARR (or 100/ARR if percentages are used rather than proportions). A large treatment effect, in the absolute scale, leads to a small number needed to treat. A treatment that will lead to one saved life for every 10 patients treated is clearly better than a competing treatment that saves one life for every 50 treated. Note that when there is no treatment effect the absolute risk reduction is zero and the number needed to treat is infinite. As we will see below, this causes problems.

As with other estimates, it is important that the uncertainty in the estimated number needed to treat is accompanied by a confidence interval. A confidence interval for the number needed to treat is obtained simply by taking reciprocals of the values defining the confidence interval for the absolute risk reduction. 1 2 When the treatment effect is significant at the 5% level, the 95% confidence interval for the absolute risk reduction will not include zero, and thus the 95% confidence interval for the number needed to treat will not include infinity (). To take an example, if the ARR is 10% with a 95% confidence interval of 5% to 15%, the NNT is 10 (that is, 100/10) and the 95% confidence interval for the NNT is 6.7 to 20 (that is, 100/15 to 100/5). The case of a treatment effect that is not significant is more difficult. The same finding of ARR=10% with a wider 95% confidence interval for the ARR of 5% to 25% gives a NNT=10 (20 to 4). There are two difficulties with this confidence interval. Firstly, the number needed to treat can only be positive, and, secondly, the confidence interval does not seem to include the best estimate of 10. To avoid such perplexing results, the number needed to treat is often given without a confidence interval when the treatments are not significantly different.

A negative number needed to treat indicates that the treatment has a harmful effect. An NNT=20 indicates that if 20 patients are treated with the new treatment, one fewer would have a good outcome than if they all received the standard treatment. A negative number needed to treat has been called the number needed to harm (NNH). 3 4

As already noted, the number needed to treat is infinity () when the absolute risk reduction is zero, so the confidence interval calculated as 20 to 4 must include . The confidence interval is therefore peculiar, apparently encompassing two disjoint regionsvalues of the NNT from 4 to and values of the NNT from 20 to (or NNH from 20 to), as shown in figure 1. This situation led McQuay and Moore to observe that in the case of a non-significant difference it is not possible to get a useful confidence interval, and so only a point estimate is available.3

It is not satisfactory for the confidence interval to be presented only when the result is significant. Indeed this goes against advice that the confidence interval is especially useful when the result of a trial is not significant.5 In this article I show how a sensible confidence interval can be quoted for any trial. I also consider the use of the number needed to treat in meta-analysis. I approach the problem initially from a graphical perspective.

Rethinking the NNT scale

The number needed to treat is calculated by taking the reciprocal of the absolute risk reduction. When we obtain the confidence interval for the number needed to treat, we take reciprocals of the values defining the confidence interval for the absolute risk reduction and we reverse their order. As noted, a difficulty arises when the confidence interval for the absolute risk reduction encompasses both positive and negative values, and hence spans zero.

In the example, the 95% confidence interval for the number needed to treat was 20 to 4, or NNH=20 to NNT=4. Before reconsidering the meaning of the confidence interval, I wish to suggest that NNT and NNH are not good abbreviations. It seems more appropriate that the number of patients needed to be treated for one additional patient to benefit or be harmed are denoted NNTB and NNTH respectively, or perhaps

NNT(benefit) and NNT(harm). Using these descriptors, the confidence interval can be rewritten as NNTH 20 to NNTB 4. As already noted, this interval does not seem to include the overall estimate of NNTB 10, although figure 1 shows that it does.

When transforming data that are all positive, the effect of taking reciprocals is to reverse the order of the observations. The reciprocal transformation can be applied to negative values too, and the order of these is also reversed, but they remain negative. The overall effect of the transformation is thus quite strange when applied to data with both positive and negative values, as figure 1 illustrates. The confidence interval is peculiar, apparently encompassing two disjoint regionsvalues of the NNTB from 4 to and values of the NNTH from 20 to . I say "apparently" because the confidence interval is rather more logical than these values suggest.

The 95% confidence interval for the absolute risk reduction includes all values from 5% to 25%, including zero. As already noted, the number needed to treat is infinity () when the absolute risk reduction is zero, so the confidence interval calculated as NNTH 20 to NNTB 4 must include infinity. Figure 2 shows the absolute risk reduction and 95% confidence interval for the same example. The left hand axis shows the absolute risk reduction and the right hand scale shows the number needed to treat. Note that the number needed to treat scale now goes from NNTH=1 to NNTB=1 via infinity. It is clear that, rather unusually, infinity is in the middle of the scale, not at the ends. We should consider NNTB=1 as an extreme and unattainable valueit corresponds to the situation in which, say, all patients die if not given the new treatment and all survive with it. The other extreme, NNTH=1, corresponds to the case in which everyone lives unless given the treatment, in which case they all die. The values NNTB=1 and NNTH=1 correspond to ARR=100% and ARR=100% respectively, and are not shown. Conversely, the midpoint on the number needed to treat scale is the case where the treatment makes no difference (ARR=0 and NNT=). We need to remember the absolute risk reduction scale when trying to interpret the number needed to treat and its confidence interval.

"When there is no treatment effect the absolute risk reduction is zero and the number needed to treat is infinite ... this causes problems"

There is an argument that one does not wish to know the number needed to treat unless there is clear evidence of effectiveness, which for convenience alone is often taken as having achieved P<0.05. This advice seems to be based, at least partly, on trying to avoid the difficulty of an infinite number needed to treat rather than statistical soundness. In fact, we might often wish to quote a confidence interval for the number needed to treat when the confidence interval for the absolute risk reduction includes zero. Though this can be done by quoting two separate intervals, such as NNTB 10 (NNTH 20 to and NNTB 4 to), I suggest that it is done as, for example, NNTB 10 (NNTH 20 to to NNTB 4), which emphasises the continuity.

Tramèr et al quoted a NNT of 12.5 (3.7 to) for a trial comparing the antiemetic efficacy of intravenous ondansetron and intravenous droperidol.6 This negative number needed to treat implies that ondansetron was less effective than droperidol and the quoted 95% confidence interval was incomplete. The ARR was0.08 (0.27 to 0.11). We can convert this finding to the number needed to treat scale as NNTH=12.5 (NNTH 3.7 to to NNTB 9.1). With this presentation we can see that an NNTB less than (better than) 9 is unlikely. Similarly incomplete confidence intervals have been presented by other researchers. 7 8

Number needed to treat in meta-analysis

In meta-analyses it is desirable to show graphically the results of all the trials with their confidence intervals. The usual type of plot is called a forest plot. When the effect size has been summarised as the relative risk or odds ratio the analysis is based on the logarithms of these values, and the plot is best shown using a log scale for the treatment effect. In this scale the confidence intervals for each trial are symmetrical around the estimate.

"We need to remember the absolute risk reduction scale when trying to interpret the number needed to treat and its confidence interval"

Much the same can be done with the number needed to treat. Once we realise that the number needed to treat should be plotted on the absolute risk reduction scale, it is simple to plot numbers needed to treat with confidence intervals for several trials, even when (as is usual) some of the trials did not show significant results. Figure 3 shows such a plot for eight randomised trials comparing coronary angioplasty with bypass surgery.9 The plot was produced using the absolute risk reduction scale, and then relabelled. Both scales could be shown in the figure. This analysis is based on use of the absolute risk reduction as the effect measure in the meta-analysis. Meta-analysis is often more suitably performed using the relative risk or odds ratio. The number needed to treat can be obtained from the pooled estimates from such analyses if one specifies the control group event rate.10

A similar approach can be used for comparing numbers needed to treat derived for different interventions (as in fig 4) or for showing treatment effects in subgroups within a large randomised trial. The number needed to treat (benefit) (NNTB) values are shown to the left and number needed to treat (harm) (NNTH) values on the right as it has become more usual to show beneficial effects on the left.

Comment

The valuable concept of the number need to treat was introduced about 10 years ago.12 Its use has increased in recent years, especially in systematic reviews and in journals of secondary publication such as *ACP Journal Club* and *Evidence-Based Medicine*. Confidence intervals are usually quoted for the results of clinical trials, and this is widely recommended. 5 13 An exception has been when the number needed to treat is quoted for trials where the treatment effect was not significant. Here confidence intervals have either been omitted or reported incompletely. In this paper I have shown how to produce sensible confidence intervals for the number needed to treat in all cases, both for numerical summary and graphical display. These should be quoted whenever a number needed to treat value is presented.

• Bayesians and frequentists

J Martin Bland, *professor of medical statistics*, a **Douglas G Altman**, *head*. b

a Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE, b ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Correspondence to: Professor Bland

There are two competing philosophies of statistical analysis: the Bayesian and the frequentist. The frequentists are much the larger group, and almost all the statistical analyses which appear in the *BMJ* are frequentist. The Bayesians are much fewer and until recently could only snipe at the frequentists from the high ground of university departments of mathematical statistics. Now the increasing power of computers is bringing Bayesian methods to the fore.

