AIDS case-definition, the criteria of persistent cough for longer than 1 month or general lymphadenopathy should not be used in patients with proven tuberculosis. With this modification, the WHO clinical definition was 93% specific, 55% sensitive, and had a positive predictive value of 80% for HIV seropositivity.

We have confirmed the strong association between a history of varicella zoster infection and HIV seropositivity in another study[11] and therefore propose that "a history of herpes zoster during the previous 5 years" replaces "recurrent herpes zoster" as a minor criterion in the clinical definition.

In other regions or countries, the diagnostic value of the provisional WHO clinical case-definition may differ. The positive predictive value of a definition for infection will be higher in a population where the prevalence of HIV infection is high. The specificity of the definition will be higher in places that have few other diseases with AIDS-like clinical features (such as tuberculosis, malnutrition, cancer, trypanosomiasis). In the department of internal medicine at Mama Yemo Hospital all conditions were present to give a high specificity and positive predictive value for a clinical case-definition of AIDS. Indeed, among hospital inpatients, HIV seroprevalence was very high and other diseases with AIDS-like symptoms were rare.

For diagnostic purposes, laboratory confirmation of HIV infection is highly desirable if it is available. Because the main use of the provisional WHO clinical case-definition will be surveillance of the disease, epidemiological criteria such as being in a risk group were not included in this definition.

Although our results encourage the use of the WHO clinical case-definition for AIDS in Africa, expected variation in HIV seroprevalence and disease expression require that a similar evaluation be carried out in other regions.

We thank the commissioners of health K. Mushubekwa and K. Ngandu; the Belgian-Zairian Cooperation; the European Economic Community; the Banque Commerciale du Zaire; the Belgolaise; the Union des banques Zairoises; the US Public Health Service; and physicians Binyingo Embonga and Robert Ryder.

Correspondence should be addressed to P. P., Department of Microbiology, Institute of Tropical Medicine, Nationalestraat 155, B-2000 Antwerp, Belgium.

REFERENCES

1. Piot P, Quinn TC, Taelman H, et al. Acquired immunodeficiency syndrome in a heterosexual population in Zaire. *Lancet* 1984; ii: 65–69.
2. Van de Perre P, Rouvroy D, Lepage P, et al. Acquired immunodeficiency syndrome in Rwanda. *Lancet* 1984; ii. 62–65.
3. Serwadda D, Mugerwa RD, Sewankambo NK, et al. Slim disease: a new disease in Uganda and its association with HTLV-III infection. *Lancet* 1985; ii: 849–52.
4. Bayley AC, Dgwning RG, Cheingsong-Popov R, et al. HTLV-III distinguishes atypical and endemic Kaposi's sarcoma in Africa. *Lancet* 1985; i· 359–61
5. Brunet JB, Ancelle RA. The international occurrence of the acquired immunodeficiency syndrome. *Ann Intern Med* 1985; **103:** 670–73.
6. World Health Organisation Workshop on AIDS in Central Africa. Bangui Central African Republic, 22–24 Oct, 1985. WHO/CDS/SIDA/85 1.
7. World Health Organisation. Acquired immunodeficiency syndrome (AIDS). *Weekly Epidemol Rec* 1986; **61:** 69–73.
8. Sandstrom EG, Schooley RT, Ho DD. Detection of human anti HTLV-III antibodies by indirect immunofluorescence using fixed cells. *Transfusion* 1985; **25:** 308–12.
9. Mann JM, Francis H, Quinn TC, et al. HTLV-III/LAV seroprevalence among hospital workers in Kinshasa, Zaire: lack of association with occupational exposure. *JAMA* 1986, **256:** 3099–102.
10. Mann JM, Snider DE, Francis H, et al Association between HTLV-III/LAV infection and tuberculosis in Zaire. *JAMA* 1986; **256:** 346.
11. Colebunders R, Mann JM, Francis H, et al. Herpes zoster and LAV/HTLV-III infection in Africa, Zaire Abstracts of International Conference on AIDS. Paris, 23–25 June, 1986.

