

BMJ 1994;308:896 (2 April)

Papers

Statistics Notes: 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.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

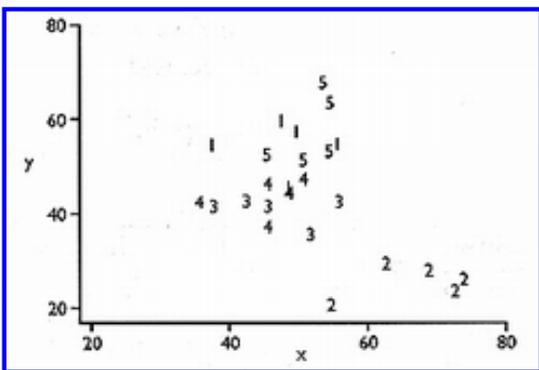
In clinical research we are often able to take several measurements on the same patient. The correct analysis of such data is more complex than if each patient were measured once. This is because the variability of measurements made on different subjects is usually much greater than the variability between 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			
55	62	48	58	63	28	38	40	51	46
51	50	56	53	74	24	56	41	46	36
54	66	49	44	69	26	46	40	36	41
46	51	38	53	55	19	43	41	49	43
55	52	50	56	73	22	52	34	46	45
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	r=0.32	r=-0.30	r=0.37				
P=0.55		P=0.97	P=0.59	P=0.63	P=0.55				
P=0.33									



Simulated data for five pairs of measurement of two uncorrelated variables (X and Y) for five subjects

View larger version (8K):

[\[in this window\]](#)

[\[in a new window\]](#)

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.

This article has been cited by other articles:

- Callow, J., Summers, L. K., Bradshaw, H., Frayn, K. N (2002). Changes in LDL particle composition after the consumption of meals containing different amounts and types of fat. *Am. J. Clin. Nutr.* 76: 345-350 [\[Abstract\]](#) [\[Full text\]](#)
- RAAIJMAKERS, E., FAES, TH. J. C., SCHOLTEN, R. J. P. M., GOOVAERTS, H. G., HEETHAAR, R. M. (1999). A Meta-analysis of Published Studies Concerning the Validity of Thoracic Impedance Cardiography. *Annals NYAS Online* 873: 121-127 [\[Abstract\]](#) [\[Full text\]](#)
- Kunst, P. W. A., Noordegraaf, A. V., Raaijmakers, E., Bakker, J., Groeneveld, A. B. J., Postmus, P. E., de Vries, P. M. J. M. (1999). Electrical Impedance Tomography in the Assessment of Extravascular Lung Water in Noncardiogenic Acute Respiratory Failure*. *Chest* 116: 1695-1702 [\[Abstract\]](#) [\[Full text\]](#)
- Persaud, R, Bland, J M, Altman, D G (1994). Correlation, regression, and repeated data. *BMJ* 308: 1510a-1510 [\[Full text\]](#)
- Hill-Smith, I (1994). Sample size. *BMJ* 308: 1304-1304 [\[Full text\]](#)
- Altman, D. G, Bland, J M. (1997). **Statistics notes:** Units of analysis. *BMJ* 314: 1874-1874 [\[Full text\]](#)
- Bland, J M., Altman, D. G (1995). **Statistics notes:** Calculating correlation coefficients with repeated observations: Part 2--correlation between subjects. *BMJ* 310: 633-633 [\[Full text\]](#)
- Bland, J M., Altman, D. G (1995). **Statistics notes:** Calculating correlation coefficients with repeated observations: Part 1--correlation within subjects. *BMJ* 310: 446-446 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond](#) to this article
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

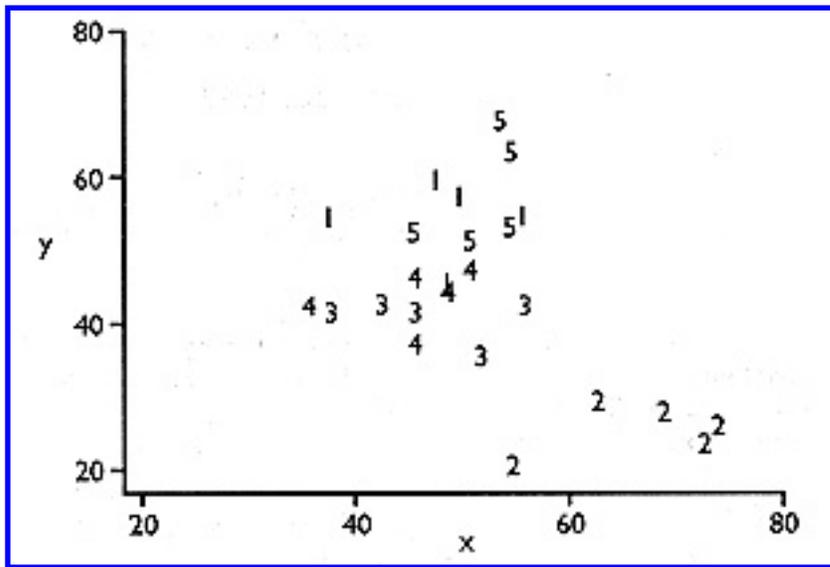
Related letters in BMJ:

Sample size

I Hill-Smith
BMJ 1994 308: 1304. [\[Letter\]](#)

Correlation, regression, and repeated data

R Persaud, J M Bland, and D G Altman
BMJ 1994 308: 1510. [\[Letter\]](#)



Simulated data for five pairs of measurement of two uncorrelated variables (X and Y) for five subjects

[\[View larger version \(10K\)\]](#)

BMJ 1994;308:1499 (4 June)

Education and debate

Statistic Notes: Regression towards the mean

J M Bland, D G **Altman**

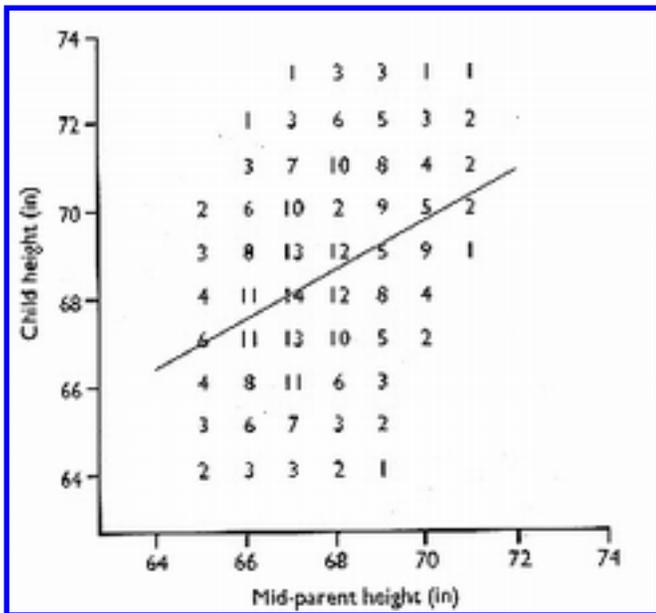
Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX.

The statistical term "regression," from a Latin root meaning "going back," was first used by Francis Galton in his paper "Regression towards Mediocrity in Hereditary Stature."¹ Galton related the heights of children to the average height of their parents, which he called the mid-parent height (figure). Children and parents had the same mean height of 68.2 inches. The ranges differed, however, because the mid-parent height was an average of two observations and thus had its range reduced. Now, consider those parents with a mid-height between 70 and 71 inches. The mean height of their children was 69.5 inches, which was closer to the mean height of all children than the mean height of their parents was to the mean height of all parents. Galton called this phenomenon "regression towards mediocrity"; we now call it "regression towards the mean." The same thing happens if we start with the children. For the children with height between 70 and 71 inches, the mean height of their parents was 69.0 inches. This is a statistical, not a genetic phenomenon.

If we take each group of mid-parents by height and calculate the mean height of their children, these means will lie close to a straight line. This line came to be called the regression line, and hence the process of fitting such lines became known as "regression."

In mathematical terms, if variables X and Y have standard deviations s_X and s_Y , and correlation r , the slope of the familiar least squares regression line can be written rs_Y/s_X . Thus a change of one standard deviation in X is associated with a change of r standard deviations in Y. Unless X and Y are exactly linearly related, so that all the points lie along a straight line, r is less than 1. For a given value of X the predicted value of Y is always fewer standard deviations from its mean than is X from its mean. Regression towards the mean occurs unless $r=1$, perfect correlation, so it always occurs in practice. We give some examples in a subsequent note.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)



Galton's original data showing the relation between the heights of children and their parents, with regression line¹

View larger version (12K):

[\[in this window\]](#)

[\[in a new window\]](#)

- Galton F. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute* 1886;15:246-63.

This article has been cited by other articles:

- Landow, L. (2002). Another Example of Regression to the Mean. *Anesth Analg* 94: 1673-1673 [\[Full text\]](#)
- Vickers, A., Irnich, D., Krauss, M. (2001). Acupuncture for treatment of chronic neck pain. *BMJ* 323: 1306-1306 [\[Full text\]](#)
- Vickers, A. J, **Altman**, D. G (2001). Statistics Notes: Analysing controlled trials with baseline and follow up measurements. *BMJ* 323: 1123-1124 [\[Full text\]](#)
- Ly, L. P., Jimenez, M., Zhuang, T. N., Celermajer, D. S., Conway, A. J., Handelsman, D. J. (2001). A Double-Blind, Placebo-Controlled, Randomized Clinical Trial of Transdermal Dihydrotestosterone Gel on Muscular Strength, Mobility, and Quality of Life in Older Men with

- [▶ Email this article to a friend](#)
- [▶ Respond to this article](#)
- [▶ PubMed citation](#)
- [▶ Related articles in PubMed](#)
- [▶ Download to Citation Manager](#)
- [▶ Search Medline for articles by: **Bland, J M** || **Altman, D G**](#)
- [▶ Alert me when: **New articles cite this article**](#)

- Partial Androgen Deficiency. *J Clin Endocrinol Metab* 86: 4078-4088 [\[Abstract\]](#) [\[Full text\]](#)
- Deeks, J. J (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 323: 157-162 [\[Full text\]](#)
 - Browne, S. M., Halligan, P. W., Wade, D. T., Taggart, D. P. (1999). COGNITIVE PERFORMANCE AFTER CARDIAC OPERATION: IMPLICATIONS OF REGRESSION TOWARD THE MEAN. *J Thorac Cardiovasc Surg* 117: 481-485 [\[Abstract\]](#) [\[Full text\]](#)
 - Richards, M., Hardy, R., Kuh, D., Wadsworth, M. E J (2001). Birth weight and cognitive function in the British 1946 birth cohort: longitudinal population based study. *BMJ* 322: 199-203 [\[Abstract\]](#) [\[Full text\]](#)
 - Coutant, R., Carel, J.-C., Letrait, M., Bouvattier, C., Chatelain, P., Coste, J., Chaussain, J.-L. (1998). Short Stature Associated with Intrauterine Growth Retardation: Final Height of Untreated and Growth Hormone-Treated Children. *J Clin Endocrinol Metab* 83: 1070-1074 [\[Abstract\]](#) [\[Full text\]](#)
 - Doepfmer, S., Guggenmoos-Holzmann, I. (1996). Plasma cholesterol response to dietary saturated fat. *BMJ* 312: 511c-512 [\[Full text\]](#)
 - Bland, J M, **Altman**, D G (1994). Statistics Notes: Some examples of regression towards the mean. *BMJ* 309: 780-780 [\[Full text\]](#)
 - Greco, L., Power, C., Peckham, C. (1995). Adult outcome of normal children who are short or underweight at age 7 years. *BMJ* 310: 696-700 [\[Abstract\]](#) [\[Full text\]](#)
 - Sharp, S. J, Thompson, S. G, **Altman**, D. G (1996). The relation between treatment benefit and underlying risk in meta-analysis. *BMJ* 313: 735-738 [\[Full text\]](#)
 - Coste, J., Letrait, M., Carel, J. C., Tresca, J. P., Chatelain, P., Rochiccioli, P., Chaussain, J. L., Job, J. C. (1997). Long term results of growth hormone treatment in France in children of short stature: population, register based study. *BMJ* 315: 708-713 [\[Abstract\]](#) [\[Full text\]](#)
 - Carel, J.-C., Ecosse, E., Nicolino, M., Tauber, M., Leger, J., Cabrol, S., Bastie-Sigeac, I., Chaussain, J.-L., Coste, J. (2002). Adult height after long term treatment with recombinant growth hormone for idiopathic isolated growth hormone deficiency: observational follow up study of the French population based registry. *BMJ* 325: 70-70 [\[Abstract\]](#) [\[Full text\]](#)

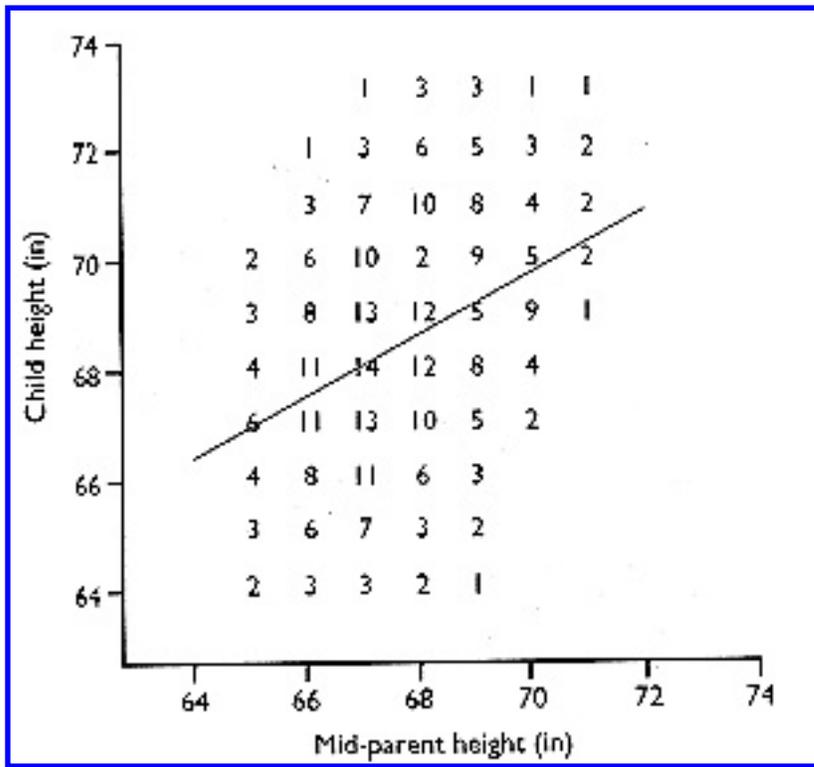
[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



Galton's original data showing the relation between the heights of children and their parents, with regression line¹

[\[View larger version \(19K\)\]](#)

BMJ 1994;308:1552 (11 June)

General practice

Statistics Notes: Diagnostic tests 1: sensitivity and specificity

D G Altman, J M Bland

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D G](#) || [Bland, J M](#)
- ▶ Alert me when:
[New articles cite this article](#)

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 diagnosis¹

Liver scan	Pathology		Total
	Abnormal (+)	Normal (-)	
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.

1. Drum DE, Christacapoulos JS. Hepatic scintigraphy in clinical decision making. *J Nucl Med* 1972;13:908-15. [\[Medline\]](#)
2. Gardner MJ, Altman DG. Calculating confidence intervals for proportions and their differences. In: Gardner MJ, Altman DG, eds. *Statistics with confidence*. London: BMJ Publishing Group, 1989:28-33.

This article has been cited by other articles:

- Maffulli, N. (1998). The Clinical Diagnosis of Subcutaneous Tear of the Achilles Tendon: A Prospective Study in 174 Patients. *Am J Sports Med* 26: 266-270 [\[Abstract\]](#) [\[Full text\]](#)
- Equi, A C, Pike, S E, Davies, J, Bush, A (2001). Use of cough swabs in a cystic fibrosis clinic. *Arch. Dis. Child.* 85: 438-439 [\[Abstract\]](#) [\[Full text\]](#)
- Hassey, A., Gerrett, D., Wilson, A. (2001). A survey of validity and utility of electronic patient

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by: [Altman, D G](#) || [Bland, J M](#)
- ▶ Alert me when: [New articles cite this article](#)

records in a general practice. *BMJ* 322: 1401-1405 [\[Abstract\]](#) [\[Full text\]](#)

- Rushforth, H., Bliss, A., Burge, D., Glasper, E. A. (2000). A pilot randomised controlled trial of medical versus nurse clerking for minor surgery. *Arch. Dis. Child.* 83: 223-226 [\[Abstract\]](#) [\[Full text\]](#)
- Altman, D G, Bland, J M (1994). **Statistics Notes:** Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 309: 188-188 [\[Full text\]](#)
- Harding, S P, Broadbent, D M, Neoh, C, White, M C, Vora, J (1995). Sensitivity and specificity of photography and direct ophthalmoscopy in screening for sight threatening eye disease: the Liverpool diabetic eye study. *BMJ* 311: 1131-1135 [\[Abstract\]](#) [\[Full text\]](#)
- McConville, J P, Craig, J J, Collinge, J., Rossor, M. N, Thomas, D., Frosh, A., Tolley, N., Otto, M., Zerr, I., Poser, S., Wiltfang, J., Schütz, E., Pfahlberg, A., Gefeller, O. (1998). Diagnosis of Creutzfeldt-Jakob disease by measurement of S100 protein in serum. *BMJ* 317: 472-472 [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1994;309:102 (9 July)

General practice

Statistics Notes: 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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

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 specificity¹ 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:

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

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

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.

This article has been cited by other articles:

- Maffulli, N. (1998). The Clinical Diagnosis of Subcutaneous Tear of the Achilles Tendon: A Prospective Study in 174 Patients. *Am J Sports Med* 26: 266-270 [[Abstract](#)] [[Full text](#)]
- Equi, A C, Pike, S E, Davies, J, Bush, A (2001). Use of cough swabs in a cystic fibrosis clinic. *Arch. Dis. Child.* 85: 438-439 [[Abstract](#)] [[Full text](#)]
- Hassey, A., Gerrett, D., Wilson, A. (2001). A survey of validity and utility of electronic patient records in a general practice. *BMJ* 322: 1401-1405 [[Abstract](#)] [[Full text](#)]
- McKeith, I. G., Ballard, C. G., Perry, R. H., Ince, P. G., O'Brien, J. T., Neill, D., Lowery, K., Jaros, E., Barber, R., Thompson, P., Swann, A., Fairbairn, A. F., Perry, E. K. (2000). Prospective validation of Consensus criteria for the diagnosis of dementia with Lewy bodies. *Neurology* 54: 1050-1058 [[Abstract](#)] [[Full text](#)]
- Foster, P. J, Devereux, J. G, Alsbirk, P. H., Lee, P. S., Uranchimeg, D., Machin, D., Johnson, G. J, Baasanhu, J. (2000). Detection of gonioscopically occludable angles and primary angle closure glaucoma by estimation of limbal chamber depth in Asians: modified grading scheme. *Br. J. Ophthalmol.* 84: 186-192 [[Abstract](#)] [[Full text](#)]
- Altman, D G, Bland, J M (1994). **Statistics Notes**: Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 309: 188-188 [[Full text](#)]
- Hussaini, S H, Sheridan, M B, Davies, M (1999). The predictive value of transabdominal ultrasonography in the diagnosis of biliary tract complications after orthotopic liver transplantation. *Gut* 45: 900-903 [[Abstract](#)] [[Full text](#)]
- Maurer, M., Shambal, S., Berg, D., Woydt, M., Hofmann, E., Georgiadis, D., Lindner, A., Becker, G. (1998). Differentiation Between Intracerebral Hemorrhage and Ischemic Stroke by Transcranial Color-Coded Duplex-Sonography. *Stroke* 29: 2563-2567 [[Abstract](#)] [[Full text](#)]
- McConville, J P, Craig, J J, Collinge, J., Rossor, M. N, Thomas, D., Frosh, A., Tolley, N., Otto, M., Zerr, I., Poser, S., Wiltfang, J., Schütz, E., Pfahlberg, A., Gefeller, O. (1998). Diagnosis of Creutzfeldt-Jakob disease by measurement of S100 protein in serum. *BMJ* 317: 472-472 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)
[Help](#)
[Search/Archive](#)
[Feedback](#)
[Search Result](#)

BMJ 1994;309:188 (16 July)

Education and debate

Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots

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 1RE.

We have previously considered diagnosis based on tests that give a yes or no answer.^{1,2} Many diagnostic tests, however, are quantitative, notably in clinical chemistry. The same statistical approach can be used only if we can select a cut off point to distinguish "normal" from "abnormal," which is not a trivial problem. Firstly, we can investigate to what extent the test results differ among people who do or do not have the diagnosis of interest. The receiver operating characteristic (ROC) plot is one way to do this. These plots were developed in the 1950s for evaluating radar signal detection. Only recently have they become commonly used in medicine.

We assume that high values are more likely among those dubbed "abnormal." Figure [1](#) shows the values of an index of mixed epidermal cell lymphocyte reactions in bone marrow transplant recipients who did or did not develop graft versus host disease.³ The usefulness of the test for predicting graft versus host disease will clearly relate to the degree of non- overlap between the two distributions.

- ▶ [Email this article to a friend](#)
- ▶ **[Respond to this article](#)**
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D G](#) || [Bland, J M](#)
- ▶ Alert me when:
[New articles cite this article](#)

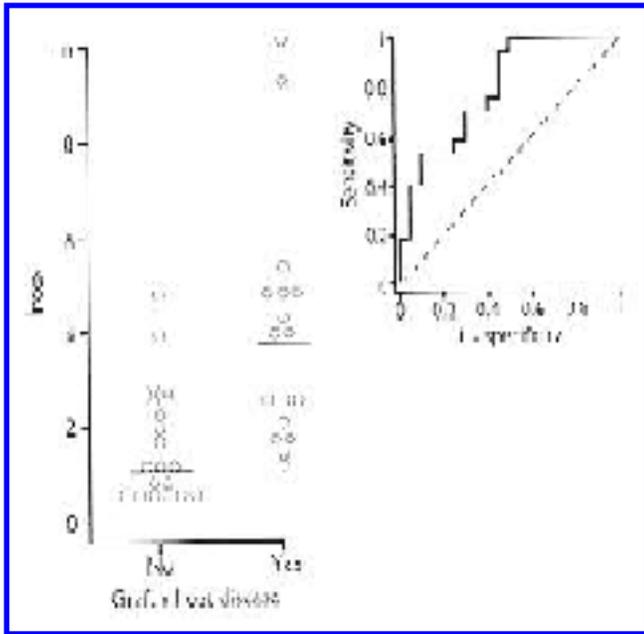


FIG 1 (left) - Distribution of values of an index of mixed epidermal cell lymphocyte reactions in patients who did or did not develop grafts versus host disease³

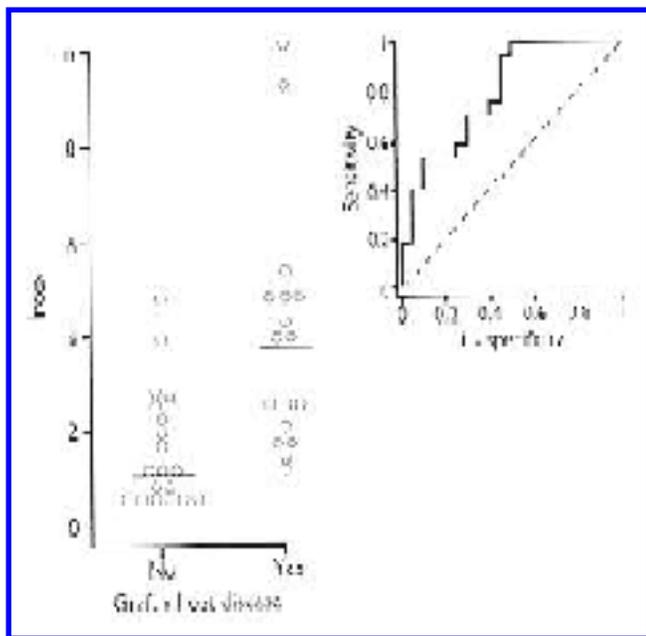
View larger version (15K):

[\[in this window\]](#)

[\[in a new window\]](#)

A receiver operating characteristic plot is obtained by calculating the sensitivity and specificity of every observed data value and plotting sensitivity against 1 - specificity, as in Figure 2. A test that perfectly discriminates between the two groups would yield a "curve" that coincided with the left and top sides of the plot. A test that is completely useless would give a straight line from the bottom left corner to the top right corner. In practice there is virtually always some overlap of the values in the two groups, so the curve will lie somewhere between these extremes.

FIG 2 (above) - Receiver operating characteristic curve for the data shown in fig 1



View larger version (15K):

[\[in this window\]](#)

[\[in a new window\]](#)

A global assessment of the performance of the test (sometimes called diagnostic accuracy⁴) is given by the area under the receiver operating characteristic curve. This area is equal to the probability that a random person with the disease has a higher value of the measurement than a random person without the disease. (This probability is a half for an uninformative test - equivalent to tossing a coin.)

No test will be clinically useful if it cannot discriminate,⁴ so a global assessment of discriminatory power is an important step. Having determined that a test does provide good discrimination the choice can be made of the best cut off point for clinical use. This requires the choice of a particular point, and is thus a local assessment. The simple approach of minimising "errors" (equivalent to maximising the sum of the sensitivity and specificity) is not necessarily best. Consideration needs to be given to the costs (not just financial) of false negative and false positive diagnoses and to the prevalence of the disease in the subjects being tested.⁴ For example, when screening the general population for cancer the cut off point would be chosen to ensure that most cases were detected (high sensitivity) at the cost of many false positives (low specificity), who could then be eliminated by a further test.

A receiver operating characteristic plot is particularly useful when comparing two or more measures. A test with a curve that lies wholly above the curve of another will be clearly better. Methods for comparing the areas under two curves for both paired and unpaired data are reviewed by Zweing and Campbell,⁴ who give a full assessment of this method.

1. Altman DG, Bland M. Diagnostic tests 1: sensitivity and specificity. *BMJ* 1994;308:1552. [\[Full](#)

[Text](#)

2. Altman DG, Bland M. Diagnostic tests 2: predictive values. *BMJ* 1994;309:102. [\[Full Text\]](#)
3. Bagot M, Mary J-Y, Heslan M, et al. The mixed epidermal cell lymphocyte - reaction is the most predictive factor of acute graft-versus- host disease in bone marrow graft recipients. *Br J Haematol* 1988;70:403-9. [\[Medline\]](#)
4. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561-77. [\[Abstract\]](#)

This article has been cited by other articles:

- Stoeber, K., Swinn, R., Prevost, A. T., de Clive-Lowe, P., Halsall, I., Dilworth, S. M., Marr, J., Turner, W. H., Bullock, N., Doble, A., Hales, C. N., Williams, G. H. (2002). Diagnosis of Genito-Urinary Tract Cancer by Detection of Minichromosome Maintenance 5 Protein in Urine Sediments. *J Natl Cancer Inst* 94: 1071-1079 [\[Abstract\]](#) [\[Full text\]](#)
- Gosche, K. M., Mortimer, J. A., Smith, C. D., Markesbery, W. R., Snowdon, D. A. (2002). Hippocampal volume as an index of Alzheimer neuropathology: Findings from the Nun Study. *Neurology* 58: 1476-1482 [\[Abstract\]](#) [\[Full text\]](#)
- Dimitriou, G, Greenough, A, Endo, A, Cherian, S, Rafferty, G F (2002). Prediction of extubation failure in preterm infants. *Arch. Dis. Child. Fetal Neonatal Ed.* 86: F32-35 [\[Abstract\]](#) [\[Full text\]](#)
- Esterhuizen, A.D., Franken, D.R., Lourens, J.G.H., van Rooyen, L.H. (2001). Clinical importance of zona pellucida-induced acrosome reaction and its predictive value for IVF. *Hum Reprod* 16: 138-144 [\[Abstract\]](#) [\[Full text\]](#)
- Renehan, A. G., Painter, J. E., O'Halloran, D., Atkin, W. S., Potten, C. S., O'Dwyer, S. T., Shalet, S. M. (2000). Circulating Insulin-Like Growth Factor II and Colorectal Adenomas. *J Clin Endocrinol Metab* 85: 3402-3408 [\[Abstract\]](#) [\[Full text\]](#)
- Esterhuizen, A.D., Franken, D.R., Lourens, J.G.H., Prinsloo, E., van Rooyen, L.H. (2000). Sperm chromatin packaging as an indicator of in-vitro fertilization rates. *Hum Reprod* 15: 657-661 [\[Abstract\]](#) [\[Full text\]](#)
- Adams, M. R., Nakagomi, A., Keech, A., Robinson, J., McCredie, R., Bailey, B. P., Freedman, S. B., Celermajer, D. S. (1995). Carotid Intima-Media Thickness Is Only Weakly Correlated With the Extent and Severity of Coronary Artery Disease. *Circulation* 92: 2127-2134 [\[Abstract\]](#) [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond](#) to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D G](#) || [Bland, J M](#)
- ▶ Alert me when:
[New articles cite this article](#)

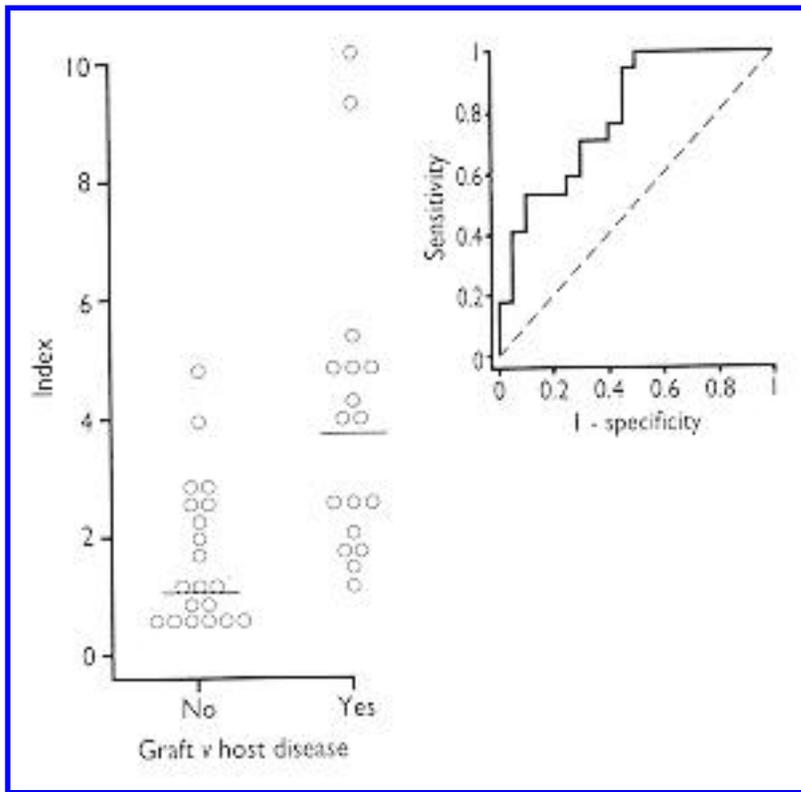


FIG 1 (left) - Distribution of values of an index of mixed epidermal cell lymphocyte reactions in patients who did or did not develop grafts versus host disease³

[\[View larger version \(8K\)\]](#)

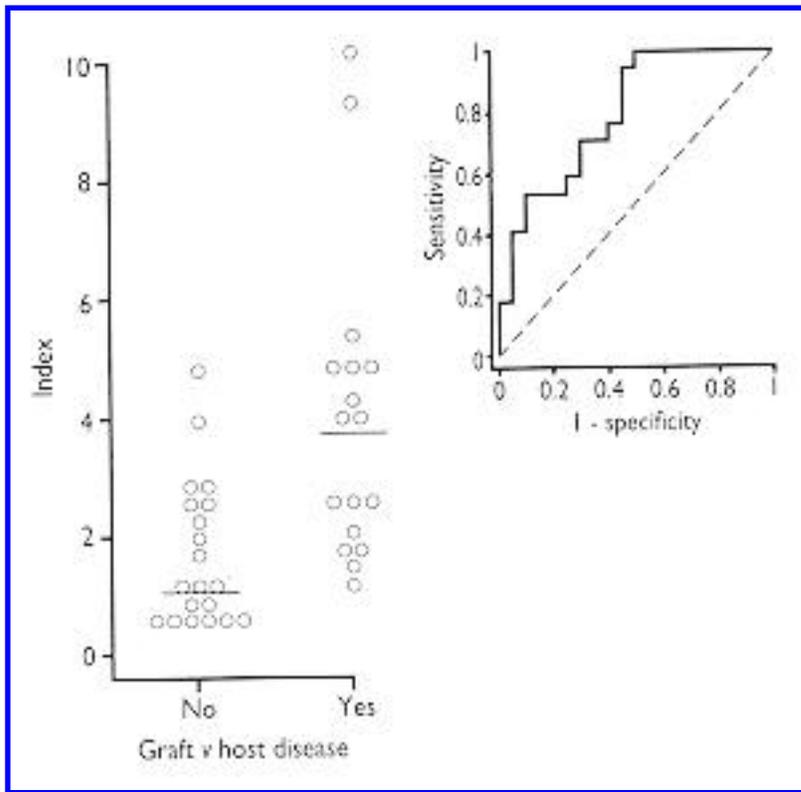


FIG 2 (above) - Receiver operating characteristic curve for the data shown in fig 1

[\[View larger version \(8K\)\]](#)

BMJ 1994;309:248 (23 July)

General practice

Statistics Notes: One and two sided tests of significance

J M Bland, D G Bland

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.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Bland, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

In some comparisons - for example, between two means or two proportions - there is a choice between two sided or one sided tests of significance (all comparisons of three or more groups are two sided).

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

When we use a test of significance to compare two groups we usually start with the null hypothesis that there is no difference between the populations from which the data come. If this hypothesis is not true the alternative hypothesis must be true - that there is a difference. Since the null hypothesis specifies no direction for the difference nor does the alternative hypothesis, and so we have a two sided test. In a one sided test the alternative hypothesis does specify a direction - for example, that an active treatment is better than a placebo. This is sometimes justified by saying that we are not interested in the possibility that the active treatment is worse than no treatment. This possibility is still part of the test; it is part of the null hypothesis, which now states that the difference in the population is zero or in favour of the placebo.

A one sided test is sometimes appropriate. Luthra et al investigated the effects of laparoscopy and hydrotubation on the fertility of women presenting at an infertility clinic.¹ After some months laparoscopy was carried out on those who had still not conceived. These women were then observed for several further months and some of these women also conceived. The conception rate in the period before laparoscopy was compared with that afterwards. The less fertile a woman is the longer it is likely to take her to conceive. Hence, the women who had the laparoscopy should have a lower conception rate (by an unknown amount) than the larger group who entered the study, because the more fertile women had conceived before their turn for laparoscopy came. To see whether laparoscopy increased fertility, Luthra et al tested the null hypothesis that the conception rate after laparoscopy was less than or equal to that

before. The alternative hypothesis was that the conception rate after laparoscopy was higher than that before. A two sided test was inappropriate because if the laparoscopy had no effect on fertility the conception rate after laparoscopy was expected to be lower.

One sided tests are not often used, and sometimes they are not justified. Consider the following example. Twenty five patients with breast cancer were given radiotherapy treatment of 50 Gy in fractions of 2 Gy over 5 weeks.² Lung function was measured initially, at one week, at three months, and at one year. The aim of the study was to see whether lung function was lowered following radiotherapy. Some of the results are shown in the table, the forced vital capacity being compared between the initial and each subsequent visit using one sided tests. The direction of the one sided tests was not specified, but it may appear reasonable to test the alternative hypothesis that forced vital capacity decreases after radiotherapy, as there is no reason to suppose that damage to the lungs would increase it. The null hypothesis is that forced vital capacity does not change or increases. If the forced vital capacity increases, this is consistent with the null hypothesis, and the more it increases the more consistent the data are with the null hypothesis. Because the differences are not all in the same direction, at least one P value should be greater than 0.5. What has been done here is to test the null hypothesis that forced vital capacity does not change or decreases from visit 1 to visit 2 (nine week), and to test the null hypothesis that it does not change or increases from visit 1 to visit 3 (three months) or visit 4 (one year). These authors seem to have carried out one sided tests in both directions for each visit and then taken the smaller probability. If there is no difference in the population the probability of getting a significant difference by this approach is 10%, not 5% as it should be. The chance of a spurious significant difference is doubled. Two sided tests should be used, which would give probabilities of 0.26, 0.064, and 0.38, and no significant differences.

In general a one sided test is appropriate when a large difference in one direction would lead to the same action as no difference at all. Expectation of a difference in a particular direction is not adequate justification. In medicine, things do not always work out as expected, and researchers may be surprised by their results. For example, Galloe et al found that oral magnesium significantly increased the risk of cardiac events, rather than decreasing it as they had hoped.³ If a new treatment kills a lot of patients we should not simply abandon it; we should ask why this happened.

Two sided tests should be used unless there is a very good reason for doing otherwise. If one sided tests are to be used the direction of the test must be specified in advance. One sided tests should never be used simply as a device to make a conventionally non-significant difference significant.

1. Lund MB, Myhre KI, Melsom H, Johansen B. The effect on pulmonary function of tangential field technique in radiotherapy for carcinoma of the breast. *Br J Radiol* 1991;64:520-3. [[Abstract](#)]
2. Luthra P, Bland JM, Stanton SL. Incidence of pregnancy after laparoscopy and hydrotubation. *BMJ* 1982;284:1013. [[Medline](#)]
3. Galloe AM, Rasmussen HS, Jorgensen LN, Aurup P, Balslov S, Cintin C, Graudal N, McNair P. Influence of oral magnesium supplementation on cardiac events among survivors of an acute

myocardial infarction. *BMJ* 1993;307:585-7. [[Medline](#)]

This article has been cited by other articles:

- Moyer, L. A., Tita, A. T.N. (2002). Defending the Rationale for the Two-Tailed Test in Clinical Research. *Circulation* 105: 3062-3065 [[Full text](#)]
- Wolterbeek, R, Enkin, M W, Bland, J M, Altman, D G (1994). One and two sided tests of significance Statistical hypothesis should be brought into line with clinical hypothesis. *BMJ* 309: 873a-874 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond](#) to this article
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Bland, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

Related letters in BMJ:

One and two sided tests of significance Statistical hypothesis should be brought into line with clinical hypothesis

R Wolterbeek, M W Enkin, J M Bland, and D G Altman
BMJ 1994 309: 873-874. [[Letter](#)]

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1994;309:780 (24 September)

General practice

Statistics Notes: Some examples of 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 $r < 1$ 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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

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 equally 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 opposite 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.

1. Bland JM, Altman DG. Regression towards the mean. *BMJ* 1994;308:1499. [\[Full Text\]](#)
2. Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol* 1976;104:493-8. [\[Abstract\]](#)
3. Hayes RJ. Methods for assessing whether change depends on initial value. *Statistics in Medicine* 1988;7:915-27. [\[Medline\]](#)
4. Barrett JFR, Jarvis GJ, Macdonald, HN, Buchan PC, Tyrrell SN, Lilford RJ. Inconsistencies in clinical decision making in obstetrics. *Lancet* 1990;336:549-51. [\[Medline\]](#)
5. Esmail A, Bland M. Caesarian section for fetal distress. *Lancet* 1990;336:819.
6. Kuskowska-Wolk A, Karlsson P, Stolt M, Rossner S. The predictive value of body mass index based on reported weight and height. *Int J Obesity* 1989;13:441-3.
7. Schilchting P, Hoilund-Carlsen, Quaade F. Comparison of self reported height and weight with controlled height and weight in women and men. *Int J Obesity* 1981;5:67-76.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307-10.
9. Rousseeuw PJ. Why the wrong papers get published. *Chance* 1991;4:41-3.

This article has been cited by other articles:

- Decensi, A., Omodei, U., Robertson, C., Bonanni, B., Guerrieri-Gonzaga, A., Ramazzotto, F., Johansson, H., Mora, S., Sandri, M. T., Cazzaniga, M., Franchi, M., Pecorelli, S. (2002). Effect of Transdermal Estradiol and Oral Conjugated Estrogen on C-Reactive Protein in Retinoid-Placebo Trial in Healthy Women. *Circulation* 106: 1224-1228 [\[Abstract\]](#) [\[Full text\]](#)
- Landow, L. (2002). Another Example of Regression to the Mean. *Anesth Analg* 94: 1673-1673 [\[Full text\]](#)
- Vickers, A. J, Altman, D. G (2001). **Statistics Notes:** Analysing controlled trials with baseline and follow up measurements. *BMJ* 323: 1123-1124 [\[Full text\]](#)
- Browne, S. M., Halligan, P. W., Wade, D. T., Taggart, D. P. (1999). COGNITIVE PERFORMANCE AFTER CARDIAC OPERATION: IMPLICATIONS OF REGRESSION TOWARD THE MEAN. *J Thorac Cardiovasc Surg* 117: 481-485 [\[Abstract\]](#) [\[Full text\]](#)
- Richards, M., Hardy, R., Kuh, D., Wadsworth, M. E J (2001). Birth weight and cognitive function

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

in the British 1946 birth cohort: longitudinal population based study. *BMJ* 322: 199-203

[\[Abstract\]](#) [\[Full text\]](#)

- Doepfmer, S., Guggenmoos-Holzmann, I. (1996). Plasma cholesterol response to dietary saturated fat. *BMJ* 312: 511c-512 [\[Full text\]](#)
- Greco, L., Power, C., Peckham, C. (1995). Adult outcome of normal children who are short or underweight at age 7 years. *BMJ* 310: 696-700 [\[Abstract\]](#) [\[Full text\]](#)
- Sharp, S. J, Thompson, S. G, Altman, D. G (1996). The relation between treatment benefit and underlying risk in meta-analysis. *BMJ* 313: 735-738 [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1994;309:996 (15 October)

General practice

Statistics Notes: 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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D G](#) || [Bland, J M](#)
- ▶ Alert me when:
[New articles cite this article](#)

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

1. Cole TJ. *Do growth charts need a face lift?* *BMJ* 1994;308:641-2.
2. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1991:419-26.
3. Campbell MJ, Gardner MJ. Calculating confidence intervals for some non parametric analyses. In: Gardner MJ, Altman DG, eds. *Statistics with confidence*. London: British Medical Journal, 1989:71-9.