Bayesian methods are based on the idea that unknown quantities, such as population means and proportions, have probability distributions. The probability distribution for a population proportion expresses our prior knowledge or belief about it, before we add the knowledge which comes from our data. For example, suppose we want to estimate the prevalence of diabetes in a health district. We could use the knowledge that the percentage of diabetics in the United Kingdom as a whole is about 2%, so we expect the prevalence in our health district to be fairly similar. It is unlikely to be 10%, for example. We might have information based on other datasets that such rates vary between 1% and 3%, or we might guess that the prevalence is somewhere between these values. We can construct a prior distribution which summarises our beliefs about the prevalence in the absence of specific data. We can do this with a distribution having mean 2 and standard deviation 0.5, so that two standard deviations on either side of the mean are 1% and 3%. (The precise mathematical form of the prior distribution depends on the particular problem.)

Suppose we now collect some data by a sample survey of the district population. We can use the data to modify the prior probability distribution to tell us what we now think the distribution of the population percentage is; this is the posterior distribution. For example, if we did a survey of 1000 subjects and found 15 (1.5%) to be diabetic, the posterior distribution would have mean 1.7% and standard deviation 0.3%. We can calculate a set of values, a 95% credible interval (1.2% to 2.4% for the example), such that there is a probability of 0.95 that the percentage of diabetics is within this set. The frequentist analysis, which ignores the prior information, would give an estimate 1.5% with standard error 0.4% and 95% confidence interval 0.8% to 2.5%. This is similar to the results of the Bayesian method, as is usually the case, but the Bayesian method gives an estimate nearer the prior mean and a narrower interval.

Frequentist methods regard the population value as a fixed, unvarying (but unknown) quantity, without a probability distribution. Frequentists then calculate confidence intervals for this quantity, or significance tests of hypotheses concerning it. Bayesians reasonably object that this does not allow us to use our wider knowledge of the problem. Also, it does not provide what researchers seem to want, which is to be able to say that there is a probability of 95% that the population value lies within the 95% confidence interval, or that the probability that the null hypothesis is true is less than 5%. It is argued that researchers want this, which is why they persistently misinterpret confidence intervals and significance tests in this way.

A major difficulty, of course, is deciding on the prior distribution. This is going to influence the conclusions of the study, yet it may be a subjective synthesis of the available information, so the same data analysed by different investigators could lead to different conclusions. Another difficulty is that Bayesian methods may lead to intractable computational problems. (All widely available statistical packages use frequentist methods.)

Most statisticians have become Bayesians or frequentists as a result of their choice of university. They did not know that Bayesians and frequentists existed until it was too late and the choice had been made. There have been subsequent conversions. Some who were taught the Bayesian way discovered that when they had huge quantities of medical data to analyse the frequentist approach was much quicker and more practical, although they may remain Bayesian at heart. Some frequentists have had Damascus road conversions to the Bayesian view. Many practising statisticians, however, are fairly ignorant of the methods used by the rival camp and too busy to have time to find out.

The advent of very powerful computers has given a new impetus to the Bayesians. Computer intensive methods of analysis are being developed, which allow new approaches to very difficult statistical problems, such as the location of geographical clusters of cases of a disease. This new practicability of the Bayesian approach is leading to a change in the statistical paradigmand a rapprochement between Bayesians and

frequentists. 1 2 Frequentists are becoming curious about the Bayesian approach and more willing to use Bayesian methods when they provide solutions to difficult problems. In the future we expect to see more Bayesian analyses reported in the *BMJ*. When this happens we may try to use Statistics notes to explain them, though we may have to recruit a Bayesian to do it.

• Generalisation and extrapolation

Douglas G Altman, *head*, a **J Martin Bland**, *professor of medical statistics*. b

a ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF, b Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

Correspondence to: Mr **Altman**.

All medical research is carried out on selected individuals, although the selection criteria are not always clear. The usefulness of research lies primarily in the generalisation of thefindings rather than in the information gained about those particular individuals. We study the patients in a trial not to find out anything about them but to predict what might happen to future patients given these treatments.

A recent randomised trial showed no benefit of fine needle aspiration over expectant management in women with simple ovarian cysts.1 The clinical question is whether the results canbe deemed to apply to a given patient. For most conditions it is widely accepted that a finding like this validly predicts the effect of treatment in other hospitals and in other countries. It would not, however, be safe to make predictions about patients with another condition, such as a breast lump. In between these extremes lie some cases where generalisability is less clear.

For example, when trials showed the benefits of blockers after myocardial infarction the studies had been carried out on middle aged men. Could the findings reasonably be extrapolated to women, or to older men? It is probably rare that treatment effectiveness truly varies by sex, and claims of this kind often arise from faulty subgroup analysis.2 Age too rarely seems to affect the benefit of a treatment, but clinical characteristics certainly do. Treatments that work in mild disease may not be equally effective in patients with severe disease, or vice versa. Likewise the mode of deliveryfor example, oral versus subcutaneousor dose may affect treatment benefit. Clinical variation is likely to affect the size of beneft of a treatment, not whether any benefit exists.

The extent to which it is wise or safe to generalise must be judged in individual circumstances, and there may not be a consensus. Arguably many studies (especially randomised controlled trials) use over-restrictive inclusion criteria, so that the degree of safe generalisability is reduced.3 Even geographical generalisation may sometimes be unwarranted. For example, BCG vaccination against tuberculosis is much less effective in India than in Europe, probably because of greater exposure in India.4 For the clinician treating a patient the question can be expressed as: "Is my patient so different from those in the trial that its results cannot help me make my treatment decision?"5

In a clinical trial we are interested in the difference in effectiveness between two treatments. There is no need to generalise the success rate of a particular treatment. In some other types of research, such as surveys to establish prevalence and prognostic or diagnostic studies, we may be trying to estimate a single population value rather than the difference between two of them. Here generalisation may be less safe. For example, the prevalence of many diseases varies across social and geographical groups. Results may not even hold up across time. For example, changes in case mix over time can affect the properties of a diagnostic test.6

Many studies use regression analysis to derive a model for predicting an outcome from one or more explanatory variables. The model, represented by an equation, is strictly valid only within the range of the observed data on the explanatory variable(s). When a measurement is included in the regression model it is possible to make predictions for patients outside the range of the original data (perhaps inadvertently). This numerical form of generalisation is called extrapolation. It can be seriously misleading.

To take an extreme example, a linear relation was found between ear size and age in men aged 30 to 93, with ear length (in mm) estimated as 55.9+0.22Å~age in years.7 The value of 55.9 corresponds to an age of zero. A baby with ears 5.6 cm long would look like Dumbo.

Extrapolating may be especially dangerous when a curved relation is found. Figure 1 shows fetal biparietal diameter (on a log scale) in relation to gestational age. Also shown are quadratic and cubic models fitted to the log biparietal diameter measurements from only those fetuses less than 30 weeks' gestation. Both curves fit the data well up to 30 weeks, but both give highly misleading predictions thereafter. The quadratic model shows a spurious maximum at around 34 weeks, while the cubic curve takes us again into elephantine regions.

When we have two explanatory variables it will not usually be apparent (unless we examine a scatter diagram) when a patient has a combination of characteristics which do not fall within the span of the original

data set. With more than two variables, such as in many prognostic models, it is not possible to be sure that the original data included any patients with the combination of values of a new patient. Nevertheless, it is reasonable to use such models to make predictions for patients whose important characteristics are within the range in the original data.

Clearly patient characteristics, including the criteria for sample selection, need to be fully reported in medical papers. Yet such basic information is not always provided.

·Regression towards the mean

J M Bland, **D G Altman**

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX.

We have previously shown that regression towards the mean occurs whenever we select an extreme group based on one variable and then measure another variable for that group (4 June, p 1499).1 The second group mean will be closer to the mean for all subjects than is the first, and the weaker the correlation between the two variables the bigger the effect will be. Regression towards the mean happens in many types of study. The study of heredity1 is just one. Once one becomes aware of the regression effect it seems to be everywhere. The following are just a few examples.