Statistical Analysis

CONFIDENCE INTERVALS

C. J. BULPITT

Department of Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT

MUCH of medical research involves the comparison of two or more groups. The p or probability value for a comparison between two groups indicates the long-term probability that the groups originate from the same population; if the p value is low it is unlikely that the groups arise from the same population. The researcher may therefore summarise the result of a study as "the difference between the groups was not significant, $p > 0{\cdot}05$" or "the difference was significant, $p < 0{\cdot}05$". The precise p value should be quoted to at least 2 decimal places, but even then it gives no information on the actual differences observed or on the range of possible differences between the groups. The confidence interval gives this range.

Medical research is concerned as much with estimation as with hypothesis testing and, since repeated studies would provide different estimates, we need to know an interval of values that is likely to contain the true result.[1] Our sample statistic—for example, the mean difference—is our best estimate of the result. However, this figure is only a single result. If a second, third, or even a hundred such comparisons were made, what would be the results? The confidence limits (CL) for one comparison encompass a range of values (the confidence interval [CI]) that is likely to cover the true population value.

DEFINITION OF A CONFIDENCE INTERVAL

The confidence interval is a range of values that is likely to cover the true but unknown value. In classic significance testing the confidence interval is based on the concept of repeated trials or studies—ie, with the 95% confidence interval, if the study was repeated 100 times, the confidence interval would be expected to include the true value on 95 occasions. However, a reported CI either does or does not include the true value.

USE OF THE CONFIDENCE INTERVAL

In 1973 Wulff[2] pointed out that confidence limits were well known to everyone interested in statistics, but rarely encountered in medical journals. Rothman provided powerful support in 1978,[3] and many workers have since advocated the use of confidence intervals and presented the results of such calculations.[4-7] Gardner and Altman[8] have formulated a policy for the *British Medical Journal*—namely, that "Confidence intervals, if appropriate to the type of study, should be used for major findings in both the main text of a paper and its abstract".

The term confidence interval rather than confidence limits is to be preferred. The term limits implies that values are not possible beyond these and ignores the fact that in 5 out of 100 studies the true values would be outside these limits. Gardner and Altman also predicted that authors would tend to selectively quote one limit rather than the other and this has already happened. The confidence interval cannot, by definition, be presented as a single figure. For example, the 95% confidence interval for a mean should

be represented as follows: mean x units, 95% CI y to z units.

In an article on the use of ketanserin in the treatment of traumatic vasospastic disease[9] the summary stated "Ninety-five percent confidence intervals for difference between the treatments, however, showed that finger systolic pressure may be 80% better and rewarming time 256 seconds faster after treatment with ketanserin than after placebo". The enthusiasm of the authors for ketanserin was apparent since they only reported the upper confidence limits in the summary. However, to be fair, I must say that they reported the full 95% confidence intervals in the results section (as −3% to +80% and −32 to +256 seconds).

THE RELATION BETWEEN HYPOTHESIS TESTING AND THE CONFIDENCE INTERVAL

Rothman emphasised that testing for statistical significance and the reporting of a "significant" or "non-significant" result may be useful in decision-making, but is of little value in medical research where the problem is one of estimation.[3] He pointed out that highly "significant" p values may, in a large study, be associated with trivial and biologically unimportant results. Similarly, in a small study a "non-significant" result may, if true, be of great biological importance. The results of medical research must be presented as a problem of estimation and the confidence interval gives us a range of values compatible with the data.

Two problems of presentation exist—firstly, when there is a statistically significant finding to report; and, secondly, when the authors judge the result to be negative. Let us assume that the authors are comparing two groups. In the first instance, when a statistically significant difference is reported, the readers require not only the mean difference between the two groups but also the range of outcomes with which the result is compatible. When the result is negative, the reader wishes to know whether the range of possible outcomes is grouped close to zero or whether the interval encompasses a wide range of positive or negative outcomes.