This article has been cited by other articles:

- Foster, J. H., Marshall, E. J., Peters, T. J. (2000). OUTCOME AFTER IN-PATIENT DETOXIFICATION FOR ALCOHOL DEPENDENCE: A NATURALISTIC COMPARISON OF 7 VERSUS 28 DAYS STAY. *Alcohol Alcohol*. 35: 580-586 [[Abstract](#)] [[Full text](#)]
- Altman, D. G, Bland, J M. (1996). **Statistics Notes**: Detecting skewness from summary information. *BMJ* 313: 1200-1200 [[Full text](#)]

BMJ 1994;309:1128 (29 October)

General practice

Statistics notes: Matching

J M Bland, D G Altman

Correspondence to: Mr Bland

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

In many medical studies a group of cases, people with a disease under investigation, are compared with a group of controls, people who do not have the disease but who are thought to be comparable in other respects. This happens in epidemiological case-control studies, where a possible risk factor is compared between cases and controls to investigate the nature of the disease. In both types of study cases and controls are sometimes matches. This means that for every case there is a control who has the same (or closely similar) values of the matching variables. Matching may be by sex, age to within five years, ethnic group, etc. Sometimes there are two or more such controls for each case.

We match to ensure that controls and cases are similar in variables which may be related to the variable we are studying but are not of interest in themselves. For example, in many epidemiological case-control studies age is an important predictor of exposure to the risk factor under investigation. There are strong cohort effects in variables such as cigarette smoking and diet. If we do not take age into account we may get spurious differences between cases and controls because, for example, cases are older than controls. Matching ensures that any difference between cases and controls cannot be a result of differences in the matching variables. However, we cannot then examine the effects of the matching variables.

Sometimes matching is ignored in the analysis of the data. If the matching variables are important, this is inefficient. Matching variables, such as age and sex, may be strongly related to the variable of interest. If we allow for the matching in the analysis the variation due to these variables is removed. If we ignore the matching the variability which is related to the variation and may obscure important differences. For example, if we compare the mean blood pressure of subjects with a disease to that of their age matched controls, the variability in blood pressure which is associated with its increase with age will be part of the residual variance and will increase the standard error of the difference between the means. Instead, we

should use the differences between individual matched cases and their controls. Appropriate simple methods include the paired t test for means, McNemar's test for proportions, and the sign test for ordinal data. Sometimes there is no suitable method of matched analysis, as in survival analysis. We can usually adjust for the matching variables, however.

It is desirable to adjust for matching when this was done to make the groups comparable for believed prognostic or confounding variables. This should be done even if in the sample the variable is not significantly prognostic or confounding. By contrast, matching is sometimes merely a convenient method of drawing the sample. For example, in studying cot deaths we might take as a control the next birth in the same hospital. This is sometimes referred to as cosmetic matching. We can ignore the matching in the analysis of such studies.

There are disadvantages to matching. If we match we can only use cases for whom we have matching controls. The more variables we match on the more difficult it may be to find such controls. Even to match on age, sex, and ethnic group we need a large population of potential controls from which to draw. A practical difficulty with matched pairs is that if we want to adjust for other, non-matched, variables the analysis required is more complex than ordinary multiple or logistic regression.

In a large study with many variables it is easier to take an unmatched control group and adjust in the analysis for the variables on which we would have matched, using ordinary regression methods.

Matching is particularly useful in small studies, where we might not have sufficient subjects to adjust for several variables at once.

Some authors use "matched" to mean that the two groups are similar in the distribution of the matching variables, but not that there is individual matching of each case to his or her own control. Such studies should not be described as matched.

This article has been cited by other articles:

- Sorensen, H. T., Gillman, M. W (1995). Matching in case-control studies. *BMJ* 310: 329d-330 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

BMJ 1995;310:170 (21 January)

Statistics notes

Multiple significance tests: the Bonferroni method

J Martin Bland, *reader in medical statistics*,^a **Douglas G Altman**, *head*^b

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

Many published papers include large numbers of significance tests. These may be difficult to interpret because if we go on testing long enough we will inevitably find something which is "significant." We must beware of attaching too much importance to a lone significant result among a mass of non-significant ones. It may be the one in 20 which we expect by chance alone.

Lee et al simulated a clinical trial of the treatment of coronary artery disease by allocating 1073 patient records from past cases into two "treatment" groups at random.¹ They then analysed the outcome as if it were a genuine trial of two treatments. The analysis was quite detailed and thorough. As we would expect, it failed to show any significant difference in survival between those patients allocated to the two treatments. Patients were then subdivided by two variables which affect prognosis, the number of diseased coronary vessels and whether the left ventricular contraction pattern was normal or abnormal. A significant difference in survival between the two "treatment" groups was found in those patients with three diseased vessels (the maximum) and abnormal ventricular contraction. As this would be the subset of patients with the worst prognosis, the finding would be easy to account for by saying that the superior "treatment" had its greatest advantage in the most severely ill patients! This approach to the comparison of subgroups is clearly flawed.

Why does this happen? If we test a null hypothesis which is in fact true, using 0.05 as the critical significance level, we have a probability of 0.95 of coming to a not significant--that is, correct--conclusion. If we test two independent true null hypotheses, the probability that neither test will be significant is $0.95 \times 0.95 = 0.90$. If we test 20 such hypotheses the probability that none will be significant is $0.95^{20} = 0.36$. This gives a probability of $1 - 0.36 = 0.64$ of getting at least one significant result--we are

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

more likely to get one than not. The expected number of spurious significant results is $20 \times 0.05 = 1$. In general, if we have (k) independent significant tests at the (α) level of null hypotheses which are all true, the probability that we will get no significant differences is $(1 - (\alpha))^k$. If we make (α) small enough we can make the probability that none of the separate tests is significant equal to 0.95. Then if any of the (k) tests has a P value less than (α) we will have a significant difference between the treatments at the 0.05 level. Since (α) will be very small, it can be shown that $(1 - (\alpha))^k \approx 1 - k(\alpha)$. If we put $k(\alpha) = 0.05$, so $(\alpha) = 0.05/k$, we will have probability 0.05 that one of the (k) tests will have a P value less than (α) if the null hypotheses are true. Thus, if in a clinical trial we compare two treatments within five subsets of patients the treatments will be significantly different at the 0.05 level if there is a P value less than 0.01 within any of the subsets. This is the Bonferroni method. Note that they are not significant at the 0.01 level, but at only the 0.05 level.

We can do the same thing by multiplying the observed P value from the significance tests by the number of tests, $(k)P$, any $(k)P$ which exceeds one being ignored. Then if any $(k)P$ is less than 0.05 the two treatments are significant at the 0.05 level.

Williams et al randomly allocated elderly patients discharged from hospital to two groups.² There were no significant differences overall between the intervention and control groups, but among women aged 75-79 living alone the control group showed significantly greater deterioration in physical score than did the intervention group ($P=0.04$), and among men aged over 80 the control group showed significantly greater deterioration in disability score than did the intervention group ($P=0.03$). Subjects were cross classified by age groups, whether living alone, and sex, so there were at least eight subgroups and three different measurement scales. Even if we considered the scales separately the corrected P values are $8 \times 0.04 = 0.32$ and $8 \times 0.03 = 0.24$.

A similar problem arises if we have multiple outcome measurements, where the tests will not in general be independent. Newnham et al randomised pregnant women to receive either standard care or a series of Doppler ultrasound blood flow measurements.³ They found a significantly higher proportion of birth weights in the Doppler group below the 10th and 3rd centiles ($P=0.006$ and $P=0.02$). Birth weight was not the primary outcome variable for the trial. These were only two of many comparisons and one suspects that there might be some spurious significant differences among so many. At least 35 tests were reported in the paper. These tests are not independent because they are all on the same subjects, using variables which may not be independent. The proportions of birth weights below the 10th and 3rd centiles are clearly not independent, for example. The probability that two correlated variables both give non-significant differences when the null hypothesis is true is now greater than $(1 - (\alpha))^2$, because if the first test is not significant the second has a probability greater than $1 - (\alpha)$ of also being not significant. A P value less than (α) for any variable, or $(k)P < 0.05$, would still mean that the treatments were significantly different. The overall P value is actually smaller than the nominal 0.05--by an unknown amount which depends on the lack of independence between the tests. The power of the test,

its ability to detect true differences in the population, is correspondingly diminished. In statistical terms, the test is conservative.

For the example, we have $(\alpha)=0.05/35=0.0014$, and so by the Bonferroni criterion the treatment groups are not significantly different. Alternatively, the P values could be adjusted to give $35 \times 0.006=0.21$ and $35 \times 0.02=0.70$.

Other multiple testing problems arise when we have more than two groups of subjects and wish to compare each pair of groups; when we have a series of observations over time, such as blood pressure every 15 minutes after administration of a drug, where there may be a temptation to test each time point separately; and when we have relations between many variables to examine, as in a survey. For all these problems the multiple tests are highly correlated and the Bonferroni method is inappropriate, as it will be highly conservative and may miss real differences. We shall deal with these types of analysis in separate notes.

1. Lee KL, McNeer JF, Starmer FC, Harris PJ, Rosati RA. Clinical judgements and statistics: lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980;61:508-15. [\[Abstract\]](#)
2. Williams EI, Greenwell J, Groom LM. The care of people over 75 years old after discharge from hospital: an evaluation of timetabled visiting by health visitor assistants. *J Pub Hlth Med* 1992;14:138-44.
3. Newnham JP, Evans SF, Con AM, Stanley F J, Landau LI. Effects of frequent ultrasound during pregnancy: a randomized controlled trial. *Lancet* 1993;342:887-91. [\[Medline\]](#)

This article has been cited by other articles:

- Johnson, A. H., Peacock, J. L., Greenough, A., Marlow, N., Limb, E. S., Marston, L., Calvert, S. A., the United Kingdom Oscillation Study Group, (2002). High-Frequency Oscillatory Ventilation for the Prevention of Chronic Lung Disease of Prematurity. *N Engl J Med* 347: 633-642 [\[Abstract\]](#) [\[Full text\]](#)
- Saraux, A, Maillefert, J F, Fautrel, B, Flipo, R M, Kaye, O, Lafforgue, P, Guillemin, F, Botton, E (2002). Laboratory and imaging studies used by French rheumatologists to determine the cause of recent onset polyarthritis without extra-articular manifestations. *Ann Rheum Dis* 61: 626-629 [\[Abstract\]](#) [\[Full text\]](#)
- Weets, I., Siraux, V., Daubresse, J.-C., De Leeuw, I. H., Fery, F., Keymeulen, B., Krzentowski,

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

- G., Letiexhe, M., Mathieu, C., Nobels, F., Rottiers, R., Scheen, A., Van Gaal, L., Schuit, F. C., Van Der Auwera, B., Rui, M., De Pauw, P., Kaufman, L., Gorus, F. K. (2002). Relation between Disease Phenotype and HLA-DQ Genotype in Diabetic Patients Diagnosed in Early Adulthood. *J Clin Endocrinol Metab* 87: 2597-2605 [[Abstract](#)] [[Full text](#)]
- Allison, D. B., Coffey, C. S. (2002). Two-Stage Testing in Microarray Analysis: What Is Gained?. *J Gerontol A Biol Sci Med Sci* 57: B189-192 [[Abstract](#)] [[Full text](#)]
 - Balkestein, E. J., Staessen, J. A., Wang, J.-G., van der Heijden-Spek, J. J., Van Bortel, L. M., Barlassina, C., Bianchi, G., Brand, E., Herrmann, S.-M., Struijker-Boudier, H. A. (2001). Carotid and Femoral Artery Stiffness in Relation to Three Candidate Genes in a White Population. *Hypertension* 38: 1190-1197 [[Abstract](#)] [[Full text](#)]
 - Flaherty, J. H., Takahashi, R., Teoh, , Kim, J.-I., Habib, S., Ito, M., Matsushita, S. (2001). Use of Alternative Therapies in Older Outpatients in the United States and Japan: Prevalence, Reporting Patterns, and Perceived Effectiveness. *J Gerontol A Biol Sci Med Sci* 56: M650-655 [[Abstract](#)] [[Full text](#)]
 - Walker-Batson, D., Curtis, S., Natarajan, R., Ford, J., Dronkers, N., Salmeron, E., Lai, J., Unwin, D. H., Feeney, D. M. (2001). A Double-Blind, Placebo-Controlled Study of the Use of Amphetamine in the Treatment of Aphasia Editorial Comment. *Stroke* 32: 2093-2098 [[Abstract](#)] [[Full text](#)]
 - Jung, H., Wensing, M, de Wilt, A, Olesen, F, Grol, R (2000). Comparison of patients' preferences and evaluations regarding aspects of general practice care. *Fam. Pract.* 17: 236-242 [[Abstract](#)] [[Full text](#)]
 - Robinson, T., Potter, J. (1997). Cardiopulmonary and Arterial Baroreflex-Mediated Control of Forearm Vasomotor Tone Is Impaired After Acute Stroke. *Stroke* 28: 2357-2362 [[Abstract](#)] [[Full text](#)]
 - Tyrberg, B., Ustinov, J., Otonkoski, T., Andersson, A. (2001). Stimulated Endocrine Cell Proliferation and Differentiation in Transplanted Human Pancreatic Islets: Effects of the ob Gene and Compensatory Growth of the Implantation Organ. *Diabetes* 50: 301-307 [[Abstract](#)] [[Full text](#)]
 - O'Dwyer, P. J., Manola, J., Valone, F. H., Ryan, L. M., Hines, J. D., Wadler, S., Haller, D. G., Arbuck, S. G., Weiner, L. M., Mayer, R. J., Benson, A. B. III (2001). Fluorouracil Modulation in Colorectal Cancer: Lack of Improvement With N -Phosphonoacetyl- 1 -Aspartic Acid or Oral Leucovorin or Interferon, But Enhanced Therapeutic Index With Weekly 24-Hour Infusion Schedule--An Eastern Cooperative Oncology Group/Cancer and Leukemia Group B Study. *J Clin Oncol* 19: 2413-2421 [[Abstract](#)] [[Full text](#)]
 - Fried, P. W., Wechsler, A. S. (2001). How to get your paper published. *J Thorac Cardiovasc Surg* 121: S3-7 [[Abstract](#)] [[Full text](#)]
 - Ye, J., Yang, L., Del Bigio, M. R., Summers, R., Jackson, D., Somorjai, R. L., Salerno, T. A., Deslauriers, R. (1997). RETROGRADE CEREBRAL PERFUSION PROVIDES LIMITED DISTRIBUTION OF BLOOD TO THE BRAIN: A STUDY IN PIGS. *J Thorac Cardiovasc Surg* 114: 660-665 [[Abstract](#)] [[Full text](#)]
 - Jeffery, K. J. M., Siddiqui, A. A., Bunce, M., Lloyd, A. L., Vine, A. M., Witkover, A. D., Izumo, S., Usuku, K., Welsh, K. I., Osame, M., Bangham, C. R. M. (2000). The Influence of HLA Class I

Alleles and Heterozygosity on the Outcome of Human T Cell Lymphotropic Virus Type I Infection. *The JJ* 165: 7278-7284 [\[Abstract\]](#) [\[Full text\]](#)

- Signoretti, S., Montironi, R., Manola, J., Altimari, A., Tam, C., Bubley, G., Balk, S., Thomas, G., Kaplan, I., Hlatky, L., Hahnfeldt, P., Kantoff, P., Loda, M. (2000). Her-2-neu Expression and Progression Toward Androgen Independence in Human Prostate Cancer. *J Natl Cancer Inst* 92: 1918-1925 [\[Abstract\]](#) [\[Full text\]](#)
- Jacobson, L., Zurakowski, D., Majzoub, J. A. (1997). Protein Malnutrition Increases Plasma Adrenocorticotropin and Anterior Pituitary Proopiomelanocortin Messenger Ribonucleic Acid in the Rat. *Endocrinology* 138: 1048-1057 [\[Abstract\]](#) [\[Full text\]](#)
- CUESTA, M. J., PERALTA, V., ZARZUELA, A. (2000). Reappraising insight in psychosis: Multi-scale longitudinal study. *Br J Psychiatry* 177: 233-240 [\[Abstract\]](#) [\[Full text\]](#)
- Butland, B. K, Fehily, A. M, Elwood, P. C (2000). Diet, lung function, and lung function decline in a cohort of 2512 middle aged men. *Thorax* 55: 102-108 [\[Abstract\]](#) [\[Full text\]](#)
- Herman, N. L., Choi, K. C., Affleck, P. J., Calicott, R., Brackin, R., Singhal, A., Andreasen, A., Gadalla, F., Fong, J., Gomillion, M. C., Hartman, J. K., Koff, H. D., Lee, S. H. R., Decar, T. K. V. (1999). Analgesia, Pruritus, and Ventilation Exhibit a Dose-Response Relationship in Parturients Receiving Intrathecal Fentanyl During Labor. *Anesth Analg* 89: 378-378 [\[Abstract\]](#) [\[Full text\]](#)
- Camp, S. J., Stevenson, V. L., Thompson, A. J., Miller, D. H., Borrás, C., Auriacombe, S., Brochet, B., Falautano, M., Filippi, M., Hérissé-Dulo, L., Montalban, X., Parricira, E., Polman, C. H., De Sa, J., Langdon, D. W. (1999). Cognitive function in primary progressive and transitional progressive multiple sclerosis: A controlled study with MRI correlates. *Brain* 122: 1341-1348 [\[Abstract\]](#) [\[Full text\]](#)
- Voss, S., George, S. (1995). Multiple significance tests. *BMJ* 310: 1073-1073 [\[Full text\]](#)
- Jeffery, K. J. M., Usuku, K., Hall, S. E., Matsumoto, W., Taylor, G. P., Procter, J., Bunce, M., Ogg, G. S., Welsh, K. I., Weber, J. N., Lloyd, A. L., Nowak, M. A., Nagai, M., Kodama, D., Izumo, S., Osame, M., Bangham, C. R. M. (1999). HLA alleles determine human T-lymphotropic virus-I (HTLV-I) proviral load and the risk of HTLV-I-associated myelopathy. *Proc. Natl. Acad. Sci. U. S. A.* 96: 3848-3853 [\[Abstract\]](#) [\[Full text\]](#)
- Fossum, E., Hoieggen, A., Moan, A., Rostrup, M., Kjeldsen, S. E. (1999). Insulin Sensitivity Is Related to Physical Fitness and Exercise Blood Pressure to Structural Vascular Properties in Young Men. *Hypertension* 33: 781-786 [\[Abstract\]](#) [\[Full text\]](#)
- Bender, R., Lange, S. (1999). Multiple test procedures other than Bonferroni's deserve wider use. *BMJ* 318: 600a-600 [\[Full text\]](#)
- Altman, D. G, Bland, J M. (1996). **Statistics Notes:** Comparing several groups using analysis of variance. *BMJ* 312: 1472-1473 [\[Full text\]](#)
- Altman, D. G, Matthews, J. N S (1996). **Statistics Notes:** Interaction 1: heterogeneity of effects. *BMJ* 313: 486-486 [\[Full text\]](#)
- Fossum, E., Hoieggen, A., Moan, A., Rostrup, M., Nordby, G., Kjeldsen, S. E. (1998). Relationship Between Insulin Sensitivity and Maximal Forearm Blood Flow in Young Men. *Hypertension* 32: 838-843 [\[Abstract\]](#) [\[Full text\]](#)
- Zaman, M. M., Ikemoto, S., Yoshiike, N., Date, C., Yokoyama, T., Tanaka, H. (1997).

Association of Apolipoprotein Genetic Polymorphisms With Plasma Cholesterol in a Japanese Rural Population : The Shibata Study. *Arterioscler Thromb* 17: 3495-3504 [[Abstract](#)] [[Full text](#)]

- Nussbaum, A. K., Dick, T. P., Keilholz, W., Schirle, M., Dietz, K., Heinemeyer, W., Groll, M., Wolf, D. H., Huber, R., Rammensee, H.-G., Schild, H. (1998). Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1. *Proc. Natl. Acad. Sci. U. S. A.* 95: 12504-12509 [[Abstract](#)] [[Full text](#)]
- Perneger, T. V (1998). What's wrong with Bonferroni adjustments. *BMJ* 316: 1236-1238 [[Full text](#)]

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1995;310:298 (4 February)

Statistics notes

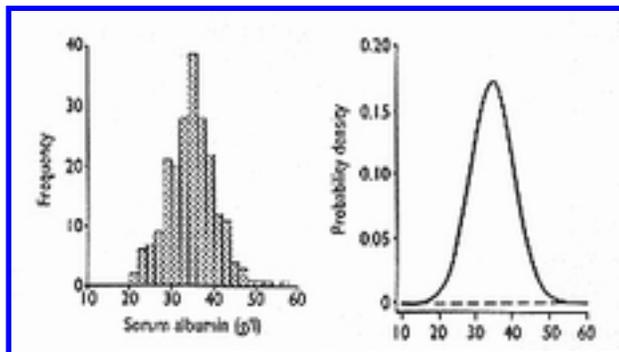
The normal distribution

Douglas G Altman, *Head*,^a **J Martin Bland**, *reader in medical statistics*^b

^a Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX, ^b Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

Correspondence to: Mr Altman.

When we measure a quantity in a large number of individuals we call the pattern of values obtained a distribution. For example, figure 1 shows the distribution of serum albumin concentration in a sample of adults displayed as a histogram. This is an empirical distribution. There are also theoretical distributions, of which the best known is the normal distribution (sometimes called the Gaussian distribution), which is shown in figure 2. Although widely referred to in statistics, the normal distribution remains a mysterious concept to many. Here we try to explain what it is and why it is important.



View larger version (15K):

[\[in this window\]](#)

[\[in a new window\]](#)

FIG 1 (left)--Serum albumin values in 248 adults
FIG 2 (right)--Normal distribution with the same mean and standard deviation as the serum albumin values

- ▶ [Email this article to a friend](#)
- ▶ **[Respond to this article](#)**
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

In this context the name "normal" causes much confusion. In statistics it is just a name; statisticians often use a capital N to emphasise this and to clarify that Normality does not necessarily imply normality. Indeed, in some medical specialties normal distributions are rare.

Various methods of analysis make assumptions about normality, including correlation, regression, t tests, and analysis of variance. It is not in fact necessary for the distribution of the observed data to be normal, but rather the sample values should be compatible with the population (which they represent) having a normal distribution. Indeed, samples from a population in which the true distribution is normal will not necessarily look normal themselves, especially if the sample is small. Figure 3 shows the distributions of samples of different sizes drawn at random from normal distributions--few of the small samples look like a normal distribution, but the similarity increases as the sample size increases.

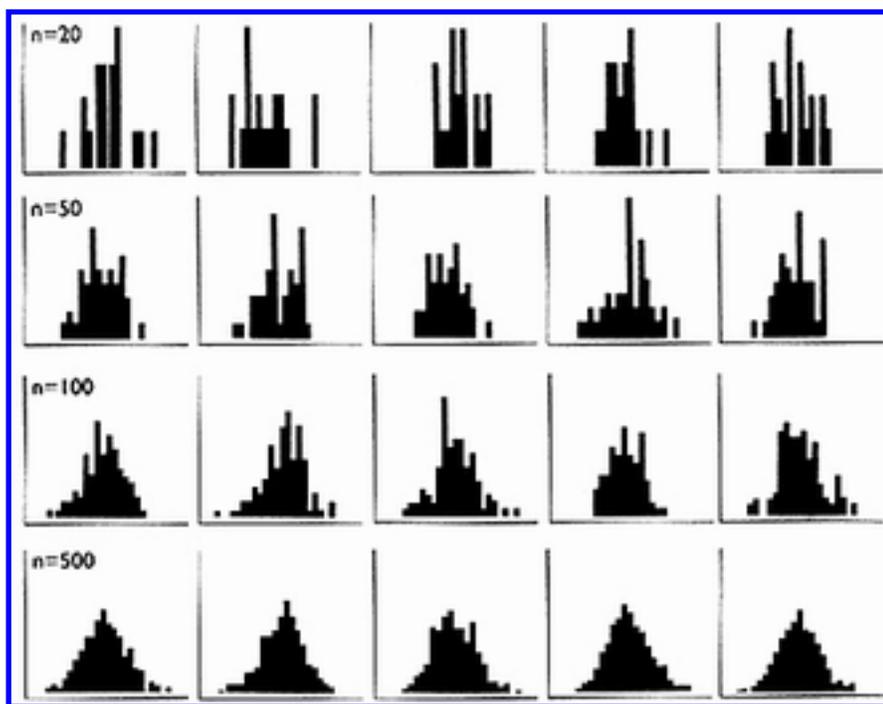


FIG 3--Random samples from normal distributions--five samples of size 20, 50, 100, and 500

View larger version (35K):

[\[in this window\]](#)

[\[in a new window\]](#)

Although some statistical methods, such as the t test, are not sensitive to moderate departures from normality, it is generally preferable not to rely on this feature. Visual inspection of the distribution may suggest whether the assumption of normality is reasonable but, as figure 3 suggests, this approach is unreliable. Significance tests and normal plots can be used to assess formally whether sample data are a plausible sample from a normal population.¹ When data do not have a normal distribution we can either

transform the data (for example, by taking logarithms) or use a method that does not require the data to be normally distributed. We consider these topics in future notes.

The normal distribution has another essential place in statistics. Just as separate samples selected at random from the same population will differ (fig 3), so will calculated statistics such as the mean blood pressure. We can think of the means from many samples as themselves also having a distribution. A key theoretical result, called the central limit theorem, underpins many methods of analysis. It states that the means of random samples from any distribution will themselves have a normal distribution. As a consequence, when we have samples of hundreds of observations we can often ignore the distribution of the data. Nevertheless, because most clinical studies are of a modest size, it is usually advisable to transform non-normal data, especially when they have a skewed distribution.

We can consider binary attributes in the same way. For example, the proportions of individuals with asthma will vary from sample to sample. If having asthma is represented by the value 1 and not having asthma by the value 0 then the mean of these values in the sample is the proportion of individuals with asthma. Thus a proportion is also a mean and will follow a normal distribution. These methods are not valid in small samples--some "exact" methods can be used.² Similar comments apply to some other statistics, such as regression coefficients or standardised mortality ratios, but for mortality ratios the sample size may have to be very large indeed.

One of the most important applications of these results is in calculating confidence intervals. The general method is based on the idea that the statistic of interest (such as the difference between two means or proportions) would have a normal distribution in repeated samples.³

1. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991:132-45.
2. Gardner MJ, Altman DG. Calculating confidence intervals for proportions and their differences. In: Gardner MJ, Altman DG, eds. Statistics with confidence. London: British Medical Journal, 1989:28-33
3. Gardner MJ, Altman DG, eds. Statistics with confidence. London: British Medical Journal, 1989:17.

This article has been cited by other articles:

- Altman, D. G, Bland, J M. (1999). **Statistics notes** Variables and parameters. *BMJ* 318: 1667-1667 [\[Full text\]](#)
- Bland, J M., Altman, D. G (1996). **Statistics Notes**: Transforming data. *BMJ* 312: 770-770 [\[Full text\]](#)

- Altman, D. G, Bland, J M. (1996). **Statistics Notes**: Detecting skewness from summary information. *BMJ* 313: 1200-1200
[\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

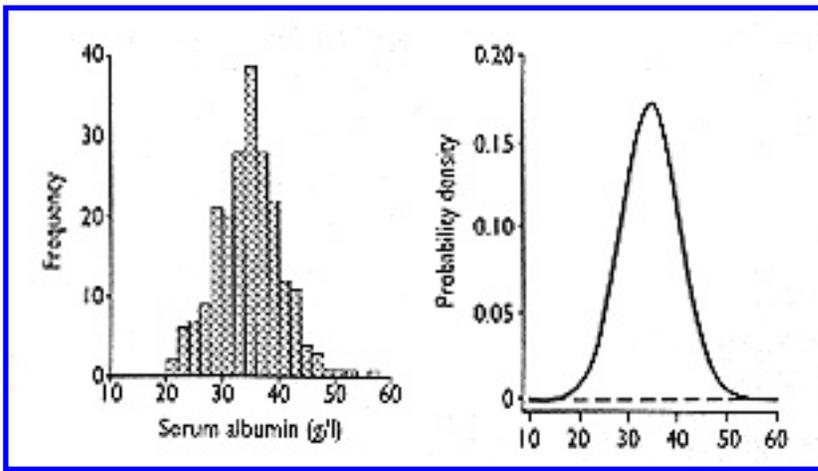


FIG 1 (left)--Serum albumin values in 248 adults FIG 2 (right)--Normal distribution with the same mean and standard deviation as the serum albumin values

[\[View larger version \(16K\)\]](#)

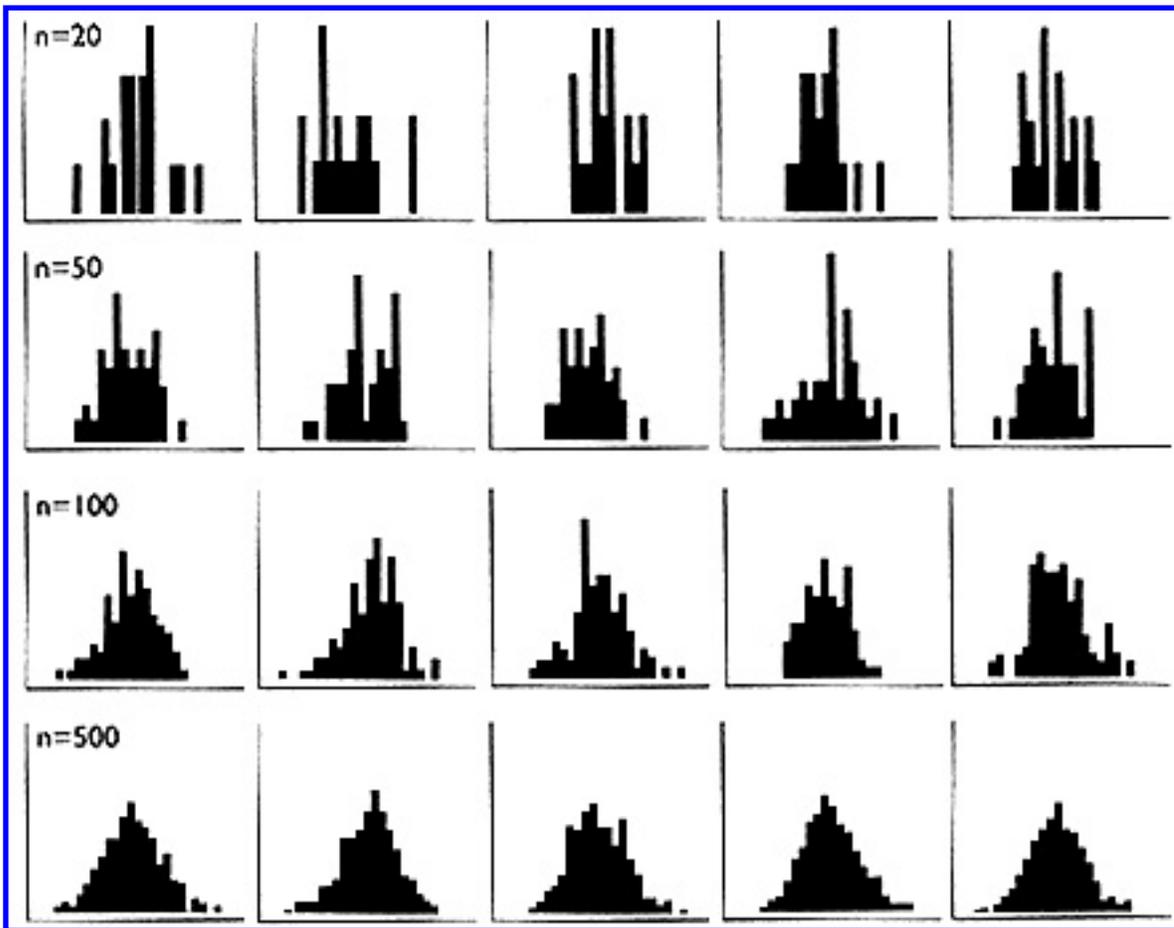


FIG 3--Random samples from normal distributions--five samples of size 20, 50, 100, and 500

[\[View larger version \(39K\)\]](#)

BMJ 1995;310:446 (18 February)

Statistics notes

Calculating correlation coefficients with repeated observations: Part 1-- correlation within subjects

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

J Martin Bland, *reader in medical statistics*,^a **Douglas G Altman**, *head*^b

^a Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE, ^b Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX

Correspondence to: Dr Bland.

In an earlier Statistics Note¹ we commented on the analysis of paired data where there is more than one observation per subject, as shown in table I. We pointed out that it could be highly misleading to analyse such data by combining repeated observations from several subjects and then calculating the correlation coefficient as if the data were a simple sample. This note is a response to several letters about the appropriate analysis for such data.

TABLE I--Repeated measurements of intramural pH and PaCO₂ for eight subjects²

Subject	pH	PaCO ₂	Subject	pH	PaCO ₂
1	6.68	3.97	5	7.30	4.32
1	6.53	4.12	5	7.37	3.23
1	6.43	4.09	5	7.27	4.46
1	6.33	3.97	5	7.28	4.72
2	6.85	5.27	5	7.32	4.75
2	7.06	5.37	5	7.32	4.99

2	7.13	5.41	6	7.38	4.78
2	7.17	5.44	6	7.30	4.73
3	7.40	5.67	6	7.29	5.12
3	7.42	3.64	6	7.33	4.93
3	7.41	4.32	6	7.31	5.03
3	7.37	4.73	6	7.33	4.93
3	7.34	4.96	7	6.86	6.85
3	7.35	5.04	7	6.94	6.44
3	7.28	5.22	7	6.92	6.52
3	7.30	4.82	8	7.19	5.28
3	7.34	5.07	8	7.29	4.56
4	7.36	5.67	8	7.21	4.34
4	7.33	5.10	8	7.25	4.32
4	7.29	5.53	8	7.20	4.41
4	7.30	4.75	8	7.19	3.69
4	7.35	5.51	8	6.77	6.09
5	7.35	4.28	8	6.82	5.58
5	7.30	4.44			

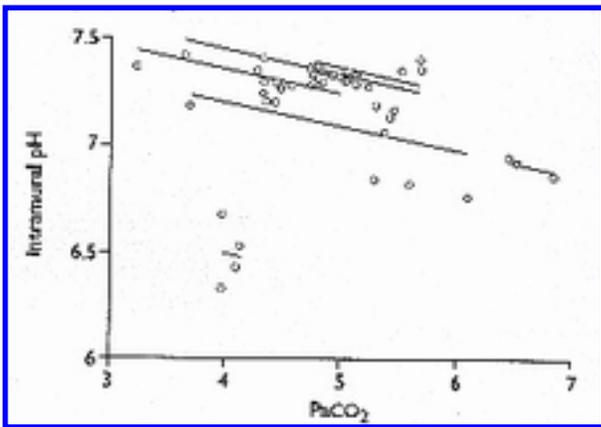
The choice of analysis for the data in table I depends on the question we want to answer. If we want to know whether subjects with high values of intramural pH also tend to have high values of PaCO₂ we are interested in whether the average pH for a subject is related to the subject's average PaCO₂. We can use the correlation between the subject means, which we shall describe in a subsequent note. If we want to know whether an increase in pH within the individual was associated with an increase in PaCO₂ we want to remove the differences between subjects and look only at changes within.

To look at variation within the subject we can use multiple regression. We make one of our variables, pH or PaCO₂, the outcome variable and the other variable and the subject the predictor variables. Subject is treated as a categorical factor using dummy variables^{3 4} and so has seven degrees of freedom. We use the analysis of variance table^{3 4} for the regression (table II), which shows how the variability in pH can be partitioned into components due to different sources. This method is also known as analysis of covariance and is equivalent to fitting parallel lines through each subject's data (see figure). The residual sum of squares in table II represents the variation about these lines. We remove the variation due to subjects (and any other nuisance variables which might be present) and express the variation in pH due to PaCO₂ as a proportion of what's left: (Sum of squares for PaCO₂)/(Sum of squares for PaCO₂ + residual sum of squares) The magnitude of the correlation coefficient within subjects is the square root of this proportion. For table II this is: (square root) 0.1153/0.1153+0.3337 = 0.51 The sign of the correlation coefficient is given by the sign of the regression coefficient for PaCO₂. Here the regression slope is -0.108, so the correlation coefficient within subjects is -0.51. The P value is found either from the F test in the associated analysis of variance table, or from the t test for the regression slope. It doesn't matter which

variable we regress on which; we get the same correlation coefficient and P value either way.

TABLE II--Analysis of variance for the data in table I

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability
Subjects	7	2.9661	0.4237	48.3	<0.0001
PaCO ₂	1	0.1153	0.1153	13.1	0.0008
Residual	38	0.3337	0.0088		
Total	46	3.3139	0.0720		



pH against PaCO₂ for eight subjects, with parallel lines fitted for each subject

View larger version (12K):

[\[in this window\]](#)

[\[in a new window\]](#)

If we incorrectly calculate the correlation coefficient ignoring the fact that we have 47 observations on only 8 subjects, we get -0.07, P=0.7. Hence the correct analysis within subjects reveals a relation which the incorrect analysis misses.

1. Bland JM, Altman DG. Correlation, regression, and repeated data. *BMJ* 1994;308:896. [\[Full Text\]](#)
2. Boyd O, Mackay CJ, Lamb G, Bland JM, Grounds RM, Bennett ED. Comparison of clinical information gained from routine blood-gas analysis and from gastric tonometry for intramural pH.

Lancet 1993;341:142-6. [[Medline](#)]

3. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991.
4. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford: Blackwell, 1994.

This article has been cited by other articles:

- Chambers, D C, Ayres, J G (2001). Effect of nebulised L- and D-arginine on exhaled nitric oxide in steroid naive asthma. *Thorax* 56: 602-606 [[Abstract](#)] [[Full text](#)]
- Schultz, C. J., Neil, H. A. W., Dalton, R. N., Bahu, T. K., Dunger, D. B. (2001). Blood Pressure Does Not Rise Before the Onset of Microalbuminuria in Children Followed From Diagnosis of Type 1 Diabetes. *Diabetes Care* 24: 555-560 [[Abstract](#)] [[Full text](#)]
- Ahmed, M. L., Ong, K. K. L., Watts, A. P., Morrell, D. J., Preece, M. A., Dunger, D. B. (2001). Elevated Leptin Levels Are Associated with Excess Gains in Fat Mass in Girls, But Not Boys, with Type 1 Diabetes: Longitudinal Study during Adolescence. *J Clin Endocrinol Metab* 86: 1188-1193 [[Abstract](#)] [[Full text](#)]
- De Marinis, L., Mancini, A., Valle, D., Bianchi, A., Milardi, D., Proto, A., Lanzone, A., Tacchino, R. (1999). Plasma Leptin Levels after Biliopancreatic Diversion: Dissociation with Body Mass Index. *J Clin Endocrinol Metab* 84: 2386-2389 [[Abstract](#)] [[Full text](#)]
- Ahmed, M. L., Ong, K. K. L., Morrell, D. J., Cox, L., Drayer, N., Perry, L., Preece, M. A., Dunger, D. B. (1999). Longitudinal Study of Leptin Concentrations during Puberty: Sex Differences and Relationship to Changes in Body Composition. *J Clin Endocrinol Metab* 84: 899-905 [[Abstract](#)] [[Full text](#)]
- Subhedar, N V, Shaw, N J (2000). Changes in pulmonary arterial pressure in preterm infants with chronic lung disease. *Arch. Dis. Child. Fetal Neonatal Ed.* 82: 243F-247 [[Abstract](#)] [[Full text](#)]
- Lovell, A. T., Marshall, A. C., Elwell, C. E., Smith, M., Goldstone, J. C. (2000). Changes in Cerebral Blood Volume with Changes in Position in Awake and Anesthetized Subjects. *Anesth Analg* 90: 372-372 [[Abstract](#)] [[Full text](#)]
- CUTTITTA, G., CIBELLA, F., VISCONTI, A., SCICILONE, N., BELLIA, V., BONSIGNORE, G. (2000). Spontaneous Gastroesophageal Reflux and Airway Patency during the Night in Adult Asthmatics. *Am J Respir Crit Care Med* 161: 177-181 [[Abstract](#)] [[Full text](#)]
- Booth, S. L, O'Brien-Morse, M. E, Dallal, G. E, Davidson, K. W, Gundberg, C. M (1999). Response of vitamin K status to different intakes and sources of phyloquinone-rich foods: comparison of younger and older adults. *Am. J. Clin. Nutr.* 70: 368-377 [[Abstract](#)] [[Full text](#)]
- Altman, D. G, Bland, J M. (1997). **Statistics notes:** Units of analysis. *BMJ* 314: 1874-1874 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

- Bland, J M., Altman, D. G (1995). **Statistics notes:** Calculating correlation coefficients with repeated observations: Part 2--correlation between subjects. *BMJ* 310: 633-633 [\[Full text\]](#)
- Lunn, P. G., Erinoso, H. O., Northrop-Clewes, C. A., Boyce, S. A. (1999). Giardia intestinalis Is Unlikely To Be a Major Cause of the Poor Growth of Rural Gambian Infants. *J. Nutr.* 129: 872-877 [\[Abstract\]](#) [\[Full text\]](#)
- Subhedar, N V, Shaw, N J (1997). Changes in oxygenation and pulmonary haemodynamics in preterm infants treated with inhaled nitric oxide. *Arch. Dis. Child. Fetal Neonatal Ed.* 77: 191F-197 [\[Abstract\]](#) [\[Full text\]](#)
- Ludwig, D. S., Majzoub, J. A., Al-Zahrani, A., Dallal, G. E., Blanco, I., Roberts, S. B. (1999). High Glycemic Index Foods, Overeating, and Obesity. *Pediatrics* 103: 26e-26 [\[Abstract\]](#) [\[Full text\]](#)

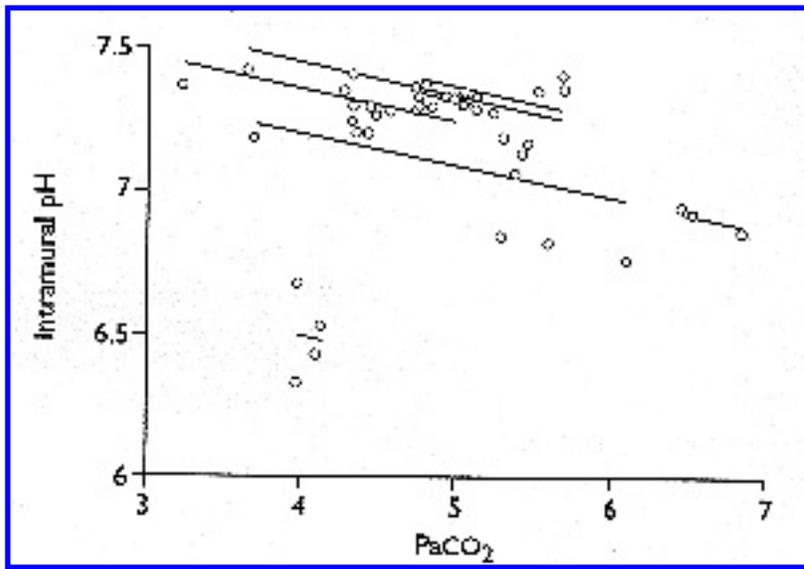
[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



pH against PaCO₂ for eight subjects, with parallel lines fitted for each subject

[\[View larger version \(15K\)\]](#)

BMJ 1995;310:633 (11 March)

Statistics notes

Calculating correlation coefficients with repeated observations: Part 2-- correlation between subjects

J Martin Bland, *reader in medical statistics*,^a **Douglas G Altman**, *head*^b

^a Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE, ^b Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX

Correspondence to: Dr Bland.