Treatment to reduce high levels of a measurement - In clinical practice there are many measurements, such as weight, serum cholesterol concentration, or blood pressure, for which particularly high or low values are signs of underlying disease or risk factors for disease. People with extreme values of the measurement, such as high blood pressure, may be treated to bring their values closer to the mean. If they are measured again we will observe that the mean of the extreme group is now closer to the mean of the whole population - that is, it is reduced. This should not be interpreted as showing the effect of the treatment. Even if subjects are not treated the mean blood pressure will go down, owing to regression towards the mean. The first and second measurement will have correlation r<l because of the inevitable measurement error and biological variation. The difference between the second mean for the subgroup and the population mean will be approximately r times the difference between the first mean and the population mean. We need to separate any genuine reductions due to treatment from the effect of regression towards the mean. This is best done by using a randomised control group, but it can be estimated directly.2

Relating change to initial value - We may be interested in the relation between the initial value of a measurement and the change in that quantity over time. In antihypertensive drug trials, for example, it may be postulated that the drug's effectiveness would be different (usually greater) for patients with more severe hypertension. This is a reasonable question, but, unfortunately, the regression towards the mean will be greater for the patients with the highest initial blood pressures, so that we would expect to observe the postulated effect even in untreated patients.3

Assessing the appropriateness of clinical decisions - Clinical decisions are sometimes assessed by asking a review panel to read case notes and decide whether they agree with the decision made. Because agreement between observers is seldom perfect the panel is sure to conclude that some decisions are "wrong." For example, Barrett et al reviewed cases of women who had had a caesarean section because of fetal distress.4 The percentage agreement between pairs of observers in the panel varied from 60% to 82.5%. They judged a caesarean section to be "appropriate" if at least four of the five observers thought a caesarean should have been done. Because there was poor agreement among the panel, judgments by panel members and the actual obstetricians doing the sections must also be poorly related and not all caesareans will be deemed appropriate by the panel. The authors concluded that 30% of all caesarean sections for fetal distress were unnecessary, but what the study actually showed was that decisions about whether women should have emergency surgery for fetal distress are difficult and that obstetricians do not always agree.5

Comparison of two methods of measurement - When comparing two methods of measuring the same quantity researchers are sometimes tempted to regress one method on the other. The fallacious argument is that if the methods agree the slope should be 1. Because of the effect of regression towards the mean we expect the slope to be less than 1 even if the two methods agree closely. For example, in two similar studies self reported weight was obtained from a group of subjects, and the subjects were then weighed.6,7 Regression analysis was done, with reported weight as the outcome variable and measured weight as the predictor variable. The regression slope was less than 1 in each study. According to the regression equation, the mean reported weight of heavy subjects was less than their mean measured weight, and the mean reported weight of light subjects was greater than their mean measured weight. We have a finding which allows a simple and attractive, but misleading, interpretation: those who are overweight tend to underestimate their weights and those who are excessively thin tend to overestimate their weights. In fact we would expect to find a slope less than 1, as a result of regression towards the mean. If self reported and measured weight were equaly good measures of the subject's true weight then the slope of the regression of reported weight on measured weight will be less than 1. But the slope of the regression of measured weight on reported weight will also be less than 1. Now we have the oppostive conclusion: people who are heavy

have overestimated their weights and people who are light have underestimated theirs. Elsewhere we describe a better approach to such data.8

Publication bias - Rousseeuw notes that referees for papers submitted for publication do not always agree which papers should be accepted.9 Because referees' judgments of the quality of papers are therefore made with error, they cannot be perfectly correlated with any measure of the true quality of the paper. Thus when an editor accepts the "best" papers for publication the average quality of these will be less than the editor thinks, and the average quality of those rejected will be higher than the editor thinks. Next time you are turned down by the BMJ do not be too despondent. It could be just another example of regression towards the mean.

• Quartiles, quintiles, centiles, and other quantiles
D G Altman, **J M Bland**

Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE Correspondence to: Mr **Altman**.

When presenting or analysing measurements of a continuous variable it is sometimes helpful to group subjects into several equal groups. For example, to create four equal groups we need the values that split the data such that 25% of the observations are in each group. The cut off points are called quartiles, and there are three of them (the middle one also being called the median). Likewise, we use two tertiles to split data into three groups, four quintiles to split them into five groups, and so on. The general term for such cut off points is quantiles; other values likely to be encountered are deciles, which split data into 10 parts, and centiles, which split the data into 100 parts (also called percentiles). Values such as quartiles can also be expressed as centiles; for example, the lowest quartile is also the 25th centile and the median is the 50th centile. We consider below some common applications of quantiles.

A common confusion is to use the terms tertiles, quartiles, quintiles, etc, not for the cut off points but for the groups so obtained, but these are properly called thirds, quarters, fifths, and so on.

Data description - The mean and standard deviation are useful to summarise a set of observations. When the data have a skewed distribution it is often preferable to quote instead the median and two outer centiles, such as the 10th and 90th. The first and third quartiles (25th and 75th centiles) are sometimes used; these define the interquartile range. The median is a useful summary statistic when some of the values are not actually measured - for example, because some values are outside the range of the measuring equipment. Similarly, the median is frequently used when summarising survival data, when it is usual for some of the survival times to be unknown.

Reference intervals and centiles - A special type of data description arises in the construction of a reference interval (normal range). A 95% reference interval is defined by the values that cut off 2/1/2% at each end of the distribution. (These values are often quite reasonably called the 2/1/2 and 97/1/2th centiles, although it is not strictly correct to have half centiles.) Reference intervals are widely used in clinical chemistry. By contrast, charts for the assessment of human size or growth usually show several centiles.1 Reference centiles are sometimes derived using the normal distribution,2 in which case any new observation can be placed at a specific centile.

Analysis of continuous variables - Continuous variables, such as serum cholesterol concentration and lung function, are often categorised in statistical analyses. It is usual to use quantiles, so that there are the same number of individuals in each group. Such grouping discards information but may allow for simpler presentation, such as in tables. The fewer groups created the greater is the loss of information. In regression analyses continuous explanatory variables are often categorised into two or more groups. Although this slightly complicates the analysis, it avoids a direct assumption that there is a linear relation between the variable and the outcome of interest. However, it leads to a model in which risk apparently jumps at certain values of the predictor variable rather than increasing smoothly.

Calculation of quantiles - The calculation of centiles and other quantiles is not as simple as it might seem. The data should be ranked from 1 to n in order of increasing size. The kth centile is obtained by calculating $q=k(n+1)/100$ and then interpolating between the two values with ranks either side of the qth. For example, for the 5th centile of a sample of 145 observations we have $q=5 \times 146/100=7.3$. We estimate the 5th centile as the value 0.3 of the way between the 7th and 8th ranked observations. If these data values are 11.4 and 14.9 the estimated centile is 12.45. Confidence intervals can be constructed for any quantile.3

• Diagnostic tests 1: sensitivity and specificity
D G Altman, **J M Bland**

Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE.

The simplest diagnostic test is one where the results of an investigation, such as an x ray examination or biopsy, are used to classify patients into two groups according to the presence or absence of a symptom or sign. For example, the table shows the relation between the results of a test, a liver scan, and the correct diagnosis based on either necropsy, biopsy, or surgical inspection.[1] How good is the liver scan at diagnosis of abnormal pathology?

Relation between results of liver scan and correct diagnosis[1]

	Pathology		
Liver scan	Abnormal (+)	Normal (-)	Total
Abnormal(+)	231	32	263
Normal(-)	27	54	81
Total	258	86	344

One approach is to calculate the proportions of patients with normal and abnormal liver scans who are correctly "diagnosed" by the scan. The terms positive and negative are used to refer to the presence or absence of the condition of interest, here abnormal pathology. Thus there are 258 true positives and 86 true negatives. The proportions of these two groups that were correctly diagnosed by the scan were 231/258=0.90 and 54/86=0.63 respectively. These two proportions have confusingly similar names.

Sensitivity is the proportion of true positives that are correctly identified by the test.

Specificity is the proportion of true negatives that are correctly identified by the test.

We can thus say that, based on the sample studied, we would expect 90% of patients with abnormal pathology to have abnormal (positive) liver scans, while 63% of those with normal pathology would have normal (negative) liver scans.

The sensitivity and specificity are proportions, so confidence intervals can be calculated for them using standard methods for proportions.[2]

Sensitivity and specificity are one approach to quantifying the diagnostic ability of the test. In clinical practice, however, the test result is all that is known, so we want to know how good the test is at predicting abnormality. In other words, what proportion of patients with abnormal test results are truly abnormal? This question is addressed in a subsequent note.

• **Diagnostic tests 2: predictive values**

Douglas G Altman, *head Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE*,a **J Martin Bland**, *reader in medical statistics* a

a Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX

The whole point of a diagnostic test is to use it to make a diagnosis, so we need to know the probability that the test will give the correct diagnosis. The sensitivity and specificity[1] do not give us this information. Instead we must approach the data from the direction of the test results, using predictive values.

Positive predictive value is the proportion of patients with positive test results who are correctly diagnosed.

Negative predictive value is the proportion of patients with negative test results who are correctly diagnosed.

Using the same data as in the previous note,[1] we know that 231 of 263 patients with abnormal liver scans had abnormal pathology, giving the proportion of correct diagnoses as 231/263 = 0.88. Similarly, among the 81 patients with normal liver scans the proportion of correct diagnoses was 54/81 = 0.59. These proportions are of only limited validity, however. The predictive values of a test in clinical practice depend critically on the prevalence of the abnormality in the patients being tested; this may well differ from the prevalence in a published study assessing the usefulness of the test.