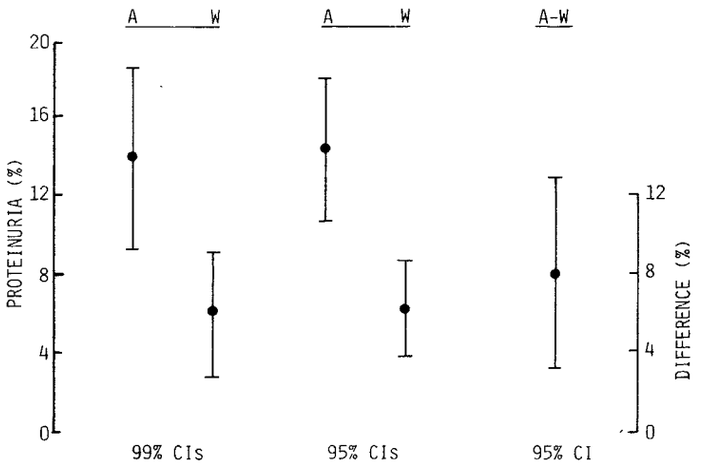
Whether the results of a study are statistically significant or not, the authors should report a range of outcomes compatible with their data—the confidence interval. When the result is significant at the 5% level, the 95% confidence interval will not encompass zero. When the result is "non-significant" the confidence interval will include zero. The interval may be wide when the data are very variable, or the numbers studied are very small, or both. (In these instances the power of the study to detect the observed difference would be said to be low.) Having a greater number of patients would narrow the CI—for example, a fourfold increase in each group would halve the width of the confidence interval.

GENERAL PRINCIPLES FOR CALCULATING THE CONFIDENCE INTERVAL

The appropriate CI and the level of confidence to be employed must be discussed especially when multiple statistical tests are employed or subgroup analysis is performed. The biological significance of the interval must also be considered.

The Appropriate Confidence Limit

In comparison of two proportions or means, confidence intervals may overlap for the individual sample proportions or means, but the difference may still be statistically significantly different. The CI for the difference should be presented when only two groups are considered. However, it is often helpful to present also the CI for separate groups of data, especially when large numbers of groups have to be compared.



Percentage of diabetics with proteinuria.

99% and 95% confidence intervals for Asians (A) and whites (W) together with the 95% CI for the difference between Asians and whites (A–W). Ref 10.

Samanta and colleagues[10] reported that 14% of 370 Asians with diabetes had proteinuria compared with 6% of 368 whites. In correspondence Cruickshank[11] argued that selection bias might be large in this study, and that confidence intervals should also be calculated. He reported the 99% CI at 9·4–18·6% for Asians and 2·8–9·2% for whites (figure) and commented that these distributions are only just apart. However, these are not the most appropriate CIs. The 95% CI for the difference of 8% ranged from 3·4% to 12·7% and did not include zero (figure).

The Level of Confidence to be Employed

There is no rule that the 95% confidence limits should be reported. A 90% or a 99% interval may be presented. However, it must be understood that the 99% interval will be much wider than the 95% interval, and the 95% interval wider than the 90% interval.

90% limits tend to be employed in small studies to limit the width of the confidence interval, and 99% limits in large studies (usually in conjunction with a definition of statistical significance at the 1% level) when the interval is narrow. However, the 90% CI may be employed to give an impression of greater than usual precision and the 99% in a critical appraisal of the data. In the example in the figure, 99% rather 95% CIs were initially calculated. This strategy tends to minimise the difference between the groups but may be appropriate when many variables are being examined. In order to avoid bias in overemphasising or underestimating a difference, consistent use of the 95% CI is to be preferred.

Subgroup Analysis

Analysis by subgroups may lead to different statistical results in each subgroup. Even if the subgroups all show the same result—for example, the same difference in prevalence of proteinuria between male and female Asians and whites—then the number of subjects and the levels of significance will be decreased in all subgroups and the confidence limits will be wider in the subgroups than in the group as a whole. In the example above, the difference

between all Asians and all whites was statistically significant, but Cruickshank's letter[11] included the statement, "For the sexes separately 99% confidence intervals overlap with those for whites"—a good example of the subgroup analysis hiding a possibly real difference.

A different problem occurs when the null hypothesis is rejected for the group as a whole but is not rejected in one of the subgroups. Analysis of subgroups provides little evidence on when to accept or reject the null hypothesis. Where possible, subgroup analysis should be avoided.

What to do when Several Statistical Tests are Employed

When many statistical tests are employed, some say that the confidence limits should be wider (eg, 99%). Multiple testing arises in three main ways—repeated looks at the data (eg, as the numbers increase during a study); examination of several variables in the course of a single study; and the performance of several independent studies. Let us consider the effect of each of these on the confidence interval.

Repeated looks at the data.—This will happen in a long-term trial when the ethical or steering committee examines the data each year. The level of significance must be adjusted for the effect of repeated looks. There is no argument over the necessity for such an adjustment and the methods have been described.[12] These methods should be used to calculate 95% CIs that take into account the repeated looks at the data.