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

In earlier **Statistics Notes**¹ ² we commented on the analysis of paired data where there is more than one observation per subject. It can be highly misleading to analyse such data by combining repeated observations from several subjects and then calculating the correlation coefficient as if the data were a simple sample.¹ The appropriate analysis depends on the question we wish to answer. If we want to know whether an increase in one variable within the individual is associated with an increase in the other we can calculate the correlation coefficient within subjects.² If we want to know whether subjects with high values of one variable also tend to have high values of the other we can use the correlation between the subject means, which we shall describe here.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ [See Correction for this article](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

Means of repeated measurements of intramural pH and Paco2 for eight subjects³

Subject	pH	Paco2	Number
1	6.49	4.04	4
2	7.05	5.37	4
3	7.36	4.83	9
4	7.33	5.31	5
5	7.31	4.40	8
6	7.32	4.92	6
7	6.91	6.60	3
8	7.12	4.78	8

The table shows the mean pH and Paco2 for each of eight subjects, with the number of pairs of observations for each. The 47 pairs of measurements from which these means were calculated were given previously.² Here we are interested in whether the average pH for a subject is related to the subject's average Paco2.

We can calculate the usual correlation coefficient for the mean pH and mean Paco2. For the data in the table this gives $r=0.09$, $P=0.8$.

This analysis does not take into account the different numbers of measurements on each subject. Whether this matters depends on how different the numbers of observations are and whether the measurements within subjects vary much compared with the means between subjects. We can calculate a weighted correlation coefficient, using the number of observations as weights. Many computer programs will calculate this, but it is not difficult to do by hand.

We denote the mean pH and Paco2 for subject i by x_i and y_i , the number of observations for subject i by m_i , and the number of subjects by n . It is fairly obvious⁴ that the weighted mean of the x_i is $(\sum m_i x_i) / (\sum m_i)$. In the usual case, where there is one observation per subject, the m_i are all one and this formula gives the usual mean $(\sum x_i) / n$.

An easy way to calculate the weighted correlation coefficient is to replace each individual observation by its subject mean. Thus the table would yield 47 pairs of observations, the first four of which would each be $pH=6.49$ and $Paco2=4.04$, and so on. If we use the usual formula for the correlation coefficient on the expanded data we will get the weighted correlation coefficient. However, we must be careful when it comes to the P value. We have only 8 observations (n in general), not 47. We should ignore any P value printed by our computer program, and use a statistical table instead.

The actual formula for a weighted correlation coefficient is:
$$\frac{(\sum_{i=1}^n m_i x_i y_i) - (\sum_{i=1}^n m_i x_i)(\sum_{i=1}^n m_i y_i) / (\sum_{i=1}^n m_i)}{\sqrt{((\sum_{i=1}^n m_i x_i)^2 / (\sum_{i=1}^n m_i) - (\sum_{i=1}^n m_i x_i^2) / (\sum_{i=1}^n m_i)) ((\sum_{i=1}^n m_i y_i)^2 / (\sum_{i=1}^n m_i) - (\sum_{i=1}^n m_i y_i^2) / (\sum_{i=1}^n m_i))}}$$
 where all summations are from $i=1$ to n . When all the m_i are equal they cancel out, giving the usual formula for a correlation coefficient.

For the data in the table the weighted correlation coefficient is $r=0.08$, $P=0.9$. There is no evidence that subjects with a high pH also have a high Paco^2 . However, as we have already shown,² within the subject a rise in pH was associated with a fall in Paco^2 .

1. Bland JM, Altman DG. Correlation, regression and repeated data. *BMJ* 1994;308:896. [\[Full Text\]](#)
2. Bland JM, Altman DG. Calculating correlation coefficients with repeated observations: Part 1-- correlation within subjects. *BMJ* 1995;310:446. [\[Full Text\]](#)
3. Boyd O, Mackay CJ, Lamb G, Bland JM, Grounds RM, Bennett ED. Comparison of clinical information gained from routine blood-gas analysis and from gastric tonometry for intramural pH. *Lancet* 1993;341:142-6. [\[Medline\]](#)
4. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford: Blackwell, 1994:215.

This article has been cited by other articles:

- Fedorcsak, P., Dale, P. O., Storeng, R., Tanbo, T., Abyholm, T. (2001). The impact of obesity and insulin resistance on the outcome of IVF or ICSI in women with polycystic ovarian syndrome. *Hum Reprod* 16: 1086-1091 [\[Abstract\]](#) [\[Full text\]](#)
- Jialal, I., Stein, D., Balis, D., Grundy, S. M., Adams-Huet, B., Devaraj, S. (2001). Effect of Hydroxymethyl Glutaryl Coenzyme A Reductase Inhibitor Therapy on High Sensitive C-Reactive Protein Levels. *Circulation* 103: 1933-1935 [\[Abstract\]](#) [\[Full text\]](#)
- Duffield, A. J, Thomson, C. D, Hill, K. E, Williams, S. (1999). An estimation of selenium requirements for New Zealanders. *Am. J. Clin. Nutr.* 70: 896-903 [\[Abstract\]](#) [\[Full text\]](#)
- Booth, S. L, O'Brien-Morse, M. E, Dallal, G. E, Davidson, K. W, Gundberg, C. M (1999). Response of vitamin K status to different intakes and sources of phylloquinone-rich foods: comparison of younger and older adults. *Am. J. Clin. Nutr.* 70: 368-377 [\[Abstract\]](#) [\[Full text\]](#)
- Altman, D. G, Bland, J M. (1997). **Statistics notes:** Units of analysis. *BMJ* 314: 1874-1874

- [▶ Email this article to a friend](#)
- [▶ Respond to this article](#)
- [▶ PubMed citation](#)
- [▶ Related articles in PubMed](#)
- [▶ Download to Citation Manager](#)
- [▶ See Correction for this article](#)
- [▶ Search Medline for articles by: Bland, J M. || Altman, D. G](#)
- [▶ Alert me when: New articles cite this article](#)

[\[Full text\]](#)

- Morrish, P K, Rakshi, J S, Bailey, D L, Sawle, G V, Brooks, D J (1998). Measuring the rate of progression and estimating the preclinical period of Parkinson's disease with [18F]dopa PET. *J. Neurol. Neurosurg. Psychiatry* 64: 314-319 [\[Abstract\]](#) [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1995;311:485 (19 August)

Statistics notes

Absence of evidence is not evidence of absence

Douglas G Altman, *head*,^a J Martin Bland, *reader in medical statistics*^b

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

Correspondence to: Mr Altman.

The non-equivalence of statistical significance and clinical importance has long been recognised, but this error of interpretation remains common. Although a significant result in a large study may sometimes not be clinically important, a far greater problem arises from misinterpretation of non-significant findings. By convention a P value greater than 5% ($P > 0.05$) is called "not significant." Randomised controlled clinical trials that do not show a significant difference between the treatments being compared are often called "negative." This term wrongly implies that the study has shown that there is no difference, whereas usually all that has been shown is an absence of evidence of a difference. These are quite different statements.

The sample size of controlled trials is generally inadequate, with a consequent lack of power to detect real, and clinically worthwhile, differences in treatment. Freiman et al¹ found that only 30% of a sample of 71 trials published in the New England Journal of Medicine in 1978-9 with $P > 0.1$ were large enough to have a 90% chance of detecting even a 50% difference in the effectiveness of the treatments being compared, and they found no improvement in a similar sample of trials published in 1988. To interpret all these "negative" trials as providing evidence of the ineffectiveness of new treatments is clearly wrong and foolhardy. The term "negative" should not be used in this context.²

A recent example is given by a trial comparing octreotide and sclerotherapy in patients with variceal bleeding.³ The study was carried out on a sample of only 100 despite a reported calculation that suggested that 1800 patients were needed. This trial had only a 5% chance of getting a statistically significant result if the stated clinically worthwhile treatment difference truly existed. One consequence

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

of such low statistical power was a wide confidence interval for the treatment difference. The authors concluded that the two treatments were equally effective despite a 95% confidence interval that included differences between the cure rates of the two treatments of up to 20 percentage points.

Similar evidence of the dangers of misinterpretation of non-significant results is found in numerous metaanalyses (overviews) of published trials, when few or none of the individual trials were statistically large enough. A dramatic example is provided by the overview of clinical trials evaluating fibrinolytic treatment (mostly streptokinase) for preventing reinfarction after acute myocardial infarction. The overview of randomised controlled trials found a modest but clinically worthwhile (and highly significant) reduction in mortality of 22%,⁴ but only five of the 24 trials had shown a statistically significant effect with $P < 0.05$. The lack of statistical significance of most of the individual trials led to a long delay before the true value of streptokinase was appreciated.

While it is usually reasonable not to accept a new treatment unless there is positive evidence in its favour, when issues of public health are concerned we must question whether the absence of evidence is a valid enough justification for inaction. A recent publicised example is the suggested link between some sudden infant deaths and antimony in cot mattresses. Statements about the absence of evidence are common--for example, in relation to the possible link between violent behaviour and exposure to violence on television and video, the possible harmful effects of pesticide residues in drinking water, the possible link between electromagnetic fields and leukaemia, and the possible transmission of bovine spongiform encephalopathy from cows. Can we be comfortable that the absence of clear evidence in such cases means that there is no risk or only a negligible one?

When we are told that "there is no evidence that A causes B" we should first ask whether absence of evidence means simply that there is no information at all. If there are data we should look for quantification of the association rather than just a P value. Where risks are small P values may well mislead: confidence intervals are likely to be wide, indicating considerable uncertainty. While we can never prove the absence of a relation, when necessary we should seek evidence against the link between A and B--for example, from case-control studies. The importance of carrying out such studies will relate to the seriousness of the postulated effect and how widespread is the exposure in the population.

1. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: survey of two sets of "negative" trials. In: Bailar JC, Mosteller F, eds. *Medical uses of statistics*. 2nd ed. Boston, MA: NEJM Books, 1992:357-73.
2. Chalmers I. Proposal to outlaw the term "negative trial." *BMJ* 1985;290:1002.
3. Sung JY, Chung SCS, Lai C-W, Chan FKL, Leung JWC, Yung M-L, Kassianides C, et al. Octreotide infusion or emergency sclerotherapy for variceal haemorrhage. *Lancet* 1993;342:637-41. [[Medline](#)]
4. Yusuf S, Collins R, Peto R, Furberg C, Stampfer MJ, Goldhaber SZ, et al. Intravenous and

intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J* 1985;6:556-85.

[\[Medline\]](#)

This article has been cited by other articles:

- Moseley, J. B., O'Malley, K., Petersen, N. J., Menke, T. J., Brody, B. A., Kuykendall, D. H., Hollingsworth, J. C., Ashton, C. M., Wray, N. P. (2002). A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee. *N Engl J Med* 347: 81-88 [\[Abstract\]](#) [\[Full text\]](#)
- Wong, S. M., Griffith, J. F., Tang, A., Hui, A. C. F. (2002). Re: The role of ultrasonography in the diagnosis and management of idiopathic plantar fasciitis. *Rheumatology* 41: 835-836 [\[Full text\]](#)
- Cummings, P, Koepsell, T D (2002). Statistical and design issues in studies of groups. *Inj Prev* 8: 6-7 [\[Full text\]](#)
- Briggs, A. H., O'Brien, B. J., Blackhouse, G. (2002). THINKING OUTSIDE THE BOX: Recent Advances in the Analysis and Presentation of Uncertainty in Cost-Effectiveness Studies. *Annu. Rev. Public Health* 23: 377-401 [\[Abstract\]](#) [\[Full text\]](#)
- Sahota, P., Rudolf, M. C J, Dixey, R., Hill, A. J, Barth, J. H, Cade, J. (2001). Randomised controlled trial of primary school based intervention to reduce risk factors for obesity. *BMJ* 323: 1029-1029 [\[Abstract\]](#) [\[Full text\]](#)
- Briggs, A. (2000). Economic evaluation and clinical trials: size matters. *BMJ* 321: 1362-1363 [\[Full text\]](#)
- Bower, P., Byford, S., Sibbald, B., Ward, E., King, M., Lloyd, M., Gabbay, M. (2000). Randomised controlled trial of non-directive counselling, cognitive-behaviour therapy, and usual general practitioner care for patients with depression. II: Cost effectiveness. *BMJ* 321: 1389-1392 [\[Abstract\]](#) [\[Full text\]](#)
- Harrison, J. E. (2000). Evidence-based Orthodontics--How do I assess the evidence?. *Br. J. Orthod.* 27: 189-197 [\[Full text\]](#)
- Rushton, L. (2000). Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. *Occup Environ Med* 57: 1-9 [\[Abstract\]](#) [\[Full text\]](#)
- Rigby, A. S. (1999). Getting past the statistical referee: moving away from P-values and towards interval estimation. *Health Educ Res* 14: 713-715 [\[Full text\]](#)
- King, R., Denne, J. (1995). Audit suggests that use of aspirin is rising in coronary heart disease. *BMJ* 311: 1504-1504 [\[Full text\]](#)
- Raynor, P., Rudolf, M. C J, Cooper, K., Marchant, P., Cottrell, D., BLAIR, M. (1999). A

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

randomised controlled trial of specialist health visitor intervention for failure to thrive •
Commentary. *Arch. Dis. Child.* 80: 500-506 [\[Abstract\]](#) [\[Full text\]](#)

- TARNOW-MORDI, W. O., HEALY, M. J R (1999). Distinguishing between "no evidence of effect" and "evidence of no effect" in randomised controlled trials and other comparisons. *Arch. Dis. Child.* 80: 210-211 [\[Full text\]](#)
- Rudolf, M C J, Lyth, N, Bundle, A, Rowland, G, Kelly, A, Bosson, S, Garner, M, Guest, P, Khan, M, Thazin, R, Bennett, T, Damman, D, Cove, V, Kaur, V (1999). A search for the evidence supporting community paediatric practice. *Arch. Dis. Child.* 80: 257-261 [\[Abstract\]](#) [\[Full text\]](#)
- Bender, R., Sawicki, P. T (1996). Interpretation of study's results is open to criticism. *BMJ* 312: 254-254 [\[Full text\]](#)
- Jones, B, Jarvis, P, Lewis, J A, Ebbutt, A F (1996). Trials to assess equivalence: the importance of rigorous methods. *BMJ* 313: 36-39 [\[Full text\]](#)
- Hatcher, S. (1996). Predicting which psychiatric patients are at risk of suicide. *BMJ* 313: 884-884 [\[Full text\]](#)
- Williams, C, Harrad, R A, Sparrow, J M, Harvey, I, Golding, J, Lee, J., Adams, G., Sloper, J., McIntyre, A., Fielder, A. R, Aylward, G W, Rahi, J., Dezateux, C. (1998). Future of preschool vision screening. *BMJ* 316: 937a-937 [\[Full text\]](#)

[Home](#)[Help](#)[Search/Archive](#)[Feedback](#)[Search Result](#)

BMJ 1995;311:1145-1148 (28 October)

Education and debate

Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons

M J Campbell, *reader in medical statistics*,^a S A Julious, *statistician programmer*,^a
D G Altman, *head*^b

^a Medical Statistics and Computing, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, ^b Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX

Correspondence to: Dr Campbell.

Sample size calculations are now mandatory for many research protocols, but the ones useful in common situations are not all easily accessible. This paper outlines the ways of calculating sample sizes in two group studies for binary, ordered categorical, and continuous outcomes. Formulas and worked examples are given. Maximum power is usually achieved by having equal numbers in the two groups. However, this is not always possible and calculations for unequal group sizes are given.

A sample size calculation is now almost mandatory in research protocols and to justify the size of clinical trials in papers.¹ Nevertheless, one of the most common faults in papers reporting clinical trials is in fact a lack of justification of the sample size, and it is a major concern that important therapeutic effects are being missed because of inadequately sized studies.² A recent paper has concluded "the reporting of statistical power and sample size needs to be improved."³ Recent articles in the BMJ have described the basis of sample size calculations,^{4 5} and explained the fundamental concepts of statistical significance (alpha), effect size ((delta)), and power (1-β). A nomogram for sample size calculations for continuous data is also available.⁶ However, there have been some recent developments in the theory of sample size calculations, which are likely to prove useful, and the purpose of this paper is to make available a collection of formulas and examples for a variety of situations likely to be encountered in practice. In particular, situations not dealt with in previous articles are two group comparisons with unequal sample sizes, and sample sizes for ordered categorical outcomes (for example categories better, same, or worse). The paper describes sample size calculations, and provides tables, for studies comparing two groups of individuals that have outcome variables that are binary (yes/no), ordered categorical, or continuous. A further paper will consider studies when the data are paired. Further examples are given by Machin and Campbell.⁷

Parameter definition

Of all the parameters that have to be specified before the sample size can be determined the most critical is the effect size. Reducing the effect size by half will quadruple the required sample size. The effect size can be interpreted as a "clinically important difference," but this is often difficult to quantify. A valuable attempt at classification was made by Burnand et al, who reviewed three major medical journals and looked for words such as "impressive difference," "important difference," "dramatic increase" and then calculated a standardised effect size.⁸ This provided a guide to the size of effect regarded as important by other authors. There are several ways of eliciting useful sample sizes: a Bayesian perspective has been given recently,⁹ along with an economic approach,¹⁰ and one based on patients' rather than clinicians' perceptions of benefit.¹¹

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ [See Correction for this article](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Campbell, M J](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

In statistical significance tests one sets up a null hypothesis and, given the observed difference of interest, calculates the probability of observing the difference (or a more extreme one) under the null hypothesis. This yields the P value. If the P value is less than some prespecified level then we reject the null hypothesis. This level is known as the significance level (α). If we reject the null hypothesis when it is true we make a type I error, and we set (α), the significance level, to control the probability of doing this. If the null hypothesis is in fact false but we fail to reject it, we make a type II error, and the probability of a type II error is denoted as β . The probability of rejecting the null hypothesis when it is false is termed the power and is defined as $1-\beta$.

Unequal numbers in each group

For a given total sample size the maximum power is achieved by having equal numbers of subjects in the two groups. Often, however, in observational studies an equal number is not expected in each group since the incidence of a particular factor may be higher in one group than in another. In clinical trials, also, the numbers of subjects taking one treatment may have to be limited, so to achieve the necessary power one has to allocate more patients to the other treatment. In this case the sample sizes should be adjusted by a factor dependent on the allocation ratio,¹² given as equation 1 in the Appendix.

If one were to maintain the same sample size as calculated for a 1:1 ratio but then allocated in the ratio 2:1 the loss in power would be quite small (around 5%). However, if the allocation ratio is allowed to exceed 2:1 with the same total sample size the power falls very quickly (a loss of around 25% in power for a ratio of 5:1) and consequently a considerably larger total sample size is required to maintain a fixed power with an imbalanced study than with a balanced one.

Continuous data

In a two group comparative study where the outcome measure is a continuous variable which is plausibly normally distributed, such as blood pressure, a two sample t test would be the statistical test used in the final analysis.

To calculate a sample size, in addition to the parameters discussed above, an estimate of the population standard deviation (σ) must be given. The sample size formula⁷ is given as equation 2 in the Appendix, and table I gives the sample size required for different values of the standardised difference d , defined as $d=(\delta)/(\sigma)$, at various levels of power at the two sided 5% significance level.

Alternatively, Lehr gives a quick formula for calculating these sample sizes.¹³ For a two sided significance level of 5% and power of 80%, the number required in each group is given approximately by $m=16/d^2$. This formula overestimates the sample sizes a little for small values of d ; otherwise it gives close approximations to the sample size.

WORKED EXAMPLE

In a recent paper, Godfrey et al¹⁴ found that 46 people who had no whorls on their fingers had a mean systolic blood pressure of 136 mm Hg compared with 93 patients with at least one whorl for whom the mean blood pressure was 144 mm Hg.

Suppose an experimenter wished to confirm these findings but suspected that the mean difference would be less than that observed, with 5 mm Hg being the clinically minimum difference accepted. The overall standard deviation of blood pressure in each group is assumed to be 17 mm Hg, the same as that published. We find $d=5/17=0.294$, which is about 0.3, and so from table I the sample size required to detect this difference with a two sided significance level of 5% and with 80% power would be 176 subjects in each group and so 352 subjects in total. Alternatively, from Lehr's quick formula we get $m=16/0.294^2=185$ patients per group. Suppose, like Godfrey et al, we would expect to recruit two people with whorls for every one person with no whorls. With $r=2$ from equation 1 we find that $m'=3 \times 176/4=132$ and so $rm'=264$, giving a modified total sample size of 396. The overall sample size is larger if the groups were unequal because the design has less power than a design of the same size with equal numbers in the two groups.

TABLE I--Sample sizes required per group at the two sided 5% significance level for different values of d and power (d=expected mean difference/standard deviation)

d	Power (1-beta)				
	99	95	90	80	50
0.10	3676	2600	2103	1571	770
0.20	920	651	527	394	194
0.30	410	290	235	176	87
0.40	231	164	133	100	49
0.50	148	105	86	64	32
0.60	104	74	60	45	23
0.70	76	54	44	33	17
0.80	59	42	34	26	13
0.90	47	34	27	21	11
1.00	38	27	22	17	9
1.10	32	23	19	14	8
1.20	27	20	16	12	7
1.30	23	17	14	11	6
1.40	20	15	12	9	5
1.50	18	13	11	8	5

Binary data

A binary outcome is a response which has just two categories. These categories may be of the form yes/no or presence/absence in relation to a given factor, for example alive/dead. Often an experimenter may wish to compare treatments by testing whether the difference in proportions responding on each treatment could be due to chance. In this case the effect size can be formulated as $(\delta)=pA-pB$, where pA and pB are the proportions expected in the two treatment groups. The statistical test used to test for the association between two binary variables is the Pearson χ^2 test.

To calculate the number of patients required in each arm of a binary trial use equation 3 in the Appendix. For proportions greater than about 0.1 this simplifies to equation 4. Table II gives the sample sizes required for various values of pA and pB for two sided significance level (alpha) and power 1-β. Note, however, that for pA in the table only values up to 0.5 are given. This is because having a success rate of 65%, say, is identical to a failure rate of 35% and so the sample sizes for comparing pA to pB are the same as those for comparing 1-pA and 1-pB.

An approximate result similar to Lehr's formula¹³ for 80% power and two sided 5% significance level is that $m=16p(1-p)/(pA-pB)^2$, where $p=(pA + pB)/2$. Like Lehr's equation given earlier, this overestimates the sample size a little.

Observational surveys such as case control studies are often summarised by an odds ratio or relative risk, rather than a difference in proportions. If pA is the proportion of cases exposed to a risk factor and pB is the proportion of controls exposed to the same risk factor, then the odds ratio of being a case given the risk factor is $\text{odds ratio}=\frac{pA(1-pB)}{pB(1-pA)}$. An approximate sample size formula using the odds ratio (OR) is given by equation 5 in the Appendix.

WORKED EXAMPLE

Tovey and Bonell stated that 52 (19%) out of 281 men found condoms too tight.¹⁵ Of these 68% had experienced their condom splitting compared with only 26% of men whose condoms were not tight. Suppose from anecdotal evidence a researcher suspected that the prevalence of reported splitting was nearer 50% in the group finding condoms too tight and wished to conduct a study to show this prevalence still to be significantly higher than in the other group.

The expectation is that the observed ratio of the frequencies of "not tight" (A) to "tight" (B) would be 4:1. Here $p_A=0.5$, $p_B=0.25$ and $r=4$. From table II the sample size required with equal allocation in each group would be 58, and using equation 2 one derives a modified sample size of just 37 subjects in the group who found condoms too tight and 148 in the other group, giving a total of 185. In the unlikely event of equal group sizes a total of 116 subjects would be required, yielding a saving of 69 subjects. Again, this arises because the equal groups case is more efficient. Note that Lehr's formula for equal sized groups gives approximately 60 per group or a total of 120 subjects required. If we specified the effect size as an odds ratio, then the postulated odds of splitting when the condom is too tight are three times that when it is not. From equation 5, we find in this case that for equal allocation we require 55 subjects per group.

TABLE II--Sample sizes to detect a difference in two proportions, p_A and p_B , at a 5% significance level with 80% power

		pB												
		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65
pA		0.70	0.75	0.80	0.85	0.90	0.95	1.00						
0.00			152	74	48	35	27	22	18	15	13	11	10	8
7	6	6	6	5	4	4	3	2						
0.05				435	141	76	49	36	27	22	18	15	12	11
9	8	7	6	5	4	4	3							
0.10					686	199	100	62	43	32	25	20	16	14
11	10	8	7	6	5	4	4							
0.15						906	250	121	73	49	36	27	22	17
14	12	10	8	7	6	5	4							
0.20							1094	294	138	82	54	39	29	23
18	15	12	10	8	7	6	5							
0.25								1251	329	152	89	58	41	31
24	19	15	12	10	8	7	6							
0.30									1377	356	163	93	61	42
31	24	19	15	12	10	8	6							
0.35										1471	376	170	96	62
43	31	24	18	14	11	9	7							
0.40											1534	388	173	97
62	42	31	23	17	14	11	8							
0.45												1565	392	173
96	61	41	29	22	16	12	10							

Ordered categorical data

A study may be undertaken where the outcome measure of interest is an ordered scale, such as a Likert scale (strongly disagree, disagree, agree and strongly agree) or a rating scale (better, same, worse). The statistical test used in this instance is the Mann-Whitney U test, with allowance for ties.¹⁶ The calculation of sample sizes when the data are ordered is not immediately straightforward. The problem becomes considerably easier, however, if one considers a number of pragmatic steps which will be described later in this section.

As before, we need to specify an effect size, and here it turns out to be easier to use the odds ratio. We must also specify the proportion of subjects expected in each category of the scale for one of the groups. Suppose we have *t* categories, with the higher ordered categories indicating worse prognosis, and the proportions expected in group A are *p*_{A1}, *p*_{A2}, ... *p*_{At} (where *p*_{A1} + *p*_{A2} + ... + *p*_{At} = 1) with similar notation for group B. Let *c*_{A1}, *c*_{A2}, ... *c*_{At}, be the cumulative probabilities, so *c*_{A1} = *p*_{A1}, *c*_{A2} = *p*_{A1} + *p*_{A2}, etc. The odds ratio is the chance of a subject being in a given category or lower in one group compared with the other. For category 1 it is given by $OR_1 = \{c_{A1}/(1-c_{A1})\} / \{c_{B1}/(1-c_{B1})\}$ and similarly *OR*₂ for category 2, up to category *t*-1. As will be shown later, the odds ratio may not necessarily be too difficult to estimate, as the proportions expected for one group may already be known through a pilot study or from previous research. The experimenter may postulate that on the new treatment a patient is only half as likely to have a score above a given level than on the old treatment and so the odds ratio would be estimated as 0.5. Alternatively, an experimenter may know the expected proportions in each category for one group and speculate that, if a proportion, *p*, were in a particular category or better, then a clinically significant difference would be for the corresponding proportion to be about 20% higher in the other group. From this information an odds ratio can be calculated and hence the other expected proportions and the sample size.

Equation 6 in the Appendix gives the formula for sample size calculations for ordered categorical data. It assumes that the odds ratio is constant for each pair of adjacent categories, that is *OR*₁ = *OR*₂ = ... = *OR*_{t-1}, and this assumption means that the Mann-Whitney U test is the best test to use. It also means that one can estimate the odds ratio from any cumulative proportion from each group. To aid the calculations table III gives values of the numerator from equation 6 for different values of odds ratio and power.

TABLE III--For ordered categorical data, values for 6
 $(z_{1-(\alpha/2)} + z_{1-\beta^2} / \log OR)^2$
 for various values of the odds ratio (OR) and power
 (1-β) at two sided 5% significance

Odds ratio	Power (1-β)				
	99	95	90	80	50
0.75	1331.97	942.09	761.77	569.03	278.50
1.25	2213.86	1565.85	1266.13	945.78	562.89
1.50	670.52	474.25	383.48	286.45	140.20
1.75	352.00	248.96	201.31	150.38	73.60
2	229.44	162.60	131.22	98.02	47.97
3	91.33	64.60	52.23	39.02	19.10
4	57.36	40.97	32.80	24.50	11.99
5	42.56	30.10	24.34	18.18	8.90
10	20.79	14.71	11.89	8.88	4.35

If the number of categories is large it is difficult to postulate the proportion of people who would fall in a given category. However, Whitehead has shown that there is little increase in power (and hence saving in number of subjects recruited) to be gained by

increasing the number of groups beyond five.¹⁷

WORKED EXAMPLE

In a randomised controlled trial of paracetamol for the treatment of feverish children, Kinmonth et al categorised playfulness as normal or slightly, moderately, or very listless.¹⁸ The results for the 43 replies are given in table V, together with the proportions and the cumulative proportions. The first odds ratio in the table is calculated from $\{0.14/(1-0.14)/(0.27/(1-0.27))\} = 0.44$, and in a similar way we get 0.287 and 0.1625 for the other two pairs. The average is about 0.3.

Suppose a new study was planned in which we wished to replicate these results. The distribution of children in the control group (group A) was expected to be about the same as was found previously and shall be used in the calculation of the sample sizes. If an odds ratio of 0.33 in favour of paracetamol (or equivalently an odds ratio of about 3 against the control) was expected, then from the definition of the odds ratio we can calculate the expected cumulative proportions in the treatment group (group B) from the formula $CB_i = CA_i / (CA_i + OR(1 - CA_i))$. Thus the proportion expected in the first category of group B is $0.14 / (0.14 + 0.33(1 - 0.14)) = 0.33$ and so on. The cumulative proportions expected in group B are 0.33, 0.65, 0.83, and 1.00, and so the actual proportions expected are 0.33, $0.32 = (0.65 - 0.33)$, $0.18 = (0.83 - 0.65)$, and $0.17 = (1.00 - 0.83)$. The average proportions p are given by 0.235, 0.280, 0.210, and 0.275. Thus $(1 - (\Sigma)p^3) = 0.935$. For 80% power and 5% significance level, from table III, the numerator is 39.02, and so the sample size is $39.02 / 0.935 = 41.7$, or about 42 patients per group.

The formula is quite complicated and we have a number of suggestions to simplify matters. If the mean proportions (π_i 's) in each category are roughly equal then the denominator in equation 6 is constant for a given number of categories, and if the number of categories exceeds five it is approximately unity. Thus for 80% power and a two sided significance of 5%, an estimate of the sample size can be obtained from $m = 47 / (\log OR)$.² If the number of categories is less than or equal to five then multiply this sample size estimate by a correction factor given in table IV. From this table, in the situation of approximately equal proportions, it is evident that having only two categories in your data for analysis may require you to recruit a third more patients than if the data were kept continuous. For our example, the correction factor from table IV is 1.067 and so $n = 1.067 \times 47 / (\log 0.33)^2 = 40.8$, or 41 patients.

TABLE IV--Correction factor to be used with table III when the number of categories is ≤ 5

No of categories	Correction factor
2	1.333
3	1.125
4	1.067
5	1.042

TABLE V--Playfulness in children

Odds Category	Numbers*		Proportions		Cumulative proportions		
	A	B	A	B	A	B	
			pAi	pBi	cAi	cBi	ratios
Normal 0.440	3	6	0.14	0.27			0.14 0.27
Slightly listless 0.287	5	9	0.24	0.41			0.38 0.68
Moderately listless 0.1625	5	5	0.24	0.23			0.62 0.91
Very listless	8	2	0.38	0.09			1.00 1.00
Total	21	22	1.00	1.00			

*A=control, B=paracetamol.

Another simplification occurs if the proportion of subjects in one category for both groups is expected to be large. We can combine categories until there are only two left and use the formula and table given previously for binary data. Combining categories reduces the amount of information available, so one would expect the required sample size to increase.

In the worked example if we had pooled those scoring 1-2 and those scoring 3-4, we would compare proportion pA=0.38 to pB=0.65. Formula 4 shows that this study would require 49.9, or about 50 patients per group. Thus, use of all four categories, rather than simply two, yields a reduction of 16% in the study size, and this might outweigh the benefit of an easier sample size calculation.

Comment

From the equations in the Appendix it is clear that the sample size, significance level, power, and effect size are all interlinked. Given any three parameters, in principle the equations can be solved for the fourth. Thus, if the sample size were limited by resources, and the significance level fixed in advance, one could arbitrarily increase the power of the study by postulating larger effect sizes. In practice, however, the estimate of the effect of an intervention often proves too optimistic, resulting in many trials which are too small. The need for sample size calculations provides an excellent opportunity to involve a statistician early in the planning of a study and not just when the analysis is required. This paper has covered only a limited range of designs, and a statistician could advise on other designs. These include comparison of more than two groups,¹⁹ comparison of survival curves,^{7 20 21} and studies to demonstrate bioequivalence.²² Computer software is available for some of the sample size calculations discussed here,^{23 24 25 26} and other reviews have been given.^{27 28}

We thank Dr D Machin for comments on an earlier manuscript.

Funding: MJC and SAJ are funded by the Higher Education Funding Council and DGA by the Imperial Cancer Research Fund.

Conflict of interest: None.

Appendix

In each of the following m is the number of subjects required in each group for a two sided significance (α) and power $1-\beta$, and $z_{1-(\alpha/2)}$ and $z_{1-\beta}$ are the appropriate values from the standard Normal distribution for the $100(1-(\alpha/2))$ and $100(1-\beta)$ percentiles respectively. Some useful values are the following: for two sided (α)=0.05, $z_{1-(\alpha/2)}$ =1.96; for two sided (α)=0.01, $z_{1-(\alpha/2)}$ =2.58; for β =0.2, $z_{1-\beta}$ =0.84; and for β =0.1, $z_{1-\beta}$ =1.28.

UNEQUAL ALLOCATION

Given m , calculated assuming equal sized groups, let m' be the sample size in the first group and rm' the sample size in the second group. Then m' is given by $m' = r+1/2rxm$, (1) where r is the allocation ratio.

CONTINUOUS DATA

To detect a difference (δ) we require⁷: $m = 2(z_{1-(\alpha/2)} + z_{1-\beta})^2 / d^2 + z_{1-(\alpha/2)}^2 / 4$ (2) where $d = (\delta) / (\sigma)$ and (σ) is the standard deviation of the measurements. The last term in the equation is a correction factor to enable Normal tables rather than t tables to be used and can be ignored except for very small sample sizes. For a 5% two sided significance level it increases the sample size by 1. Table I gives the sample size required for different values of d and power from 50% to 99%.

BINARY OUTCOME

Suppose the expected proportions in groups A and B were p_A and p_B .

$m = [z_{1-(\alpha/2)}(\sqrt{2p(1-p)}) + z_{1-\beta}(\sqrt{p_A(1-p_A) + p_B(1-p_B)})]^2 / (\delta)^2$ (3) where $(\delta) = p_A - p_B$, and $p = (p_A + p_B) / 2$. An approximate, simpler formula, is: $m = (z_{1-(\alpha/2)} + z_{1-\beta})^2 [p_A(1-p_A) + p_B(1-p_B)] / (\delta)^2$ (4) which is sufficiently accurate except when p_A , p_B are small (say < 0.05). Table II gives the sample size required per group at 5% significance level and 80% power for values of p_A between 0 and 0.45 and p_B between 0.05 and 1.00.

If the effect size is specified as an odds ratio $OR = p_A(1-p_B) / p_B(1-p_A)$, then an approximate formula is given by $m = 2(z_{1-(\alpha/2)} + z_{1-\beta})^2 / \log(OR)^2 p(1-p)$ (5) Ordered categorical data $m = 6(z_{1-(\alpha/2)} + z_{1-\beta})^2 / (\log OR)^2 [1 - (\sum_{i=1}^k \pi_i^3)]$, (6) where OR is the odds ratio of a patient being in category i or less for one treatment compared to the other, k is the number of categories and π_i is the mean proportion expected in category i —that is, $\pi_i = (p_{Ai} + p_{Bi}) / 2$ where p_{Ai} and p_{Bi} are the proportions expected in category i for the two groups A and B respectively.

1. Gardner MJ, Machin D and Campbell MJ. Use of checklists in assessing the statistical content of medical studies. *BMJ* 1986;292:810-12. [\[Medline\]](#)
2. Freiman JA, Thomas AB, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;299:690-4. [\[Abstract\]](#)
3. Moher D, Dulberg CS, Wells GA. Statistical power, sample size and their reporting in randomised controlled trials. *JAMA* 1994;272:122-4. [\[Medline\]](#)
4. Florey C du V. Sample sizes for beginners. *BMJ* 1993;306:1181-4. [\[Medline\]](#)
5. Daly L. Confidence intervals and sample sizes: don't throw out all your old sample size tables. *BMJ* 1991;302:333-6. [\[Medline\]](#)
6. Gore SM, Altman DG. Statistics in practice. BMA: London, 1982.
7. Machin D, Campbell MJ. Statistical tables for the design of clinical trials. Oxford: Blackwell Scientific, 1987.
8. Burnand B, Kernan WN, Feinstein AR. Indexes and boundaries for "quantitative significance" in statistical decisions. *J Clin Epidemiol* 1990;43:1273-84. [\[Medline\]](#)

9. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. *J R Statistic Soc A* 1994;157:357-416.
10. Drummond M, O'Brian B. Clinical importance, statistical significance and the assessment of economic and quality-of-life outcomes. *Health Economics* 1993;2:205-12. [[Medline](#)]
11. Naylor CD, Llewellyn-Thomas HA. Can there be a more patients-centred approach to determining clinically important effect size for randomised treatments? *J Clin Epidemiol* 1994;47:787-95.
12. Woodward M. Formulas for sample-size, power and minimum detectable relative risk in medical studies. *Statistician* 1992;41:185-96.
13. Lehr R. Sixteen s squared over d squared: a relation for crude sample size estimates. *Statistics in Medicine* 1992;11:1099-1102. [[Medline](#)]
14. Godfrey KM, Barker DJP, Peace J, Cloke J, Osmond C. Relationship of fingerprints and shape of the palm to fetal growth and adult blood pressure. *BMJ* 1993;307:405-9. [[Medline](#)]
15. Tovey SJ, Bonell CP. Condoms: a wider range needed. *BMJ* 1993;307:987. [[Medline](#)]
16. Conover WJ. Practical nonparametric statistics. 2nd ed. New York: John Wiley, 1980.
17. Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine* 1993;12:2257-72. [[Medline](#)]
18. Kinmonth A-L, Fulton Y, Campbell MJ. Management of feverish children at home. *BMJ* 1992;305:1134-6. [[Medline](#)]
19. Day SJ, Graham DF. Sample size estimation for comparing two or more groups. *Statistics in Medicine* 1991;10:33-43. [[Medline](#)]
20. Borenstein M. Planning for precision in survival studies. *J Clin Epidemiol* 1994;47:1277-85. [[Medline](#)]
21. Fayers PM, Machin D. Sample size: How many patients are necessary? *Br J Cancer* 1995;72:1-9.
22. Diletti E, Hauschke D, Steinijans VW. Sample size determination for bioequivalence assessment by means of confidence intervals. *Int J Clin Pharm Therapy Toxicol* 1991;29:1-8.
23. Goldstein R. Power and sample size via MS/PC-DOS computers. *American Statistician* 1989;43:253-60.
24. Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. *Controlled Clinical Trials* 1990;11:116-28. [[Medline](#)]
25. Rahlfs V. N Handbook. Munchen: IDV-Datenanalyse und Versuchplanung, 1987.
26. Pinol A. SAMPLE. Geneva: HRP, World Health Organisation, 1995.
27. Lachin JM. Introduction to sample size determination and power analysis in clinical trials. *Controlled Clinical Trials* 1981;2:93-113. [[Medline](#)]
28. Donner A. Approaches to sample size estimation in the design of clinical trials--a review. *Statistics in Medicine* 1984;3:199-214. [[Medline](#)]

(Accepted 21 July 1995)

This article has been cited by other articles:

- Piper, S. N., Fent, M. T., Rohm, K. D., Maleck, W. H., Suttner, S. W., Boldt, J. (2001). Urapidil does not prevent postanesthetic shivering: a dose-ranging study : [L'urapidil n'empêche pas le frisson postanesthésique : une étude de dosage]. *Can. J. Anaesth* 48: 742-747 [[Abstract](#)] [[Full text](#)]
- Lam, C. L., Lauder, I. J (2000). The impact of chronic diseases on the health-related quality of life (HRQOL) of Chinese patients in primary care. *Fam. Pract.* 17: 159-166 [[Abstract](#)] [[Full text](#)]
- Wade, A. (2000). Fear or favour? Statistics in pathology. *J Clin Pathol* 53: 16-18 [[Full text](#)]
- Pandit, J. J., Bree, S., Dillon, P., Elcock, D., McLaren, I. D., Crider, B. (2000). A Comparison of Superficial Versus Combined (Superficial and Deep) Cervical Plexus Block for Carotid Endarterectomy: A Prospective, Randomized Study. *Anesth Analg* 91: 781-786 [[Abstract](#)] [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ [See Correction for this article](#)
- ▶ Search Medline for articles by:
[Campbell, M J](#) || [Altman, D G](#)
- ▶ Alert me when:
[New articles cite this article](#)

- Balogh, I., Szôke, G., Kárpáti, L., Wartiovaara, U., Katona, E., Komáromi, I., Haramura, G., Pfliegler, G., Mikkola, H., Muszbek, L. (2000). Val34Leu polymorphism of plasma factor XIII: biochemistry and epidemiology in familial thrombophilia. *Blood* 96: 2479-2486 [\[Abstract\]](#) [\[Full text\]](#)
- Rowat, A. M., Wardlaw, J. M., Dennis, M. S., Warlow, C. P. (2000). Does Feeding Alter Arterial Oxygen Saturation in Patients With Acute Stroke?. *Stroke* 31: 2134-2140 [\[Abstract\]](#) [\[Full text\]](#)
- Joynt, G. M., Kew, J., Gomersall, C. D., Leung, V. Y. F., Liu, E. K. H. (2000). Deep Venous Thrombosis Caused by Femoral Venous Catheters in Critically Ill Adult Patients*. *Chest* 117: 178-183 [\[Abstract\]](#) [\[Full text\]](#)
- Peveler, R., George, C., Kinmonth, A.-L., Campbell, M., Thompson, C. (1999). Effect of antidepressant drug counselling and information leaflets on adherence to drug treatment in primary care: randomised controlled trial. *BMJ* 319: 612-615 [\[Abstract\]](#) [\[Full text\]](#)
- Jones, B, Jarvis, P, Lewis, J A, Ebbutt, A F (1996). Trials to assess equivalence: the importance of rigorous methods. *BMJ* 313: 36-39 [\[Full text\]](#)

[Home](#)[Help](#)[Search/Archive](#)[Feedback](#)[Search Result](#)

BMJ 1996;312:572 (2 March)

Education and debate

Statistics Notes: Presentation of numerical data

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

^a IRCF Medical Statistics Group, Centre for Statistics in Medicine,

Institute of Health Sciences, PO Box 777, Oxford OX3 7LF, ^b Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

Correspondence to: Mr Altman.