This is the fourth in a series of occasional notes on medical statistics.

In the liver scan study the prevalence of abnormality was 0.75. If the same test was used in a different clinical setting where the prevalence of abnormality was 0.25 we would have a positive predictive value of 0.45 and a negative predictive value of 0.95. The rarer the abnormality the more sure we can be that a negative test indicates no abnormality, and the less sure that a positive result really indicates an abnormality. Predictive values observed in one study do not apply universally.

The positive and negative predictive values (PPV and NPV) can be calculated for any prevalence as follows:

$$PPV = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

$$NPV = \frac{\text{specificity} \times (1 - \text{prevalence})}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})}$$

If the prevalence of the disease is very low, the positive predictive value will not be close to 1 even if both the sensitivity and specificity are high. Thus in screening the general population it is inevitable that many people with positive test results will be false positives.

The prevalence can be interpreted as the probability before the test is carried out that the subject has the disease, known as the prior probability of disease. The positive and negative predictive values are the revised estimates of the same probability for those subjects who are positive and negative on the test, and are known as posterior probabilities. The difference between the prior and posterior probabilities is one way of assessing the usefulness of the test.

For any test result we can compare the probability of getting that result if the patient truly had the condition of interest with the corresponding probability if he or she were healthy. The ratio of these probabilities is called the likelihood ratio, calculated as sensitivity/ (1 - specificity).

The likelihood ratio indicates the value of the test for increasing certainty about a positive diagnosis. For the liver scan data the prevalence of abnormal pathology was 0.75, so the pre-test odds of disease were 0.75/(1 -0.75) = 3.0. The sensitivity was 0.895 and the specificity was 0.628. The post-test odds of disease given a positive test is 0.878/(1 -0.878) = 7.22, and the likelihood ratio is 0.895/(1 - 0.628) = 2.41. The posttest odds of having the disease is the pre-test odds multiplied by the likelihood ratio.

A high likelihood ratio may show that the test is useful, but it does not necessarily follow that a positive test is a good indicator of the presence of disease.

1 **Altman** DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. *BMJ* 1994;000:00000.

• Correlation, regression, and repeated data
J M Bland, **D G Altman**

Department of Public Health Sciences, St George's Hospital Medical School, London SW 17 0RE Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX Correspondence to: Dr Bland.

In clinical research we are often able to take several measurements on the same patient. The correct analysis of such data is more complex than if each patient were measured once. This is because the variability of measurements made on different subjects is usually much greater than the variability between measurements on the same subject, and we must take both kinds of variability into account. For example, we may want to investigate the relation between two variables and take several pairs of readings from each of a group of subjects. Such data violate the assumption of independence inherent in many analyses, such as t tests and regression.

Researchers sometimes put all the data together, as if they were one sample. Most statistics textbooks do not warn the researcher not to do this. It is so ingrained in statisticians that this is a bad idea that it never occurs to them that anyone would do it.

Consider the following example. The data were generated from random numbers, and there is no relation between X and Y at all. Firstly, values of X and Y were generated for each "subject," then a further random number was added to make the individual "observation." The data are shown in the table and figure. For each subject separately the correlation between X and Y is not significant. We have only five subjects and so only five points. Using each subject's mean values, we get the correlation coefficient r=-0.67, df=3, P=0.22. However, if we put all 25 observations together we get r=-0.47, df=23, P=0.02. Even though this correlation coefficient is smaller than that between means, because it is based on 25 pairs of observations rather than five it becomes significant. The calculation is performed as if we have 25 subjects, and so the number of degrees of freedom for the significance test is increased incorrectly and a spurious significant difference is produced. The extreme case would occur if we had only two subjects, with repeated pairs of observations on each. We would have two separate clusters of points centred at the subjects' means. We

would get a high correlation coefficient, which would appear significant despite there being no relation whatsoever.

Simulated data showing five pairs of measurements of two uncorrelated variables for subjects 1, 2, 3, 4, and 5

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5	
	48	58	63	28	38	40	51	46	55	62
	56	53	74	24	56	41	46	36	51	50
	49	44	69	26	46	40	36	41	54	66
	38	53	55	19	43	41	49	43	46	51
	50	56	73	22	52	34	46	45	55	52
Subject mean	48.2	52.8	66.8	23.8	47.0	39.2	45.6	42.2	52.2	56.2
Correlation coefficient	r=-0.02 P=0.97		r=0.32 P=0.59		r=-0.30 P=0.63		r=0.37 P=0.55		r=0.55 P=0.33	

There are two simple ways to approach these types of data. If we want to know whether subjects with a high value of X tend also to have a high value of Y we can use the subject means and find the correlation between them. For different numbers of observations for each subject, we can use a weighted analysis, weighting by the number of observations for the subject. If we want to know whether changes in one variable in the same subject are paralleled by changes in the other we can estimate the relation within subjects using multiple regression. In either case we should not mix observations from different subjects indiscriminately, whether using correlation or the closely related regression analysis.

• Variables and parameters

Douglas G Altman, *professor of statistics in medicine* a, **J Martin Bland**, *professor of medical statistics* b.

Like all specialist areas, statistics has developed its own language. As we have noted before,[1] much confusion may arise when a word in common use is also given a technical meaning. Statistics abounds in such terms, including normal, random, variance, significant, etc. Two commonly confused terms are variable and parameter; here we explain and contrast them.

Information recorded about a sample of individuals (often patients) comprises measurements such as blood pressure, age, or weight and attributes such as blood group, stage of disease, and diabetes. Values of these will vary among the subjects; in this context blood pressure, weight, blood group and so on are variables. Variables are quantities which vary from individual to individual.

By contrast, parameters do not relate to actual measurements or attributes but to quantities defining a theoretical model. The figure shows the distribution of measurements of serum albumin in 481 white men aged over 20 with mean 46.14 and standard deviation 3.08 g/l. For the empirical data the mean and SD are called sample estimates. They are properties of the collection of individuals. Also shown is the normal[1] distribution which fits the data most closely. It too has mean 46.14 and SD 3.08 g/l. For the theoretical distribution the mean and SD are called parameters. There is not one normal distribution but many, called a family of distributions. Each member of the family is defined by its mean and SD, the parameters[1] which specify the particular theoretical normal distribution with which we are dealing. In this case, they give the best estimate of the population distribution of serum albumin if we can assume that in the population serum albumin has a normal distribution.

Most statistical methods, such as *t* tests, are called parametric because they estimate parameters of some underlying theoretical distribution. Non-parametric methods, such as the Mann-Whitney U test and the log rank test for survival data, do not assume any particular family for the distribution of the data and so do not estimate any parameters for such a distribution.

Another use of the word parameter relates to its original mathematical meaning as the value(s) defining one of a family of curves. If we fit a regression model, such as that describing the relation between lung function and height, the slope and intercept of this line (more generally known as regression coefficients) are the parameters defining the model. They have no meaning for individuals, although they can be used to predict an individual's lung function from their height.

In some contexts parameters are values that can be altered to see what happens to the performance of some system. For example, the performance of a screening programme (such as positive predictive value or cost effectiveness) will depend on aspects such as the sensitivity and specificity of the screening test. If we

look to see how the performance would change if, say, sensitivity and specificity were improved, then we are treating these as parameters rather than using the values observed in a real set of data.

Parameter is a technical term which has only recently found its way into general use, unfortunately without keeping its correct meaning. It is common in medical journals to find variables incorrectly called parameters (but not in the *BMJ* we hope[2]). Another common misuse of parameter is as a limit or boundary, as in "within certain parameters." This misuse seems to have arisen from confusion between parameter and perimeter.

Misuse of medical terms is rightly deprecated. Like other language errors it leads to confusion and the loss of valuable distinction. Misuse of non-medical terms should be viewed likewise.

• Measurement error

J Martin Bland, *professor of medical statistics*,a **Douglas G Altman**, *head* b

a Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE, b IRCF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, PO Box 777, Oxford OX3 7LF

Correspondence to: Professor Bland.

Several measurements of the same quantity on the same subject will not in general be the same. This may be because of natural variation in the subject, variation in the measurement process, or both. For example, table 1 shows four measurements of lung function in each of 20 schoolchildren (taken from a larger study[1]). The first child shows typical variation, having peak expiratory flow rates of 190, 220, 200, and 200 l/min.