Several variables in a single study.—The examination of several variables in one study may lead to a demand for a higher level of significance for any one variable. Alternatively, multivariate methods may be employed to see whether there is an overall difference between two groups.[13] If such a difference is statistically significant, then the individual variables may be compared: you can either use Bonferroni's adjustment[14] or require a fixed high level of significance for each variable (eg, 1%). In this case it seems best to calculate the 99% confidence interval for one of many variables.

Many similar studies considered at once.—The performance of more than one study also theoretically increases the chance of one false positive result. When a series of studies or trials is being examined we can expect one or more to provide a false positive result. (In reality, the tendency to perform small studies means that falsely negative results are more likely.) When the type 1 (α) error exactly equals 5%, then the probablity of a false positive result is 0·05, and the probability of a "correct" result is 0·95. For twenty studies the probability of a false positive result is $1-0{\cdot}95^{20}=0{\cdot}64$. It is often assumed that we expect one false positive result in 20. In fact, the likelihood of at least one false positive result in 20 studies is 0·64. Fortunately, positive results have type 1 errors often appreciably less than $p=0{\cdot}05$ (eg, $p=0{\cdot}01$), so in well-designed experimental studies such as randomised controlled trials false positive results are very rare.[15] False positive results are frequent in observational studies, such as case-control studies, but there the false results are usually due to bias from confounding variables and the underlying statistical theory is still valid. The 95% CI is still the most appropriate interval to report since a powerful bias cannot be allowed for by adjusting the level of significance.

Biological importance of the confidence limits

The *British Medical Journal*'s campaign for confidence intervals started in a controversial way. In his editorial Langman[17] quoted a mean difference in systolic blood pressure between diabetics and non-diabetics of 6 mm Hg ($p<0{\cdot}02$, 95% CI 1·1–10·9 mm Hg, higher in diabetics) and, asserted, "Not only is the mean difference of 6 mm Hg rather small, however, and so unlikely to have practical relevance—but we also say using the confidence interval that there is only a 2·5% chance that the true difference in the population at large is greater than 10·9 mm Hg. Again, this is a figure which is unlikely to be of clinical importance; so the conclusion must be that we are unlikely to be missing a large and clinically important difference." Many will not agree with Langman that 10 or even 6 mm Hg is unimportant. However, the confidence interval allows the reader to assess the matter for himself or herself.

GRAPHICAL PRESENTATION OF THE CONFIDENCE INTERVAL

Many figures provide mean results with bars sticking out above and below (and often sideways as well). Sometimes the reader is not informed what these represent. Usually they provide a graphical illustration of one standard deviation (SD) either side of the mean or one standard error of the mean (SEM) on either side. What can be understood from the figure depends on the amount of overlap or separation between the lines, the relative length of the intervals, and the sample sizes. Browne[16] concluded that the only truly universal rule is that when SEM intervals overlap, the means are never significantly different.

When two means are compared the best illustration is that of the mean difference and bars showing the 95% CI for the difference (figure). However, when we compare two separate means with more than ten observations in each group, if the confidence intervals do not overlap then the means are significantly different. It would greatly simplify the understanding of graphical presentation if authors presented the 95% CI in a standard bar format.

CALCULATION OF CONFIDENCE INTERVALS

The following sections give brief descriptions of how to calculate the confidence intervals for a mean result, the difference between two mean results, a proportion, the difference between two proportions, a regression coefficient, and an estimate of relative risk. References are given for the computation of other confidence intervals such as those for a median, correlation coefficient, and so on.

The 95% CI for a single mean.—For normally distributed quantitative data, let SD be the standard deviation, SE the standard error, n the sample size, and $\bar{x}$ the mean. Then $SE=SD/\sqrt{n}$ and for samples over 30 the 95% CI ranges from $\bar{x}-(1{\cdot}96\times SE)$ to $\bar{x}+(1{\cdot}96\times SE)$. (For samples less than 30, the multiplier is not 1·96 and must be taken from tables of the t distribution with $n-1$ degrees of freedom.[8,18] The multiplier 2 is used below as an approximate to 1·96.)