The purpose of a scientific paper is to communicate, and within the paper this applies especially to the presentation of data.

Continuous data, such as serum cholesterol concentration or triceps skinfold thickness, can be summarised numerically either in the text or in tables or plotted in a graph. When numbers are given there is the problem of how precisely to specify them. As far as possible the numerical precision used should be consistent throughout a paper and especially within a table. In general, summary statistics such as means should not be given to more than one extra decimal place over the raw data. The same usually applies to measures of variability or uncertainty such as the standard deviation or standard error, though greater precision may be warranted for these quantities as they are often used in further calculations. Similar comments apply to the results of regression analyses, where spurious precision should be avoided. For example, the regression equation¹

birth weight = $-3.0983527 + 0.142088 \times \text{chest circumf} + 0.158039 \times \text{midarm circumf}$, purports to predict birth weight to 1/1000000 g.

Categorical data, such as disease group or presence or absence of symptoms, can be summarised as frequencies and percentages. It can be confusing to give percentages alone, as the denominator may be unclear. Also, giving frequencies allows percentages to be given as integers, such as 22%, rather than more precisely. Percentages to one decimal place may sometimes be reasonable, but not in small

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

samples; greater precision is unwarranted. Such data rarely need to be shown graphically.

Test statistics, such as values of t or χ^2 , and correlation coefficients should be given to no more than two decimal places. Confidence intervals are better presented as, say, "12.4 to 52.9" because the format "12.4-52.9" is confusing when one or both numbers are negative. P values should be given to one or two significant figures. P values are always greater than zero. Because computer output is often to a fixed number of decimal places $P=0.0000$ really means $P<0.00005$ --such values should be converted to $P<0.0001$. P values always used to be quoted as $P<0.05$, $P<0.01$, and so on because results were compared with tabulated values of statistical distributions. Now that most P values are produced by computer they should be given more exactly, even for non-significant results--for example, $P=0.2$. Values such as $P=0.0027$ can be rounded up to $P=0.003$, but not in general to $P<0.01$ or $P<0.05$. In particular, the use of $P<0.05$ (or, even worse, $P=NS$) may conceal important information: there is minimal difference between $P=0.06$ and $P=0.04$. In tables, however, it may be necessary to use symbols to denote degrees of significance; a common system is to use *, **, and *** to mean $P<0.05$, 0.01, and 0.001 respectively. Mosteller gives a more extensive discussion of numerical presentation.²

The choice between using a table or figure is not easy, nor is it easy to offer much general guidance. Tables are suitable for displaying information about a large number of variables at once, and graphs are good for showing multiple observations on individuals or groups, but between these cases lie a wide range of situations where the best format is not obvious. One point to consider when contemplating using a figure is the amount of numerical information contained. A figure that displays only two means with their standard errors or confidence intervals is a waste of space as a figure; either more information should be added, such as the raw data (a really useful feature of a figure), or the summary values should be put in the text.

In tables information about different variables or quantities is easier to assimilate if the columns (rather than the rows) contain like information, such as means or standard deviations. Interpretation of tables showing data for individuals (or perhaps for many groups) is aided by having the data ordered by one of the variables--for example, by the baseline value of the measurement of interest or by some important prognostic characteristic.

1. Bhargava SK, Ramji S, Kumar A, Mohan MAN, Marwah J, Sachdev HPS. Mid-arm and chest circumferences at birth as predictors of low birth weight and neonatal mortality in the community. *BMJ* 1985;291:1617-9. [[Medline](#)]
2. Mosteller F. Writing about numbers. In: Bailar JC, Mosteller F. eds. *Medical uses of statistics*. 2nd ed. Boston: NEJM Books, 1992:375-89.

This article has been cited by other articles:

- Rushton, L. (2000). Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. *Occup Environ Med* 57: 1-9
[\[Abstract\]](#) [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1996;312:700 (16 March)

Education and debate

Statistics Notes: Logarithms

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, PO Box 777, Oxford OX3 7LF

Correspondence to: Professor Bland.

Logarithms (or logs for short) are much used in statistics. We often analyse the logs of measurements rather than the measurements themselves, and some widely used methods of analysis, such as logistic and Cox regression, produce coefficients on a logarithmic scale. Here we shall give a brief summary of the properties of logarithms which make them so useful.

We shall start with logarithms to base 10. These are the common logarithms formerly widely used to do calculations for which we now use calculators and computers. The log to base 10 of a number a is b where $a=10^b$. We write $b=\log_{10}a$. Thus for example $\log_{10}(10)=1$, $\log_{10}(100)=2$, $\log_{10}(1000)=3$, $\log_{10}(10000)=4$, and so on. It is common to omit the brackets and write $\log_{10}a$, but we are using them for clarity.

If we multiply two numbers, the log of the product is the sum of their logs: $\log(ab)=\log(a)+\log(b)$. For example, $100 \times 1000 = 10^2 \times 10^3 = 10^{2+3} = 10^5 = 100000$. Or in log terms: $\log_{10}(100 \times 1000) = \log_{10}(100) + \log_{10}(1000) = 2 + 3 = 5$. Hence $100 \times 1000 = 10^5 = 100000$. It follows that any multiplicative relationship of the form $y=axbxcxd$ can be made additive by a log transformation: $\log(y)=\log(a)+\log(b)+\log(c)+\log(d)$. Likewise, the difference between two logs is the log of the ratio: $\log(a)-\log(b)=\log(a/b)$. As statistical methods cope with additive relationships much more easily than with multiplicative ones, logarithms have many uses. As we shall see in future **Statistics Notes**, working with the logarithms of data rather than the data themselves may have several advantages. Multiplicative relationships may become additive, skewed distributions may become symmetrical, and curves may become straight lines.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

Most scientific calculators have a LOG key, which will give the logarithm of the number in the display. They usually have a 10^x key, too, which gives us the number of which the display is the logarithm. This is called the antilogarithm or antilog and is useful when dealing with the results of calculations on the log scale.

Many statistical computer programs do not use logs to base 10, but logs to the base e, called natural logarithms. Here $e=2.7183\dots$ is a mathematical constant, in much the same way that $(\pi)=3.1412\dots$. Mathematicians, and hence statisticians, almost always use logs to the base e because it simplifies many formulae. On a calculator this is usually given by the LN key, and the antilog by the e^x key. The numerical relation between logs to base e and base 10 is that $\log_{10}(e) \times \log_e(x) = \log_{10}(x)$. Natural logarithms are also written as $\ln(x)$, or often simply as $\log(x)$.

The base which is used for logarithms is a matter of convenience, depending only on the particular application. The base chosen affects the values of the logs themselves, but nothing else. Provided we use the correct antilog to return to the natural scale, it does not matter what base we use.

This article has been cited by other articles:

- Bland, J M., Altman, D. G (1996). **Statistics Notes:** Transforming data. *BMJ* 312: 770-770 [[Full text](#)]
- Bland, J M., Altman, D. G (1996). **Statistics notes:** Transformations, means, and confidence intervals. *BMJ* 312: 1079-1079 [[Full text](#)]
- Bland, J M., Altman, D. G (1996). **Statistics Notes:** The use of transformation when comparing two means. *BMJ* 312: 1153-1153 [[Full text](#)]
- Bland, J M., Altman, D. G (1996). **Statistics Notes:** Measurement error proportional to the mean. *BMJ* 313: 106-106 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)
[Help](#)
[Search/Archive](#)
[Feedback](#)
[Search Result](#)

BMJ 1996;312:770 (23 March)

Education and debate

Statistics Notes: Transforming data

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, PO Box 777, Oxford OX3 7LF

Correspondence to: Professor Bland.

We often transform data by taking the logarithm, square root, reciprocal, or some other function of the data. We then analyse the transformed data rather than the untransformed or raw data. We do this because many statistical techniques, such as t tests, regression, and analysis of variance, require that data follow a distribution of a particular kind. The observations themselves must come from a population which follows a normal distribution,¹ and different groups of observations must come from populations which have the same variance or standard deviation. We need this uniform variance because we estimate the variance within the groups, and we can do this well only if we can assume it to be the same in each group. Many biological variables do follow a normal distribution with uniform variance. Many of those which do not can be made to do so by a suitable transformation. Fortunately, a transformation which makes data follow a normal distribution often makes the variance uniform as well, and vice versa. In this note we shall try to explain why this is the case.

Firstly, the normal distribution and uniform variance go together. It can be shown mathematically that if we take random samples from a population the means and standard deviations of these samples will be independent (and thus uncorrelated) if the population has a normal distribution. In other words, the standard deviation of the samples will not be related to the mean. Furthermore, if the mean and standard deviation are independent the distribution must be normal. This is harder to credit, but it is true.

Secondly, if we add together many variables we usually get a normal distribution. For example, the central limit theorem shows that the means of large samples will follow a normal distribution, whatever

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

the distribution of the observations themselves.¹ Similarly, if a biological variable is the result of the sum of many influences, it will follow a normal distribution. Human height is an example. Many biological measurements are not like this, however, but are the product of several factors. Substances in blood, for example, may be removed at a rate depending on the level of some other substance, which in turn is produced at a rate which depends on something else, and so on. We have the product of several influences multiplied together, rather than the sum. If we take the logarithm of the product of several variables, we get the sum of their logarithms.² So a variable which is the product of several factors has a logarithm which is the sum of several factors and so will follow a normal distribution.

Thirdly, any relation between variance and mean over several groups is usually fairly simple. The variance may be proportional to the group mean, the mean squared, the mean to the fourth power, etc. For such relations simple transformations can be found which will make the variance independent of the mean. If the variance is proportional to the mean we can use the square root transformation. This is often the case for data which are counts of things or events--for example, the number of cells of a particular type in a given volume of blood or number of deaths from AIDS in a geographical area over one year. Such data tend to follow a Poisson distribution, which has its variance equal to its mean. If the variance is proportional to the mean squared--that is, the standard deviation is proportional to the mean--we use the logarithmic transformation. This is the most frequent case in practice, suitable for variables such as serum cholesterol. If the variance is proportional to the mean to the fourth power--that is, the standard deviation is proportional to the mean squared--we use a reciprocal transformation, used for highly variable quantities such as serum creatinine. Thus we can transform the data to make the variance unrelated to the mean, in which case the data are likely to follow a normal distribution.

Some people ask whether the use of a transformation is cheating. There is no reason why the "natural" scale should be the only, or indeed the best, way to present measurements. pH, for example, is always presented as a logarithmic measure, $\text{pH} = -\log_{10}(\text{H}^+)$, where H^+ is the concentration of hydrogen ions in moles per cubic decimetre. Thus the "natural" scale is $10^{-\text{pH}}$. This natural scale is very awkward to use, and the logarithm is always used instead.

If we can transform data to follow a normal distribution with variance independent of the mean, valid analyses can be carried out on this transformed scale. There is one drawback, however, as confidence intervals on the transformed scale may be difficult to interpret. We shall deal with this in a subsequent note.

1. Altman DG, Bland JM. The normal distribution. *BMJ* 1995;310:298. [\[Full Text\]](#)
2. Bland JM, Altman DG. Logarithms. *BMJ* 1996;312:700. [\[Full Text\]](#)

This article has been cited by other articles:

- Bland, J M., Altman, D. G (2000). **Statistics Notes**: The odds ratio. *BMJ* 320: 1468-1468 [[Full text](#)]
- Myles, P. S., Troedel, S., Boquest, M., Reeves, M. (1999). The Pain Visual Analog Scale: Is It Linear or Nonlinear?. *Anesth Analg* 89: 1517-1517 [[Abstract](#)] [[Full text](#)]
- Bland, J M., Altman, D. G (1996). **Statistics notes**: Transformations, means, and confidence intervals. *BMJ* 312: 1079-1079 [[Full text](#)]
- Bland, J M., Altman, D. G (1996). **Statistics Notes**: The use of transformation when comparing two means. *BMJ* 312: 1153-1153 [[Full text](#)]
- Bland, J M., Altman, D. G (1996). **Statistics Notes**: Measurement error proportional to the mean. *BMJ* 313: 106-106 [[Full text](#)]
- Altman, D. G, Bland, J M. (1996). **Statistics Notes**: Detecting skewness from summary information. *BMJ* 313: 1200-1200 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1996;312:1079 (27 April)

Statistics notes

Transformations, means, and confidence intervals

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, PO Box 777, Oxford OX3 7LF

Correspondence to: Professor Bland.

When we use transformed data in analyses,¹ this affects the final estimates that we obtain. Figure [1](#) shows some serum triglyceride measurements, which have a skewed distribution. A logarithmic transformation is often useful for data which have positive skewness like this, and here the approximation to a normal distribution is greatly improved. For the untransformed data the mean is 0.51 mmol/l and the standard deviation 0.22 mmol/l. The mean of the log₁₀ transformed data is -0.33 and the standard deviation is 0.17. If we take the mean on the transformed scale and back transform by taking the antilog, we get $10^{-0.33}=0.47$ mmol/l. We call the value estimated in this way the geometric mean. The geometric mean will be less than the mean of the raw data.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

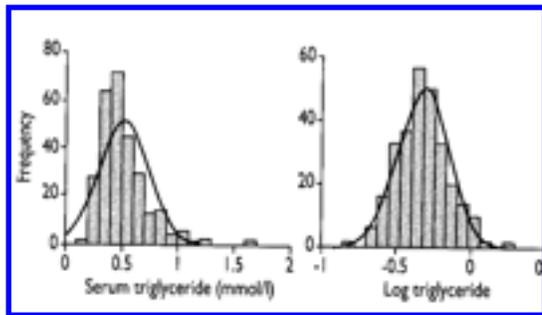


Fig 1--Serum triglyceride and log₁₀ serum triglyceride concentrations in cord blood for 282 babies, with best fitting normal distribution

View larger version (28K):

[\[in this window\]](#)

[\[in a new window\]](#)

When triglyceride is measured in mmol/l the log of a single observation is the log of a measurement in mmol/l. The average of n such transformed measurements is also the log of a number in mmol/l, so the antilog is back in the original units, mmol/l.

The antilog of the standard deviation, however, is not measured in mmol/l. Calculation of the standard deviation of the log transformed data requires taking the difference between each log observation and the log geometric mean. The difference between the log of two numbers is the log of their ratio.² As a ratio is a dimensionless pure number, the units in which serum triglyceride was measured would not matter; the standard deviation on the log scale would be the same. As a result, we cannot transform the standard deviation back to the original scale.

If we want to use the standard deviation or standard error it is easiest to do all calculations on the transformed scale and transform back, if necessary, at the end. For example, the 95% confidence interval for the mean on the log scale is -0.35 to -0.31. To get back to the original scale we antilog the confidence limits on the log scale to give a 95% confidence interval for the geometric mean on the natural scale (0.47) of 0.45 to 0.49 mmol/l. For comparison, the 95% confidence interval for the arithmetic mean using the raw, untransformed data is 0.48 to 0.54 mmol/l. These limits are wider than those for the geometric mean. This is because with highly skewed data the extreme observations have a large influence on the arithmetic mean, making it more prone to sampling error. Lessening this influence is one advantage of using transformed data.

If we use another transformation, such as the reciprocal or the square root,¹ the same principle applies. We carry out all calculations on the transformed scale and transform back once we have calculated the confidence interval. This works for the sample mean and its confidence interval. Things become more complicated if we look at the difference between two means. We shall look at this in another Statistics Note.

1. Bland JM, Altman DG. Transforming data. *BMJ* 1996;312:770. [\[Full Text\]](#)
2. Bland JM, Altman DG. Logarithms. *BMJ* 1996;312:700. [\[Full Text\]](#)

This article has been cited by other articles:

- Decensi, A., Omodei, U., Robertson, C., Bonanni, B., Guerrieri-Gonzaga, A., Ramazzotto, F., Johansson, H., Mora, S., Sandri, M. T., Cazzaniga, M., Franchi, M., Pecorelli, S. (2002). Effect of Transdermal Estradiol and Oral Conjugated Estrogen on C-Reactive Protein in Retinoid-Placebo Trial in Healthy Women. *Circulation* 106: 1224-1228 [\[Abstract\]](#) [\[Full text\]](#)
- Bland, J M., Altman, D. G (1996). **Statistics Notes:** The use of transformation when comparing two means. *BMJ* 312: 1153-1153 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

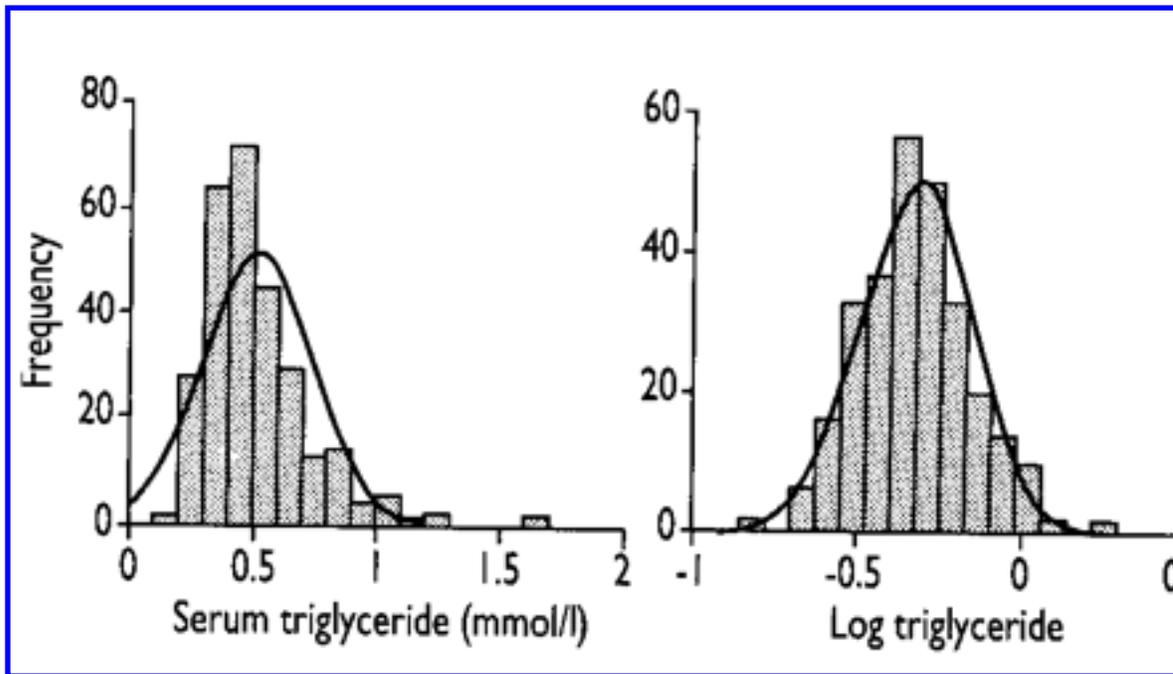


Fig 1--Serum triglyceride and log₁₀ serum triglyceride concentrations in cord blood for 282 babies, with best fitting normal distribution

[\[View larger version \(193K\)\]](#)

BMJ 1996;312:1153 (4 May)

Education and debate

Statistics Notes: The use of transformation when comparing two means

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

- ▶ [Email this article to a friend](#)
- ▶ **Respond** to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

^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, PO Box 777, Oxford OX3 7LF

Correspondence to: Professor Bland.

The usual statistical technique used to compare the means of two groups is a confidence interval or significance test based on the t distribution. For this we must assume that the data are samples from normal distributions with the same variance. Table 1 shows the biceps skinfold measurements for 20 patients with Crohn's disease and nine patients with coeliac disease.

Table 1--Biceps skinfold thickness (mm) in two groups of patients

Crohn's disease				Coeliac disease	
1.8	2.8	4.2	6.2	1.8	3.8
2.2	3.2	4.4	6.6	2.0	4.2
2.4	3.6	4.8	7.0	2.0	5.4
2.5	3.8	5.6	10.0	2.0	7.6
2.8	4.0	6.0	10.4	3.0	
Mean=4.72				Mean=3.53	
SD=2.42				SD=1.96	

The data have been put into order of magnitude, and it is fairly obvious that the distribution is skewed and far from normal. When, as here, the assumption of normality is wrong we can often transform the data to another scale where the assumption of normality is reasonable. The transformation which achieves a normal distribution should also give us similar variances.¹ Table 2 shows the results of analyses using the square root, logarithmic, and reciprocal transformations. The log transformation gives the most similar variances and so gives the most valid test of significance. It also gives a reasonable approximation to a normal distribution.

Table 2--Biceps skinfold thickness compared for two groups of patients, using different transformations

Transformation	Two sample		95% Confidence interval for difference on transformed scale	Variance ratio, larger/smaller
	ttest, 27 df	P		
None, raw data	1.28	0.21	-0.71 mm to 3.07 mm	1.52
Square root	1.38	0.18	-0.140 to 0.714	1.16
Logarithm	1.48	0.15	-0.114 to 0.706	1.10
Reciprocal	-1.65	0.11	-0.203 to 0.022	1.63

Confidence intervals for transformed data are more difficult to interpret, however. Unlike the case of a single sample,² the confidence limits for the difference between means cannot be transformed back to the original scale. If we try to do this the square root and reciprocal limits give ludicrous results. The lower limit for the square root transformation is negative. If we square this we get a positive lower limit and the confidence interval does not contain zero, even though the difference is not significant. If the observed difference were exactly zero the confidence limits would be equal in magnitude but opposite in sign. Transforming back by squaring would make them equal. For the reciprocal transformation the upper limit is very small (0.022) and transforming back by taking the reciprocal again gives 45.5. There is no way that the difference between mean skinfold in these two groups could be 45.5 mm. Thus the confidence interval for a difference cannot be interpreted on the untransformed scale for these transformations.

Only the log transformation gives interpretable (and thus useful) results after we transform back. Using the antilog transformation, we get a confidence interval of 0.89 to 2.03, but these are not limits for the difference in millimetres. How could they be, for they do not contain zero, yet the difference is not significant? They are in fact the 95% confidence limits for the ratio of the geometric mean² for patients with Crohn's disease to the geometric mean for patients with coeliac disease. If there were no difference the expected value of this ratio would be 1, not 0, and so lie within the limits. This procedure works because when we take the difference between the logarithms of the two geometric means we get the logarithm of their ratio, not of their

difference.³ We thus have the logarithm of a pure number and we antilog this to give the dimensionless ratio of the two geometric means. The logarithmic transformation is strongly preferable to other transformations for this reason. Fortunately, for medical measurements it often achieves the desired effect.

1. Bland JM, Altman DG. Transforming data. *BMJ* 1996;312:770. [\[Full Text\]](#)
2. Bland JM, Altman DG. Transformations, means, and confidence intervals. *BMJ* 1996;312:1079. [\[Full Text\]](#)
3. Bland JM, Altman DG. Logarithms. *BMJ* 1996;312:700. [\[Full Text\]](#)

This article has been cited by other articles:

- Vickers, A. J, Altman, D. G (2001). **Statistics Notes**: Analysing controlled trials with baseline and follow up measurements. *BMJ* 323: 1123-1124 [\[Full text\]](#)
- Azizi, M., Ezan, E., Nicolet, L., Grognet, J.-M., Menard, J. (1997). High Plasma Level of N-Acetyl-Seryl-Aspartyl-Lysyl-Proline : A New Marker of Chronic Angiotensin-Converting Enzyme Inhibition. *Hypertension* 30: 1015-1019 [\[Abstract\]](#) [\[Full text\]](#)
- Kerry, S. M, Bland, J M. (1998). Analysis of a trial randomised in clusters. *BMJ* 316: 54-54 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)
[Help](#)
[Search/Archive](#)
[Feedback](#)
[Search Result](#)

BMJ 1996;312:1472-1473 (8 June)

Education and debate

Statistics Notes: Comparing several groups using analysis of variance

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

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

Correspondence to: Mr Altman.

Many studies, including most controlled clinical trials, contrast data from two different groups of subjects. Observations which are measurements are often analysed by the t test, a method which assumes that the data in the different groups come from populations where the observations have a normal distribution and the same variances (or standard deviations). While the t test is well known, many researchers seem unaware of the correct method for comparing three or more groups. For example, table 1 shows measurements of galactose binding for three groups of patients. A common error is to compare each pair of groups using separate two sample t tests¹ with the consequent problem of multiple testing.² The correct approach is to use one way analysis of variance (also called ANOVA), which is based on the same assumptions as the t test. We compare the groups to evaluate whether there is evidence that the means of the populations differ. Why then is the method called analysis of variance?

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by: [Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when: [New articles cite this article](#)

Table 1--Measurements of galactose binding in three groups of patients (data from M Weldon)

	Crohn's disease	Ulcerative colitis	Controls	
	1343	1264	1809	2850
	1393	1314	1926	2964
	1420	1399	2283	2973
	1641	1605	2384	3171
	1897	2385	2447	3257
	2160	2511	2479	3271
	2169	2514	2495	3288
	2279	2767	2525	3358
	2890	2827	2541	3643
		2895	2769	3657
		3011		
		3013		
		3355		
Mean	1910.2	2373.8	2804.5	

SD	515.7	727.1	526.8
----	-------	-------	-------

We can partition the variability of the individual data values into components corresponding to within and between group variation. Table 2 shows the analysis of variance table for the data in table 1. Fuller details about the calculations can be found in textbooks³ (although a computer would generally be used). The first column shows the "sum of squares" associated with each source of variation; these add to give the total sum of squares. The second column shows the corresponding degrees of freedom. For the comparison of k groups there are k-1 degrees of freedom. The third column gives the sums of squares divided by the degrees of freedom, which are the variances associated with each component (perhaps confusingly called mean squares). When there are two groups the residual variance is the same as the pooled variance used in the two sample t test.

Table 2--Analysis of variance table for the data in table 1

Source of variation (P)	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)
Between groups 0.002	2	5 174 310.0	2 587 155.0	7.34
Residual (within groups)	39	13 743 776.2	352 404.5	
Total	41	18 918 086.2		

Analysis of variance assesses whether the variability of the group means--that is, the between group variance--is greater than would be expected by chance. Under the null hypothesis that all the population means are the same the between and within group variances will be the same, and so their expected ratio would be 1. The test statistic is thus the ratio of the between and within group variances, denoted F in table 2. The larger the value of F the more evidence there is that the means of the groups differ. The observed value of F is compared with a table of values of the F distribution using the degrees of freedom for both the numerator and denominator--this value is sometimes written as F*RF [2,39]*. For the data in table 1 and F value greater than 3.24 would be significant with P<0.05. The observed value is far larger than this, giving strong evidence that the three populations of patients differ. With two groups one way analysis of variance is exactly equivalent to the usual two sample t test, and we have F=t².

When the groups are significantly different we will often wish to explore further to see where the differences lie. When we compare more than two groups we need a clear idea of which comparisons we are interested in. Very often we are not equally interested in all possible comparisons. Many statistical procedures are available, their appropriateness depending on the question one wishes to answer. One simple method is to use the residual variance as the basis for modified t tests comparing each pair of groups. Here we get: group 1 v group 2, P=0.12; 1 v 3, P=0.0002; 2 v 3, P=0.06. The main difference is thus between groups 1 and 3, as can be seen from table 1. This procedure is an improvement on simply performing three two sample t tests in the first place because we proceed to comparing pairs of groups only if there is evidence of significant variability among all the groups, and also because we use a more reliable estimate of the variance within groups. Investigation of all pairs of groups often does not yield a simple interpretation, which is the price we can pay for not having a specific hypothesis. When the overall F test is not significant it is generally unwise to explore differences between pairs of groups. If the groups have a natural ordering--for example, representing

patients with different stages of a disease--it is preferable to examine directly evidence for a (linear) trend in means across the groups.¹ We will consider such data in a subsequent statistics note.

This type of analysis can be extended to more complex data sets with two classifying variables, using two way analysis of variance, and so on. Analysis of variance is a special type of regression analysis, and most data sets for which analysis of variance is appropriate can be analysed by regression with the same results.

1. Godfrey K. Comparing the means of several groups. In: Bailar JC, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston, MA: NEJM Books, 1992: 233-57.
2. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170. [\[Full Text\]](#)
3. Armitage P, Berry G. Statistical methods for research workers. 3rd ed. Oxford: Blackwell, 1994.

This article has been cited by other articles:

- Kjaergard, L. L., Als-Nielsen, B. (2002). Association between competing interests and authors' conclusions: epidemiological study of randomised clinical trials published in the *BMJ*. *BMJ* 325: 249-249 [\[Abstract\]](#) [\[Full text\]](#)
- Salinas, M., Rosas, J., Iborra, J., Manero, H., Pascual, E. (1997). Comparison of manual and automated cell counts in EDTA preserved synovial fluids. Storage has little influence on the results. *Ann Rheum Dis* 56: 622-626 [\[Abstract\]](#) [\[Full text\]](#)
- Bland, J M., Altman, D. G (1996). **Statistics notes**: Measurement error. *BMJ* 312: 1654-1654 [\[Full text\]](#)
- Bland, J M., Altman, D. G (1996). **Statistics Notes**: Measurement error. *BMJ* 313: 744-744 [\[Full text\]](#)
- Ferro, C. J., Spratt, J. C., Haynes, W. G., Webb, D. J. (1998). Inhibition of Neutral Endopeptidase Causes Vasoconstriction of Human Resistance Vessels In Vivo. *Circulation* 97: 2323-2330 [\[Abstract\]](#) [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

BMJ 1996;312:1654 (29 June)

Statistics notes

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 study¹). The first child shows typical variation, having peak expiratory flow rates of 190, 220, 200, and 200 l/min.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

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

Child No	PEFR (l/min)				Mean	SD
	1st	2nd	3rd	4th		
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 $(\zeta)_w = (\text{square root})460.5 = 21.5$ 1/min. The calculation is easier using a program that performs one way analysis of variance² (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 $(\tau) = 0.16$, $P = 0.3$).

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

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

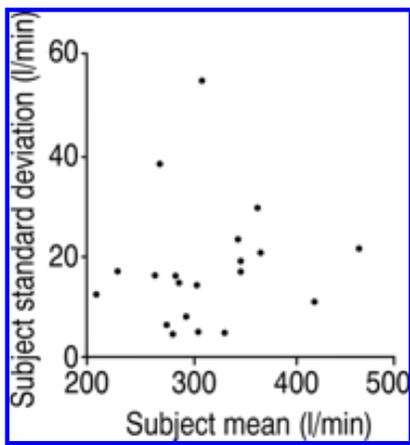


Fig 1--Individual subjects' standard deviations plotted against their means

View larger version (19K):

[\[in this window\]](#)

[\[in a new window\]](#)

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 d_i the within-subject standard deviation (ζ) 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 (ζ) 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 $(\text{square root})2 \times 1.96 (\zeta)w$ 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 $2.77 \times 2.5 = 60$ 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**.

1. Bland JM, Holland WW, Elliott A. The development of respiratory symptoms in a cohort of Kent schoolchildren. *Bull Physio-Path Resp* 1974;10:699-716.
2. Altman DG, Bland JM. Comparing several groups using analysis of variance. *BMJ* 1996;312:1472. [\[Full Text\]](#)

This article has been cited by other articles:

- Molyneux, P D, Tofts, P S, Fletcher, A, Gunn, B, Robinson, P, Gallagher, H, Moseley, I F, Barker, G J, Miller, D H (1998). Precision and reliability for measurement of change in MRI lesion volume in multiple sclerosis: a comparison of two computer assisted techniques. *J. Neurol. Neurosurg. Psychiatry* 65: 42-47 [\[Abstract\]](#) [\[Full text\]](#)
- Gawne-Cain, M L, O'Riordan, J I, Coles, A, Newell, B, Thompson, A J, Miller, D H (1998). MRI lesion volume measurement in multiple sclerosis and its correlation with disability: a comparison of fast fluid attenuated inversion recovery (fFLAIR) and spin echo sequences. *J. Neurol. Neurosurg. Psychiatry* 64: 197-203 [\[Abstract\]](#) [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

- Childs, C, Goldring, S, Tann, W, Hillier, V F (1998). Suprasternal Doppler ultrasound for assessment of stroke distance. *Arch. Dis. Child.* 79: 251-255 [\[Abstract\]](#) [\[Full text\]](#)
- Childs, C., Harrison, R., Hodkinson, C. (1999). Tympanic membrane temperature as a measure of core temperature. *Arch. Dis. Child.* 80: 262-266 [\[Abstract\]](#) [\[Full text\]](#)
- Bland, J M., Altman, D. G (1996). **Statistics Notes:** Measurement error and correlation coefficients. *BMJ* 313: 41-42 [\[Full text\]](#)

[Home](#)[Help](#)[Search/Archive](#)[Feedback](#)[Search Result](#)

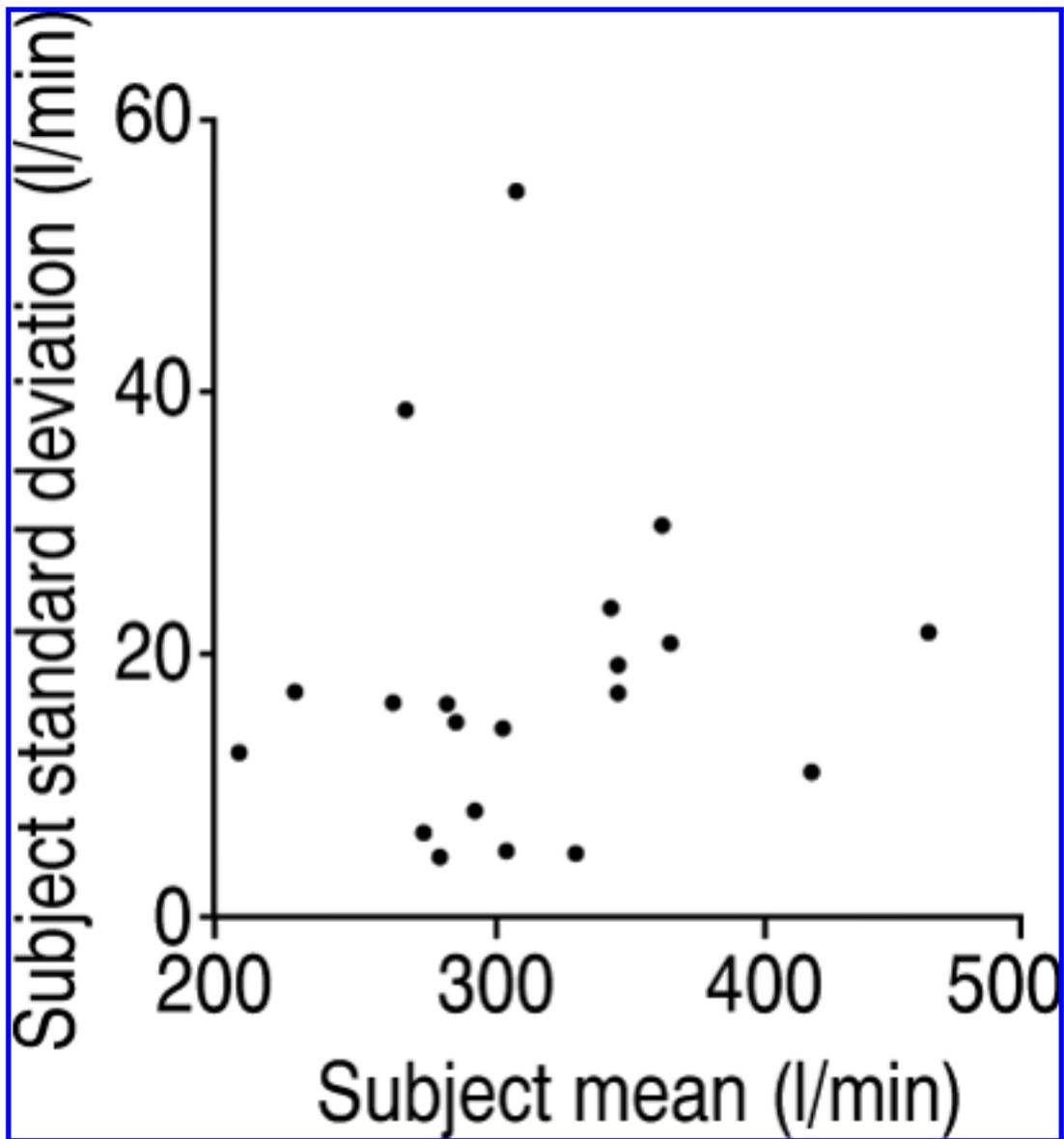


Fig 1--Individual subjects' standard deviations plotted against their means

[\[View larger version \(79K\)\]](#)

BMJ 1996;313:41-42 (6 July)

Education and debate

Statistics Notes: Measurement error and correlation coefficients

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, PO Box 777, Oxford OX3 7LF

Correspondence to: Professor Bland.

- ▶ **extra:** [Correction to Table 1](#)
- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Read](#) responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

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.

Table 1--Pairs of measurements of FEV₁ (litres) a few weeks apart from 20 Scottish schoolchildren, taken from a larger study (D Strachan, personal communication)

Subject No	Measurement		Subject No	Measurement	
	1st	2nd		1st	2nd
1	1.19	1.37	11	1.54	1.57
2	1.33	1.32	12	1.59	1.60
3	1.35	1.40	13	1.61	1.53
4	1.36	1.25	14	1.61	1.61
5	1.38	1.29	15	1.62	1.68
6	1.38	1.37	16	1.78	1.76
7	1.38	1.40	17	1.80	1.82
8	1.40	1.38	18	1.85	1.89
9	1.43	1.38	19	1.94	2.10
10	1.43	1.51	20	2.10	2.20

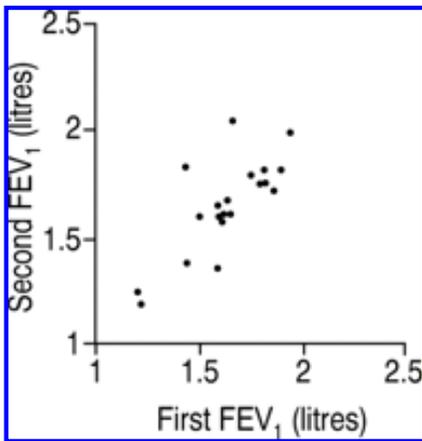


Fig 1--Measurements from pairs of observations plotted against each other

View larger version (15K):

[\[in this window\]](#)

[\[in a new window\]](#)

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

$$rI = 2 \times 1.52981 - 1.74651 / (2 - 1) \times 1.74651 = 0.75$$

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

Probability Source of variation (P)	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)
Children <0.0001	19	1.52981	0.08052	7.4
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 correlations. 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

1. Bland JM, Altman DG. Measurement error. *BMJ* 1996;312:1654. [\[Full Text\]](#)
2. Altman DG, Bland JM. *Comparing several groups using a analysis of variance* *BMJ* 1996;312:1472-3.

This article has been cited by other articles:

- Halligan, S (2002). Reproducibility, repeatability, correlation and measurement error. *Br J Radiol* 75: 193-194 [\[Full text\]](#)
- Bijur, P. E., Silver, W., Gallagher, E. J. (2001). Reliability of the Visual Analog Scale for Measurement of Acute Pain. *Acad Emerg Med* 8: 1153-1157 [\[Abstract\]](#) [\[Full text\]](#)
- Stone, B. D., Elias-Todd, T., Parrino, J., Ward, C., Walters, E. H., Faul, J. L., Burke, C. M., Poulter, L. W. (2001). EG-1 POSITIVE EOSINOPHILS IN ASTHMA. *Am J Respir Crit Care Med* 164: 171a-172 [\[Full text\]](#)
- Talvik, M., Nordstrom, A.-L., Nyberg, S., Olsson, H., Halldin, C., Farde, L. (2001). No Support for Regional Selectivity in Clozapine-Treated Patients: A PET Study With [11C]Raclopride and [11C]FLB 457. *Am. J. Psychiatry* 158: 926-930 [\[Abstract\]](#) [\[Full text\]](#)
- Nirmalan, M., Willard, T., Columb, M. O., Nightingale, P. (2001). Effect of changes in arterial-mixed venous oxygen content difference ($C(a-v)O_2$) on indices of pulmonary oxygen transfer in a model ARDS lung. *Br J Anaesth* 86: 477-485 [\[Abstract\]](#) [\[Full text\]](#)
- FAUL, J. L., DEMERS, E. A., BURKE, C. M., POULTER, L. W. (1999). The Reproducibility of Repeat Measures of Airway Inflammation in Stable Atopic Asthma. *Am J Respir Crit Care Med* 160: 1457-1461 [\[Abstract\]](#) [\[Full text\]](#)
- SALOME, C. M., ROBERTS, A. M., BROWN, N. J., DERMAND, J., MARKS, G. B., WOOLCOCK, A. J. (1999). Exhaled Nitric Oxide Measurements in a Population Sample of Young Adults. *Am J Respir Crit Care Med* 159: 911-916 [\[Abstract\]](#) [\[Full text\]](#)
- Bland, J M., Altman, D. G (1996). **Statistics Notes:** Measurement error proportional to the mean. *BMJ* 313: 106-106 [\[Full text\]](#)

- ▶ [extra: Correction to Table 1](#)
- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Read responses to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by: [Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when: [New articles cite this article](#)

Rapid Responses:

Read all [Rapid Responses](#)

Is the formula for the intraclass correlation coefficient correct?