Table 1--Repeated peak expiratory flow rate (PEFR) measurements for 20 schoolchildren

Child No	PEFR (l/min) 1st	2nd	3rd	4th	Mean	SD
1	190	220	200	200	202.50	12.58
2	220	200	240	230	222.50	17.08
3	260	260	240	280	260.00	16.33
4	210	300	280	265	263.75	38.60
5	270	265	280	270	271.25	6.29
6	280	280	270	275	276.25	4.79
7	260	280	280	300	280.00	16.33
8	275	275	275	305	282.50	15.00
9	280	290	300	290	290.00	8.16
10	320	290	300	290	300.00	14.14
11	300	300	310	300	302.50	5.00
12	270	250	330	370	305.00	55.08
13	320	330	330	330	327.50	5.00
14	335	320	335	375	341.25	23.58
15	350	320	340	365	343.75	18.87
16	360	320	350	345	343.75	17.02
17	330	340	380	390	360.00	29.44
18	335	385	360	370	362.50	21.02
19	400	420	425	420	416.25	11.09
20	430	460	480	470	460.00	21.60

Let us suppose that the child has a "true" average value over all possible measurements, which is what we really want to know when we make a measurement. Repeated measurements on the same subject will vary around the true value because of measurement error. The standard deviation of repeated measurements on the same subject enables us to measure the size of the measurement error. We shall assume that this standard deviation is the same for all subjects, as otherwise there would be no point in estimating it. The main exception is when the measurement error depends on the size of the measurement, usually with measurements becoming more variable as the magnitude of the measurement increases. We deal with this case in a subsequent statistics note. The common standard deviation of repeated measurements is known as the within-subject standard deviation, which we shall denote by (zeta)w.

To estimate the within-subject standard deviation, we need several subjects with at least two measurements for each. In addition to the data, table 1 also shows the mean and standard deviation of the four readings for each child. To get the common within-subject standard deviation we actually average the variances, the squares of the standard deviations. The mean within-subject variance is 460.52, so the estimated within-subject standard deviation is (zeta)w=(square root)460.5 = 21.5 1/min. The calculation is easier using a program that performs one way analysis of variance[2] (table 2). The value called the residual mean square is

the within-subject variance. The analysis of variance method is the better approach in practice, as it deals automatically with the case of subjects having different numbers of observations. We should check the assumption that the standard deviation is unrelated to the magnitude of the measurement. This can be done graphically, by plotting the individual subject's standard deviations against their means (see fig 1). Any important relation should be fairly obvious, but we can check analytically by calculating a rank correlation coefficient. For the figure there does not appear to be a relation (Kendall's (lau) = 0.16, P = 0.3).

Table 2--One way analysis of variance for the data of table 1

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability (P)
Children	19	285318.44	15016.78	32.6	<0.0001
Residual	16	27631.25	460.52		
Total	79	312949.69			

View larger version (19K):
[in this window]
[in a new window]

Fig 1--Individual subjects' standard deviations plotted against their means

A common design is to take only two measurements per subject. In this case the method can be simplified because the variance of two observations is half the square of their difference. So, if the difference between the two observations for subject I is di the within-subject standard deviation (zeta)w is given by when n is the number of subjects. We can check for a relation between standard deviation and mean by plotting for each subject the absolute value of the difference--that is, ignoring any sign--against the mean.

The measurement error can be quoted as (zeta)w. The difference between a subject's measurement and the true value would be expected to be less than 1.96 (zeta)w for 95% of observations. Another useful way of presenting measurement error is sometimes called the repeatability, which is (square root)2 x 1.96 (zeta)w or 2.77 (zeta)w. The difference between two measurements for the same subject is expected to be less than 2.77 (zeta)w for 95% of pairs of observations. For the data in table 1 the repeatability is 2.77 x 2.5 = 60 l/min. The large variability in peak expiratory flow rate is well known, so individual readings of peak expiratory flow are seldom used. The variable used for analysis in the study from which table 1 was taken was the mean of the last three readings.1

Other ways of describing the repeatability of measurements will be considered in subsequent statistics notes.

• **Measurement error and correlation coefficients**
J Martin Bland, *professor of medical statistics*,a **Douglas G Altman**, *head* b
Measurement error is the variation between measurements of the same quantity on the same individual.1 To quantify measurement error we need repeated measurements on several subjects. We have discussed the within-subject standard deviation as an index of measurement error,1 which we like as it has a simple clinical interpretation. Here we consider the use of **correlation** coefficients to quantify measurement error.

A common design for the investigation of measurement error is to take pairs of measurements on a group of subjects, as in table 1. When we have pairs of observations it is natural to plot one measurement against the other. The resulting scatter diagram (see figure 1) may tempt us to calculate a **correlation** coefficient between the first and second measurement. There are difficulties in interpreting this **correlation** coefficient. In general, the **correlation** between repeated measurements will depend on the variability between subjects. Samples containing subjects who differ greatly will produce larger **correlation** coefficients than will samples containing similar subjects. For example, suppose we split this group in whom we have measured forced expiratory volume in one second (FEV1) into two subsamples, the first 10 subjects and the second 10 subjects. As table 1 is ordered by the first FEV1 measurement, both subsamples vary less than does the whole sample. The **correlation** for the first subsample is r = 0.63 and for the second it is r = 0.31, both less than r = 0.77 for the full sample. The **correlation** coefficient thus depends on the way the sample is chosen, and it has meaning only for the population from which the study subjects can be regarded as a random sample. If we select subjects to give a wide range of the measurement, the natural approach when investigating measurement error, this will inflate the **correlation** coefficient.

The **correlation** coefficient between repeated measurements is often called the reliability of the measurement method. It is widely used in the validation of psychological measures such as scales of anxiety and depression, where it is known as the test-retest reliability. In such studies it is quoted for different populations (university students, psychiatric outpatients, etc) because the **correlation** coefficient differs

between them as a result of differing ranges of the quantity being measured. The user has to select the **correlation** from the study population most like the user's own.

Another problem with the use of the **correlation** coefficient between the first and second measurements is that there is no reason to suppose that their order is important. If the order were important the measurements would not be repeated observations of the same thing. We could reverse the order of any of the pairs and get a slightly different value of the **correlation** coefficient between repeated measurements. For example, reversing the order of the even numbered subjects in table 1 gives r = 0.80 instead of r = 0.77. The intra-class **correlation** coefficient avoids this problem. It estimates the average **correlation** among all possible orderings of pairs. It also extends easily to the case of more than two observations per subject, where it estimates the average **correlation** between all possible pairs of observations.

Few computer programs will calculate the intra-class **correlation** coefficient directly, but when the number of observations is the same for each subject it can be found from a one way analysis of variance table2 such as table 2. We need the total sum of squares, SST, and the sum of squares between subjects, SSB.

Then

rI = mSSB - SST/(m - 1) SST

where m is the number of observations per subject. For table II, m = 2 and

rI = 2 x 1.52981 - 1.74651/(2 - 1) x 1.74651 = 0.75

Table 2--One way analysis of variance for the data in table 1

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability (P)
Children	19	1.52981	0.08052	7.4	<0.0001
Residual	20	0.21670	0.01086		
Total	39	1.74651			

In practice, there will usually be little difference between r and rI for true repeated measurements. If, however, there is a systematic change from the first measurement to the second, as might be caused by a learning effect, rI will be much less than r. If there was such an effect the measurements would not be made under the same conditions and so we could not measure reliability.

The **correlation** coefficient can be used to compare measurements of different quantities, such as different scales for measuring anxiety. We could make repeated measurements of all the quantities on the same subjects and calculate intra-class **correlation**s. The measures with the highest **correlation** between repeated measurements would discriminate best between individuals; in other words they would carry the most information. For most applications, however, we prefer the within-subjects standard deviation as an index of measurement error, as it has a more direct interpretation which can be applied to individual measurements.1

•The Legend of the *P* Value
Zeev N. Kain, MD, MBA
Anesth Analg 2005;101:1454-1456

Although there is a growing body of literature criticizing the use of mere statistical significance as a measure of clinical impact, much of this literature remains out of the purview of the discipline of anesthesiology. Currently, the magical boundary of $P < 0.05$ is a major factor in determining whether a manuscript will be accepted for publication or a research grant will be funded. Similarly, the Federal Drug Administration does not currently consider the magnitude of an advantage that a new drug shows over placebo. As long as the difference is statistically significant, a drug can be advertised in the United States as "effective" whether clinical trials proved it to be 10% or 200% more effective than placebo. We submit that if a treatment is to be useful to our patients, it is not enough for treatment effects to be statistically significant; they also need to be large enough to be clinically meaningful.

Unfortunately, physicians often misinterpret statistically significant results as showing clinical significance as well. One should realize, however, that with a large sample it is quite possible to have a statistically significant result between groups despite a minimal impact of treatment (i.e., small effect size). Also, study outcomes with lower *P* values are typically misinterpreted by physicians as having stronger effects than those with higher *P* values. That is, most clinicians agree that a result with a $P = 0.002$ has a much greater treatment effect than a result of $P = 0.045$. Although this is true if the sample size is the same in both studies, it is not true if the sample size is larger in the study with the smaller *P* value. This is of particular

concern when one realizes that most pharmaceutically funded studies have very large sample sizes and effect sizes are typically not reported in these types of studies. In the following editorial I highlight some of issues related to this complex problem. Please note that a detailed discussion of the underlying statistics involved in this topic is beyond the scope of this editorial.