The 95% CI for the difference between two means.—The calculation is the same as above when normally distributed quantitative data are compared. The 95% CI ranges from mean difference $-(2\times SED)$ to mean difference $+(2\times SED)$ where SED is the standard error of the difference and the number of comparisons is more than 30. However, the SED differs according to whether the differences are paired, as in a within-patient or paired analysis, or whether the average difference is simply the difference between two means (an unpaired analysis). Gardner and Altman[8] have clearly shown how to make these calculations and have provided examples.

The 95% CI for a proportion.—A proportion, such as the proportion of subjects with proteinuria, does not represent normally distributed data and the calculations differ from those for a single mean. Again, let SE be standard error, n be the number in the sample, and p be the proportion with a certain characteristic: then $SE=\sqrt{p(1-p)/n}$ and the 95% confidence interval ranges from $p-(1{\cdot}96\times SE)$ to $p+(1{\cdot}96\times SE)$. The multiplying factor (1·96) is independent of sample size in this instance, but the calculation is only approximate for small values of $n\times p$.

The 95% CI for a difference between two proportions.—Gardner and Altman[8] have again provided a worked example of how to derive the CI for the difference between two proportions, both in the paired and in the unpaired case.

The 95% CI in regression analysis.—Armitage[18] has described how to calculate the confidence interval for a regression coefficient

(slope, b). Again, the CI for a sample over 30 is given by $b \pm 2 \times SE(b)$ where SE(b) is the standard error of the slope b. For smaller numbers, t tables have to be consulted and the degrees of freedom are $n-2$. Useful sections in Snedecor and Cochran[19] indicate how to get CIs for correlation coefficients and provide the basis for determining the CI of the differences between two correlation or regression coefficients.

The 95% CI for a reduction in mortality or morbidity rates in a randomised controlled trial.—During a randomised controlled trial, events may be reduced by no more than 100% but can be increased without limit. The confidence limits may therefore extend more into a positive increase than a negative reduction and are said not to be symmetrical around the observed result. For example, a −3% reduction in mortality may have a 95% CI from −42% to +62%. Katz and colleagues[20] have provided the necessary methods of calculation.

The 95% CI for relative risk.—The relative risk (RR) from having a certain factor (eg, smoking) may be determined either as the true relative rate measured in a longitudinal study

$$RR = \frac{\text{incidence rate/100/yr in persons who are factor positive}}{\text{incidence rate/100/yr in persons who are factor negative}}$$

or as an approximate result derived from a case-control study. Katz et al[20] give the method for calculating the CI in the former case and Armitage[18] provides a method of determining the CI for an estimate of relative risk in a case-control study.

The 95% CI when non-parametric statistics are employed.—It is not always possible, or appropriate, to calculate confidence intervals for quantitative data that do not have a normal distribution.[21] However, sometimes the data can be transformed so that they assume a fairly normal distribution. It is also possible to calculate confidence intervals for the median.[22]

GUIDELINES

1. The 95% confidence interval of the mean should be reported as follows:

mean x units, 95% CI y to z units.

2. When two groups are compared the appropriate confidence limits are usually the confidence limits for the difference between the two results.

3. It is neither misleading nor incorrect to report 90% or 99% confidence limits. However, confusion may be avoided by using 95% and bias will be avoided from over-emphasising or underestimating any difference.

4. Bias in reporting occurs if only the upper or lower confidence limit is reported; therefore the full confidence interval should always be reported.

REFERENCES

1 Bhattacharyya GK, Johnson RA. Statistical concepts and methods. New York: Wiley, 1977.
2. Wulff HR. Confidence limits in evaluating controlled therapeutic trials. *Lancet* 1973; ii: 969–70
3. Rothman KJ. A show of confidence. *N Engl J Med* 1978; **299:** 1362–63.
4. Rose GA Beta-blockers in immediate treatment of myocardial infarction. *Br Med J* 1980; **280:** 1088
5. Editorial. Interpreting clinical trials. *Br Med J* 1978; ii. 1318.
6. Baber NS, Lewis JA. Beta-blockers in treatment of myocardial infarction. *Br Med J* 1980; ii: 59.
7 Gore SM. Statistics in question. Assessing methods—confidence intervals. *Br Med J* 1981; **283:** 660–62.
8 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986; **292:** 746–50.
9 Larsen VH, Fabricus J, Nielsen GL, Hansen KS. Ketanserin in the treatment of traumatic vasospastic disease. *Br Med J* 1986; **293:** 650–52.
10 Samanta A, Burden AC, Feehally J, Walls J. Diabetic renal disease: differences between Asian and white patients. *Br Med J* 1986; **293:** 366–67.
11. Cruickshank JK. Diabetic renal disease: differences between Asian and white patients. *Br Med J* 1986; **293:** 696.
12 McPherson K. Statistics: The problem of examining accumulating data more than once. *N Engl J Med* 1974, **290:** 501–02.
13 Cupples L, Heeren T, Schatzkin A, Colton T. Multiple testing of hypotheses in comparing two groups. *Ann Intern Med* 1984, **100:** 122–29.
14 Wallenstein S, Zucker CL, Fleiss JL. Some statistical methods useful in circulation research *Circ Res* 1980; **47:** 1–9.
15 Bulpitt CJ. Randomised controlled clinical trials. The Hague: Martinus Nijhoff, 1983: 241