Peter Schuck

bmj.com, 1 Sep 2000 [\[Full text\]](#)

precedence rules in formula

Michael D McStephen

bmj.com, 2 May 2002 [\[Full text\]](#)

Re: precedence rules in formula

J Martin Bland

bmj.com, 3 May 2002 [\[Full text\]](#)

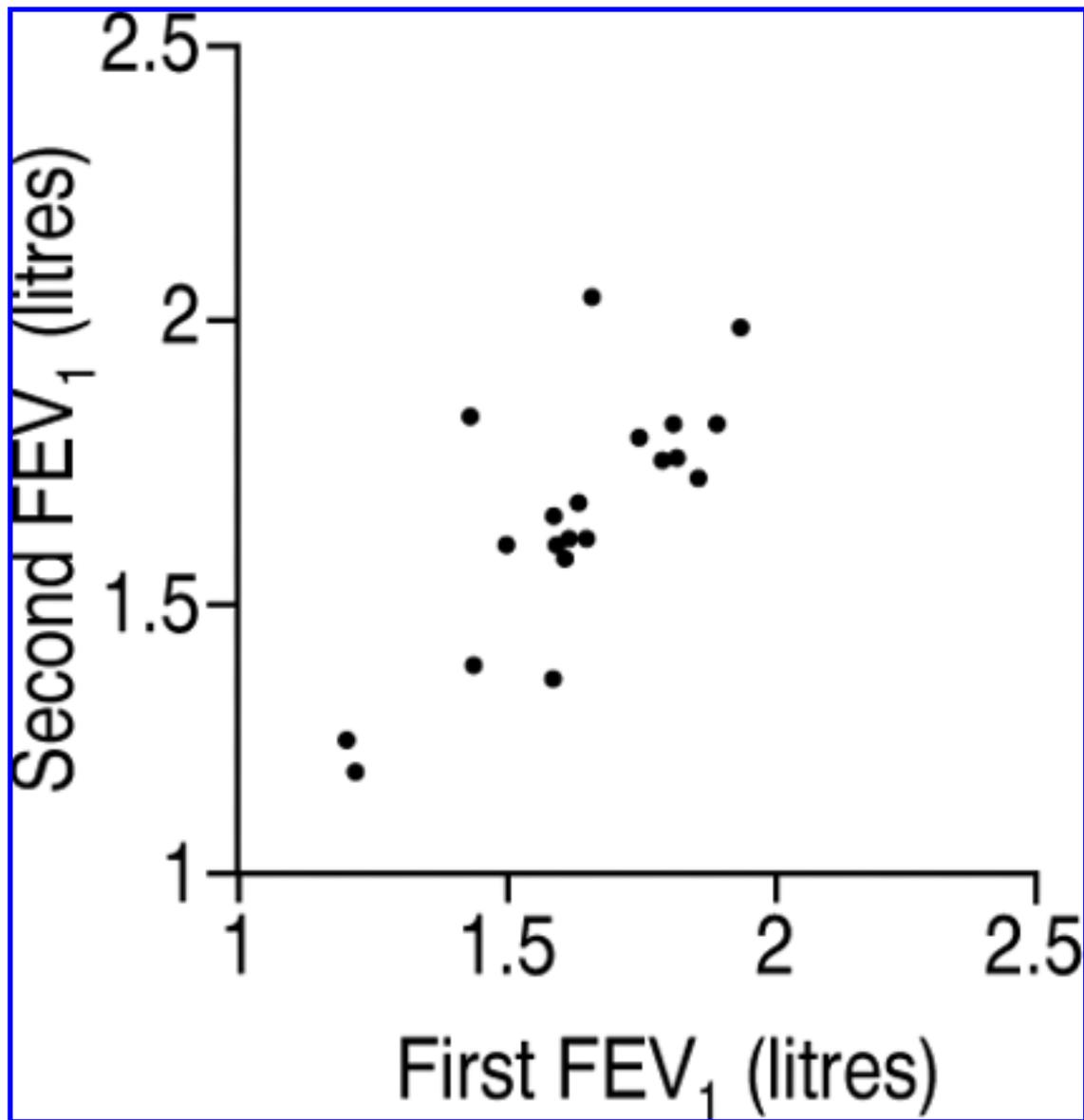


Fig 1--Measurements from pairs of observations plotted against each other

[\[View larger version \(63K\)\]](#)

Correct data for Statistics Note

There was an error in Statistics Note 22, Measurement error and correlation coefficients, Bland JM and Altman DG, 1996, *British Medical Journal* [313, 41-2](#).

The wrong table of data was printed. The correct data are:

Sub.	1st	2nd	Sub.	1st	2nd
1	1.20	1.24	11	1.62	1.68
2	1.21	1.19	12	1.64	1.61
3	1.42	1.83	13	1.65	2.05
4	1.43	1.38	14	1.74	1.80
5	1.49	1.60	15	1.78	1.76
6	1.58	1.36	16	1.80	1.76
7	1.58	1.65	17	1.80	1.82
8	1.59	1.60	18	1.85	1.73
9	1.60	1.58	19	1.88	1.82
10	1.61	1.61	20	1.92	2.00

- ▶ [Full Text of this article](#)
- ▶ [Email this article to a friend](#)
- ▶ **[Respond to this article](#)**
- ▶ Alert me when:
[New articles cite this article](#)

- ▶ [Full Text of this article](#)
- ▶ [Email this article to a friend](#)
- ▶ **[Respond to this article](#)**
- ▶ Alert me when:
[New articles cite this article](#)

BMJ 1996;313:106 (13 July)

Education and debate

Statistics Notes: Measurement error proportional to the mean

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, PO Box 777, Oxford OX3 7LF

Correspondence to: Professor Bland.

We often need to know the error with which measurements are made--for example, so that we can decide whether the change in a clinical observation represents a real change in a patient's condition. We have discussed previously the within-subject standard deviation as a practical index of measurement error.¹ We said that this approach should be used when the measurement error was not related to the magnitude of the measurement and recommended that we plot the subject standard deviation against the subject mean to check this. Table 1 shows some duplicate salivary cotinine measurements taken from a larger study. Figure 1 shows absolute subject difference against subject mean, which is equivalent to a standard deviation versus mean plot when we have only two measurements per subject.¹ If we are to use the within-subject standard deviation as an index of measurement error we need the subject standard deviation to be independent of the subject mean. Here, there is a clear relationship, the variability increasing with the magnitude. We can test this using a rank correlation coefficient if we wish; here Kendall's (tau) = 0.62, P = 0.0001. Under these circumstances a logarithmic transformation of the data almost always solves the problem, but we can check by plotting log standard deviation against log mean. For these data the slope is 0.9; as this is very close to 1 the subject standard deviation is roughly proportional to the subject mean and a log transformation is indicated.² Figure 2 shows the plot of absolute difference versus subject mean for the log transformed data. There is now no evidence of a relationship (Kendall's (tau) = 0.07, P = 0.7).

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

Table 1--Duplicate salivary cotinine measurements for a group of Scottish schoolchildren (ng/ml) (D Strachan, personal communication)

Subject No	Measurement 1st	Measurement 2nd	Subject No	Measurement 1st	Measurement 2nd
1	0.1	0.1	11	1.2	0.9
2	0.2	0.1	12	1.9	2.8
3	0.2	0.3	13	2.0	1.4
4	0.3	0.4	14	2.7	1.4
5	0.3	0.4	15	2.8	6.8
6	0.4	0.3	16	3.2	2.9
7	0.4	1.4	17	4.7	4.5
8	0.8	0.5	18	4.9	1.4
9	1.0	1.6	19	4.9	3.9
10	1.1	0.9	20	7.0	4.0

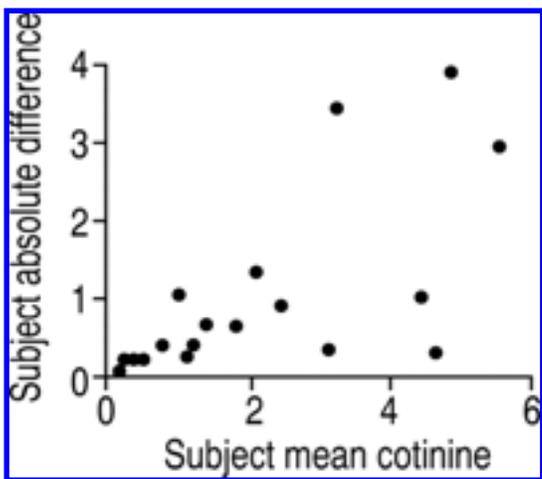


Fig 1--Absolute difference against mean for data in table [1](#)

View larger version (21K):

[\[in this window\]](#)

[\[in a new window\]](#)

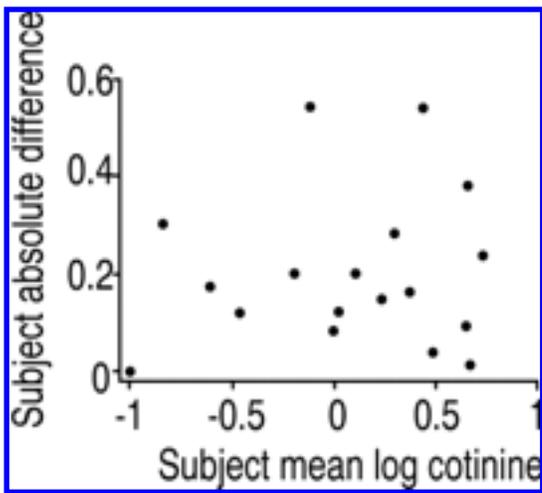


Fig 2--Absolute difference against mean after log10 transformation

View larger version (19K):

[\[in this window\]](#)

[\[in a new window\]](#)

As the variability is now independent of the magnitude of the measurement, we can calculate the within-subject standard deviation¹ as $(\sigma)_w = 0.175$. This is a standard deviation on the logarithmic scale, so we need to antilog it before we can interpret it easily. We will denote the antilog of $(\sigma)_w$ by $(\alpha)(\sigma_w)$. For the cotinine data, $(\alpha)(\sigma_w) = 1.496$.

When we antilog $(\sigma)_w$ we have a ratio, not a quantity measured in the units of the original data. This is because to calculate a standard deviation we subtract the mean from each observation. Subtracting on the logarithmic scale is equivalent to dividing on the natural scale.³ Dividing the observation by the mean in this way produces a dimensionless ratio. Hence $(\alpha)(\sigma_w)$ is not a standard deviation in the original units, but a related quantity sometimes referred to as the geometric standard deviation.

To estimate one standard deviation on either side of the observed value, we should multiply and divide by $(\alpha)(\sigma_w)$. The difference between a subject's measurement and the true value would be expected to be less than $1.96(\sigma)_w$ for 95% of observations.¹ To get the equivalent of 1.96 log scale standard deviations on either side of an observed value we would multiply and divide on the natural scale by $(\alpha)^{1.96}(\sigma_w)$ or approximately $(\alpha)^2(\sigma_w)$. This procedure gives limits which should include the subject's mean for 95% of observations. Thus for the cotinine data we would divide and multiply by $1.496^2 = 2.238$. A measurement of 2 ng/ml would tell us that the person's true value probably lies somewhere between $2/2.238 = 0.9$ and $2 \times 2.238 = 4.5$ ng/ml.

Multiplying on the natural scale is equivalent to adding on the log scale. Multiplying a subject's actual measurement by $(\alpha)(\sigma_w)$ is equivalent to adding one standard deviation on the log scale. Provided the standard deviation is not large compared with the level of the measurement, $(\alpha)(\sigma_w)-1$ is approximately equal to the standard deviation expressed as a proportion of the

measurement. The ratio of standard deviation to mean is called a coefficient of variation, and here $(\alpha)(\sigma/\mu)-1$ is the within-subject coefficient of variation.

For the cotinine data the estimated coefficient of variation is $1.496-1 = 0.496$ or 49.6%. This is rather too large for the approximation to be reliable.

The within-subject variability for salivary cotinine seems very large, but the possible range of values, from these very lightly exposed children to heavy smokers, is very wide and salivary cotinine is sufficiently precise to distinguish between many different levels of exposure. The precision of a method of measurement must be interpreted in the light of its intended use.

1. Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ* 1996;313:41-2. [\[Full Text\]](#)
2. Bland JM, Altman DG. Transforming data. *BMJ* 1996;312:770. [\[Full Text\]](#)
3. Bland JM, Altman DG. Logarithms. *BMJ* 1996;312:700. [\[Full Text\]](#)

This article has been cited by other articles:

- Salinas, M., Rosas, J., Iborra, J., Manero, H., Pascual, E. (1997). Comparison of manual and automated cell counts in EDTA preserved synovial fluids. Storage has little influence on the results. *Ann Rheum Dis* 56: 622-626 [\[Abstract\]](#) [\[Full text\]](#)
- Gawne-Cain, M L, O'Riordan, J I, Coles, A, Newell, B, Thompson, A J, Miller, D H (1998). MRI lesion volume measurement in multiple sclerosis and its correlation with disability: a comparison of fast fluid attenuated inversion recovery (fFLAIR) and spin echo sequences. *J. Neurol. Neurosurg. Psychiatry* 64: 197-203 [\[Abstract\]](#) [\[Full text\]](#)
- Gøtzsche, P. C, Hammarquist, C., Burr, M. (1998). House dust mite control measures in the management of asthma: meta-analysis. *BMJ* 317: 1105-1110 [\[Abstract\]](#) [\[Full text\]](#)
- Massé, J., Bland, J M, Doyle, J R, Doyle, J M (1997). Measurement error. *BMJ* 314: 147-147 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

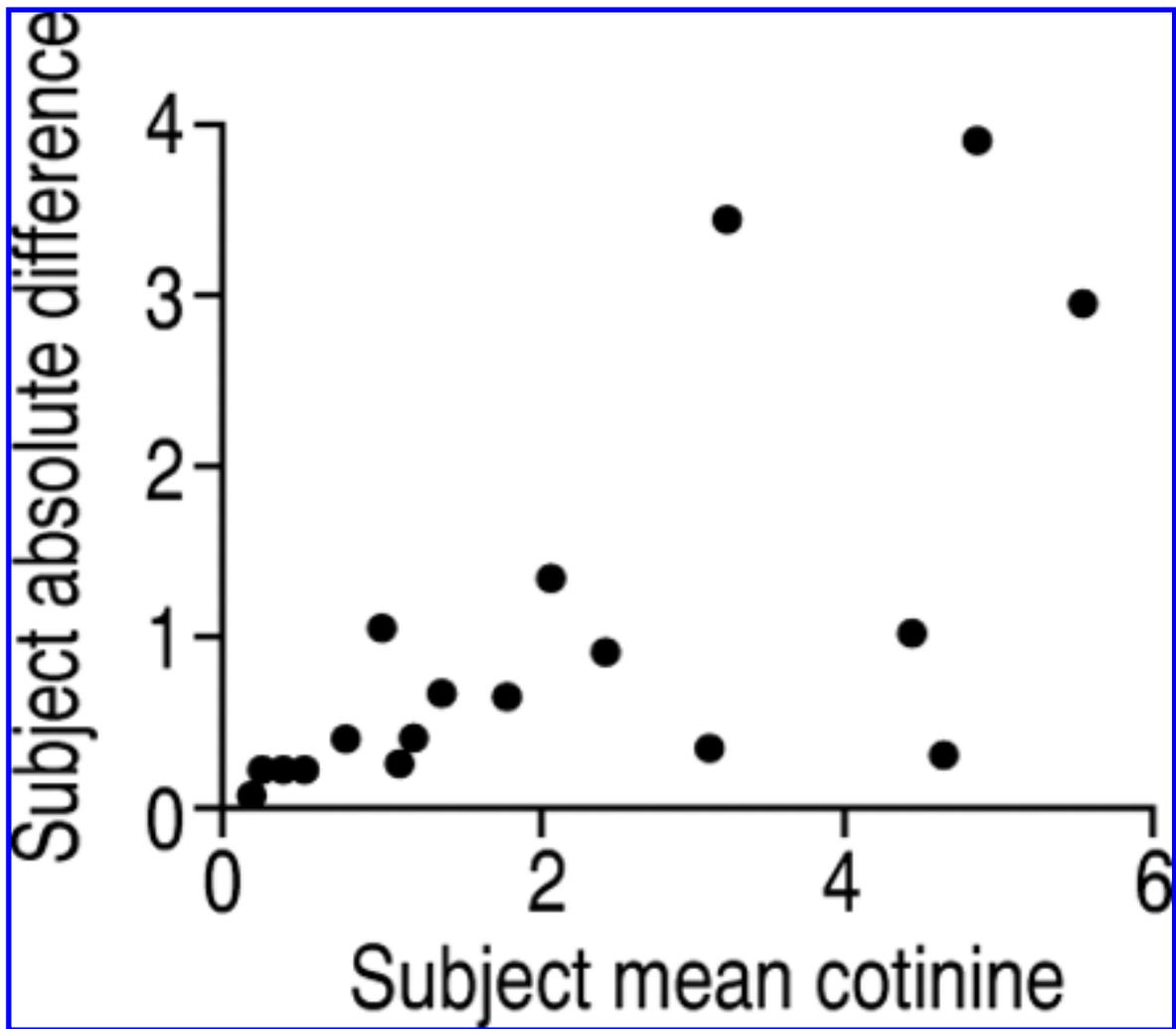


Fig 1--Absolute difference against mean for data in table [1](#)

[\[View larger version \(76K\)\]](#)

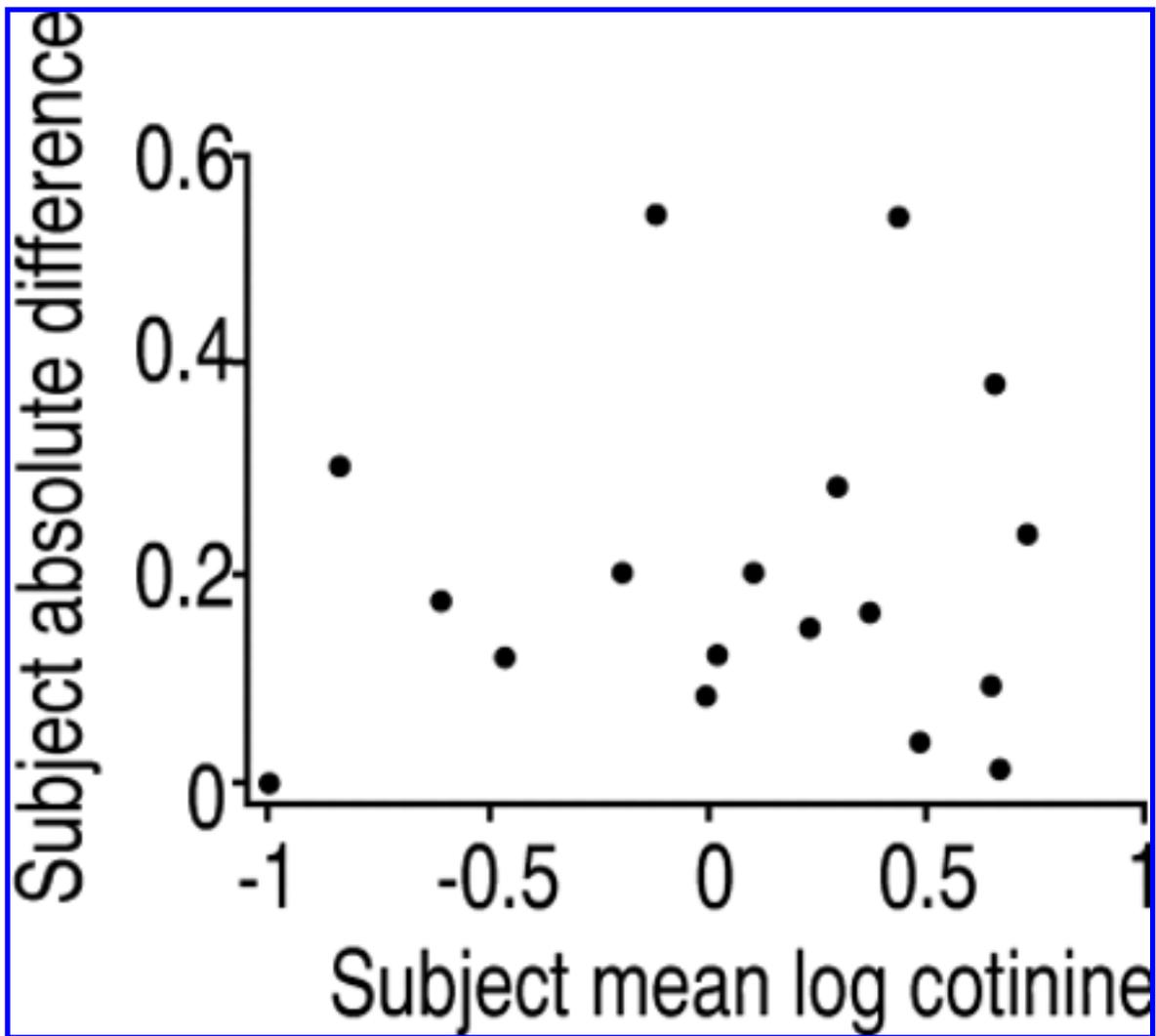


Fig 2--Absolute difference against mean after log10 transformation

[\[View larger version \(69K\)\]](#)

BMJ 1996;313:744 (21 September)

Correction

Statistics Notes: Measurement error proportional to the mean

A typesetting error occurred in Note 23 (13 July, p 106). Throughout the text the symbol (sigma) should have been s , to be consistent with the previous two notes. Also the first reference should have been to note 21 (on measurement error, republished above), not note 22 (on measurement error and correlation coefficients, 6 July, p 41).

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
 ▶ [New articles cite this article](#)
- ▶ Alert me when:
 ▶ [New articles cite this article](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
 ▶ [New articles cite this article](#)
- ▶ Alert me when:
 ▶ [New articles cite this article](#)

[Home](#)[Help](#)[Search/Archive](#)[Feedback](#)[Search Result](#)*BMJ* 1996;313:486 (24 August)

Education and debate

Statistics Notes: Interaction 1: heterogeneity of effects

Douglas G Altman, *head*,^a **John N S Matthews**, *senior lecturer in medical statistics*^b

^a ICRF Medical Statistics Group, Centre for Statistics in Medicine,

Institute of Health Sciences, PO Box 777, Oxford OX3 7LF, ^b Department of Medical Statistics, University of Newcastle, Newcastle upon Tyne NE2 4HH

Correspondence to: Mr Altman.

In several types of study we may want to examine the consistency of an observed relation across two or more subgroups of the individuals studied. For example, in a clinical trial we might want to know if the observed treatment difference is the same for young and old patients or for different stages of disease at presentation. In an epidemiological study we might want to know whether the observed relation between an exposure and disease is different among smokers and non-smokers

In such cases we are interested in examining whether one effect is modified by the value of another variable. This may be viewed as the examination of the heterogeneity of an observed effect, such as treatment benefit in a clinical trial, across subsets of individuals. The statistical term for heterogeneity of this type is interaction; the medical concept of synergy is the same thing.

While it may well be of interest to look for heterogeneity of effect, this is not always wise. In a controlled trial there are numerous subgroups which might be compared by splitting the patients according to sociodemographic or clinical categories at the start of the trial. In addition, for continuous variables such as age or blood pressure there are many ways of creating groups. Exploratory examination of many such subgroups is almost certain to throw up some spurious significant interactions, and in practice we cannot tell if a specific interaction is real or spurious. For example, in a randomised controlled trial comparing dexamethasone phosphate with placebo for preventing neonatal respiratory distress syndrome, the researchers found unexpectedly that the overall beneficial effect of the active treatment was present only in female infants.¹ Further studies would be needed to confirm the finding (or not). The refutation of such

▶ [Email this article to a friend](#)

▶ [Respond to this article](#)

▶ [PubMed citation](#)

▶ [Related articles in PubMed](#)

▶ [Download to Citation Manager](#)

▶ This article has been cited by [other articles](#)

▶ Search Medline for articles by:

[Altman, D. G](#) || [Matthews, J. N S](#)

▶ Alert me when:

[New articles cite this article](#)

unexpected observations is common, and indeed in this case the finding was not replicated in further studies.²

Likewise, we can investigate the interaction between any pair of variables in a regression model. With 10 variables there are 45 such potential interactions and much scope for being misled. So, although we do not necessarily believe that all effects are truly independent, in many cases it is reasonable not to examine any possible interactions. For example, Pocock et al found a negative association between tooth lead concentrations and IQ (intelligence quotient) in children aged 6.³ Exploratory analysis revealed a strong association among boys and little association among girls. They were rightly cautious in their interpretation as there had been no prior hypothesis about such an effect.

By contrast, when there is a specific prior suspicion of the existence of a particular interaction it is perfectly reasonable and desirable to examine it. A common example already mentioned is the interest in a possible difference of risk between smokers and non-smokers. For example, a study of Danish porcelain painters found that the adverse effects of cobalt exposure on lung function were more severe among non-smokers than smokers.⁴

Results of tests for interactions are likely to be convincing only if they were specified at the start of the study. In any study that presents subgroup analyses it is important to specify when and why the subgroups were chosen. Studies which present analyses without such justification can be difficult to interpret. For example, Penttinen found a significant excess of ischaemic heart disease in relation to back pain in farmers aged 30-49 and a non-significant difference in the opposite direction among those aged 50-66.⁵ He did not explain why this age division was made, nor did he note that there was no relation when the two age groups were considered together. Studies where subgroup definition has been guided by the data, for example concentrating on males born in October,⁶ should be based on statistical tests that account for any multiple comparisons that have been made⁷ and should be scientifically sensible; even then they should be treated with scepticism until confirmed in subsequent studies.

Problems of interpretation are exacerbated by incorrect analysis. We consider right and wrong ways to examine possible interactions in two subsequent **statistics notes**.

1. Collaborative Group on Antenatal Steroid Therapy. Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *Am J Obstet Gynecol* 1981;141:276-87. [[Medline](#)]
2. Crowley P, Chalmers I, Keirse MJNC. The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. *Br J Obstet Gynaecol* 1990;97:11-25. [[Medline](#)]
3. Pocock SJ, Ashby D, Smith MA. Lead exposure and children's intellectual performance. *Int J Epidemiol* 1987;16:57-67. [[Abstract](#)]
4. Raffn E, Mikkelsen S, Altman DG, Christensen JM, Groth S. Health effects due to occupational

exposure to cobalt blue dye among plate painters in a porcelain factory in Denmark. *Scand J Work Environ Health* 1988;14:378-84. [\[Medline\]](#)

5. Penttinen J. Back pain and risk of fatal ischaemic heart disease: 13 year follow up of Finnish farmers. *BMJ* 1994;309:1267-8. [\[Full Text\]](#)
6. Helgason T, Jonasson MR. Evidence for a food additive as a cause of ketosis-prone diabetes. *Lancet* 1981;ii:716-20.
7. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170. [\[Full Text\]](#)

This article has been cited by other articles:

- Matthews, J. N S, Altman, D. G (1996). **Statistics Notes:** Interaction 2: compare effect sizes not P values. *BMJ* 313: 808-808 [\[Full text\]](#)
- Matthews, J. N S, Altman, D. G (1996). **Statistics notes:** Interaction 3: How to examine heterogeneity. *BMJ* 313: 862-862 [\[Full text\]](#)
- Altman, D. G, Bland, J M. (1998). **Statistics Notes:** Generalisation and extrapolation. *BMJ* 317: 409-410 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Matthews, J. N S](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1996;313:744 (21 September)

Education and debate

Statistics Notes:
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 study¹). 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)			Mean	SD
1	190	220	200	200	12.58
2	220	200	240	230	17.08
3	260	260	240	280	16.33
4	210	300	280	265	38.60
5	270	265	280	270	6.29
6	280	280	270	275	4.79

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

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 sw .

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 $sw = (\text{square root})460.52 = 21.5$ 1/min. The calculation is easier using a program that performs one way analysis of variance² (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 $(\tau) = 0.16$, $P = 0.3$).

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	60	27631.25	460.52		
Total	79	312949.69			

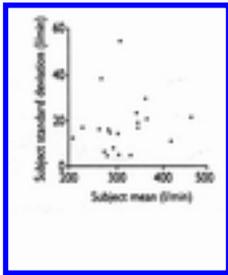


Fig 1--Individual subjects' standard deviations plotted against their means

View larger version (6K):

[\[in this window\]](#)

[\[in a new window\]](#)

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 d_i the within-subject standard deviation s_w is given by $s_w^2 = 1/2n(\sum d_i^2)$, where 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 s_w . The difference between a subject's measurement and the true value would be expected to be less than $1.96 s_w$ for 95% of observations. Another useful way of presenting measurement error is sometimes called the repeatability, which is $(\text{square root})^2 \times 1.96 s_w$ or $2.77 s_w$. The difference between two measurements for the same subject is expected to be less than $2.77 s_w$ for 95% of pairs of observations. For the data in table 1 the repeatability is $2.77 \times 21.5 = 60$ 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**.

1. Bland JM, Holland WW, Elliott A. The development of respiratory symptoms in a cohort of Kent schoolchildren. *Bull Physio-Path Resp* 1974;10:699-716.
2. Altman DG, Bland JM. Comparing several groups using analysis of variance. *BMJ* 1996;312:1472. [\[Full Text\]](#)

This article has been cited by other articles:

- Halligan, S (2002). Reproducibility, repeatability, correlation and measurement error. *Br J Radiol* 75: 193-194 [\[Full text\]](#)
- Dyer, C A E, Singh, S J, Stockley, R A, Sinclair, A J, Hill, S L (2002). The incremental shuttle walking test in elderly people with chronic airflow limitation. *Thorax* 57: 34-38 [\[Abstract\]](#) [\[Full text\]](#)
- Auleley, G.-R., Duche, A., Drape, J.-L., Dougados, M., Ravaud, P. (2001). Measurement of joint space width in hip osteoarthritis: influence of joint positioning and radiographic procedure. *Rheumatology* 40: 414-419 [\[Abstract\]](#) [\[Full text\]](#)
- Silvestri, M., Spallarossa, D., Battistini, E., Brusasco, V., Rossi, G. A (2000). Dissociation between exhaled nitric oxide and hyperresponsiveness in children with mild intermittent asthma. *Thorax* 55: 484-488 [\[Abstract\]](#) [\[Full text\]](#)
- Walsh, T. S., McLellan, S., Mackenzie, S. J., Lee, A. (1999). Hyperlactatemia and Pulmonary Lactate Production in Patients With Fulminant Hepatic Failure. *Chest* 116: 471-476 [\[Abstract\]](#) [\[Full text\]](#)
- Watt, V, Pickering, M, Wales, J K H (1998). A comparison of ultrasonic and mechanical stadiometry. *Arch. Dis. Child.* 78: 269-270 [\[Abstract\]](#) [\[Full text\]](#)
- Massé, J., Bland, J M, Doyle, J R, Doyle, J M (1997). Measurement error. *BMJ* 314: 147-147 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

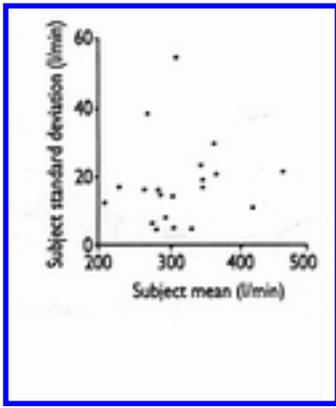


Fig 1--Individual subjects' standard deviations plotted against their means

[\[View larger version \(0K\)\]](#)

BMJ 1996;313:808 (28 September)

Education and debate

Statistics Notes: Interaction 2: compare effect sizes not P values

John N S Matthews, *senior lecturer in medical statistics*,^a
Douglas G Altman, *head*^b

^a Department of Medical Statistics, University of Newcastle, Newcastle upon Tyne NE2 4HH, ^b ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, PO Box 777, Oxford OX3 7LF

Correspondence to: Dr Matthews.

As we have previously described,¹ the statistical term interaction relates to the non-independence of the effects of two variables on the outcome of interest. For example, in a controlled trial comparing a new treatment with a standard treatment we may want to examine whether the observed benefit was the same for different subgroups of patients. A common approach to answering this question is to analyse the data separately in each subgroup. Here we illustrate this approach and explain why it is incorrect.

One of several subgroup analyses in a trial of antenatal steroids for preventing neonatal respiratory distress syndrome² was performed to see whether the effect of treatment was different in mothers who did or did not develop pre-eclampsia. Among mothers with preeclampsia 21.2% (7/33) of babies whose mothers were given dexamethasone developed neonatal respiratory distress syndrome compared with 27.3% (9/33) of babies whose mothers received placebo, giving $P = 0.57$. Among mothers who did not have pre-eclampsia 7.9% (21/267) of babies in the steroid group and 14.1% (37/262) of babies in the placebo group developed neonatal respiratory distress syndrome, giving $P = 0.021$.

There is a temptation to claim that the difference in P values establishes a difference between subgroups because "there is a treatment effect in mothers without pre-eclampsia but not in those with pre-eclampsia." This argument is false: the key to realising this is to recall that a statement such as $P = 0.57$ does not mean there is no difference, merely that we have found no evidence that there is a difference. A P value is a composite which depends not only on the size of an effect but also on how precisely the

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Matthews, J. N S](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

effect has been estimated (its standard error). So differences in P values can arise because of differences in effect sizes or differences in standard errors or a combination of the two.

This is well illustrated by the present example. If we measure treatment effect by the difference in percentages developing neonatal respiratory distress syndrome in the placebo and steroid groups, then the treatment effect among mothers with pre-eclampsia, namely $27.3 - 21.2 = 6.1\%$, is very close to the effect among mothers without pre-eclampsia, which is $14.1 - 7.9 = 6.2\%$. The difference in P values has arisen because only a small proportion of mothers had pre-eclampsia (66 out of 595), so the former treatment effect is estimated much less precisely than the latter.

Another example can be found in a study of the effect of vitamin D supplementation for preventing neonatal hypocalcaemia: expectant mothers were given either supplements or placebo and the serum calcium concentration of the baby was measured at one week.³ The benefit of supplementation was investigated separately for breast and bottle fed infants, and t tests to compare the treatment groups gave $P = 0.40$ in the breast fed group and $P = 0.0006$ in the bottle fed group.

As we have seen, it would be wrong to infer that vitamin D supplementation had a different effect on breast and bottle fed babies on the basis of these two P values: the correct way to proceed is to compare directly the sizes of the treatment effects. The effect of vitamin D supplementation can be measured by the difference in mean serum calcium concentrations between supplement and placebo groups and this gives effects of 0.04 mmol/l in the breast fed babies and 0.10 mmol/l in bottle fed babies. In order to interpret the difference in effect sizes, namely 0.06 mmol/l, we need to construct a confidence interval or perform a test of the null hypothesis that the true effect sizes are the same in each subgroup. A 95% confidence interval for the difference in effect sizes is - 0.05 to 0.17 mmol/l and a test of the null hypothesis gives $P = 0.28$. There is thus no evidence that the effect of vitamin D supplementation differs between breast and bottle fed infants. Comparing P values alone can be misleading.

Details of how to construct relevant confidence intervals and carry out associated tests are contained in a subsequent Statistics Note.

1. Altman DG, Matthews JNS. Interaction 1: heterogeneity of effects. *BMJ* 1996;313:486. [\[Full Text\]](#)
2. Collaborative Group on Antenatal Steroid Therapy. Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *Am J Obstet Gynecol* 1981;141:276-87. [\[Medline\]](#)
3. Cockburn F, Belton NR, Purvis RJ, Giles MM, Brown JK, Turner TL, et al. Maternal vitamin D intake and mineral metabolism in mothers and their newborn infants. *BMJ* 1980;281:11-4. [\[Medline\]](#)

This article has been cited by other articles:

- Matthews, J N S (1999). Sponsored trials do not necessarily give more-favourable results. *BMJ* 318: 1762a-1762 [\[Full text\]](#)
- Matthews, J. N S, Altman, D. G (1996). **Statistics notes:** Interaction 3: How to examine heterogeneity. *BMJ* 313: 862-862 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond](#) to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Matthews, J. N S](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1996;313:862 (5 October)

Statistics notes

Interaction 3: How to examine heterogeneity

John N S Matthews, senior lecturer in medical statistics,^a **Douglas G Altman**, head^b

^a Department of Medical Statistics, University of Newcastle, Newcastle upon Tyne NE2 4HH, ^b ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, PO Box 777, Oxford OX3 7LF

Correspondence to: Dr Matthews.

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by: [Matthews, J. N S](#) || [Altman, D. G](#)
- ▶ Alert me when: [New articles cite this article](#)

In preceding **Statistics Notes** we introduced the concept of interaction¹ and explained why a common approach to the assessment of interaction is incorrect.² In this note we give details of the correct approach using the same two examples.

In a study of the effect of maternal vitamin D supplementation on neonatal serum calcium concentrations³ the researchers were interested in the possible difference between the effect of supplementation on breast and bottle fed babies. We define the treatment effect in each feeding group to be the difference in the mean serum calcium concentration of babies receiving supplements and those receiving placebo in that group: the treatment means and observed effects in the feeding groups are given in table [1](#).

Table 1--Serum calcium concentrations (mmol/l) at 1 week in babies born to mothers given vitamin D supplements or placebo and analysed according to whether they were breast fed or bottle fed

	Breast fed		Bottle fed	
Serum calcium	Supplement	Placebo	Supplement	Placebo
Treatment mean	2.45	2.41	2.30	2.20
Standard error	0.036	0.032	0.022	0.019
No	64	102	169	285
Treatment effect		0.04		0.10
Standard error		0.048		0.029
P value		0.40		0.0006

The first step is to compute the difference between the two treatment effects--that is, $0.10 - 0.04 = 0.06$ mmol/l. The standard error of this difference is 0.056 mmol/l, found from the standard errors of the separate effects using the usual method for the standard error of a difference.⁴ This is the same method that provides the standard error of a treatment effect from the standard errors of the treatment means. The P value can found from the ratio of the difference to its standard error, namely $0.06/0.056 = 1.07$, again using standard methods,⁴ which gives $P = 0.28$, showing there is no evidence that the effects are different between the two feeding groups. An approximate 95% confidence interval can be found for the difference in the treatment effects in the usual way,⁴--that is,

as 0.06 +/- 1.96 x 0.056, or - 0.05 to 0.17 mmol/l.

A similar approach is adopted with a binary outcome measure. In a controlled trial of antenatal steroid therapy for neonatal respiratory distress syndrome 27.3% (9/33) of babies born to mothers with pre-eclampsia and 14.1% (37/262) of babies born to mothers without pre-eclampsia in the control group developed neonatal respiratory distress syndrome; the corresponding figures in the steroid group were 21.2% (7/33) and 7.9% (21/267) respectively.⁵ Once standard errors of each of these percentages have been found in the usual way⁴ the method for assessing an interaction between steroid therapy and mother's pre-eclampsia is the same as for continuous outcomes. The treatment effect in babies of mothers with pre-eclampsia is 27.3 - 21.2 = 6.1% (standard error 10.5%) and in babies born to unaffected mothers it is 14.1 - 7.9 = 6.2% (standard error 2.7%), so the difference in treatment effects is 6.2 - 6.1 = 0.1% (standard error 10.9%), from which the P value for the difference in treatment effects is P = 0.99. Thus there is no evidence in this trial that the effect of antenatal steroids depends on whether the mother suffered from pre-eclampsia: the 95% confidence interval for the difference in the treatment effects can also be constructed as before, giving 0.1 +/- 1.96 x 10.9 or - 21.3% to 21.5%.

1. Altman DG, Matthews JNS. Interaction 1: heterogeneity of effects. *BMJ* 1996;313:486. [[Full Text](#)]
2. Matthews JNS, Altman DG. Interaction 2: compare effect sizes not P values. *BMJ* 1996;313:808. [[Full Text](#)]
3. Cockburn F, Belton NR, Purvis RJ, Giles MM, Brown JK, Turner TL, et al. Maternal vitamin D intake and mineral metabolism in mothers and their newborn infants. *BMJ* 1980;281:11-4. [[Medline](#)]
4. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991:160-7.
5. Collaborative Group on Antenatal Steroid Therapy. Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *Am J Obstet Gynecol* 1981;141:276-87. [[Medline](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Matthews, J. N S](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)
[Help](#)
[Search/Archive](#)
[Feedback](#)
[Search Result](#)

BMJ 1996;313:1200 (9 November)

Education and debate

Statistics Notes: Detecting skewness from summary information

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, PO Box 777, Oxford OX3 7LF, ^b Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

Correspondence to: Mr Altman.

As we have noted before, many statistical methods of analysis assume that the data have a normal distribution.¹ When the data do not they can often be transformed to make them more normal.² Readers of published papers may wish to be reassured that the authors have carried out an appropriate analysis. When authors present data in the form of a histogram or scatter diagram then readers can see at a glance whether the distributional assumption is met. If, however, only summary statistics are presented--as is often the case--this is much more difficult. If the summary statistics include the range of the data then some idea of the distribution may be gained. For example, a range from 7 to 41 around a mean of 15 suggests that the data have positive skewness. However, as the range is based on the two most extreme (and hence atypical) values this inference is not reliable. Similar asymmetry affecting the lower and upper quartiles³ would be much more convincing evidence of a skewed distribution. Usually, however, the only summary statistics presented are the mean and either the standard deviation or standard error. Such information cannot show that the data are near to a normal distribution, but they can sometimes show that they are not.

There are two useful tricks. The normal distribution extends beyond two standard deviations either side of the mean. It follows that for measurements which must be positive (like most of those encountered in medicine) if the mean is smaller than twice the standard deviation the data are likely to be skewed. Table [1](#) shows urinary cotinine levels related to number of cigarettes smoked daily. Clearly the data must be highly skewed, as the mean is smaller than the standard deviation in each group. This aspect of the data was not apparent in the original paper, which gave just the means and standard errors. (We added the

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

standard deviations, derived simply as standard error \times (square root) n .) As a consequence, the use of t tests was not easily seen to be incorrect.

Table 1--Urinary cotinine excretion ($\mu\text{g}/\text{mg}$ creatinine) related to number of cigarettes smoked daily⁴

Cigarettes smoked per day	No in group	Mean	SE	SD
1-9	25	0.31	0.08	0.40
10-19	57	0.42	0.10	0.75
20-29	99	0.87	0.19	1.89
30-39	38	1.03	0.25	1.54
>40	28	1.56	0.57	3.02
Unspecified	25	0.56	0.16	0.80

The second indicator of skewness can be used when, as in table 1, there are data for several groups of individuals. As we have noted,² deviations from the normal distribution and a relation between the standard deviation and mean across groups often go together. If the standard deviation increases as the mean increases then this is a good indication that the data are positively skewed, and specifically that a log transformation may be needed.² There is a clear relation between mean and standard deviation for the cotinine data. As we have noted, log transformation often removes skewness and makes the standard deviations more similar.

In this example we can detect skewness from summary statistics, but we cannot tell what the effect of log transformation would have been. That requires the raw data.