When examining the report of a clinical trial investigating a new treatment, clinicians should be interested in answering the following three basic questions:

1. Could the findings of the clinical trial be solely a result of a chance occurrence? (i.e., statistical significance)
2. How large is the difference between the primary end-points of the study groups? (i.e., impact of treatment, effect size)
3. Is the difference of primary end-points between groups meaningful to a patient? (i.e., clinical significance)

It was Sir Ronald A. Fisher, an extraordinarily influential British statistician, who first suggested the use of a boundary to accept or reject a null hypothesis, and he arbitrarily set this boundary at $P = 0.05$; where "*P*" stands for probability related to chance (1,2). That is, the level of statistical significance as defined by Fisher in 1925 and as used today refers to the probability that the difference between two groups would have occurred solely by chance (i.e., probability of <5 in 100 is reported as $P < 0.05$). Fisher's emphasis on significance testing and the arbitrary boundary of $P < 0.05$ has been widely criticized over the past 80 yr. This criticism was based on the rationale that focusing on the *P* value does not take into account the size and clinical significance of the observed effect. That is, a small effect in a study with large sample size has the same *P* value as a large effect in a study with a small sample size. Also, *P* value is commonly misinterpreted when there are multiple comparisons, in which case a traditional level of statistical significance of $P < 0.05$ is no longer valid. Fisher himself indicated some 25 yr after his initial publication that "If *P* is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05..." (3). Indeed, this issue has been addressed in multiple recent review articles and editorials in the general medical and psychological literature (4–8).

In an attempt to address some of the limitations of the *P* value, the use of the confidence intervals (CI) has been advocated by some clinicians (9). One should realize, however, that these two definitions of statistical significance are essentially reciprocal (10). That is, getting a $P < 0.05$ is the same as having a 95% CI that does not overlap zero. CIs can also, however, be used to estimate the size of difference between groups in addition to merely indicating the existence or absence of statistical significance (11). This later approach, however, is not widely used in the medical and psychological literature, and today CIs are mostly used as surrogates for the hypothesis test rather than considering the full range of likely effect size.

The group of statistics called "effect sizes" designate indices that measure the magnitude of difference between groups, controlling for variation within the groups; effect sizes can be thought of as a standardized difference. In other words, although a *P* value denotes whether the difference between two groups in a particular study is likely to occur solely by chance, the effect size quantifies the amount of difference between the two groups. Quantification of effect size does not rely on sample size but instead relies on the strength of the intervention. There are a number of different types of effect sizes and a description of these various types and formulae is beyond the scope of this editorial. We refer the interested reader to review articles that describe the various types of effect sizes and their calculation methodology (12,13). Effect sizes of the d type are the most commonly used in the medical literature, as they are primarily used to compare two treatment groups. D type effect size is defined as the magnitude of difference between two means, divided by the sd [(Mean of control group – Mean of treatment group)/sd of the control group]. Thus, the d effect size is dependent on variation within the control group and the differences between the control and intervention groups. Values of the d type effect sizes range from – to +, where zero denotes no effect and values less than or more than zero are treated as absolute values when interpreting magnitude. Conventionally, d type effect sizes that are near 0.20 are interpreted as small, effect sizes near 0.50 are considered "medium," and effect sizes in the range of 0.80 are considered "large" (14). However, interpretation of the magnitude of an effect size depends on the type of data gathered and the discipline involved. Effect sizes of another type—the risk potency type—include likelihood ratios such as odds ratio, risk ratio, risk difference, and relative risk reduction. Clinicians are probably more familiar with these less abstract statistics and it may be helpful to realize that likelihood statistics are a type of effect size.

Clinicians should be cautioned to not interpret magnitude of change (effect size) as an indication of clinical significance. The clinical significance of a treatment should be based on external standards provided by patients and clinicians. That is, a small effect size may still be clinically significant and, likewise, a large effect size may not be clinically significant, depending on what is being studied. Indeed, there is a growing recognition that traditional methods used, such as statistical significance tests and effect sizes, should be supplemented with methods for determining clinically significant changes. Although there is little consensus about the criteria for these efficacy standards, the most prominent definitions of clinically significant change include: 1) treated patients make a statistically reliable improvement in the change scores; 2) treated patients are empirically indistinguishable from a normal population after treatment, or 3) changes of at least one sd. The most frequently used method for evaluating the reliability of change scores is the Jacobson-

Truax method in combination with clinical cutoff points (15). Using this method, change is considered reliable, or unlikely to be the product of measurement error, if the reliable change index (RCI) is more than 1.96. That is, when the individual has a change score more than 1.96, one can reasonably assume that the individual has improved.

Unfortunately, most of the methods above are difficult to adopt in the perioperative arena, as comparison with a normal population is not an option in most trials, and the RCI, which controls for statistical issues involving the assessment tool, is a somewhat complicated and controversial technique. Thus, clinical significance in the perioperative arena may be best assessed by posing a particular question such as "is a change of 8.5% reduction in intraoperative bleed clinically significant?" or "how many sd does this change represent?" Obviously, both of these questions have a subjective component in them and although it is traditionally agreed that at least a 1-sd change is generally needed for clinical significance, this boundary has no scientific underpinning. The validity of a clinical cutoff for these last two methods can be improved by establishing external validity (e.g., patient perspective) for the decision. For example, Flor et al. (16) have conducted a large meta-analysis that was aimed at evaluating the effectiveness of multidisciplinary rehabilitation for chronic pain. The investigators found that pain among the patients who received the intervention was indeed reduced by 25%. This reduction was certainly statistically significant and had an effect size of 0.7. Colvin et al. (17), however, reported earlier that patients would consider only a 50% improvement in their pain levels as a treatment "success." Thus, in this example, a reduction of 25% in pain scores may be statistically, but not clinically, significant. Clearly this is a developing area that warrants further discussion.

In conclusion, we suggest that reporting of perioperative medical research should continue beyond reporting results consisting primarily of descriptive and statistically significant or nonsignificant findings. The interpretation of findings should occur in the context of the magnitude of change that occurred and the clinical significance of the findings.

References

• **Sample size calculations in randomised trials: mandatory and mystical**
The Lancet 2005; 365:1348-1353

Kenneth F Schulz email address a Corresponding Author Information and David A Grimes

Summary

Investigators should properly calculate sample sizes before the start of their randomised trials and adequately describe the details in their published report. In these a-priori calculations, determining the effect size to detect—eg, event rates in treatment and control groups—reflects inherently subjective clinical judgments. Furthermore, these judgments greatly affect sample size calculations. We question the branding of trials as unethical on the basis of an imprecise sample size calculation process. So-called underpowered trials might be acceptable if investigators use methodological rigor to eliminate bias, properly report to avoid misinterpretation, and always publish results to avert publication bias. Some shift of emphasis from a fixation on sample size to a focus on methodological quality would yield more trials with less bias. Unbiased trials with imprecise results trump no results at all. Clinicians and patients deserve guidance now.

Back to top

Sample size calculations for randomised trials seem unassailable. Indeed, investigators should properly calculate sample sizes and adequately describe the key details in their published report. Research methodologists describe the approaches in books and articles. Protocol committees and ethics review boards require adherence. CONSORT reporting guidelines clearly specify the reporting of sample size calculations.1,2 Almost everyone agrees.

An important impetus to this unanimity burst on the medical world more than a quarter of a century ago. A group of researchers, led by Tom Chalmers, published a landmark article detailing the lack of statistical power in so-called negative randomised trials published in premier general medical journals.3 In Chalmers' long illustrious career, he published hundreds of articles. This article on sample size and power received many citations. Paradoxically, that troubled him.4 He regarded it as the most damaging paper that he had ever coauthored. Why? We will describe his concerns later, so stay tuned.

Components of sample size calculations

Calculating sample sizes for trials with dichotomous outcomes (eg, sick vs well) requires four components: type I error (?), power, event rate in the control group, and a treatment effect of interest (or analogously an event rate in the treatment group). These basic components persist through calculations with other types of outcomes, except other assumptions can be necessary. For example, with quantitative outcomes and a typical statistical test, investigators might assume a difference between means and a variance for the means.

In clinical research, hypothesis testing risks two fundamental errors (panel 1). First, researchers can conclude that two treatments differ when, in fact, they do not. This type I error (?) measures the probability of making this false-positive conclusion. Conventionally, ? is most frequently set at 0·05, meaning that

investigators desire a less than 5% chance of making a false-positive conclusion. Second, researchers can conclude that two treatments do not differ when, in fact, they do—ie, a false-negative conclusion. This type II error (?) measures the probability of this false-negative conclusion. Conventionally, investigators set ? at 0·20, meaning that they desire less than a 20% chance of making a false-negative conclusion.