References continued at foot of next column

Conference

HUMAN GENETICS: FRAGILE SITES STILL A MYSTERY

FRAGILE sites are a feature of human chromosomes which has puzzled cytogeneticists for a long time. They are visible as elongated, lightly staining regions between two chromosome bands which are normally adjacent (see accompanying figure). Such regions are visible only in metaphases of cells that have been subjected to various treatments known or suspected to interfere with DNA replication. There are two types of fragile sites–rare and common. Common sites are seen in cells from any individual, but only in a small proportion of metaphases (typically a few per cent); more than sixty such sites have been mapped on the various human chromosomes. Rare fragile sites are present only in some individuals (and inherited in families) and are seen in a higher proportion of metaphases (typically 20% and sometimes as high as 90%). Nineteen rare fragile sites have been reported; much of the interest in this subject is focused on the site close to the end of the long arm of the X chromosome, in view of its very strong association with the X-linked mental retardation syndrome.[1,2]

An EMBO workshop held in Marseille last September provided an opportunity for medical geneticists, cytogeneticists, and recombinant DNA scientists to exchange views and pool knowledge on this intriguing subject. The molecular nature of fragile sites remains unknown: none of them has been cloned so far, and in spite of the considerable number of probes mapped, in particular to the q27–q28 bands of the X chromosome, none of these has been shown to map in the X fragile site which lies between these two bands. Fragile sites are a chromosomal phenotype which requires a particular environment (culture

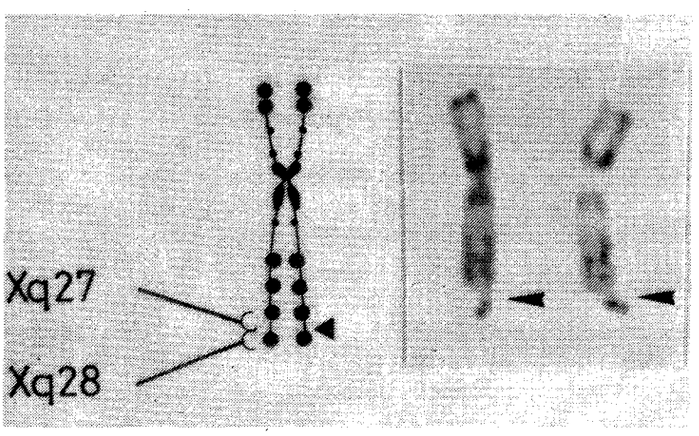


Diagram of the human X chromosome as seen by R-banding (left) and two X chromosomes (right) displaying the Xq27 fragile site (courtesy of M-G. Mattei).

C. J. BULPITT: REFERENCES—*continued*

16. Browne RH. On visual assessment of the significance of a mean difference. *Biometrics* 1979; **35:** 657–65.
17. Langman MJS. Towards estimation and confidence limits. *Br Med J* 1986; **292:** 716.
18. Armitage P. Statistical methods in medical research. Oxford: Blackwell, 1971.
19. Snedecor GW, Cochran WG. Statistical methods. Ames, Iowa: State University Press, 1967.
20. Katz D, Baptista J, Azer SP, Pike MC. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 1978; **34:** 469–71.
21. Evans C. Confidence intervals, medical housing need, and inappropriate statistics. *Br Med J* 1986; **293:** 756.
22. Diem K, Lentner C, eds. Documenta Geigy Scientific tables, 7th ed. Basel Geigy, 1970.