1. Altman DG, Bland JM. The normal distribution. *BMJ* 1995;310:298. [\[Full Text\]](#)
2. Bland JM, Altman DG. Transforming data. *BMJ* 1996;312:770. [\[Full Text\]](#)
3. Altman DG, Bland JM. Quartiles, quintiles, centiles and other quantiles. *BMJ* 1994;309:996. [\[Full Text\]](#)
4. Matsukura S, Taminato T, Kitano N, Seino Y, Hamada H, Uchihashi M, et al. Effects of environmental tobacco smoke on urinary cotinine excretion in nonsmokers. *N Engl J Med* 1984;311:828-32. [\[Abstract\]](#)

This article has been cited by other articles:

- MAMMEN, P., GEORGE, C., THARYAN, P. (2001). Questions About Reasons for Living. *Am. J. Psychiatry* 158: 1331-1332 [[Full text](#)]
- SIMMONDS, S., COID, J., JOSEPH, P., MARRIOTT, S., TYRER, P. (2001). Community mental health team management in severe mental illness: a systematic review. *Br J Psychiatry* 178: 497-502 [[Abstract](#)] [[Full text](#)]
- BARBUI, C., HOTOPF, M. (2001). Amitriptyline v. the rest: still the leading antidepressant after 40 years of randomised controlled trials. *Br J Psychiatry* 178: 129-144 [[Abstract](#)] [[Full text](#)]
- Wahlbeck, K., Cheine, M., Essali, A., Adams, C. (1999). Evidence of Clozapine's Effectiveness in Schizophrenia: A Systematic Review and Meta-Analysis of Randomized Trials. *Am. J. Psychiatry* 156: 990-999 [[Abstract](#)] [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1997;314:572 (22 February)

General practice

Statistics notes: Cronbach's alpha

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 PO Box 777 Oxford OX3 7LF

Correspondence to: Professor Bland

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Read responses to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

Article

Many quantities of interest in medicine, such as anxiety or degree of handicap, are impossible to measure explicitly. Instead, we ask a series of questions and combine the answers into a single numerical value. Often this is done by simply adding a score from each answer. For example, the mini-HAQ is a measure of impairment developed for patients with cervical myelopathy.¹ This has 10 items (table 1) recording the degree of difficulty experienced in carrying out daily activities. Each item is scored from 1 (no difficulty) to 4 (can't do). The scores on the 10 items are summed to give the mini-HAQ score.

- ▲ [Top](#)
- [Article](#)
- ▼ [References](#)

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Table 1 Mini-HAQ scale in 249 severely impaired subjects

When items are used to form a scale they need to have internal consistency. The items should all measure the same thing, so they should be correlated with one another. A useful coefficient for assessing internal consistency is Cronbach's alpha.² The formula is:

[This figure is not available.]

where k is the number of items, si^2 is the variance of the i th item and sT^2 is the variance of the total score formed by summing all the items. If the items are not simply added to make the score, but first multiplied by weighting coefficients, we multiply the item by its coefficient before calculating the variance si^2 . Clearly, we must have at least two items—that is $k > 1$, or α will be undefined.

The coefficient works because the variance of the sum of a group of independent variables is the sum of their variances. If the variables are positively correlated, the variance of the sum will be increased. If the items making up the score are all identical and so perfectly correlated, all the si^2 will be equal and $sT^2 = k^2 si^2$, so that $\sum si^2 / sT^2 = 1/k$ and $\alpha = 1$. On the other hand, if the items are all independent, then $sT^2 = \sum si^2$ and $\alpha = 0$. Thus α will be 1 if the items are all the same and 0 if none is related to another.

For the mini-HAQ example, the standard deviations of each item and the total score are shown in the table. We have $\sum si^2 = 11.16$, $sT^2 = 77.44$, and $k = 10$. Putting these into the equation, we have

[This figure is not available.]

which indicates a high degree of consistency.

For scales which are used as research tools to compare groups, α may be less than in the clinical situation, when the value of the scale for an individual is of interest. For comparing groups, α values of 0.7 to 0.8 are regarded as satisfactory. For the clinical application, much higher values of α are needed. The minimum is 0.90, and $\alpha=0.95$, as here, is desirable.

In a recent example, McKinley *et al* devised a questionnaire to measure patient satisfaction with calls made by general practitioners out of hours.³ This included eight separate scores, which they interpreted as measuring constructs such as satisfaction with communication and management, satisfaction with doctor's attitude, etc. They quoted α for each score, ranging from 0.61 to 0.88. They conclude that the questionnaire has satisfactory internal validity, as five of the eight scores had $\alpha > 0.7$. In this issue Bosma *et al* report similar values, from 0.67 to 0.84, for assessments of three characteristics of the work

environment.⁴

Cronbach's alpha has a direct interpretation. The items in our test are only some of the many possible items which could be used to make the total score. If we were to choose two random samples of k of these possible items, we would have two different scores each made up of k items. The expected correlation between these scores is α .

References

[▲ Top](#)
[▲ Article](#)
 ■ [References](#)

1. Casey ATH, Crockard HA, Bland JM, Stevens J, Moskovich R, Ransford AO. Development of a functional scoring system for rheumatoid arthritis patients with cervical myelopathy *Ann Rheum Dis* (in press).
2. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297-333.
3. McKinley RK, Manku-Scott T, Hastings AM, French DP, Baker R. Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the United Kingdom: development of a patient questionnaire. *BMJ* 1997;314:193-8. [[Abstract/Full Text](#)]
4. Bosma H, Marmot MG, Hemingway H, Nicholson AC, Brunner E, Stansfield SA. Low job control and risk of coronary heart disease in Whitehall II (prospective cohort) study. *BMJ* 1997;314:558-65.

This article has been cited by other articles:

- Meakin, R., Weinman, J. (2002). The 'Medical Interview Satisfaction Scale' (MISS-21) adapted for British general practice. *Fam. Pract.* 19: 257-263 [[Abstract](#)] [[Full text](#)]
- Bland, J M., Altman, D. G (2002). **Statistics Notes**: Validating scales and indexes. *BMJ* 324: 606-607 [[Full text](#)]
- Marinus, J, Ramaker, C, van Hilten, J J, Stiggelbout, A M (2002). Health related quality of life in Parkinson's disease: a systematic review of disease specific instruments. *J. Neurol. Neurosurg. Psychiatry* 72: 241-248 [[Abstract](#)] [[Full text](#)]
- Immer, F. F., Krahenbuhl, E., Immer-Bansi, A. S., Berdat, P. A., Kipfer, B., Eckstein, F. S., Saner, H., Carrel, T. P. (2002). Quality of life after interventions on the thoracic aorta with deep

- ▶ [Email this article to a friend](#)
- ▶ [Respond](#) to this article
- ▶ [Read](#) responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Bland, J M.](#) || [Altman, D. G](#)
- ▶ Alert me when:
[New articles cite this article](#)

hypothermic circulatory arrest. *Eur J Cardiothorac Surg* 21: 10-14 [\[Abstract\]](#) [\[Full text\]](#)

- Iglesias, C.P., Birks, Y.F., Torgerson, D.J. (2001). Improving the measurement of quality of life in older people: the York SF-12. *QJ Med* 94: 695-698 [\[Abstract\]](#) [\[Full text\]](#)
- Marshall, M. N (1999). How well do GPs and hospital consultants work together? A survey of the professional relationship. *Fam. Pract.* 16: 33-38 [\[Abstract\]](#) [\[Full text\]](#)
- Olsson, C., Thelin, S. (1999). Quality of life in survivors of thoracic aortic surgery. *Ann. Thorac. Surg.* 67: 1262-1267 [\[Abstract\]](#) [\[Full text\]](#)
- Grunfeld, E. A., Morland, A. B., Bronstein, A. M., Gresty, M. A. (2000). Adaptation to oscillopsia: A psychophysical and questionnaire investigation. *Brain* 123: 277-290 [\[Abstract\]](#) [\[Full text\]](#)
- Hautvast, J. L., Tolboom, J. J., Kafwembe, E. M, Musonda, R. M, Mwanakasale, V., van Staveren, W. A, van 't Hof, M. A, Sauerwein, R. W, Willems, J. L, Monnens, L. A. (2000). Severe linear growth retardation in rural Zambian children: the influence of biological variables1. *Am. J. Clin. Nutr.* 71: 550-559 [\[Abstract\]](#) [\[Full text\]](#)
- MAZEIKA, P K (2000). Quality of life four years after myocardial infarction: short form 36 scores compared with a normal population. *Heart* 83: 103b-103 [\[Full text\]](#)
- Terwee, C. B, Gerding, M. N, Dekker, F. W, Prummel, M. F, Wiersinga, W. M (1998). Development of a disease specific quality of life questionnaire for patients with Graves' ophthalmopathy: the GO-QOL. *Br. J. Ophthalmol.* 82: 773-779 [\[Abstract\]](#) [\[Full text\]](#)

Rapid Responses:

Read all [Rapid Responses](#)

No formulas in electronic version

Stefan Lange

bmj.com, 12 Jan 2001 [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

Table 1 Mini-HAQ scale in 249 severely impaired subjects

Item	Mean score	SD of score <i>si</i>
Stand	2.96	1.04
Get out of bed	2.57	1.11
Cut meat	2.91	1.12
Hold cup	2.41	1.06
Walk	2.64	1.04
Climb stairs	3.06	1.04
Wash	3.25	1.01
Use toilet	2.59	1.09
Open a jar	2.86	1.02
Enter/leave car	2.80	1.03
Mini-HAQ	28.06	$sT = 8.80$

[Home](#)[Help](#)[Search/Archive](#)[Feedback](#)[Search Result](#)*BMJ* 1997;314:1874 (28 June)

General practice

Statistics Notes: Units of analysis

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ Article

In clinical studies the focus of interest is almost always the patient. If we carry out a randomised trial to compare two treatments we are interested in comparing the outcomes of patients who received each of the treatments. In some conditions several measurements will be taken on the same patient, but the focus of interest remains the patient. Failure to recognise this fact results in multiple counting of individual patients and can seriously distort the results. We explain this error below. Its frequency in medical research is indicated by the whole chapter devoted to it in Andersen's classic compilation.¹

- ▲ [Top](#)
- [Article](#)
- ▼ [References](#)

The simplest case is when researchers study a part of the human anatomy which is, so to speak, in duplicate: eyes, ears, arms, etc. At the other extreme very many measurements can be taken on a single patient. Such data arise frequently in dentistry, with measurements made on each tooth, or even each face of each tooth, and in rheumatology, in which pain or mobility may be assessed for each joint of each finger. In statistical terminology the patient is the sampling unit (or unit of investigation) and thus should be the unit of analysis.

There are two related consequences of ignoring the fact that the data include multiple observations on the same individuals. Firstly, this procedure violates the widespread assumption of statistical analyses that the separate data values should be independent. Secondly, the sample size is inflated, sometimes dramatically so, which may lead to spurious statistical significance.

Inflated samples

To take a simple case, we may wish to compare the blood pressures of two groups of 30 patients. If we measured blood pressure on each arm of each patient we could double the number of observations but not the amount of information, as the two pressures from each patient will be very similar. The use of the t test to compare the two sets of 60 observations is invalid. Andersen¹ presented data from a randomised double blind crossover trial of ketoprofen and aspirin in the treatment of rheumatoid arthritis. An impressive P value of 0.00000001 was obtained from an analysis of 3944 observations, but these were obtained from only 58 patients. Such errors are not rare. In a review of 196 randomised trials of non-steroidal anti-inflammatory agents Gøtzsche found that 63% of reports used the wrong units of analysis.²

We previously discussed a similar fallacy arising in the use of correlation coefficients, when multiple observations from each individual produced a spurious increase in the sample size and a corresponding spurious "significant" relationship.³ We suggested techniques to analyse such data when the focus was either the variation within subjects⁴ or between subjects.⁵

There is nothing wrong in collecting such data; indeed the use of multiple observations can often improve the statistical power of a study. But such studies need to be analysed correctly. The simplest approach is to collapse all the data for an individual into a summary measure.⁶ For example, we could validly analyse the mean of the two blood pressure values for each patient. Alternatively, we can use a statistical method which explicitly takes account of the multiplicity. With well designed studies we may be able to use analysis of variance. A more complex general approach is multilevel modelling,⁷ which is not available in standard statistical software and may be difficult to apply and interpret.

Take account of multiplicity

The same objection applies to the use of multiple measurements made on different occasions. Here too the sampling unit is the patient, and thus the unit of analysis should also be the patient.² A further feature of this type of study is that in some situations the number of measurements made on a patient may itself carry prognostic information. For example, repeat measurements may be made only if there is some clinical concern—for example, fetal ultrasound measurements in pregnancy. To treat all these measurements as independent is clearly wrong, but bias is introduced too when those with more data are systematically different from those with single observations. An extreme example of this phenomenon occurs when analysing multiple hospital admissions for a potentially fatal condition.¹ Those with more than one admission must have survived the first admission.

Failure to carry out the correct analysis can lead to problems of interpretation too. Commenting on one

trial, Andersen observed, "This trial resulted in the apparent conclusion that after 1 year 22% of the patients, but only 16% of the legs, have expired."¹

Similar problems arise when we cannot sample individual patients directly but choose a sample of hospitals, wards, or general practices and then obtain data for all or a subsample of the patients within these groups. Here analysis of data for individual patients leads to the errors described above. We consider this type of study in forthcoming **Statistics Notes**.

References

1. Andersen B. *Methodological errors in medical research*. Oxford: Blackwell, 1990.
2. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clin Trials* 1989;10:31-56.
3. Bland JM, Altman DG. Correlation, regression, and repeated data. *BMJ* 1994;308:896. [\[Full Text\]](#)
4. Bland JM, Altman DG. Calculating correlation coefficients with repeated observations. Part 1: correlation within subjects. *BMJ* 1995;310:446. [\[Full Text\]](#)
5. Bland JM, Altman DG. Calculating correlation coefficients with repeated observations. Part 2: correlation between subjects. *BMJ* 1995;310:633. [Correction *BMJ* 1996;312:572] [\[Full Text\]](#)
6. Matthews JNS, Altman DG, Campbell MJ, Royston JP. Analysis of serial measurements in medical research. *BMJ* 1990;300:230-235.
7. Goldstein H. *Multi-level statistical models*. 2nd ed. London: Edward Arnold, 1995.

[▲ Top](#)
[▲ Article](#)
 ▪ [References](#)

This article has been cited by other articles:

- Wears, R. L. (2002). Advanced Statistics: Statistical Methods for Analyzing Cluster and Cluster-randomized Data. *Acad Emerg Med* 9: 330-341 [\[Abstract\]](#) [\[Full text\]](#)
- Stead, L. F, Lancaster, T. (2000). A systematic review of interventions for preventing tobacco sales to minors. *Tob Control* 9: 169-176 [\[Abstract\]](#) [\[Full text\]](#)
- Pandit, J. J., Bree, S., Dillon, P., Elcock, D., McLaren, I. D., Crider, B. (2000). A Comparison of Superficial Versus Combined (Superficial and Deep) Cervical Plexus Block for Carotid

▶ [Email this article to a friend](#)
 ▶ [Respond to this article](#)
 ▶ [PubMed citation](#)
 ▶ [Related articles in PubMed](#)
 ▶ [Download to Citation Manager](#)
 ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
 ▶ Alert me when:
[New articles cite this article](#)

Endarterectomy: A Prospective, Randomized Study. *Anesth Analg* 91: 781-786

[\[Abstract\]](#) [\[Full text\]](#)

- Kerry, S. M, Bland, J M. (1998). Analysis of a trial randomised in clusters. *BMJ* 316: 54-54 [\[Full text\]](#)
- Altman, D. G (1997). Study to predict which elderly patients will fall shows difficulties in deriving and validating a model. *BMJ* 315: 1309-1309 [\[Full text\]](#)
- Bland, J M., Kerry, S. M (1997). Trials randomised in clusters. *BMJ* 315: 600-600 [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

BMJ 1998;316:549 (14 February)

Education and debate

Statistics notes: Sample size in cluster randomisation

Sally M Kerry, *statistician*,^a J Martin Bland,
professor of medical statistics^b

^a Division of General Practice and Primary Care, St George's Hospital Medical School, London SW17 0RE,

^b Department of Public Health Sciences

Correspondence to: Mrs Kerry

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Read](#) responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Kerry, S. M](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

- ▶ Collections under which this article appears:
[Other Statistics and Research Methods: descriptions](#)

Abstract

Techniques for estimating sample size for randomised trials are well established,^{1 2} but most texts do not discuss sample size for trials which randomise groups (clusters) of people rather than individuals. For example, in a study of different preparations to control head lice all children in the same class were allocated to receive the same preparation. This was done to avoid contaminating the treatment groups through contact with control children in the same class.³ The children in the class cannot be considered independent of one another and the analysis should take this into account.^{4 5} There will be some loss of power due to randomising by cluster rather than individual and this should be reflected in the sample size calculations. Here we describe sample size calculations for a cluster randomised trial.

For a conventional randomised trial assessing the difference between two sample means the number of subjects required in each group, n , to detect a difference of d using a significance level of 5% and a power of 90% is given by $n=21s^2/d^2$ where s is the standard deviation of the outcome measure. Other values of power and significance can be used.¹

- ▲ [Top](#)
- [Abstract](#)
- ▼ [References](#)

For a trial using cluster randomisation we need to take the design into account. For a continuous outcome measurement such as serum cholesterol values, a simple method of analysis is based on the mean of the observations for all subjects in the cluster and compares these means between the treatment groups. We will denote the variance of observations within one cluster by sw^2 and assume that this variance is the same for all clusters. If there are m subjects in each cluster then the variance of a single sample mean is sw^2/m . The true cluster mean (unknown) will vary from cluster to cluster, with variance sc^2 . The observed variance of the cluster means will be the sum of the variance between clusters and the variance within clusters—that is, variance of outcome= sc^2+sw^2/m . Hence we can replace s^2 by sc^2+sw^2/m in the formula for sample size above to obtain the number of clusters required in each intervention group. To do this we need estimates of sc^2 and sw^2 .

For example, in a proposed study of a behavioural intervention in general practice to lower cholesterol concentrations practices were to be randomised into two groups, one to offer intensive dietary intervention by practice nurses using a behavioural approach and the other to offer usual general practice care. The outcome measure would be mean cholesterol values in patients attending each practice one year later. Estimates of between practice variance and within practice variance were obtained from the Medical Research Council thrombosis prevention trial⁶ and were $sc^2=0.0046$ and $sw^2=1.28$ respectively. The minimum difference considered to be clinically relevant was 0.1 mmol/l. If we recruit 50 patients per practice, we would have $s^2=sc^2+sw^2/m=0.0046+1.28/50=0.0302$. The number of practices is given by $n=21 \times 0.0302 / 0.1^2 = 63$ in each group. We would require 63 practices in each group to detect a difference of 0.1 mmol/l with a power of 90% using a 5% significance level—a total of 3150 patients in each group.

It can be seen from the formula for the variance of the outcome that when the number of patients within a practice, m , is very large, sw^2/m will be very small and so the overall variance is roughly the same as the variance between practices. In this situation, increasing the number of patients per practice will not increase the power of the study. The [1](#) shows the number of practices required for different values of m , the number of subjects per practice. In all situations the total number of subjects required is greater than if simple random allocation had been used.

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Total number of practices required to detect a difference of 0.1 mmol/l cholesterol with 90% power at 5% significance level

The ratio of the total number of subjects required using cluster randomisation to the number required

using simple randomisation is called the design effect. Thus a cluster randomised trial which has a large design effect will require many more subjects than a trial of the same intervention which randomises individuals. As the number of patients per practice increases so does the design effect. In the [1](#), the design effect is very small when m is less than 10. This would involve recruiting a total of 558 practices, and the nature of the intervention and difficulties in recruiting practices made this impractical. Thus it was decided to recruit fewer practices. The design effect of using 126 practices with 50 patients from each practice was 1.17. This design requires the total sample size to be inflated by 17%. If the study involves training practice based staff it may be cost effective to reduce the number of practices even further. If we chose to use 32 practices then we would need 500 patients from each practice and the design effect would be 2.98. Thus the cluster design with 32 practices would require the total sample size to be trebled to maintain the same level of power.

We shall discuss the use of the intracluster correlation coefficient in these calculations in a future statistics note.

References

1. Florey C du V. Sample size for beginners. *BMJ* 1993;306:1181-4. [\[Medline\]](#)
2. Machin D, Campbell MJ. *Statistical tables for the design of clinical trials*. Oxford: Blackwell, 1987.
3. Chosidow O, Chastang C, Brue C, Bouvet E, Izri M, Monteny N, et al. Controlled study of malathion and d-phenothrin lotions for *Pediculus humanus* var *capitis*-infested schoolchildren. *Lancet* 1994;344:1724-7.
4. Bland JM, Kerry SM. Trials randomised in clusters. *BMJ* 1997;315:600. [\[Full Text\]](#)
5. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. *BMJ* 1998;316:54. [\[Full Text\]](#)
6. Meade TW, Roderick PJ, Brennan PJ, Wilkes HC, Kelleher CC. Extracranial bleeding and other symptoms due to low dose aspirin and low intensity oral anticoagulation. *Thromb Haemostasis* 1992;68:1-6. [\[Medline\]](#)

[▲ Top](#)
[▲ Abstract](#)
 ▪ **References**

This article has been cited by other articles:

- Wears, R. L. (2002). Advanced Statistics: Statistical Methods for Analyzing Cluster and Cluster-randomized Data. *Acad Emerg Med* 9: 330-341 [\[Abstract\]](#) [\[Full text\]](#)
- Peters, T. J, Graham, A., Salisbury, C., Moore, L., Underwood, M., Eldridge, S., Gibson, P. G, Shah, S., Sindhusake, D., Wang, H., Peat, J. K, Henry, R. L (2001). Peer led programme for

asthma education in adolescents. *BMJ* 323: 110-110 [\[Full text\]](#)

- Tai, S. S., Iliffe, S. (2000). Considerations for the design and analysis of experimental studies in physical activity and exercise promotion: advantages of the randomised controlled trial. *Br J Sports Med* 34: 220-224 [\[Full text\]](#)
- Wilson, S., Delaney, B. C, Roalfe, A., Roberts, L., Redman, V., Wearn, A. M, Hobbs, F D R. (2000). Randomised controlled trials in primary care: case study. *BMJ* 321: 24-27 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Read](#) responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Kerry, S. M](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

- ▶ Collections under which this article appears:
[Other Statistics and Research Methods: descriptions](#)

Rapid Responses:

Read all [Rapid Responses](#)

Intracluster correlation underestimated?

Johannes C van der Wouden

bmj.com, 26 Apr 1998 [\[Full text\]](#)

Re: Intracluster correlation underestimated?

J Martin Bland, et al.

bmj.com, 5 Jun 1998 [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

Total number of practices required to detect a difference of 0.1 mmol/l cholesterol with 90% power at 5% significance level

No of patients per practice (<i>m</i>)	Standard deviation	No of practices	No of patients	Design effect
10	0.364	558	5 580	1.04
25	0.236	234	5 850	1.09
50	0.173	126	6 300	1.17
100	0.132	74	7 400	1.38
500	0.085	32	16 000	2.98
No needed with individual randomisation			5 364	1.00

BMJ 1998;316:1455-1460 (9 May)

Education and debate

Statistics notes

The intracluster correlation coefficient in cluster randomisation

Sally M Kerry, *statistician*,^a J Martin Bland, *professor of medical statistics*.^b

^a Division of General Practice and Primary Care, St George's Hospital Medical School, London SW17 0RE, ^b Department of Public Health Sciences

Correspondence to: Mrs Kerry

We have described the calculation of sample size when subjects are randomised in groups or clusters in terms of two variances—the variance of observations taken from individuals in the same cluster, s_w^2 , and the variance of true cluster means, s_c^2 .¹ We described how such a study could be analysed using the sample cluster means. The variance of such means would be $s_c^2 + s_w^2/m$, where m is the number of subjects in a cluster. We used this to estimate the sample size needed for a cluster randomised trial.

This sum of two components of variance is analogous to what happens with measurement error, where we have the variance within the subject, also denoted by s_w^2 , and between subjects (s_b^2).² One way of summarising the relation between these two components is the intraclass correlation coefficient, the correlation which we expect between pairs of observations made on the same subject. This is equal to $s_b^2/(s_b^2 + s_w^2)$.² We can calculate a similar intraclass correlation coefficient between our clusters, $r_I = s_c^2/(s_c^2 + s_w^2)$. This is also called the intracluster correlation coefficient.

For cholesterol concentration in the Medical Research Council thrombosis prevention trial the two

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Kerry, S. M](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

- ▶ Collections under which this article appears:
[Randomized Controlled Trials: descriptions](#)
[Other Statistics and Research Methods: descriptions](#)

components of variance were $s_w^2=1.28$ and $s_c^2=0.0046$.^{1 3} This gives the intraclass correlation coefficient $r_I=0.0046/(0.0046+1.28)=0.0036$. Such intraclass correlations are typically small. This trial had an intervention aimed directly at the patient and an outcome measurement for which the variance between practices is low compared with the variability between patients within a practice. Studies where the intervention is aimed at changing the doctor's behaviour may have a greater intraclass correlation. For example, in a trial of guidelines to improve the appropriateness of general practitioners' referrals for x ray examinations, the intraclass correlation was 0.0190.^{4 5} We might expect the intraclass correlation to be higher in a trial where the intervention is directed at the doctor rather than the patient, because it includes the variation in the doctors' responses.

The design effect is the ratio of the total number of subjects required using cluster randomisation to the number required using individual randomisation.¹ It can be presented neatly in terms of the intraclass correlation and the number in a single cluster, m : $D=1+(m-1)r_I$. If there is only one observation per cluster, $m=1$ and the design effect is 1.0 and the two designs are the same. Otherwise, the larger the intraclass correlation—that is, the more important the variation between clusters is, the bigger the design effect and the more subjects we will need to get the same power as a simply randomised study. Even a small intraclass correlation will have an impact if the cluster size is large. A trial with the same intraclass correlation as the x ray guidelines study, 0.019, and $m=50$ referrals per practice, would have design effect $D=1+(50-1)\times 0.019=1.93$. Thus it would require almost twice as many subjects as a trial where patients were randomised to treatment individually.

The main difficulty in calculating sample size for cluster randomised studies is obtaining an estimate of the between cluster variation or intraclass correlation. Estimates of variation between individuals can often be obtained from the literature but even studies that use the cluster as the unit of analysis may not publish their results in such a way that the between practice variation can be estimated. Recognising this problem, Donner recommended that authors should publish the cluster specific event rates observed in their trial. This would enable other workers to use this information to plan further studies.

In some trials, where the intervention is directed at the individual subjects and the number of subjects per cluster is small, we may judge that the design effect can be ignored. On the other hand, where the number of subjects per cluster is large, an estimate of the variability between clusters will be important.

References

1. Kerry SM, Bland JM. Sample size in cluster randomisation. *BMJ* 1998; 316: 549 [[Abstract/Full Text](#)]
2. Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ* 1996; 313: 41-42 [[Full Text](#)].
3. Meade TW, Roderick PJ, Brennan PJ, Wilkes HC, Kelleher CC. Extracranial bleeding and other

symptoms due to low dose aspirin and low intensity oral anticoagulation. *Thromb Haemostasis* 1992; 68: 1-6[[Medline](#)].

4. Oakeshott P, Kerry SM, Williams JE. Randomised controlled trial of the effect of the Royal College of Radiologists' guidelines on general practitioners' referral for radiographic examination. *Br J Gen Pract* 1994; 44: 197-200[[Medline](#)].
5. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. *BMJ* 1998; 316: 54[[Full Text](#)].
6. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomisation 1979-1989. *Int J Epidemiol* 1990; 19: 795-800[[Abstract](#)].

© [BMJ 1998](#)

This article has been cited by other articles:

- Wears, R. L. (2002). Advanced Statistics: Statistical Methods for Analyzing Cluster and Cluster-randomized Data. *Acad Emerg Med* 9: 330-341
[\[Abstract\]](#) [\[Full text\]](#)
- GILBODY, S., WHITTY, P. (2002). Improving the delivery and organisation of mental health services: beyond the conventional randomised controlled trial. *Br J Psychiatry* 180: 13-18
[\[Abstract\]](#) [\[Full text\]](#)
- Peters, T. J, Graham, A., Salisbury, C., Moore, L., Underwood, M., Eldridge, S., Gibson, P. G, Shah, S., Sindhusake, D., Wang, H., Peat, J. K, Henry, R. L (2001). Peer led programme for asthma education in adolescents. *BMJ* 323: 110-110 [\[Full text\]](#)
- Torgerson, D. J (2001). Contamination in trials: is cluster randomisation the answer?. *BMJ* 322: 355-357 [\[Full text\]](#)
- Reading, R., Harvey, I., Mclean, M. (2000). Cluster randomised trials in maternal and child health: implications for power and sample size. *Arch. Dis. Child.* 82: 79-83 [\[Abstract\]](#) [\[Full text\]](#)
- Campbell, M. K, Grimshaw, J. M (1998). Cluster randomised trials: time for improvement. *BMJ* 317: 1171-1172 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Kerry, S. M](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

- ▶ Collections under which this article appears:
[Randomized Controlled Trials: descriptions](#)
[Other Statistics and Research Methods: descriptions](#)

BMJ 1998;317:409-410 (8 August)

Education and debate

Statistics Notes

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 the findings 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 can be 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 β 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 delivery—for example, oral versus

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

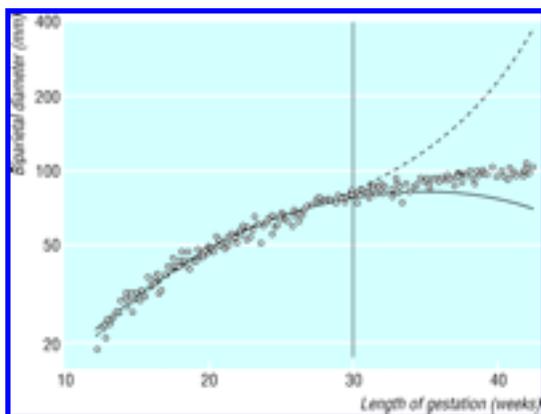
- ▶ Collections under which this article appears:
[Other Statistics and Research Methods: descriptions](#)

subcutaneous—or dose may affect treatment benefit. Clinical variation is likely to affect the size of benefit 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.



Fetal biparietal diameter (on log scale) in relation to gestational age⁸ with quadratic (solid line) and cubic (broken line) regression models fitted to data from only those fetuses less than 30 weeks' gestation (n=119)

View larger version (18K):

[\[in this window\]](#)

[\[in a new window\]](#)

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 $55.9+0.22\times\text{age}$ 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.

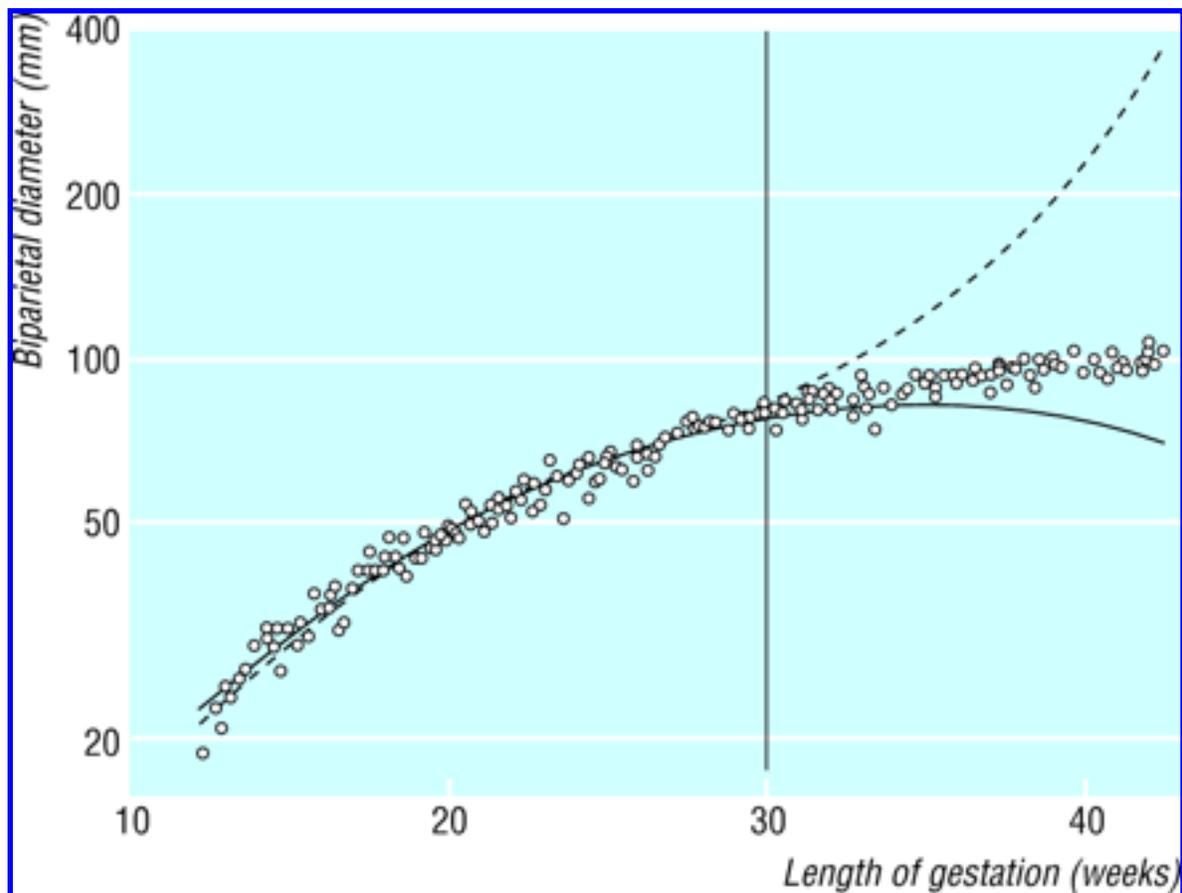
When we have two explanatory variables it will not usually be apparent (unless we examine a scatter diagram) when a patient has a combination of characteristics which do not fall within the span of the original data set. With more than two variables, such as in many prognostic models, it is not possible 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.

References

1. Zanetta G, Lissoni A, Torri V, Dalle Valle C, Trio D, Rangoni G, Mangioni C. Role of puncture and aspiration in expectant management of simple ovarian cysts: a randomised study. *BMJ* 1996; 31: 1110-1113.
2. Matthews JNS, Altman DG. Interaction 1: Heterogeneity of effects. *BMJ* 1996; 313: 486[[Full Text](#)].
3. Ellenberg JH. Selection bias in observational and experimental studies. *Statistics in Medicine* 1994; 13: 557-567[[Medline](#)].
4. Anonymous. BCG: bad news from India. *Lancet* 1980; i: 73-74.
5. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine. How to practice and teach EBM*. London: Churchill-Livingstone, 1997:167.
6. Begg CB. Biases in the assessment of diagnostic tests. *Statistics in Medicine* 1987; 6: 411-423[[Medline](#)].
7. Heathcote JA. Why do old men have big ears? *BMJ* 1995; 311: 1668[[Full Text](#)].
8. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 2. Head measurements. *Br J Obstet Gynaecol* 1994; 101: 35-43[[Medline](#)].

[\[View Larger Version of this Image \(108K JPEG file\)\]](#)



Fetal biparietal diameter (on log scale) in relation to gestational age⁸ with quadratic (solid line) and cubic (broken line) regression models fitted to data from only those fetuses less than 30 weeks' gestation (n=119)

[Home](#)[Help](#)[Search/Archive](#)[Feedback](#)[Search Result](#)*BMJ* 1998;317:468-469 (15 August)

Education and debate

Statistics Notes

Time to event (survival) data

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

In many medical studies an outcome of interest is the time to an event. Such events may be adverse, such as death or recurrence of a tumour; positive, such as conception or discharge from hospital; or neutral, such as cessation of breast feeding. It is conventional to talk about survival data and survival analysis, regardless of the nature of the event. Similar data also arise when measuring the time to complete a task, such as walking 50 metres.

The distinguishing feature of survival data is that at the end of the follow up period the event will probably not have occurred for all patients. For these patients the survival time is said to be censored, indicating that the observation period was cut off before the event occurred. We do not know when (or, indeed, whether) the patient will experience the event, only that he or she has not done so by the end of the observation period.

Censoring may also occur in other ways. Patients may be lost to follow up during the study, or they may experience a "competing" event which makes further follow up impossible. For example, patients being followed to a cardiac event may die from some other disease or in an accident.

In most survival studies patients are recruited over a period and followed up to a fixed date beyond the end of recruitment. Thus the last patients recruited will be observed for a shorter period than those recruited first and will be less likely to experience the event. An important assumption, therefore, is that patients' survival prospects (prognosis) stay the same throughout the study (although this will not matter

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Altman, D. G](#) || [Bland, J M.](#)
- ▶ Alert me when:
[New articles cite this article](#)

- ▶ Collections under which this article appears:
[Other Statistics and Research Methods: descriptions](#)

too much in a randomised trial). We also assume that patients lost to follow up have the same prognosis as those remaining in the study.

Table 1 shows the survival times of 44 patients in a randomised trial. Several patients in each group were still alive at the end of the study, while one was lost to follow up. In such a study we wish to compare the survival times of the two groups of patients. Statistical methods such as *t* tests cannot cope with the uncertainty in the data caused by censoring. Patients with censored data contribute valuable information and they should not be omitted from the analysis. It would also be wrong to treat the observed time (at censoring) as the survival time. We cannot tell, for example, whether the patient in the control group who was still alive at 127 months would have lived longer than the patient in the prednisolone group who died after 143 months. Rather we need recourse to a specialised set of statistical methods that have been developed for handling such data. We shall consider methods for graphical display and analysis of survival data in subsequent **Statistics Notes**.

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Table 1. Survival times (months) of 44 patients with chronic active hepatitis randomised to receive prednisolone or no treatment¹

Implicit in the preceding discussion is that survival should be evaluated in a cohort of patients followed forwards in time from a particular time point, such as diagnosis or randomisation, even if the cohort is identified retrospectively. An alternative, and potentially highly misleading, approach is to take a group of people experiencing the event of interest, perhaps in a certain time interval, and ascertain the elapsed time since the start of the relevant preceding time span. For example, we might take all newly diagnosed diabetics and find out when they first experienced certain symptoms. Similarly we might take birth as the start of the time period of interest for a group of individuals who have died and investigate associations between age at death and other variables.

Analyses of such data can cause serious problems. A good example is the highly dubious finding that left handed people die on average seven years younger than right handed people.² In this study those dying at old ages were survivors from a cohort born 70 or more years ago while those dying young may have been born at any time, and so on average will have been born later. Such studies make strong implicit assumptions—in essence that the prevalence of the risk factor(s), the characteristics of the population at risk, and the survival (prognosis) remain unchanged over many decades.³ These assumptions will usually be untenable and may also be untestable. Using this study design we would certainly find that people who use electric guitars or even personal computers die much younger than those who do not. The differing longevity in relation to handedness² would have arisen if the prevalence of left handedness had increased over the past 80 years. Proper prospective studies have found no evidence of an effect of handedness on

lifespan. [4](#) [5](#)

The same design was used in a study of long term survival in prostate cancer. All patients dying in a three year period who had been treated with palliative intent were "followed from death to diagnosis,"⁶ a period of up to 30 years. The authors reported that the proportion of deaths due to cancer increased with length of survival. This finding cannot be trusted because of the problems noted above, which are common to all such studies.³ Subjects with long survival times must have been diagnosed decades ago, whereas those with short survival times may include some patients diagnosed recently. The observed association could be a spurious consequence of improved treatment, earlier diagnosis, or some other change over time. The same error was seen recently in the *BMJ*.⁷

Retrospective studies can be valuable, but this design should be avoided when studying survival times. Whenever possible times to an event of interest should be studied in a definable cohort of individuals followed forwards in time.

References

1. Kirk AP, Jain S, Pocock S, Thomas HC, Sherlock S. Late results of the Royal Free Hospital prospective controlled trial of prednisolone therapy in hepatitis B surface antigen negative chronic active hepatitis. *Gut* 1980; 21: 78-83 [[Abstract](#)].
2. Halpern DF, Coren S. Handedness and life span. *N Engl J Med* 1991; 324: 998.
3. Abrahamsson PA, Adami HO, Taube A, Kim K, Zelen M, Kulldorff M. Re: Long-term survival and mortality in prostate cancer treated with noncurative intent. *J Urol* 1996; 155: 296-297.
4. Cerhan JR, Folsom AR, Potter JD, Prineas RJ. Handedness and mortality risk in older women. *Am J Epidemiol* 1994; 140: 368-374 [[Abstract](#)].
5. Aggleton JP, Bland JM, Kentridge RW, Neave NJ. Handedness and longevity: an archival study of cricketers. *BMJ* 1994; 309: 1681-1684 [[Abstract/Full Text](#)].
6. Aus G, Hugosson J, Norlén L. Long-term survival and mortality in prostate cancer treated with noncurative intent. *J Urol* 1995; 154: 460-465 [[Medline](#)].
7. MacManus I. Which doctors die first? *BMJ* 1997; 314: 1132 [[Full Text](#)].

© [BMJ 1998](#)

This article has been cited by other articles:

- Sobolev, B, Brown, P, Zelt, D (2001). Potential for bias in waiting time studies: events between enrolment and admission. *J Epidemiol Community Health* 55: 891-894 [[Abstract](#)] [[Full text](#)]

Table 1. Survival times (months) of 44 patients with chronic active hepatitis randomised to receive prednisolone or no treatment¹

Prednisolone (n=22)	Control (n=22)
2	2
6	3
12	4
54	7
56†	10
68	22
89	28
96	29
96	32
125*	37
128*	40
131*	41
140*	54
141*	61
143	63
145*	71
146	127*
148*	140*
162*	146*
168	158*
173*	167*
181*	182*

* Still alive at time of analysis.

† Lost to follow up.

*Statistics notes***Bayesians and frequentists**

J Martin Bland, Douglas G Altman,

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 paradigm—and 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.

We thank David Spiegelhalter for comments on a draft.

- 1 Breslow N. Biostatistics and Bayes (with discussion). *Statist Sci* 1990;5: 269-98.
- 2 Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). *J R Statist Soc A* 1994;157:357-416.

Correction

North of England evidence based guidelines development project: guideline for the primary care management of dementia

An editorial error occurred in this article by Martin Eccles and colleagues (19 September, pp 802-8). In the list of authors the name of Moira Livingston [not Livingstone] was misspelt.