Panel 1: Errors defined

Type I error (?)

The probability of detecting a statistically significant difference when the treatments are in reality equally effective—ie, the chance of a false-positive result.

Type II error (?)

The probability of not detecting a statistically significant difference when a difference of a given magnitude in reality exists—ie, the chance of a false-negative result.

Power (1–?)

The probability of detecting a statistically significant difference when a difference of a given magnitude really exists.

Power derives from ? error. Mathematically, it is the complement of ? (1–?) and represents the probability of avoiding a false-negative conclusion. For example, for ?=0·20, the power would be 0·80, or 80%. Stated alternatively, power represents the likelihood of detecting a difference (as significant, with p<?), assuming a difference of a given magnitude exists. For example, a trial with a power of 80% has an 80% chance of detecting a difference between two treatments if a real difference of assumed magnitude exists in the population.

Admittedly, understanding ? error, ? error, and power can be a challenge. Convention, however, usually guides investigators for inputs into sample size calculations. The other inputs cause lesser conceptual difficulties, but produce pragmatic problems. Investigators estimate the true event rates in the treatment and control groups as inputs. Usually, we recommend estimating the event rate in the population and then determining a treatment effect of interest. For example, investigators estimate an event rate of 10% in the controls. They then would estimate an absolute change (eg, an absolute reduction of 3%), a relative change (a relative reduction of 30%), or simply estimate a 7% event rate in the treatment group. Using these assumptions, investigators calculate sample sizes. Standard texts describe the procedures encompassing, for example, binary, continuous, and time-to-event measures.5–7 Commonly, investigators use sample size and power software (preferably with guidance from a statistician). Most hand calculations diabolically strain human limits, even for the easiest formula, such as we offer in panel 2

Panel 2: The simplest, approximate sample size formula for binary outcomes, assuming ?=0·05, power=0·90, and equal sample sizes in the two groups

n=the sample size in each of the groups

p1=event rate in the treatment group (not in formula but implied when R and p2 are estimated)

p2=event rate in the control group

R=risk ratio (p1/p2)
For example, we estimate a 10% event rate in the control group (p2=0·10) and determine that the clinically important difference to detect is a 40% reduction (R=0·60) with the new treatment at ?=0·05 and power=0·90. (Note: R=0·60 equates to an event rate in the treatment group of p1=0·06, ie, R=6%/10%)
n=961·665?962 in each group (PASS software version 6.0 [NCSS, Kaysville, UT, USA] with a more accurate formula yields 965)

This formula accommodates alternate ? levels and power by replacing 10·51 with the appropriate value from the table below.

	Power (1–β)		
	0·80	0·90	0·95
α (type I error)			
0·05	7·85	10·51	13·00
0·01	11·68	14·88	17·82

	Power (1–β)			
	0·50	0·80	0·90	0·99
α (type I error)				
0·05	100	200	270	480
0·01	170	300	390	630
0·001	280	440	540	820

Effect of selecting ? error and power

The conventions of ?=0·05 and power=0·80 usually suffice. However, other assumptions make sense based on the topic studied. For example, if a standard prophylactic antibiotic for hysterectomy is effective with few side-effects, in a trial of a new antibiotic we might set ? error lower (eg, 0·01) to reduce the chances of a false-positive conclusion. We might even consider lowering the power below 0·80 because of our reduced concern about missing an effective treatment—an effective safe treatment already exists. By contrast, if an investigator tests a standard prophylactic antibiotic against a cheap safe vitamin supplement the balance changes. Little harm could come from making an ? error so setting it at 0·10 might make sense.7 However, if this cheap easy intervention produced benefit, we would not want to miss it. Thus, investigators might increase power to 0·99.

Different assumptions about ? error and power directly change sample sizes. Reducing ? and increasing power both increase the sample: for example, reducing ? from 0·05 to 0·01 generates about a 70% increase in trial size at power=0·50 and a 50% increase at power=0·80 (table). At ?=0·05, increasing power from 0·50 to 0·80 yields a two-fold increase in trial size and from 0·50 to 0·99 almost a five-fold increase (table). Choices of ? and power thus produce different sample sizes and trial costs.

	Power (1–β)		
	0·80	0·90	0·95
α (type I error)			
0·05	7·85	10·51	13·00
0·01	11·68	14·88	17·82

Some investigators use one-sided tests for ? error to reduce estimated sample sizes. We discourage that approach. While we have assumed two-sided tests thus far, one-sided tests might indeed make sense in view of available biological knowledge. However, that decision should not affect sample size estimation. We suggest the same standard of evidence irrespective of whether a one-sided or two-sided test is assumed.7 Thus, a one-sided ?=0·025 yields the same level of evidence as a two-sided ?=0·05. Using a one-sided test in sample size calculations to reduce required sample sizes stretches credulity.

Estimation of population parameters

For some investigators, estimation of population parameters—eg, event rates in the treatment and control groups—has mystical overtones. Some researchers scoff at this notion, since estimating the parameters is the aim of the trial: needing to do it before the trial seems ludicrous. The key point, however, is that they are not estimating the population parameters per se but the treatment effect they deem worthy of detecting. That is a big difference.

Usually, investigators start by estimating the event rate in the control group. Sometimes scant data lead to unreliable estimates. For example, we needed to estimate an event rate for pelvic inflammatory disease in users of intrauterine devices in a family planning population in Nairobi, Kenya. Government officials estimated 40%; the clinicians at the medical centre thought that estimate was much too high and instead suggested 12%. We conservatively planned on 6%, but the placebo group in the actual randomised trial yielded 1·9%.8 The first estimate was off by more than 20-fold, which enormously affects sample size calculations.

Published reports can provide an estimate of the endpoint in the control group. Usually, however, they incorporate a host of differences, such as dissimilar locations, eligibility criteria, endpoints, and treatments. Nevertheless, some information on the control group usually exists. That becomes the starting point.

In a trial on prevention of fever after hysterectomy, data assumed to be reasonably good show that 10% of women have febrile morbidity after the standard prophylactic antibiotic. That becomes the event rate for the control group. Estimation of the effect size of interest should reflect both clinical acumen and the potential public-health effect. This important aspect should not default to a statistician. The decision process proceeds by accumulating clinical background knowledge. Assume the standard antibiotic costs US$10 for prophylaxis, incurs few side-effects, and is administered orally. The new antibiotic costs US$200 for prophylaxis, has more side-effects, is administered intravenously, but has a broader range of coverage. All

these pragmatic and clinical factors bear on the decision process. In view of the 10% event rate for fever in the control group, and knowing the clinical background, would we be interested in detecting a 10% reduction to 9%; a 20% reduction to 8%; a 30% reduction to 7%; a 40% reduction to 6%; a 50% reduction to 5%; and so forth? Determining the difference to detect reflects inherently subjective clinical judgments. No right answer exists. We could say that a 30% reduction is worthwhile to detect, but another investigator might decide on a 50% reduction.

These parameter assumptions enormously affect sample size calculations. Keeping the assumptions for the control group constant, halving the effect size necessitates a greater than four-fold increase in trial size. Similarly, quartering the effect size requires a greater than 16-fold increase in trial size. Stated alternatively, sample sizes rise by the inverse square of the effect size reduction (which statisticians call a quadratic relation). For example, in view of our initial parameter estimates of 10% in the control group and 6% in the intervention group, and ?=0·05 and power=0·90, about 965 participants would be necessary in each group Halving the effect size, thereby altering the intervention group estimate to 8%, requires a more than four-fold increase in sample size to 4301. Quartering the effect size, thereby altering the intervention group estimate to 9%, necessitates a more than 18-fold increase in trial size to 18066 per group. Small changes in effect size generate large changes in trial size.

The need for huge trial sizes with low event rates frustrates investigators. That frustration partly stems from a lack of understanding that, with binary endpoints, numerator events drive trial power rather than denominators. For example, assume ?=0·05 and a desired 40% reduction in the outcome event rate. A trial of 2000 participants (1000 assigned to the treatment group and 1000 to the control) with a control group event rate of 10% would provide similar power to a trial of 20 000 participants (10 000 ssigned to each group) with a control group event rate of 1%. Both trials would need a similar number of numerator events—about 160—for roughly 90% power.

Low power with limited available participants

What happens when sample size software—in view of an investigator's diligent estimates—yields a trial size that exceeds the number of available participants? Frequently, investigators then calculate backwards and estimate that they have low power (eg, 0·40) for their available participants. This practice may be more the rule than the exception.9

Some methodologists advise clinicians to abandon such a low-power study. Many ethics review boards deem a low power trial unethical.10–12 Chalmers' early paper on the lack of power in published trials contributed to this response, which brings us back to our opening paragraphs. He felt his group's article fuelled these over-reactions.4

Chalmers eventually stated that so-called underpowered trials can be acceptable because they could ultimately be combined in a meta-analysis.4,13 This view seems unsupported by many statisticians, surprisingly even those in favour of small trials.9 Nevertheless, we agree with Chalmers' view, which undoubtedly will draw the ire of many statisticians and ethicists. Our support attaches three caveats.