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

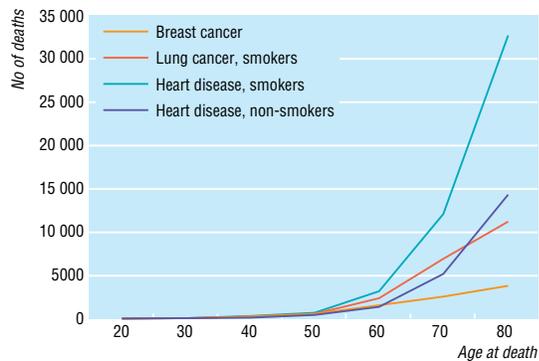
J Martin Bland, professor of medical statistics

ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Douglas G Altman, head

Correspondence to: Professor Bland

BMJ 1998;317:1151



Cumulative number of deaths in 1995 from breast cancer, lung cancer in smokers, heart disease in smokers, and heart disease in non-smokers per 100 000 women in England and Wales. A risk ratio of 12.5:1 for lung cancer and 2.3:1 for heart disease for smokers v non-smokers was assumed¹²

the cumulative probability of dying of lung cancer matches that of dying of breast cancer when women reach their early 50s; this probability doubles by age 65 and triples by age 75 (figure). Although there has been a modest fall in the number of women who smoke (mainly among older women), there is little evidence that the fear of developing lung cancer matches the fear of developing breast cancer. Ironically, lung cancer has a cure rate of <5% and can be almost entirely prevented by avoiding tobacco but, on average, 70% of patients treated for breast cancer can expect to survive for 10 years. In contrast to lung cancer there is comparatively little that can be done to prevent breast cancer.

Conclusion

The statistic that 1 in 12 women will develop breast cancer is thus correct only for women who have escaped a number of equally serious but more likely threats to life at an earlier age. For most women the lifetime risk of dying of breast cancer is only 1 in 26; the other 25 women will die of something else. Life table analyses show that the incidence of breast cancer and mortality from the disease are much lower among younger women and these risks should be understood in the context of other serious threats to life.

- 1 Assessing the odds [editorial]. *Lancet* 1997;350:1563.
- 2 Harris JR, Lippman ME, Veronesi U, Willett W. Breast cancer. *New Engl J Med* 1992;327:319-28.
- 3 Feuer EJ, Wun L-M. How much of the recent rise in breast cancer incidence can be explained by increases in mammography utilization? A dynamic population model approach. *Am J Epidemiol* 1992;136:1423-36.
- 4 Office for National Statistics. *1991 cancer statistics registrations*. London: Stationery Office, 1997. (Series MB1, No 24.)
- 5 Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: further results. *Contraception* 1996;54(suppl 3):1-106S.
- 6 Longnecker MP, Berlin JA, Orza MJ, Chalmers TC. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988;260:652-6.
- 7 Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding results of clinical trials and systematic reviews. *BMJ* 1995;311:1056-60.
- 8 McColl A, Smith H, White P, Field J. General practitioners' perceptions of the route to evidence based medicine: a questionnaire survey. *BMJ* 1998;316:361-5.
- 9 Tabar L, Fagerberg CJG, Gad A, Baldetorp L, Holmberg LH, Grontoft O, et al. Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet* 1985;i:829-32.
- 10 Office for National Statistics. *1995 mortality statistics: cause*. London: Stationery Office, 1997. (Series DH2, No 22.)
- 11 Pilote L, Hlatky MA. Attitudes of women toward hormone therapy and prevention of heart disease. *Am Heart J* 1995;129:1237-8.
- 12 Peto R, Lopez AD, Boreham J, Thun M, Heath C. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet* 1992;339:1268-78. (Accepted 26 June 1998)

Confidence intervals for the number needed to treat

Douglas G Altman

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.

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 (p_N) and standard treatment (p_S) groups is called the absolute risk reduction ($ARR = p_N - p_S$). 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

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

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

Imperial Cancer Research Fund Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF
Douglas G Altman, professor of statistics in medicine
d.altman@icrf.icnet.uk

BMJ 1998;317:1309-12



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 regions—values of the NNT from 4 to ∞ and values of the NNT from -20 to $-\infty$ (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 regions—values 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%,

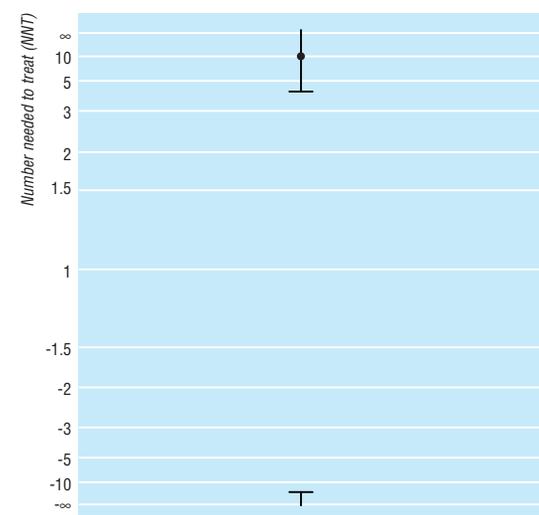


Fig 1 95% confidence interval for NNT=10

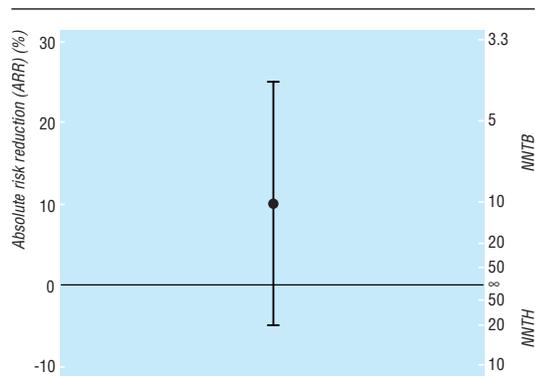


Fig 2 Relation between the absolute risk reduction (ARR) and number needed to treat and their confidence intervals (NNTB=number needed to treat (benefit); NNTH=number needed to treat (harm)) for the same example as in figure 1

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 value—it 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 was -0.08 (-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

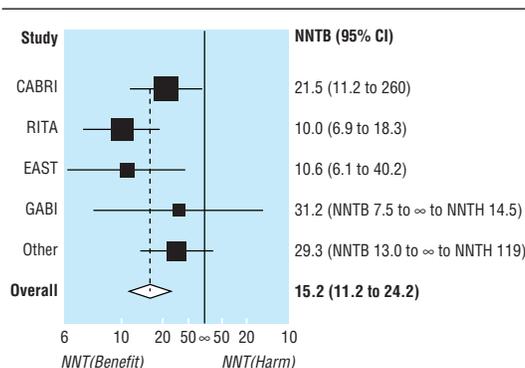


Fig 3 Forest plot for meta-analysis of data from eight randomised trials comparing bypass surgery with coronary angioplasty in relation to angina in one year.⁹ A number needed to treat (benefit) (NNTB) for coronary artery bypass grafting and its 95% confidence interval for each trial and for the overall estimate is shown

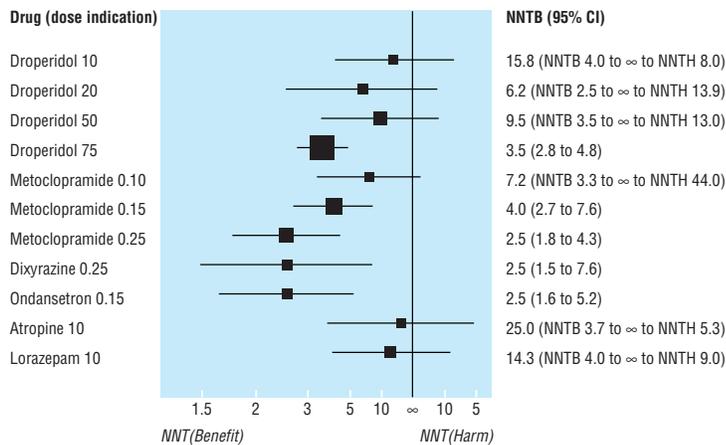


Fig 4 Summaries of meta-analyses of trials of prophylactic antiemetics in surgery for strabismus in children, showing the number needed to treat (benefit) (NNTB) value (95% confidence interval, NNTB to ∞ to number needed to harm (NNTH)) for each drug. (From Tramèr *et al*¹¹)

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.

I am grateful to Henry McQuay, Andrew Moore, and David Sackett for helpful discussions about these ideas. I thank the reviewer for suggesting figure 1.

Funding: None.

Conflicts of interest: None.

- 1 Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;310:452-4.
- 2 Daly LE. Confidence limits made easy: interval estimation using a substitution method. *Am J Epidemiol* 1998;147:783-90.
- 3 McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Ann Intern Med* 1997;126:712-20.
- 4 Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine. How to practice and teach EBM*. London: Churchill-Livingstone, 1997:208.
- 5 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG, eds. *Statistics with confidence*. London: BMJ, 1989:83-100.
- 6 Tramèr MR, Moore RA, Reynolds DJM, McQuay HJ. A quantitative systematic review of ondansetron in treatment of established postoperative nausea and vomiting. *BMJ* 1997;314:1088-92.
- 7 Silagy C, Mant D, Fowler G, Lodge M. Meta-analysis on efficacy of nicotine replacement therapies in smoking cessation. *Lancet* 1994;343:139-42.
- 8 Miller DB. Secondary prevention for ischemic heart disease. *Arch Intern Med* 1997;157:2045-52.
- 9 Pocock SJ, Henderson RA, Rickards AF, Hampton JR, King SB, Hamm CW, et al. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet* 1995;346:1184-9.
- 10 Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *Evidence-Based Med* 1996;1:164-6.
- 11 Tramèr MR, Moore RA, McQuay HJ. Prevention of vomiting after paediatric strabismus surgery: a systematic review using the numbers-needed-to-treat method. *Br J Anaesth* 1995;75:556-61.
- 12 Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;318:1728-33.
- 13 Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276:637-9.

(Accepted 27 May 1998)

One hundred years ago

Exercise and over-exercise

Dr Lauder Brunton opened the session of the York Medical Society last week by an address on Exercise and Over-exercise, in which, as was to be expected, he said a great many wise things with which every physician will agree. He said, for instance, that exercise which put into action every muscle of the body, but did not put any one into action for too great a length of time at once, or in too violent a manner, was exceedingly beneficial, but in applying this excellent principle he had the temerity to compare unfavourably with lawn tennis the three most popular physical recreations of the day—cricket, golf, and cycling. Moreover, he classed together croquet, cricket, and golf—rather a curious collocation—on the ground that in playing them there was not the same general movement of the whole body that was necessary in lawn tennis or polo. As to croquet all will probably be ready to agree, but as to cricket and golf, it is not likely that their devotees will be disposed to accept Dr Brunton's rather sweeping assertion. What muscles of the body are brought into play in lawn tennis

which are not brought into play by, say, a fast bowler, we should be rather curious to know; and as to golf, the distribution of the stiffness after a day's play in a man out of condition and practice leads at least to the suspicion that very few muscles in the body have not been called into action. As to cycling, Dr Lauder Brunton said that it tended to narrow the chest and to cause more or less a permanent stoop. He added that, as it had become so very general an amusement, its effects on the body as compared with those of other physical exercises must be very carefully watched. Like most of us, Dr Brunton has been struck by the fact that the girl of the period tends to be most divinely tall, and he seems disposed to put this down to the great popularity of lawn tennis a few years ago. It is certainly a pity that this very excellent game appears to be going out of fashion owing to the great popularity of cycling, which we should be disposed to agree with Dr Brunton is not an exercise so well calculated to produce an all-round development of the muscular system. (*BMJ* 1898;ii:1272)

Statistics Notes

Survival probabilities (the Kaplan-Meier method)

J Martin Bland, Douglas G Altman

Department of
Public Health
Sciences, St
George's Hospital
Medical School,
London SW17 0RE
J Martin Bland,
professor of medical
statistics

ICRF Medical
Statistics Group,
Centre for Statistics
in Medicine,
Institute of Health
Sciences, Oxford
OX3 7LF

Douglas G Altman,
head

Correspondence to:
Professor Bland

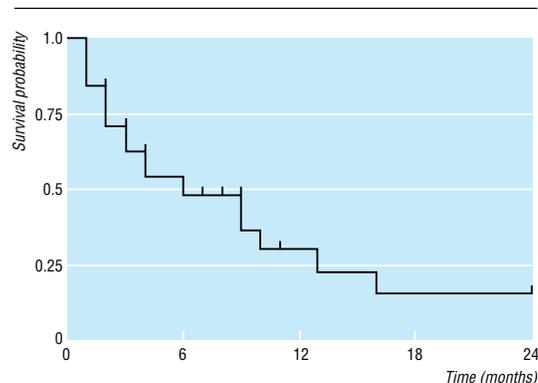
BMJ 1998;317:1572

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.

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 $0.842 \times 0.844 = 0.711$. 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.



Survival curve showing probability of not conceiving among 38 subfertile women after laparoscopy and hydrotubation²

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.

1 Altman DG, Bland JM. Time to event (survival) data. *BMJ* 1997;317:468-9.

2 Luthra P, Bland JM, Stanton SL. Incidence of pregnancy after laparoscopy and hydrotubation. *BMJ* 1982;284:1013-4.

3 Parmar MKB, Machin D. *Survival analysis: a practical approach*. Chichester: Wiley, 37, 47-9.

Time (months) to
conception or
censoring in 38
sub-fertile women
after laparoscopy
and hydrotubation²

Conceived	Did not conceive
1	2
1	3
1	4
1	7
1	7
1	8
2	8
2	9
2	9
2	9
2	11
3	24
3	24
3	
4	
4	
4	
6	
6	
9	
9	
9	
10	
13	
16	

*Statistics notes***Treatment allocation in controlled trials: why randomise?**

Douglas G Altman, J Martin Bland

Since 1991 the *BMJ* has had a policy of not publishing trials that have not been properly randomised, except in rare cases where this can be justified.¹ Why?

The simplest approach to evaluating a new treatment is to compare a single group of patients given the new treatment with a group previously treated with an alternative treatment. Usually such studies compare two consecutive series of patients in the same hospital(s). This approach is seriously flawed. Problems will arise from the mixture of retrospective and prospective studies, and we can never satisfactorily eliminate possible biases due to other factors (apart from treatment) that may have changed over time. Sacks et al compared trials of the same treatments in which randomised or historical controls were used and found a consistent tendency for historically controlled trials to yield more optimistic results than randomised trials.² The use of historical controls can be justified only in tightly controlled situations of relatively rare conditions, such as in evaluating treatments for advanced cancer.

The need for contemporary controls is clear, but there are difficulties. If the clinician chooses which treatment to give each patient there will probably be differences in the clinical and demographic characteristics of the patients receiving the different treatments. Much the same will happen if patients choose their own treatment or if those who agree to have a treatment are compared with refusers. Similar problems arise when the different treatment groups are at different hospitals or under different consultants. Such systematic differences, termed bias, will lead to an overestimate or underestimate of the difference between treatments. Bias can be avoided by using random allocation.

A well known example of the confusion engendered by a non-randomised study was the study of the possible benefit of vitamin supplementation at the time of conception in women at high risk of having a baby with a neural tube defect.³ The investigators found that the vitamin group subsequently had fewer babies with neural tube defects than the placebo control group. The control group included women ineligible for the trial as well as women who refused to participate. As a consequence the findings were not widely accepted, and the Medical Research Council later funded a large randomised trial to answer to the question in a way that would be widely accepted.⁴

The main reason for using randomisation to allocate treatments to patients in a controlled trial is to prevent biases of the types described above. We want to compare the outcomes of treatments given to groups of patients which do not differ in any systematic way. Another reason for randomising is that statistical theory is based on the idea of random sampling. In a study with random allocation the differences between treatment groups behave like the differences between random samples from a single population. We know

how random samples are expected to behave and so can compare the observations with what we would expect if the treatments were equally effective.

The term random does not mean the same as haphazard but has a precise technical meaning. By random allocation we mean that each patient has a known chance, usually an equal chance, of being given each treatment, but the treatment to be given cannot be predicted. If there are two treatments the simplest method of random allocation gives each patient an equal chance of getting either treatment; it is equivalent to tossing a coin. In practice most people use either a table of random numbers or a random number generator on a computer. This is simple randomisation. Possible modifications include block randomisation, to ensure closely similar numbers of patients in each group, and stratified randomisation, to keep the groups balanced for certain prognostic patient characteristics. We discuss these extensions in a subsequent *Statistics* note.

Fifty years after the publication of the first randomised trial⁵ the technical meaning of the term randomisation continues to elude some investigators. Journals continue to publish "randomised" trials which are no such thing. One common approach is to allocate treatments according to the patient's date of birth or date of enrolment in the trial (such as giving one treatment to those with even dates and the other to those with odd dates), by the terminal digit of the hospital number, or simply alternately into the different treatment groups. While all of these approaches are in principle unbiased—being unrelated to patient characteristics—problems arise from the openness of the allocation system.¹ Because the treatment is known when a patient is considered for entry into the trial this knowledge may influence the decision to recruit that patient and so produce treatment groups which are not comparable.

Of course, situations exist where randomisation is simply not possible.⁶ The goal here should be to retain all the methodological features of a well conducted randomised trial⁷ other than the randomisation.

ICRF Medical Statistics Group,
Centre for Statistics in Medicine,
Institute of Health Sciences, Oxford
OX3 7LF

Douglas G Altman,
professor of statistics in medicine

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

J Martin Bland,
professor of medical statistics

Correspondence to: Professor Altman.

BMJ 1999;318:1209

1 Altman DG. Randomisation. *BMJ* 1991;302:1481-2.

2 Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233-40.

3 Smithells RW, Sheppard S, Schorah CJ, Seller MJ, Nevin NC, Harris R, et al. Possible prevention of neural-tube defects by periconceptional vitamin supplementation. *Lancet* 1980;i:339-40.

4 MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council vitamin study. *Lancet* 1991;338:131-7.

5 Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 1948;2:769-82.

6 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-8.

7 Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT Statement. *JAMA* 1996;276:637-9.

- 15 Van den Hoogen HJM, Koes BW, van Eijk JT, Bouter LM, Devillé W. On the course of low back pain in general practice: a one year follow up study. *Ann Rheum Dis* 1998;57:13-9.
- 16 Croft PR, Papageorgiou AC, Ferry S, Thomas E, Jayson MIV, Silman AJ. Psychological distress and low back pain: Evidence from a prospective study in the general population. *Spine* 1996;20:2731-7.
- 17 Papageorgiou AC, Macfarlane GJ, Thomas E, Croft PR, Jayson MIV, Silman AJ. Psychosocial factors in the work place—do they predict new episodes of low back pain? *Spine* 1997;22:1137-42.
- 18 Main CJ, Wood PL, Hollis S, Spanswick CC, Waddell G. The distress and risk assessment method. A simple patient classification to identify distress and evaluate the risk of poor outcome. *Spine* 1992;17:42-52.
- 19 Coste J, Delecoeuillerie G, Cohen de Lara A, Le Parc JM, Paolaggi JB. Clinical course and prognostic factors in acute low back pain: an inception cohort study in primary care practice. *BMJ* 1994;308:577-80.
- 20 Dionne CE, Koepsell TD, Von Korff M, Deyo RA, Barlow WE, Checkoway H. Predicting long-term functional limitations among back pain patients in primary care. *J Clin Epidemiol* 1997;50:31-43.
- 21 Macfarlane GJ, Thomas E, Papageorgiou AC, Schollum J, Croft PR. The natural history of chronic pain in the community: a better prognosis than in the clinic? *J Rheumatol* 1996;23:1617-20.
- 22 Troup JDG, Martin JW, Lloyd DCEF. Back pain in industry. A prospective survey. *Spine* 1981;6:61-9.
- 23 Burton AK, Tillotson KM. Prediction of the clinical course of low-back trouble using multivariable models. *Spine* 1991;16:7-14.
- 24 Pope MH, Rosen JC, Wilder DG, Frymoyer JW. The relation between biomechanical and psychological factors in patients with low-back pain. *Spine* 1980;5:173-8.

(Accepted 31 March 1999)

Statistics notes

Variables and parameters

Douglas G Altman, J Martin Bland

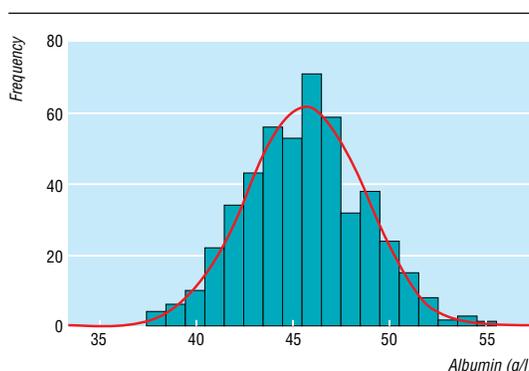
Like all specialist areas, statistics has developed its own language. As we have noted before,¹ much confusion may arise when a word in common use is also given a technical meaning. Statistics abounds in such terms, including normal, random, variance, significant, etc. Two commonly confused terms are 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 normal¹ 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 parameters¹ 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



Measurements of serum albumin in 481 white men aged over 20 (data from Dr W G Miller)

(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 hope²). 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.

1 Altman DG, Bland JM. The normal distribution. *BMJ* 1995;310:298.
2 Endpiece: What's a parameter? *BMJ* 1998;316:1877.

ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Douglas G Altman, professor of statistics in medicine

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

J Martin Bland, professor of medical statistics

Correspondence to: Professor Altman.

BMJ 1999;318:1667

of course not new. What is unusual is the espousal of this argument by a Labour government and its apparent willingness to challenge the power of its traditional support base in the trade unions and entrenched interests of the health professions, including doctors. Yet if the assumptions that lay behind the first and second ways encompassed elements of truth without seeing the whole picture, so too the critique of the forces of conservatism risks turning an accurate perception of part of the problem confronting the NHS into a programme that is applied without discrimination. If this were to happen, it would alienate managers and clinicians who support the direction of travel that has been set out by the government and whose continuing commitment is needed to deliver the modernisation agenda.

These observations take on added force because, in the life cycle of governments, Labour is moving from a preoccupation with policy development to a focus on implementation and delivery. Its impatience to see the delivery of service improvements is manifested in the prime minister's close personal involvement in domestic policy priorities and the stated commitment of ministers to increase rather than reduce the pace of change. In this context, the limited direct management experience of politicians in power may explain the approach they are pursuing, and their failure to appreciate the scale of the task that has been taken on in turning around major public services like education and health. An appeal to the altruism of those working in the NHS and recognition of the key role they have to play in delivering the modernisation programme are just as likely to succeed as an attack on their conservatism, and unless this is taken on board health policy will once more become a battleground between politicians and NHS staff.

Recognising the forces of innovation

What, then, should be done? The priority of the new health secretary, Alan Milburn, should be to add to the instruments at his disposal by recognising the forces of innovation within the NHS and providing them with the resources required to implement the government's vision. Delivering NHS modernisation depends fundamentally on ministers acknowledging this fact and not losing the support of those who are committed to providing a modern and dependable service. No amount of guidance from the NHS Executive or hectoring by politicians can substitute for a drive to improve performance that comes from within and is acknowledged and valued by those steering the process of change.

Above all, ministers should champion entrepreneurial managers and clinicians who are leading the modernisation drive within the NHS, and they should support the more rapid dissemination of good practices as they are identified. These measures may not be sufficient but they are certainly necessary in enabling the third way to be realised. And who knows, they may ultimately give credence to the claim that New Labour's approach really is different.

The thinking behind this article was stimulated by the work of Julian Le Grand and his analysis of the assumptions that lie behind policies towards the welfare state.

Competing interests: None declared.

- 1 Ham C. Managed markets in health care: the UK experiment. *Health Policy* 1996;35:279-92.
- 2 Ham C. The third way in health care reform: does the emperor have any clothes? *J Health Serv Res Policy* 1999;4:168-73.
- 3 Le Grand J, Mays N, Mulligan J-A, eds. *Learning from the NHS internal market: a review of the evidence*. London: King's Fund, 1998.

(Accepted 4 November 1999)

Calculating the number needed to treat for trials where the outcome is time to an event

Douglas G Altman, Per Kragh Andersen

Imperial Cancer Research Group Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Headington, Oxford OX3 7LF
Douglas G Altman
professor of statistics in medicine

continued over

BMJ 1999;319:1492-5

The number of patients who need to be treated to prevent one additional event (number needed to treat; NNT) has become a widely used measure of treatment benefit derived from the results of randomised controlled trials with a binary outcome.^{1,2} We show how to obtain a number needed to treat for studies where the primary outcome is the time to an event. We consider primarily the situation where there is no access to raw data, for example, when reviewing a published study, and also how to proceed when given the raw data.

Time to event data

As noted previously, for studies with binary outcome the number needed to treat will vary according to the length of follow up.³ For studies of survival this relation with time is more explicit. There is no single number needed to treat; rather it can be calculated at any time point after the start of treatment. Often there are one or two time points of particular clinical interest.

Summary points

The number needed to treat is the number of patients who need to be treated to prevent one additional adverse outcome

This number (with confidence interval) is a clinically useful way to report the results of controlled trials

For any trial which has reported a binary outcome, the number needed to treat can be obtained as the reciprocal of the absolute difference in proportions of patients with the outcome of interest

In studies where the outcome of interest is the time to an event, calculations can be extended to show the number needed to treat at any time after the start of treatment

A time specific number needed to treat represents the number of patients who need to be given the treatment in question for one additional patient to survive to that time point—that is, to benefit from the treatment. To obtain an estimate of the number needed to treat together with a confidence interval, one of the following is needed: (a) an estimate of the survival probability in each group at one fixed time point, and either the number of patients “at risk” at that time—that is, not yet having experienced the event of interest—or the standard errors of the survival probabilities; or (b) the estimated hazard ratio and its standard error, and the estimated survival probability in the control group at a fixed time. Unfortunately, the reporting of results is often inadequate in studies of survival,⁴ and the required information is often not provided.

Methods and examples

We will assume there are two treatment groups. The calculations relate to survival probabilities at a fixed time point after the start of the follow up period—that is, from the start of treatment. We consider three cases.

Only survival probabilities available

Suppose, firstly, that only a simple survival analysis has been performed, and that Kaplan-Meier survival curves have been generated. We denote the estimated survival probabilities in the active and control treatment groups at a chosen time point as S_a and S_c and will assume that the active drug is effective, so that $S_a > S_c$. The absolute risk reduction is estimated as $S_a - S_c$. If necessary, S_a and S_c can be estimated by careful measurement of a graph of the Kaplan-Meier survival curves. The number needed to treat is obtained simply as $1/(S_a - S_c)$, just as for trials with binary data.

The 95% confidence interval for the absolute risk reduction (ARR) is $ARR \pm 1.96 SE(ARR)$, where $SE(ARR)$ is the standard error of the absolute risk reduction. If the limits of this confidence interval are A_u and A_l , then the 95% confidence interval for the number needed to treat is $1/A_u$ to $1/A_l$.

When neither the standard error nor confidence interval for the absolute risk reduction is given, there are three options:

1. If confidence intervals for S_a and S_c are given, each standard error can be taken as one quarter of the width of the relevant confidence interval.
2. If the standard errors of S_a and S_c are given, $SE(ARR)$ can be calculated as $\sqrt{[SE(S_a)]^2 + [SE(S_c)]^2}$.
3. If standard errors or confidence intervals are not given, we need the numbers of patients still at risk (alive) at the time corresponding to the estimated probabilities, which we will call n_a and n_c . These numbers are sometimes shown in the graph of survival; if not, they will have to be inferred. If there is little loss to follow up, the numbers at risk will be close to $S_a N_a$ and $S_c N_c$, where N_a and N_c are the numbers randomised to each group. Information about loss to follow up is, however, often missing.⁴ The standard error of the absolute risk reduction is $\sqrt{[S_a^2(1 - S_a)/n_a + S_c^2(1 - S_c)/n_c]}$, and a 95% confidence interval is obtained as above. If none of the preceding calculations is possible, then a confidence interval cannot be obtained for the number needed to treat.

Example

Overall, 279 patients with locally advanced rectal cancer were randomised to receive radiotherapy followed by surgery compared with surgery alone.⁵ The sample size calculation was on the basis of survival for 3 years. From figure 2 in the paper the three year survival rates were 62.2% and 46.8% for the two groups, with 59 and 43 patients still alive respectively. The above formula gives $ARR = 0.622 - 0.468 = 0.154$, and $SE(ARR) = \sqrt{[0.622^2(1 - 0.622)/59 + 0.468^2(1 - 0.468)/43]} = 0.072$, giving a 95% confidence interval for the absolute risk reduction as 0.013 to 0.295. The number needed to treat at 3 years is thus $1/0.154 = 6.49$ and its 95% confidence interval is $1/0.295$ to $1/0.013$, or 3.4 to 77.6. We thus estimate that giving patients radiotherapy before surgery would lead to one extra survivor at 3 years for every 6.5 patients treated. The confidence interval is very wide, however.

When the treatment effect is not statistically significant ($P > 0.05$) the 95% confidence interval for the absolute risk reduction spans zero, and one limit of the confidence interval for the number needed to treat will be negative. In this case the inverse of the absolute risk reduction is often termed the number needed to harm (NNH).⁶ It is, however, more accurate to refer to the number needed to treat to benefit (NNTB) or to the number needed to treat to harm (NNTH).⁷ Difficulties in graphing the confidence interval are avoided by plotting the absolute risk reduction at suitable values and relabelling the axis,⁷ as illustrated below.

Survival probabilities and estimate of hazard ratio available

The hazard ratio is quite like a relative risk rather than an odds ratio,⁴ but it is not the same as a relative risk. Customary methods of analysis assume that this ratio is the same at all times after the start of treatment.

The log rank test provides the observed and expected numbers of events in each group. The hazard ratio is estimated as the ratio of the ratios of observed to expected numbers for the active and control groups. If the treatment is beneficial, the hazard ratio will be less than 1. Unfortunately, few authors provide the observed and expected numbers from this analysis.

The hazard ratio is more often available from a Cox regression, which is used in controlled trials to adjust

Department of
Biostatistics,
University of
Copenhagen,
DK-2200
Copenhagen N,
Denmark
Per Kragh
Andersen
professor of
biostatistics

Correspondence to:
D G Altman
d.altman@icrf.icnet.uk



SUE SHARPLES

the trial results for other prognostic variables. Here the regression coefficient for treatment (often denoted b or β) is the log hazard ratio. It follows that the hazard ratio is estimated as e^b . Either the regression coefficient (b) or the hazard ratio ($h=e^b$) may be quoted in a published paper.

If at some specified time, t , the survival probability in the control group is $S_c(t)$ then the survival probability in the active group is $[S_c(t)]^h$, where h is the hazard ratio comparing the treatment groups. The number needed to treat is estimated as:

$$NTT = 1 / \{ [S_c(t)]^h - S_c(t) \} \text{ (equation 1)}$$

where $S_c(t)$ is obtained in one of the ways previously described. The number of patients at risk is not needed (the information is incorporated into the standard error of h). Note that h and the number needed to treat may depend on which other variables are included in the regression model and how they are coded, although in a randomised trial the differences should be small.

The 95% confidence interval for the number needed to treat is obtained from equation 1 by replacing h in turn by the two limits of the 95% confidence interval for h . If not given explicitly, the values can be obtained from the regression coefficient b (recall that $h=e^b$) and its standard error as $e^{b-1.96set(b)}$ and $e^{b+1.96set(b)}$. The resulting confidence interval may be too narrow as it ignores the imprecision in the estimate of $S_c(t)$. We return to this issue later. If we have results of a regression analysis but do not have any estimate of the control group survival probability $S_c(t)$, we cannot estimate the number needed to treat.

Example

We use data from a randomised trial comparing intensive versus standard insulin treatment in patients with diabetes mellitus and acute myocardial infarction.⁸ From figure 1 in the paper, the control group mortality rates at 2 and 4 years were 0.33 and 0.49 respectively. The reported hazard ratio was $h=0.72$ with 95% confidence interval 0.55 to 0.92. The number needed to treat at 2 years is thus estimated as $1/(0.33^{0.72}-0.33)=8.32$. The 95% confidence interval for the number needed to treat is obtained from equation 1 setting h to 0.55 and then 0.92, giving 4.7 to 32.7.

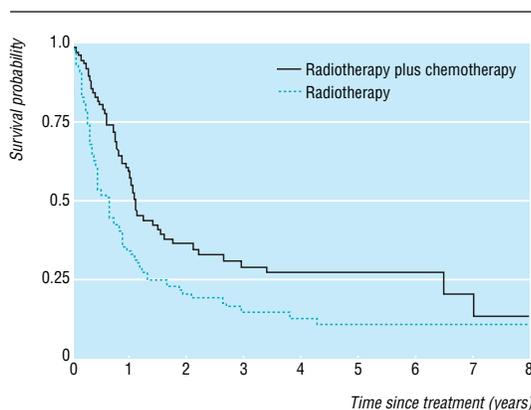


Fig 1 Kaplan-Meier plots of survival for 164 patients with non-small lung cancer treated with radiotherapy plus chemotherapy versus radiotherapy alone⁹

Number needed to treat at various times after treatment for 164 patients with non-small cell lung cancer treated with radiotherapy plus chemotherapy versus radiotherapy alone⁹

Time from treatment	Number needed to treat (95% CI)	No of patients still at risk
6 months	3.6 (2.4 to 7.4)	105
1 year	4.0 (2.5 to 10.2)	67
2 years	6.4 (3.3 to 74.3)	38
3 years	7.0 (3.6 to 128.5)	27
4 years	7.1 (3.6 to 117.0)	23
5 years	6.3 (3.5 to 37.1)	18
6 years	6.3 (3.5 to 37.1)	13

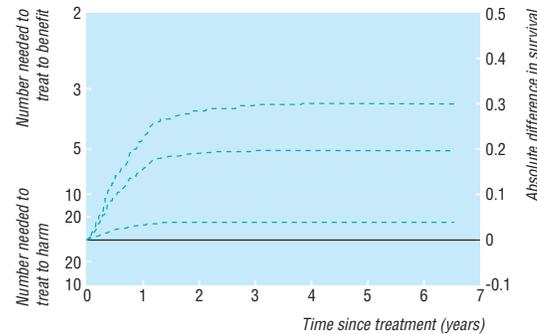


Fig 2 Number needed to treat to either benefit or harm with 95% confidence interval by time since treatment for 164 patients with non-small cell lung cancer treated with radiotherapy plus chemotherapy versus radiotherapy alone⁹ on basis of Cox regression model including only treatment

Raw data available

For researchers reporting the results of a trial, all the raw data will be available. Clearly it is possible to use any of the above methods to calculate a number needed to treat, either unadjusted or adjusted, as all of the statistics mentioned can be generated easily. We can also extend the method quite simply to generate a plot showing number needed to treat as a function of time rather than at a single time point.

Example

One hundred and seventy two patients with non-small cell lung cancer were randomised to receive either radiotherapy alone or in combination with chemotherapy.⁹ The raw data (with somewhat longer follow up) are given by Piantadosi.¹⁰ Figure 1 shows Kaplan-Meier curves of disease free survival for the two treatment groups, while the table shows the estimated number needed to treat, with 95% confidence intervals.

The table is based on simple comparison of the two treatment groups. Adjusted survival curves can be produced, often by Cox regression, to adjust a treatment comparison for various baseline variables. The number needed to treat can also be obtained from these adjusted analysis, again using equation 1. An example is shown in figure 2. If, as here, the treatment effect is statistically significant with $P<0.05$, the 95% confidence interval for the number needed to treat will exclude harmful effects at all times.

Even though the model assumes a constant hazard ratio (relative risk) for the comparison of two treatments, it is important to recognise that the number needed to treat will differ for subsets of

patients with varying prognosis. It may be valuable to construct graphs like figure 2 for important subsets of patients, such as by stage or cell type in the example.

Discussion

The need for absolute as well as relative measures of effect is increasingly recognised.² The number needed to treat has recently become a quite popular way of reporting the results of clinical trials.¹ The number needed to treat will usually tend to fall as the time from start of treatment increases. Sackett et al suggested a simple correction for length of follow up, in which the observed number needed to treat is multiplied by the ratio of the actual average duration of follow up to the duration of interest.³ This calculation assumes that the effect of treatment (relative risk reduction) is constant over time, and that events occur at a constant rate over time. Under these strong assumptions a number needed to treat of, say, 6 derived from a study in which patients were followed on average for 2 years would imply a number needed to treat of 3 if patients were followed for 4 years. Following this approach, Miller presented for several trials numbers needed to treat per year, calculated as the overall number needed to treat multiplied by the average length of follow up in years.¹¹

When actual times to an event of interest are recorded, numbers needed to treat can be obtained as a function of follow up time. For many published papers it will be possible to use these methods to obtain numbers needed to treat, perhaps adjusted for other variables. This measure should be valuable for those reviewing papers for journals of secondary publication, with the number needed to treat calculated for one or two specific time points.

The confidence interval for the number needed to treat on the basis of the Cox model may be too narrow ("conservative") because the method ignores the uncertainty in the estimate of the survival probability. This deficiency applies equally to the confidence interval obtained for the number needed to treat derived

from the log odds ratio estimated from a logistic regression model. There is no way around this problem when describing the number needed to treat from information given in a published paper. An unbiased confidence interval can be obtained from the raw data, but the method is rather complex and we have not presented it here.

The number needed to treat is valuable additional information that can be provided in reports of randomised trials where the outcome of interest was time to an event. We have shown how to calculate the number needed to treat for such studies in several ways. In general, it will better to make such calculations directly, rather than making the strong assumption that the risk reduction is constant over follow up time.

Funding: Activities of the Danish Epidemiology Service Centre are supported by a grant from the Danish National Research Foundation.

Competing interests: None declared.

- 1 Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;310:452-4.
- 2 Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine. How to practice and teach EBM*. London: Churchill Livingstone, 1997:136-41, 168-70.
- 3 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*, 2nd ed. Boston: Little Brown, 1991:208.
- 4 Altman DG, De Stavola BL, Love SB, Stepniowska KA. Review of survival analyses published in cancer journals. *Br J Cancer* 1995;72:511-8.
- 5 Medical Research Council Rectal Cancer Working Party. Randomised trial of surgery alone versus radiotherapy followed by surgery for potentially operable locally advanced rectal cancer. *Lancet* 1996;348:1605-10.
- 6 McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Ann Intern Med* 1997;126:712-20.
- 7 Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998;317:1309-12.
- 8 Malmberg K for the Diabetes Mellitus Insulin Glucose Infusion in Acute Myocardial Infarction (DIGAMI) Study Group. Prospective randomised study of intensive insulin treatment on long term survival after acute myocardial infarction in patients with diabetes mellitus. *BMJ* 1997;314:1512-5.
- 9 Lad T, Rubinstein L, Sadeghi A. The benefit of adjuvant treatment for resected locally advanced non-small-cell lung cancer. *J Clin Oncol* 1988;6:9-17.
- 10 Piantadosi S. *Clinical trials. A methodologic approach*. Chichester: John Wiley, 1997.
- 11 Miller DB. Secondary prevention for ischemic heart disease. Relative numbers needed to treat with different therapies. *Arch Intern Med* 1997;157:2045-52.

(Accepted 5 July 1999)

A memorable dream

Plagiarism

Renewed plans for reforming the upper house of parliament reminded me of the old story of the peer who dreamt that he was making a speech in the House of Lords and woke up to find that he was. My dream was similar but different. After three years at Oxford on Wednesday afternoons spent postprandially in the warmth and darkness of histology instruction I often close my eyes in lecture theatres to concentrate better.

At international conferences I try at least to go to state of the art lectures in my field. At one congress I dreamt that I was lecturing on my particular area and was showing my favourite series of slides solving, at least to my satisfaction, a critical pathophysiological problem. And then I woke to find that the lecture was being given not by me but by a Ruritanian professor who was showing as his work slide after slide of mine. Years later I again dreamt that I was lecturing, on a different pet topic, and woke to find the speaker using a series of slides in the same order as in one of my papers.

I did not reproach either lecturer. However, when I read in an authoritative monograph consecutive paragraphs with a graph which seemed cogently and convincingly to solve several specific

scientific issues, I suddenly realised that the illustration and these paragraphs had been lifted word for word, without acknowledgment or citation, from one of my articles. I did write to the eminent publishers who wrote that the author had indeed transcribed my paragraphs but unfortunately and inexplicably had omitted to place them within quotation marks or to attribute them to me or to cite my paper.

I know that imitation is said to be the sincerest form of flattery, but I still find plagiarism galling.

Jeremy Hugh Baron *honorary professorial lecturer, New York*

We welcome articles of up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake*, or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for "Endpieces," consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.

Much work still needs to be done to achieve this. To be useful in health policy at this level, all the targets need to be elaborated further and clear, practical statements must be made on their operation—especially the four targets on health policy and sustainable health systems. The WHO should stimulate the discussion of these important targets, but it should also be careful about being too prescriptive about health systems since this could be counterproductive.

In addition, more attention should be given to the usefulness of the targets in member states. One way of doing this is to rank the countries by target and to divide them into three groups. A specific level could be set for each group. For example, for target 2, three such groups could be distinguished as follows:

- Countries that have already achieved this target
- Countries for which the global target is achievable and challenging
- Countries that find the global target hard to achieve and therefore “demotivating.”

The first group needs stricter target levels, and the third group less stringent ones. If a breakdown of this kind is made for each target, some countries may be classified in different groups for different targets. In this way, the targets will provide an insight into the health status of the population and could be useful for policy makers in member states in encouraging action and allocating their resources.

We thank Dr J Visschedijk and Professor L J Gunning-Schepers and other referees of this article for their helpful comments.

Funding: This study was commissioned by Policy Action Coordination at the WHO and supported by an unrestricted educational grant from Merck & Co Inc, New Jersey, USA.

Competing interests: None declared.

- 1 World Health Assembly. Resolution WHA51.7. *Health for all policy for the twenty-first century*. Geneva: World Health Organisation, 1998.
- 2 World Health Association. *Health for all in the 21st century*. Geneva: WHO, 1998.