First, the trial should be methodologically strong, thus eliminating bias. Unfortunately, the adequate-power mantra frequently overwhelms discussion on other methodological aspects. For example, inadequate randomisation usually yields biased results. Those biased results cannot be salvaged even if a huge sample size generates great precision.14–16 By contrast, if investigators design and implement a trial properly, that trial essentially yields an unbiased estimate of effect, even if it has lower power (and precision). Moreover, because the results are unbiased, the trial could be combined with similar unbiased trials in a meta-analysis. Indeed, this idea, especially when incorporated into prospective meta-analyses,17 is akin to multicentre trials.

Second, authors must report their methods and results properly to avoid misinterpretation. If they report the trial results properly using interval estimation, the wide confidence intervals around the estimated treatment effect would accurately depict the low power. Reporting of confidence intervals represents a worthwhile contribution and avoids “the absence of evidence is not evidence of absence” problem wrought by simplistic p>0·05 conclusions.18–20

Third, low-powered trials must be published irrespective of their results, thereby becoming available for meta-analysis. Publication bias constitutes the strongest argument against underpowered trials.21,22 Publication bias emerges when published trials do not represent all trials undertaken, usually because statistically significant results tend to be submitted and published more frequently than indeterminate results. Low-powered trials contribute to the problem because they more generally yield an indeterminate result. Condemnation of all underpowered trials and prevention of their conduct, however, thwarts important research. We need to directly tackle the real culprit of publication bias, and the scientific community has made great strides. Not publishing completed trials is called both unscientific and unethical in the scientific literature.23–25 Trial registration schemes catalogue ongoing trials such that their results will not be lost. Furthermore, large systematic review enterprises, most notably the Cochrane Collaboration, scour unpublished work to reduce publication bias.

Proclamations of underpowered trials being unethical strike us as a bit odd for at least two reasons. First, preoccupation with sample size overshadows the more pertinent concerns of elimination of bias. Second,

how can a process rife with subjectivity fuel a black-white decision on its ethics? With that subjectivity, basing trial ethics on statistical power seems simplistic and misplaced. Indeed, since investigators estimate sample size on the basis of rough guesses, if deeming the implementation of low power trials as unethical is taken to a logical extreme, then the world will have no trials because sample size determination would always be open to question. “Statements that it is unethical to embark on controlled trials unless an arbitrarily defined level of statistical power can be assured make no sense if the alternative is acquiescence in ignorance of the effects of healthcare interventions.”24 Edicts that underpowered trials are unethical challenge reason and, furthermore, disregard that sometimes potential participants desire involvement in trials.26
Back to top

Sample size samba

Investigators sometimes perform a “sample size samba” to achieve adequate power.27,28 The dance involves retrofitting of the parameter estimates (in particular, the treatment effect worthy of detection) to the available participants. This practice seems fairly common in our experience and in that of others.27 Moreover, funding agencies, protocol committees, and even ethics review boards might encourage this backward process. It represents an operational solution to a real problem. In view of the circumstances, we do not judge harshly the samba, because it probably has facilitated the conduct of many important studies. Moreover, it truly depicts estimates of the sample sizes necessary given the provided assumptions. Nevertheless, the process emphasises the inconsistencies in the “underpowered trials are unethical” argument: a proposed trial is unethical before the “samba” and becomes ethical thereafter simply by shifting the estimate of effect size. All trials have an infinite number of powers, and low power is relative.
Back to top

Sample size modification

With additional available participants and resource flexibility, investigators could consider a sample size modification strategy, which would alleviate some of the difficulties with rough guesses used in the initial sample size calculations. Usually, modifications lead to increased sample sizes,29 so investigators should have access to the participants and the funding to accommodate the modifications.

Approaches to modification rely on revision of the event rate, the variance of the endpoint, or the treatment effect.30–33 Importantly, any sample size modifications at an interim stage of a trial should hinge on a prespecified plan that avoids bias. The sponsor or steering committee should describe in the protocol a comprehensible plan for the timing and method of the potential modifications.31
Back to top

Futility of post hoc power calculations

A trial yields a treatment effect and confidence interval for the results. The power of the trial is expressed in that confidence interval. Hence, the power is no longer a meaningful concern.7,27,34 Nevertheless, after trial completion, some investigators do power calculations on statistically non-significant trials using the observed results for the parameter estimates. This exercise has specious appeal, but tautologically yields an answer of low power.7,27 In other words, this ill-advised exercise answers an already answered question.
Back to top

What should readers look for in sample size calculations?

Readers should find the a-priori estimates of sample size. Indeed, in trial reports, confidence intervals appropriately indicate the power. However, sample size calculations still provide important information. First, they specify the primary endpoint, which safeguards against changing outcomes and claiming a large effect on an outcome not planned as the primary outcome.35 Second, knowing the planned size alerts readers to potential problems. Did the trial encounter recruitment difficulties? Did the trial stop early because of a statistically significant result? If so, the authors should provide a formal statistical stopping rule.36 If they did not use a formal rule, then multiple looks at the data inflated ?.5,29 Similar problems can be manifested in larger than planned sample sizes. Providing planned sizes, however arbitrary, lays the groundwork for transparent reporting.

Low reported power or unreported sample size calculations usually are not a fatal flaw. Low power can reflect a lack of methodological knowledge, but it may just indicate an inadequate number of potential participants. Sample size calculations, even with low power, still provide the vital information described above. What if authors neglect mentioning a-priori sample size calculations? Readers should cautiously interpret the results because of the missing information on primary outcome and on stopping clues. Moreover, neglecting to report sample size calculations suggests a methodological naiveté that might portend other problems.

Nevertheless, readers should be most concerned with systematic errors (bias) hidden by investigators. Authors failing to report poor randomisation, inadequate allocation concealment, deficient blinding, or defective participant retention hide inadequacies that could cause major bias.37–41 Thus, readers should ascribe less concern to perceived inadequate sample size for two substantial reasons: first, it does not cause bias and, second, any random error produced transparently surfaces in the confidence intervals and p

values. The severest problems for readers are the systematic errors that are not revealed. In other words, readers should not totally discount a trial simply because of low power, but they should carefully weigh its value accordingly. The value resides in the context of other research, either past or future.42

Readers should find all assumptions underlying any sample size calculation: type I error (?), power (or ?), event rate in the control group, and a treatment effect of interest (or analogously, an event rate in the treatment group). A statement that "we calculated necessary sample sizes of 120 in each group at ?=0·05 and power=0·90" is almost meaningless, because it neglects the estimates for the effect size and control group event rate. Even small trials have high power to detect huge treatment effects.

Readers should also examine the assumptions for the sample size calculation. For example, they might believe that a smaller effect size is more worthy than the planned effect size. Therefore, the reader would be aware of the lower power of the trial relative to their preferred effect size.
Back to top

Conclusions

Statistical power is an important notion, but it should be stripped of its ethical bellwether status. We question the branding of trials as unethical based solely on an inherently subjective, imprecise sample size calculation process. We endorse planning for adequate power, and we salute large multicentre trials of the ISIS-2 ilk;43 indeed, more such studies should be undertaken. However, if the scientific world insisted solely on large trials, many unanswered questions in medicine would languish unanswered. Some shift of emphasis from a fixation on sample size to a focus on methodological quality would yield more trials with less bias. Unbiased trials with imprecise results trump no results at all.

• **Incidence and prevalence (epidemiology)**
From Wikipedia, the free encyclopedia

The **incidence** of disease is defined as the number of *new* ? of disease occurring in a ? during a defined time interval. The number is useful to ? because it is a measure of the ? of disease.

- The *incidence rate* is defined as the incidence divided by the sum of the different times each individual was at risk of the disease.
- The *incidence per 1,000* is defined as follows:

Including the number of individuals at risk in the ? makes this measure the most common way to express incidence, although other coefficients such as 100,000 are often used.

Incidence and incidence rate are not to be confused with ?, which is defined as the number of individuals with a certain disease in a population at a specified time divided by the number of individuals in the population at that time. This measure differs from incidence in that it does not convey information about risk.

To illustrate, a disease with a long duration that was spread widely in a community in ? will have a high prevalence in ? (remembering that it has a long duration) but it might have a low incidence rate in 2003. Conversely, a disease that is easily transmitted but has a short duration may have a low prevalence and a high incidence. Incidence rate is useful for talking about diseases like chickenpox, which have a lifetime ? of almost one, since it is measured per unit time so can tell us when infections are likely to occur.
Incidence of disease can also be referred to as absolute risk.