- 3 World Health Association. *Global strategy for health for all by the year 2000*. Geneva: WHO, 1981. (WHO Health for All series No 3.)
- 4 Visschedijk J, Siméant S. Targets for health for all in the 21st century. *World Health Stat Q* 1998;51:56-67.
- 5 Van de Water HPA, van Herten LM. *Never change a winning team? Review of WHO's new global policy: health for all in the 21st century*. Leiden: TNO Prevention and Health, 1999.
- 6 World Health Organisation. *Bridging the gaps*. Geneva: WHO, 1995. (World health report.)
- 7 World Health Organisation. *Fighting disease, fostering development*. Geneva: WHO, 1996. (World health report.)
- 8 World Health Organisation. 1997: *Conquering suffering, enriching humanity*. Geneva: WHO, 1997. (World health report.)
- 9 Murray CJL, Lopez AD, eds. *The global burden of disease*. Boston: Harvard University Press, 1996.
- 10 United Nations. *The world population prospects*. New York: UN, 1998.
- 11 United Nations Development Programme. *Human development report 1997*. New York: Oxford University Press, 1997.
- 12 World Bank. *Poverty reduction and the World Bank: progress and challenges in the 1990s*. New York: World Bank, 1996.
- 13 World Health Organisation. *Third evaluation of health for all by the year 2000*. Geneva: WHO, 1999. (In press.)
- 14 Ad Hoc Committee on Health Research Relating to Future Intervention Options. *Investing in health research and development*. Geneva: WHO, 1996. (Document TDR/Gen/96.1.)
- 15 Taylor CE. Surveillance for equity in primary health care: policy implications from international experience. *Int J Epidemiol* 1992;21:1043-9.
- 16 Frerichs RR. Epidemiologic surveillance in developing countries. *Amu Rev Public Health* 1991;12:257-80.
- 17 World Health Organisation. *Health for all renewal—building sustainable health systems: from policy to action. Report of meeting on 17-19 November 1997 in Helsinki, Finland*. Geneva: WHO, 1998.
- 18 World Health Organisation. *EMC annual report 1996*. Geneva: WHO: 1996.
- 19 World Health Organisation. *Physical status: the use and interpretation of anthropometry of a WHO expert committee*. Geneva: WHO, 1995. (WHO technical report series No 834.)
- 20 World Health Organisation. *Global database on child growth and malnutrition*. Geneva: WHO, 1997.
- 21 World Health Organisation. *Tobacco or health: a global status report*. Geneva: WHO, 1997.
- 22 Erkens C. *Cost-effectiveness of 'short course chemotherapy' in smear-negative tuberculosis*. Utrecht: Netherlands School of Public Health, 1996.
- 23 Van de Water HPA, van Herten LM. *Bull's eye or Achilles' heel: WHO's European health for all targets evaluated in the Netherlands*. Leiden: Netherlands Association for Applied Scientific Research (TNO) Prevention and Health, 1996.
- 24 Van de Water HPA, van Herten LM. *Health policies on target? Review of health target and priority setting in 18 European countries*. Leiden: Netherlands Association for Applied Scientific Research (TNO) Prevention and Health, 1998.

(Accepted 4 May 1999)

Statistics notes

How to randomise

Douglas G Altman, J Martin Bland

We have explained why random allocation of treatments is a required feature of controlled trials.¹ Here we consider how to generate a random allocation sequence.

Almost always patients enter a trial in sequence over a prolonged period. In the simplest procedure, simple randomisation, we determine each patient's treatment at random independently with no constraints. With equal allocation to two treatment groups this is equivalent to tossing a coin, although in practice coins are rarely used. Instead we use computer generated random numbers. Suitable tables can be found in most statistics textbooks. The table shows an example²: the numbers can be considered as either random digits from 0 to 9 or random integers from 0 to 99.

For equal allocation to two treatments we could take odd and even numbers to indicate treatments A and B respectively. We must then choose an arbitrary

place to start and also the direction in which to read the table. The first 10 two digit numbers from a starting place in column 2 are 85 80 62 36 96 56 17 17 23 87, which translate into the sequence A B B B B A A A A for the first 10 patients. We could instead have taken each digit on its own, or numbers 00 to 49 for A and 50 to 99 for B. There are countless possible strategies; it makes no difference which is used.

We can easily generalise the approach. With three groups we could use 01 to 33 for A, 34 to 66 for B, and 67 to 99 for C (00 is ignored). We could allocate treatments A and B in proportions 2 to 1 by using 01 to 66 for A and 67 to 99 for B.

At any point in the sequence the numbers of patients allocated to each treatment will probably differ, as in the above example. But sometimes we want to keep the numbers in each group very close at all times. Block randomisation (also called restricted

ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Douglas G Altman, professor of statistics in medicine

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE

J Martin Bland, professor of medical statistics

Correspondence to: Professor Altman.

BMJ 1999;319:703-4

Excerpt from a table of random digits.² The numbers used in the example are shown in bold

89	11	77	99	94
35	83	73	68	20
84	85	95	45	52
56	80	93	52	82
97	62	98	71	39
79	36	13	72	99
34	96	98	54	89
69	56	88	97	43
09	17	78	78	02
83	17	39	84	16
24	23	36	44	14
39	87	30	20	41
75	18	53	77	83
33	93	39	24	81
22	52	01	86	71

randomisation) is used for this purpose. For example, if we consider subjects in blocks of four at a time there are only six ways in which two get A and two get B:

1: A A B B 2: A B A B 3: A B B A 4: B B A A 5: B A B A 6: B A A B

We choose blocks at random to create the allocation sequence. Using the single digits of the previous random sequence and omitting numbers outside the range 1 to 6 we get 5 6 2 3 6 6 5 6 1 1. From these we can construct the block allocation sequence B A B A / B A A B / A B A B / A B B A / B A A B, and so on. The numbers in the two groups at any time can never differ by more than half the block length. Block size is normally a multiple of the number of treatments. Large blocks are best avoided as they control balance less well. It is possible to vary the block length, again at random, perhaps using a mixture of blocks of size 2, 4, or 6.

While simple randomisation removes bias from the allocation procedure, it does not guarantee, for example, that the individuals in each group have a similar age distribution. In small studies especially some chance imbalance will probably occur, which might complicate the interpretation of results. We can use stratified randomisation to achieve approximate balance of important characteristics without sacrificing the advantages of randomisation. The method is to produce a separate block randomisation list for each subgroup (stratum). For example, in a study to

compare two alternative treatments for breast cancer it might be important to stratify by menopausal status. Separate lists of random numbers should then be constructed for premenopausal and postmenopausal women. It is essential that stratified treatment allocation is based on block randomisation within each stratum rather than simple randomisation; otherwise there will be no control of balance of treatments within strata, so the object of stratification will be defeated.

Stratified randomisation can be extended to two or more stratifying variables. For example, we might want to extend the stratification in the breast cancer trial to tumour size and number of positive nodes. A separate randomisation list is needed for each combination of categories. If we had two tumour size groups (say ≤ 4 and > 4 cm) and three groups for node involvement (0, 1-4, > 4) as well as menopausal status, then we have $2 \times 3 \times 2 = 12$ strata, which may exceed the limit of what is practical. Also with multiple strata some of the combinations of categories may be rare, so the intended treatment balance is not achieved.

In a multicentre study the patients within each centre will need to be randomised separately unless there is a central coordinated randomising service. Thus "centre" is a stratifying variable, and there may be other stratifying variables as well.

In small studies it is not practical to stratify on more than one or perhaps two variables, as the number of strata can quickly approach the number of subjects. When it is really important to achieve close similarity between treatment groups for several variables minimisation can be used—we discuss this method in a separate Statistics note.³

We have described the generation of a random sequence in some detail so that the principles are clear. In practice, for many trials the process will be done by computer. Suitable software is available at <http://www.sghms.ac.uk/phs/staff/jmb/jmb.htm>.

We shall also consider in a subsequent note the practicalities of using a random sequence to allocate treatments to patients.

1 Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *BMJ* 1999;318:1209.

2 Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1990: 540-4.

3 Treasure T, MacRae KD. Minimisation: the platinum standard for trials? *BMJ* 1998;317:362-3.

One hundred years ago

Generalisation of salt infusions

The subcutaneous infusion of salt solution has proved of great benefit in the treatment of collapse after severe operations. The practice, it may be said, developed from two sources: the new method of transfusion where water, instead of another person's blood, is injected into the patient's veins; and flushing of the peritoneum, introduced by Lawson Tait. After flushing, much of the fluid left in the peritoneum is absorbed into the circulation, greatly to the patient's advantage. Dr. Clement Penrose has tried the effect of subcutaneous salt infusions as a last extremity in severe cases of pneumonia. He continues this treatment with inhalations of oxygen. He has had experience of three cases, all considered hopeless, and succeeded in saving one. In the other two the prolongation of life and the relief of symptoms were so marked that Dr. Penrose regretted that the treatment had not

been employed earlier. Several physicians have adopted Dr. Penrose's method, and with the most gratifying results. The cases are reported fully in the *Bulletin of the Johns Hopkins Hospital* for July last. The infusions of salt solution were administered just as after an operation. The salt solution, at a little above body temperature, is poured into a graduated bottle connected by a rubber tube with a needle. The pressure is regulated by elevating the bottle, or by means of a rubber bulb with valves; the needle is introduced into the connective tissue under the breast or under the integuments of the thighs. There can be no doubt that subcutaneous saline infusions are increasing in popularity, and little doubt that their use will be greatly extended in medicine as well as surgery.

(*BMJ* 1899;ii:933)

Statistics Notes

The odds ratio

J Martin Bland, Douglas G Altman

Department of
Public Health
Sciences,
St George's
Hospital Medical
School, London
SW17 0RE

J Martin Bland
*professor of medical
statistics*

ICRF Medical
Statistics Group,
Centre for Statistics
in Medicine,
Institute of Health
Sciences, Oxford
OX3 7LF

Douglas G Altman
*professor of statistics
in medicine*

Correspondence to:
Professor Bland

BMJ 2000;320:1468

In recent years odds ratios have become widely used in medical reports—almost 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, $141/561 - 928/14\ 453 = 0.187$ (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

Eczema	Hay fever		Total
	Yes	No	
Yes	141	420	561
No	928	13 525	14 453
Total	1069	13 945	15 014

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 $141/1069 - 420/13\ 945 = 0.102$ (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

$$\frac{141/420}{928/13\ 525} \text{ and } \frac{141/928}{420/13\ 525}$$

which can both be rearranged to give

$$\frac{141 \times 13\ 525}{928 \times 420}$$

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,

$$SE(\log OR) =$$

$$\sqrt{\frac{1}{141} + \frac{1}{420} + \frac{1}{928} + \frac{1}{13\ 525}} = 0.103.$$

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 $\log(4.89) = 1.588$ and the confidence interval is $1.588 \pm 1.96 \times 0.103$, 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 χ^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.

We thank Barbara Butland for providing the data.

1 Strachan DP, Butland BK, Anderson HR. Incidence and prognosis of asthma and wheezing illness from early childhood to age 33 in a national British cohort. *BMJ* 1996;312:1195-9.

2 Bland JM, Altman DG. Transforming data. *BMJ* 1996;312:770.

3 Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *Evidence-Based Med* 1996;1:164-6.

*Statistics Notes***Blinding in clinical trials and other studies**

Simon J Day, Douglas G Altman

Leo
Pharmaceuticals,
Princes Risborough,
Buckinghamshire
HP27 9RR

Simon J Day
*manager, clinical
biometrics*

ICRF Medical
Statistics Group,
Institute of Health
Sciences, Oxford
OX3 7LF

Douglas G Altman
*professor of statistics
in medicine*

Correspondence to:
S J Day

BMJ 2000;321:504

Human behaviour is influenced by what we know or believe. In research there is a particular risk of expectation influencing findings, most obviously when there is some subjectivity in assessment, leading to biased results. Blinding (sometimes called masking) is used to try to eliminate such bias.

It is a tenet of randomised controlled trials that the treatment allocation for each patient is not revealed until the patient has irrevocably been entered into the trial, to avoid selection bias. This sort of blinding, better referred to as allocation concealment, will be discussed in a future statistics note. In controlled trials the term blinding, and in particular "double blind," usually refers to keeping study participants, those involved with their management, and those collecting and analysing clinical data unaware of the assigned treatment, so that they should not be influenced by that knowledge.

The relevance of blinding will vary according to circumstances. Blinding patients to the treatment they have received in a controlled trial is particularly important when the response criteria are subjective, such as alleviation of pain, but less important for objective criteria, such as death. Similarly, medical staff caring for patients in a randomised trial should be blinded to treatment allocation to minimise possible bias in patient management and in assessing disease status. For example, the decision to withdraw a patient from a study or to adjust the dose of medication could easily be influenced by knowledge of which treatment group the patient has been assigned to.

In a double blind trial neither the patient nor the caregivers are aware of the treatment assignment. Blinding means more than just keeping the name of the treatment hidden. Patients may well see the treatment being given to patients in the other treatment group(s), and the appearance of the drug used in the study could give a clue to its identity. Differences in taste, smell, or mode of delivery may also influence efficacy, so these aspects should be identical for each treatment group. Even colour of medication has been shown to influence efficacy.¹

In studies comparing two active compounds, blinding is possible using the "double dummy" method. For example, if we want to compare two medicines, one presented as green tablets and one as pink capsules, we could also supply green placebo tablets and pink placebo capsules so that both groups of patients would take one green tablet and one pink capsule.

Blinding is certainly not always easy or possible. Single blind trials (where either only the investigator or only the patient is blind to the allocation) are sometimes unavoidable, as are open (non-blind) trials. In trials of different styles of patient management, surgical procedures, or alternative therapies, full blinding is often impossible.

In a double blind trial it is implicit that the assessment of patient outcome is done in ignorance of

the treatment received. Such blind assessment of outcome can often also be achieved in trials which are open (non-blinded). For example, lesions can be photographed before and after treatment and assessed by someone not involved in running the trial. Indeed, blind assessment of outcome may be more important than blinding the administration of the treatment, especially when the outcome measure involves subjectivity. Despite the best intentions, some treatments have unintended effects that are so specific that their occurrence will inevitably identify the treatment received to both the patient and the medical staff. Blind assessment of outcome is especially useful when this is a risk.

In epidemiological studies it is preferable that the identification of "cases" as opposed to "controls" be kept secret while researchers are determining each subject's exposure to potential risk factors. In many such studies blinding is impossible because exposure can be discovered only by interviewing the study participants, who obviously know whether or not they are a case. The risk of differential recall of important disease related events between cases and controls must then be recognised and if possible investigated.² As a minimum the sensitivity of the results to differential recall should be considered. Blinded assessment of patient outcome may also be valuable in other epidemiological studies, such as cohort studies.

Blinding is important in other types of research too. For example, in studies to evaluate the performance of a diagnostic test those performing the test must be unaware of the true diagnosis. In studies to evaluate the reproducibility of a measurement technique the observers must be unaware of their previous measurement(s) on the same individual.

We have emphasised the risks of bias if adequate blinding is not used. This may seem to be challenging the integrity of researchers and patients, but bias associated with knowing the treatment is often subconscious. On average, randomised trials that have not used appropriate levels of blinding show larger treatment effects than blinded studies.³ Similarly, diagnostic test performance is overestimated when the reference test is interpreted with knowledge of the test result.⁴ Blinding makes it difficult to bias results intentionally or unintentionally and so helps ensure the credibility of study conclusions.

1 De Craen AJM, Roos PJ, de Vries AL, Kleijnen J. Effect of colour of drugs: systematic review of perceived effect of drugs and their effectiveness. *BMJ* 1996;313:1624-6.

2 Barry D. Differential recall bias and spurious associations in case/control studies. *Stat Med* 1996;15:2603-16.

3 Schulz KF, Chalmers I, Hayes R, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.

4 Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.

Irrationality, the market, and quality of care

Consider the irrationality of a person who pays extra so as not to share a hotel room with a colleague while on a business trip. He does this because he values privacy but he also scoffs at taking out long term care insurance to guarantee a private room in a nursing home. Why is he willing to risk sharing a room for the rest of his life with a person he does not like? This common irrationality is often masked by rationalisations such as "I would rather die than have to live in a nursing home." Yet we know that when the time comes most prefer the limited pleasures of life in a nursing home to suicide

their feet. There are even more fundamental reasons why depending on the rationality of the market will never work well for quality of care (box). Sensible policy for providing nursing home care requires a larger welfare state, a larger regulatory state, and encouragement of public, non-profit providers. Australia's recent experience shows that to head in the opposite direction is medically, economically, and politically irrational.

Competing interests: None declared.

- 1 McCallum J, Geiselhart K. *Australia's new aged: issues for young and old*. Sydney: Allen and Unwin, 1996.
- 2 Osborne D, Gaebler T. *Reinventing government*. New York: Addison-Wesley, 1992.
- 3 Jost T. The necessary and proper role of regulation to assure the quality of health care. *Houston Law Review* 1988;25:525-98.
- 4 Tingle L. Moran the big winner as aged care goes private. *Sydney Morning Herald* 16 March 2001:2.
- 5 Lohr R, Head M. Kerosene baths reveal systemic aged care crisis in Australia. World Socialist Web Site. www.wsws.org/articles/2000/mar2000/aged-m10.shtml (accessed 10 Mar 2000).
- 6 Jenkins A, Braithwaite J. Profits, pressure and corporate lawbreaking. *Crime, Law and Social Change* 1993;20:221-32.
- 7 Braithwaite J, Makkai T, Braithwaite V, Gibson D. *Raising the standard: resident centred nursing home regulation in Australia*. Canberra: Department of Community Services and Health, 1993.
- 8 Braithwaite J, Braithwaite V. The politics of legalism: rules versus standards in nursing home regulation. *Social and Legal Studies* 1995;4:307-41.
- 9 Black J. *Rules and regulators*. Oxford: Clarendon Press, 1997.
- 10 Braithwaite J, Makkai T. Can resident-centred inspection of nursing homes work with very sick residents? *Health Policy* 1993;24:19-33.
- 11 Makkai T, Braithwaite J. Praise, pride and corporate compliance. *Int J Sociology Law* 1993;21:73-91.
- 12 Braithwaite J. *Restorative justice and responsive regulation*. New York: Oxford University Press (in press).
- 13 McKibbin H. Accreditation: the on-site audit. *The Standard (Newsletter of the Aged Care Standards Agency)* 1999;2(2):2.
- 14 Power M. *The audit society*. Oxford: Oxford University Press, 1997.

*Statistics Notes***Concealing treatment allocation in randomised trials**

Douglas G Altman, Kenneth F Schulz

ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Douglas G Altman
professor of statistics in medicine

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA

Kenneth F Schulz
vice president, Quantitative Sciences

Correspondence to: D G Altman

BMJ 2001;323:446-7

We have previously explained why random allocation of treatments is a required design feature of controlled trials¹ and explained how to generate a random allocation sequence.² Here we consider the importance of concealing the treatment allocation until the patient is entered into the trial.

Regardless of how the allocation sequence has been generated—such as by simple or stratified randomisation²—there will be a prespecified sequence of treatment allocations. In principle, therefore, it is possible to know what treatment the next patient will get at the time when a decision is taken to consider the patient for entry into the trial.

The strength of the randomised trial is based on aspects of design which eliminate various types of bias. Randomisation of patients to treatment groups eliminates bias by making the characteristics of the patients in two (or more) groups the same on average, and stratification with blocking may help to reduce chance imbalance in a particular trial.² All this good work can be undone if a poor procedure is adopted to implement the allocation sequence. In any trial one or more people must determine whether each patient is eligible for the trial, decide whether to invite the patient to participate, explain the aims of the trial and the details of the treatments, and, if the patient agrees to participate, determine what treatment he or she will receive.

Suppose it is clear which treatment a patient will receive if he or she enters the trial (perhaps because

there is a typed list showing the allocation sequence). Each of the above steps may then be compromised because of conscious or subconscious bias. Even when the sequence is not easily available, there is strong anecdotal evidence of frequent attempts to discover the sequence through a combination of a misplaced belief that this will be beneficial to patients and lack of understanding of the rationale of randomisation.³

How can the allocation sequence be concealed? Firstly, the person who generates the allocation sequence should not be the person who determines eligibility and entry of patients. Secondly, if possible the mechanism for treatment allocation should use people not involved in the trial. A common procedure, especially in larger trials, is to use a central telephone randomisation system. Here patient details are supplied, eligibility confirmed, and the patient entered into the trial before the treatment allocation is divulged (and it may still be blinded⁴). Another excellent allocation concealment mechanism, common in drug trials, is to get the allocation done by a pharmacy. The interventions are sealed in serially numbered containers (usually bottles) of equal appearance and weight according to the allocation sequence.

If external help is not available the only other system that provides a plausible defence against allocation bias is to enclose assignments in serially numbered, opaque, sealed envelopes. Apart from neglecting to mention opacity, this is the method used in the famous 1948 streptomycin trial (see box). This

Description of treatment allocation in the MRC streptomycin trial⁵

“Determination of whether a patient would be treated by streptomycin and bed-rest (S case) or by bed-rest alone (C case) was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each centre by Professor Bradford Hill; the details of the series were unknown to any of the investigators or to the co-ordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel, and before admission to the streptomycin centre, the appropriate numbered envelope was opened at the central office; the card inside told if the patient was to be an S or a C case, and this information was then given to the medical officer of the centre.”

method is not immune to corruption,³ particularly if poorly executed. However, with care, it can be a good mechanism for concealing allocation. We recommend that investigators ensure that the envelopes are opened sequentially, and only after the participant's name and other details are written on the appropriate envelope.³ If possible, that information should also be transferred to the assigned allocation by using pressure sensitive paper or carbon paper inside the envelope. If an investigator cannot use numbered containers, envelopes represent the best available allocation concealment mechanism without involving outside parties, and may sometimes be the only feasible option. We suspect, however, that in years to come we will see greater use of external “third party” randomisation.

The desirability of concealing the allocation was recognised in the streptomycin trial⁵ (see box). Yet the importance of this key element of a randomised trial has not been widely recognised. Empirical evidence of the bias associated with failure to conceal the allocation^{6,7} and explicit requirement to discuss this issue in the CONSORT statement⁸ seem to be leading to wider recognition that allocation concealment is an essential aspect of a randomised trial.

Allocation concealment is completely different from (double) blinding.⁴ It is possible to conceal the randomisation in every randomised trial. Also, allocation concealment seeks to eliminate selection bias (who gets into the trial and the treatment they are assigned). By contrast, blinding relates to what happens after randomisation, is not possible in all trials, and seeks to reduce ascertainment bias (assessment of outcome).

- 1 Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *BMJ* 1999;318:1209.
- 2 Altman DG, Bland JM. How to randomise. *BMJ* 1999;319:703-4.
- 3 Schulz KF. Subverting randomization in controlled trials. *JAMA* 1995;274:1456-8.
- 4 Day SJ, Altman DG. Blinding in clinical trials and other studies. *BMJ* 2000;321:504.
- 5 Medical Research Council. Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *BMJ* 1948;2:769-82.
- 6 Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses. *Lancet* 1998;352:609-13.
- 7 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.
- 8 Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276:637-9.

The public health benefits of mobile phones

The bread and butter of public health on call is identifying contacts in the case of suspected meningococcal disease. On the whole this is straightforward but can occasionally cause difficulties. Most areas that I have worked in include several universities, and during October it is common to experience the problem of contact tracing in the student population.

There are two main problems. The first is how to define household contacts when the index patient lives in a hall of residence containing several hundred students. Finding the appropriate university protocol and not being too concerned about the different approaches adopted by neighbouring universities can reduce the number of sleepless nights. The second problem is harder. “Close kissing contacts” among 18 year olds who have been set free from parental control for the first time is a minefield. My experience suggests that it is best to assume there will be lots and that names and contact details will not necessarily have been obtained. By the end of a weekend on call, you will feel like a cross between a detective and an “agony aunt”

One year I volunteered to cover Christmas weekend in the belief that at least the students would be gone by then. I could not have been more mistaken. To add a further difficulty, the index patient presented to hospital on the night of the last day of term, and all contacts had already set off to the far reaches of the country. I could not believe my luck when the friend

accompanying the patient produced both their mobile phones and confidently reassured me that between the two of them they would have the mobile numbers of all 15 “household” contacts. She was right, and in just over two hours all of them had been contacted.

There has been much coverage in the medical and popular press about the potential health hazards of mobile phones, and if these fears are realised the 100% ownership among this small sample of students is worrying. However, in terms of contact tracing for suspected meningococcal disease, mobile phones have potential health benefits not just for their owners but also for the mental health of public health doctors. Of course, this may not solve the “close kissing contact” problem.

Debbie Lawlor *senior lecturer in epidemiology and public health, University of Bristol*

We welcome articles up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake*, or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for “Endpieces,” consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.

Statistics Notes

Analysing controlled trials with baseline and follow up measurements

Andrew J Vickers, Douglas G Altman

In many randomised trials researchers measure a continuous variable at baseline and again as an outcome assessed at follow up. Baseline measurements are common in trials of chronic conditions where researchers want to see whether a treatment can reduce pre-existing levels of pain, anxiety, hypertension, and the like.

Statistical comparisons in such trials can be made in several ways. Comparison of follow up (post-treatment) scores will give a result such as “at the end of the trial, mean pain scores were 15 mm (95% confidence interval 10 to 20 mm) lower in the treatment group.” Alternatively a change score can be calculated by subtracting the follow up score from the baseline score, leading to a statement such as “pain reductions were 20 mm (16 to 24 mm) greater on treatment than control.” If the average baseline scores are the same in each group the estimated treatment effect will be the same using these two simple approaches. If the treatment is effective the statistical significance of the treatment effect by the two methods will depend on the correlation between baseline and follow up scores. If the correlation is low using the change score will add variation and the follow up score is more likely to show a significant result. Conversely, if the correlation is high using only the follow up score will lose information and the change score is more likely to be significant. It is incorrect, however, to choose whichever analysis gives a more significant finding. The method of analysis should be specified in the trial protocol.

Some use change scores to take account of chance imbalances at baseline between the treatment groups. However, analysing change does not control for baseline imbalance because of regression to the mean^{1 2}: baseline values are negatively correlated with change because patients with low scores at baseline generally improve more than those with high scores. A better approach is to use analysis of covariance (ANCOVA), which, despite its name, is a regression method.³ In effect two parallel straight lines (linear regression) are obtained relating outcome score to baseline score in each group. They can be summarised as a single regression equation:

$$\text{follow up score} = \text{constant} + a \times \text{baseline score} + b \times \text{group}$$

where a and b are estimated coefficients and group is a binary variable coded 1 for treatment and 0 for control. The coefficient b is the effect of interest—the estimated difference between the two treatment groups. In effect an analysis of covariance adjusts each patient’s follow up score for his or her baseline score, but has the advantage of being unaffected by baseline differences. If, by chance, baseline scores are worse in the treatment group, the treatment effect will be underestimated by a follow up score analysis and overestimated by looking at change scores (because of regression to the mean). By contrast, analysis of covariance gives the same answer whether or not there is baseline imbalance.

As an illustration, Kleinhenz et al randomised 52 patients with shoulder pain to either true or sham acupuncture.⁴ Patients were assessed before and after treatment using a 100 point rating scale of pain and function, with lower scores indicating poorer outcome. There was an imbalance between groups at baseline, with better scores in the acupuncture group (see table). Analysis of post-treatment scores is therefore biased. The authors analysed change scores, but as baseline and change scores are negatively correlated (about $r = -0.25$ within groups) this analysis underestimates the effect of acupuncture. From analysis of covariance we get:

$$\text{follow up score} = 24 + 0.71 \times \text{baseline score} + 12.7 \times \text{group}$$

(see figure). The coefficient for group (b) has a useful interpretation: it is the difference between the mean change scores of each group. In the above example it can be interpreted as “pain and function score improved by an estimated 12.7 points more on average in the treatment group than in the control group.” A 95% confidence interval and P value can also be calculated for b (see table).⁵ The regression equation provides a means of prediction: a patient with a baseline score of 50, for example, would be predicted to have a follow up score of 72.2 on treatment and 59.5 on control.

An additional advantage of analysis of covariance is that it generally has greater statistical power to detect a treatment effect than the other methods.⁶ For example, a trial with a correlation between baseline and follow

Integrative
Medicine Service,
Biostatistics Service,
Memorial
Sloan-Kettering
Cancer Center, New
York, New York
10021, USA

Andrew J Vickers
assistant attending
research methodologist

ICRF Medical
Statistics Group,
Centre for Statistics
in Medicine,
Institute of Health
Sciences, Oxford
OX3 7LF

Douglas G Altman
professor of statistics
in medicine

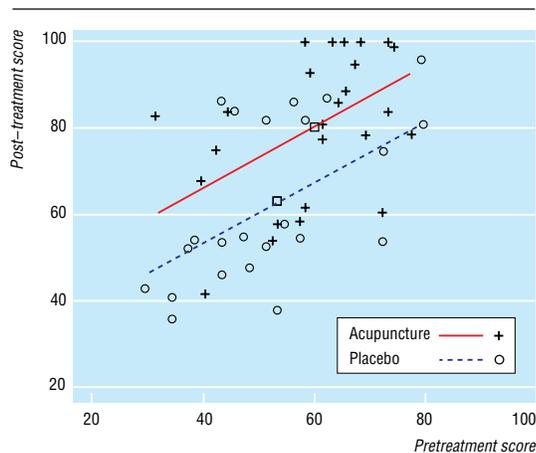
Correspondence to:
Dr Vickers
vickersa@mskccc.org

BMJ 2001;323:1123-4

Results of trial of acupuncture for shoulder pain⁴

	Pain scores (mean and SD)		Difference between means (95% CI)	P value
	Placebo group (n=27)	Acupuncture group (n=25)		
Baseline	53.9 (14)	60.4 (12.3)	6.5	
Analysis				
Follow up	62.3 (17.9)	79.6 (17.1)	17.3 (7.5 to 27.1)	0.0008
Change score*	8.4 (14.6)	19.2 (16.1)	10.8 (2.3 to 19.4)	0.014
ANCOVA			12.7 (4.1 to 21.3)	0.005

*Analysis reported by authors.⁴



Pretreatment and post-treatment scores in each group showing fitted lines. Squares show mean values for the two groups. The estimated difference between the groups from analysis of covariance is the vertical distance between the two lines

up scores of 0.6 that required 85 patients for analysis of follow up scores, would require 68 for a change score analysis but only 54 for analysis of covariance.

The efficiency gains of analysis of covariance compared with a change score are low when there is a high correlation (say $r > 0.8$) between baseline and follow up measurements. This will often be the case, particularly in stable chronic conditions such as obesity. In these

situations, analysis of change scores can be a reasonable alternative, particularly if restricted randomisation is used to ensure baseline comparability between groups.⁷ Analysis of covariance is the preferred general approach, however.

As with all analyses of continuous data, the use of analysis of covariance depends on some assumptions that need to be tested. In particular, data transformation, such as taking logarithms, may be indicated.⁸ Lastly, analysis of covariance is a type of multiple regression and can be seen as a special type of adjusted analysis. The analysis can thus be expanded to include additional prognostic variables (not necessarily continuous), such as age and diagnostic group.

We thank Dr J Kleinhenz for supplying the raw data from his study.

- 1 Bland JM, Altman DG. Regression towards the mean. *BMJ* 1994;308:1499.
- 2 Bland JM, Altman DG. Some examples of regression towards the mean. *BMJ* 1994;309:780.
- 3 Senn S. Baseline comparisons in randomized clinical trials. *Stat Med* 1991;10:1157-9.
- 4 Kleinhenz J, Streiberger K, Windeler J, Gussbacher A, Mavridis G, Martin E. Randomised clinical trial comparing the effects of acupuncture and a newly designed placebo needle in rotator cuff tendonitis. *Pain* 1999;83:235-41.
- 5 Altman DG, Gardner MJ. Regression and correlation. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with confidence*. 2nd ed. London: BMJ Books, 2000:73-92.
- 6 Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Med Res Methodol* 2001;1:16.
- 7 Altman DG, Bland JM. How to randomise. *BMJ* 1999;319:703-4.
- 8 Bland JM, Altman DG. The use of transformation when comparing two means. *BMJ* 1996;312:1153.

A memorable patient Informed consent

I first met Ivy three years ago when she came for her 29th oesophageal dilatation. She was an 86 year old spinster, deaf without speech from childhood, and the only sign language she knew was thumbs up, which she would use for saying good morning or for showing happiness. She had no next of kin and had lived in a residential home for the past 50 years. She developed a benign oesophageal stricture in 1992 and came to the endoscopy unit for repeated dilatations. The carers in the residential home used to say that she enjoyed her "days out" at the endoscopy unit.

We would explain the procedure to her in sign language. She would use the thumbs up sign and make a cross on the dotted line on the consent form. She would enter the endoscopy room smiling, put her left arm out to be cannulated, turn to her left side for endoscopy, and when fully awake would show her thumbs up again. Every time after her dilatation the nursing staff would question why an expandable oesophageal stent was not being considered. We would conclude that the indications for an expandable stent in benign strictures are not well established.

Her need for dilatation was becoming more frequent, and so on her 46th dilatation we decided to refer her to our regional centre for the insertion of a stent. She had an expandable stent inserted, and in his report the endoscopist mentioned the risk of the stent migrating down in the stomach beyond the stricture. Six weeks later she developed a bolus obstruction. At endoscopy it was noted that the stent had indeed migrated down. She consented to another stent. Four weeks later she had another bolus obstruction that could not be completely removed at the first attempt,

and she was brought back the following day for removal of the bolus by endoscopy.

She came to the endoscopy room but did not have her familiar smile. She looked around for a minute, got off her trolley, and walked out. Everyone in the endoscopy room understood that she was trying to say, "I've had enough."

She did not come back for a repeat endoscopy, and she stayed nil by mouth on intravenous fluids. Two weeks later she died of an aspiration pneumonia. We think she understood all the procedures she had agreed to. We also think it was informed consent. I hope we were right. She gave us a very clear message without saying a word on her last visit to the endoscopy room.

Do we really understand what aphasic patients are trying to tell us when we get informed consent for invasive procedures? We should try to read the non-verbal messages very carefully.

I Tiwari *associate specialist in gastroenterology, Broomfield Hospital, Chelmsford*

We welcome articles of up to 600 words on topics such as *A memorable patient*, *A paper that changed my practice*, *My most unfortunate mistake*, or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for "Endpieces," consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.

The quality and reliability of health information on the internet remains of paramount concern in Europe, as elsewhere. Self regulatory codes of ethics for health websites abound, yet the quality and practices of many are highly questionable.

Little progress seems to have been made, moreover, in assuring consumers that the information they share with health websites will not be misused. Several US studies have already concluded that websites' privacy practices do not match their proclaimed policies.⁵ In an attempt to counter this erosion of trust in Europe, the European Commission's guidelines for quality criteria for health related websites have recognised that there is no shortage of legislation in the field of privacy and security.⁶ They have drawn specific attention to a new recommendation regarding online data collection adopted in May 2001 that explains how European directives on issues such as data protection should be applied to the most common processing tasks carried out via the internet.⁷

The challenge facing Europe's health professionals and policymakers is to carefully craft the development of new approaches to the supervision of medical and pharmaceutical practice. Their ultimate goal is to raise

consumers' confidence in online healthcare. They must ensure that the mechanisms are put in place whereby health professionals themselves can benefit from using the internet, while still ensuring the highest standards of medical practice.

Avienda was formerly known as the Centre for Law Ethics and Risk in Telemedicine.

Competing interests: None declared.

- 1 http://news.bbc.co.uk/1/hi/english/uk/england/newsid_1752000/1752670.stm (accessed 5 Feb 2002).
- 2 Case C-322/01: Reference for a preliminary ruling by the Landgericht Frankfurt am Main by order of that court of 10 August 2001 in the case of Deutscher Apothekerverband e.V. against DocMorris NV and Jacques Waterval. *Official Journal of the European Communities No C* 2001 December 8:348/10.
- 3 Council Directive 1992/28/EEC of 31 March 1992 on the advertising of medicinal products for human use. (Articles 1(3) and 3(1)). *Official Journal of the European Communities No L* 1995 11 February:32/26.
- 4 Directive 2000/31/EC on mutual recognition of primary medical and specialist medical qualifications and minimum standards of training. *Official Journal of the European Communities No L* 2001 July 31:206/1-51.
- 5 Schwartz J. Medical websites faulted on privacy. *Washington Post* 2000 February 1.
- 6 http://europa.eu.int/information_society/eeurope/ehealth/quality/draft_guidelines/index_en.htm (accessed 5 Feb 2002).
- 7 European Commission. Recommendation 2/2001 on certain minimum requirements for collecting personal data on-line in the European Union. Adopted on 17 May 2001. http://europa.eu.int/comm/internal_market/en/dataprot/wpdocs/wp43en.htm (accessed 25 Jan 2002).

Statistics Notes

Validating scales and indexes

J Martin Bland, Douglas G Altman

Papers p 569

Department of
Public Health
Sciences, St
George's Hospital
Medical School,
London SW17 0RE
J Martin Bland
*professor of medical
statistics*

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

Douglas G Altman
*professor of statistics
in medicine*

Correspondence to:
Professor Bland
mbland@sghms.ac.uk

BMJ 2002;324:606-7

An index of quality is a measurement like any other, whether it is assessing a website, as in today's *BMJ*,¹ a clinical trial used in a meta-analysis,² or the quality of a life experienced by a patient.³ As with all measurements, we have to decide whether it measures what we want it to measure, and how well.

The simplest measurements, such as length and distance, can be validated by an objective criterion. The earliest criteria must have been biological: the length of a pace, a foot, a thumb. The obvious problem, that the criterion varies from person to person, was eventually solved by establishing a fundamental unit and defining all others in terms of it. Other measurements can then be defined in terms of a fundamental unit. To define a unit of weight we find a handy substance which appears the same everywhere, such as water. The unit of weight is then the weight of a volume of water specified in the basic unit of length, such as 100 cubic centimetres. Such measurements have *criterion validity*, meaning that we can take some known quantity and compare our measurement with it.

For some measurements no such standard is possible. Cardiac stroke volume, for example, can be measured only indirectly. Direct measurement, by collecting all the blood pumped out of the heart over a series of beats, would involve rather drastic interference with the system. Our criterion becomes agreement with another indirect measurement. Indeed, we sometimes have to use as a standard a method which we know produces inaccurate measurements.

Some quantities are even more difficult to measure and evaluate. Cardiac stroke volume does at least have an objective reality; a physical quantity of blood is pumped out of the heart when it beats. Anxiety and depression do not have a physical reality but are useful artificial constructs. They are measured by questionnaire scales, where answers to a series of questions related to the concept we want to measure are combined to give a numerical score. Website quality is similar. We are measuring a quantity which is not precisely defined, and there is no instrument with which we can compare any measure we might devise. How are we to assess the validity of such a scale?

The relevant theory was developed in the social sciences in the context of questionnaire scales.⁴ First we might ask whether the scale looks right, whether it asks about the sorts of thing which we think of as being related to anxiety or website quality. If it appears to be correct, we call this *face validity*. Next we might ask whether it covers all the aspects which we want to measure. A phobia scale which asked about fear of dogs, spiders, snakes, and cats but ignored height, confined spaces, and crowds would not do this. We call appropriate coverage of the subject matter *content validity*.

Our scale may look right and cover the right things, but what other evidence can we bring to the question of validity? One question we can ask is whether our score has the relationships with other variables that we would expect. For example, does an anxiety measure

distinguish between psychiatric patients and medical patients? Do we get different anxiety scores from students before and after an examination? Does a measure of depression predict suicide attempts? We call the property of having appropriate relationships with other variables *construct validity*.

We can also ask whether the items which together compose the scale are related to one another: does the scale have *internal consistency*? If not, do the items really measure the same thing? On the other hand, if the items are too similar, some may be redundant. Highly correlated items in a scale may make the scale over-long and may lead to some aspects being overemphasised, impairing the content validity. A handy summary measure for this feature is Cronbach's alpha.⁵

A scale must also be repeatable and be sufficiently objective to give similar results for different observers. If a measurement is repeatable, in that someone who has a high score on one occasion tends to have a high score on another, it must be measuring something. With physical measurements, it is often possible for the same observer (or different observers) to make repeated measurements in quick succession. When there is a subjective element in the measurement the observer can be blinded from their first measurement,

and different observers can make simultaneous measurements. In assessing the reliability of a website quality scale, it is easy to get several observers to apply the scale independently. With websites, repeat assessments need to be close in time because their content changes frequently (as does *bmj.com*). With questionnaires, either self administered or recorded by an observer, repeat measurements need to be far enough apart in time for the earlier responses to be forgotten, yet not so far apart that the underlying quantity being measured might have changed. Such data enable us to evaluate *test-retest reliability*. If two measures have comparable face, content, and construct validity the more repeatable one may be preferred for the study of a given population.

- 1 Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. *BMJ* 2002;324:569-73.
- 2 Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-6.
- 3 Muldoon MF, Barger SD, Flory JD, Manuck B. What are quality of life measurements measuring? *BMJ* 1998;316:542-5.
- 4 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 2nd ed. Oxford: Oxford University Press, 1996.
- 5 Bland JM, Altman DG. Cronbach's alpha. *BMJ* 1997;314:572.

Honour a physician with the honour due unto him

A few years ago my general practitioner told me that anyone aged over 40 with upper abdominal discomfort needed investigating. At the local teaching hospital, a pleasant young doctor did a gastroscopy, which showed a mass in my stomach wall. I was sent for a barium meal. A consultant radiologist took the x ray films, instructing me briskly to turn this way and that but not otherwise paying me any attention. He told me to wait a few minutes while he checked the films to see if all the views were satisfactory. I sat alone in the room for about five minutes.

From the moment the consultant re-entered I could see that he was slightly agitated. "I'm terribly sorry," he called out as he came through the door at the far end. And then again, "I'm terribly sorry." Perhaps these words of regret, coupled with the concern on his face, might not have had the effect they did had I not been a man with an abdominal mass on his mind. At this moment of truth and reckoning, certain visions swam before my eyes.

Three strides later, he was in front of me and looking me full in the face: "I'm terribly sorry, I hadn't realised you were a doctor." In his hand was the request form, and I could see that my general practitioner had written "ex-SR here" in one corner. He must have spotted this when checking the form as he looked at the preliminary plates. Though no further x rays were needed, he proceeded a little breathlessly to deliver three or four minutes of almost a caricature of caring, empathic interest in a patient. What branch of medicine was I in, and where did I work? Good heavens, that must be tough. Is that an Australian accent I hear? A St Mary's old boy, ah yes. What did I think about...?

I don't mean to imply that this was insincere, merely splendidly different from his earlier matter of factness and economy of word. I had thought nothing of this at the time: in such a bread and butter procedure I had

no more reason to expect the doctor to engage with me as a person than I would the phlebotomist taking a routine blood sample. Clearly, this consultant saw things similarly as a rule, but when the patient was a doctor the aesthetics of the encounter changed. He had apologised three times for what he felt was a lapse on his part, arising from his failure to notice what was written in the corner of the request form. Perhaps he thought I knew that my general practitioner had written this and that I expected this of a medical referral, and thus expected to be recognised by him not just as a patient but also as a colleague. He seemed to see this as my due. (As it happens, I didn't.)

I had forgotten this incident, but it was brought back to me by the aftermath of the Bristol cardiac surgery debacle, and by the publicity surrounding other recent medical scandals. These have all put a spotlight on relations between doctors, who seem to offer each other acknowledgement and empathy, as my consultant had sought belatedly to do to me. The general public may be coming to suspect that this collegiate solidarity is somehow not in their interests, associating it with mutual protectiveness and thus with cover-ups of medical malpractice. It is too soon to say how the profession will react, but my consultant was an older man and my guess is that, with younger generations of doctors, we will see the waning of a tradition whose roots lie with Hippocrates. For it was his oath that bound doctors to look well on each other (and not charge each other for their services).

It's another story, but I found out later that the mass was the gastroscopy instrument itself distorting the stomach wall, misdiagnosed by an inexperienced registrar. No special treatment there, anyway.

Derek Summerfield *consultant psychiatrist, CASCAID, South London and Maudsley NHS Trust, London*