STATISTICS IN MEDICINE, VOL. 8, 1183–1190 (1989)

ADJUSTING FOR BASELINE: CHANGE OR PERCENTAGE CHANGE?

LEE KAISER

Biometrics Department, Sterling-Winthrop Research Institute, Rensselaer, New York 12144, U.S.A.

SUMMARY

Clinical trials data often include baseline and response measurements on each patient. Comparisons of treatments commonly employ either change from baseline or percentage change from baseline as the analysis variable. This paper provides guidance on the choice between these two by means of plots of these variables versus baseline. I provide two examples that illustrate the problem and I discuss the impact of use of the wrong adjusted response.

KEY WORDS Baseline Change Difference Percentage change Post-treatment Pre-treatment Relative change Ratio

1. INTRODUCTION

Clinical trials often include a baseline, measured on a continuous scale, obtained before treatment and a response obtained after treatment (or during treatment in a long-term trial). Examples include: exercise tolerance time in congestive heart failure,[1] vital signs in hypertension,[2] grip strength in arthritis,[3] pulmonary function in asthma,[4] Hamilton depression scores in depression,[5] and number of cysts in acne.[6] Analysis of data from such trials often involves adjustment of the response for the baseline, most commonly *change* (= response − baseline) and *percentage change* (= 100 × change/baseline).

The purpose of this article is to provide guidance on the choice between these. Section 2 contains reasons for adjusting a response and a recommendation on how to choose between change and percentage change. The recommendation is illustrated in Section 3. Section 4 deals with the impact of use of the wrong adjusted response.

I do not discuss in this article the choice of baseline and response variables. Typically, multiple observations will be obtained both before and after treatment. The investigator and analyst must decide whether to use as response the mean, minimum, maximum, last observation, or some other summary variable of the post-treatment observations. They need to decide similarly on the choice of baseline. Also, perhaps neither change nor percentage change is appropriate for analysis. Another possible adjustment for baseline is analysis of covariance. For a recent discussion of when such analysis is appropriate see References 7, 8, and 9.

2. CHOOSING BETWEEN CHANGE AND PERCENTAGE CHANGE

From a statistical viewpoint, the primary reason to adjust a response for baseline is to remove concomitant variation in the response and improve precision of treatment comparisons. A response, properly adjusted for baseline, will show no concomitant variation with baseline, that

0277–6715/89/101183–08$05.00

Received April 1987
Revised February 1989

is, it is independent of baseline. In addition to the comparison of treatment groups, one often compares results across other groups of patients for which randomization does not assure baseline balance, for example across centres or across subgroups based on severity of disease. In these cases it is useful to analyse an adjusted response unconfounded by baseline differences between groups. Again, this argues for use of the adjusted response which shows little dependence on baseline. Finally, summary statistics (such as means or medians) of an adjusted response that has little dependence on baseline are more relevant for the prediction of a new patient's response to treatment.

Thus, for a number of reasons, it is advantageous to adjust a response to be nearly independent of baseline. Scatterplots constitute an informative way to study the relationship of two variables. For each treatment group one should plot change and percentage change versus baseline and choose the one, if either, which displays little dependence on baseline.

I claim no originality for this recommendation. It has not, however, received wide recognition. I reviewed fifty non-randomly selected books on statistical methods in biology and medicine and none discussed the problem of how to adjust a response for baseline or the implications of analysis of an improperly adjusted response. Only one[10] discussed the appropriateness of ratios in general.

The above recommendation depends on a subjective interpretation of scatterplots. An objective rule appears in Appendix I based on the ratio of the likelihoods of the data under two normal theory models – one in which change and the other in which percentage change is appropriate. This ratio appears as a supplement to the plots. I anticipate that analysts will find the plots, in general, satisfactory for discriminating between change and percentage change. In those cases where the choice is not clear, it will probably make little difference which one chooses for analysis and one can base the choice on other considerations, such as historical precedent.

3. EXAMPLES ILLUSTRATING THE RECOMMENDATION

The first example consists of two treatments in antiasthma trials. The baseline is the last pre-treatment measurement of percentage of predicted normal forced expiratory volume in one second (FEV_1). The response is the maximum percentage of predicted normal FEV_1 in the first 4 hours post-treatment. Figures 1(a) and (b) plot change versus baseline for the two treatments, and Figures 1(c) and (d) are the corresponding plots for percentage change. For both treatments, change shows little dependence on baseline while percentage change shows dependence on baseline in both the level and spread: patients with low baselines tend to have larger and more variable percentage changes than those with higher baselines. Thus, change is an appropriate adjustment for baseline.

The second example consists of data from another antiasthma trial. The baseline is the percentage of days during a pre-treatment phase of approximately 3 weeks duration in which the patient experienced at least one symptom of asthma. The response is the same variable measured during a 6 week treatment phase. Figures 2(a)–(d) are plots of change and percentage change versus baseline for the two treatments. Change shows dependence on baseline in both level and spread while percentage change displays little dependence on baseline. Thus, percentage change is an appropriate adjustment.

4. IMPACT OF USE OF THE WRONG ADJUSTMENT FOR BASELINE

The problems with use of change when percentage change is appropriate, and vice versa, include:

1. The analysis has less sensitivity in estimating treatment differences.
2. Summary statistics are less useful for comparing groups of patients.
3. Summary statistics are less useful for predicting individual patient responses.

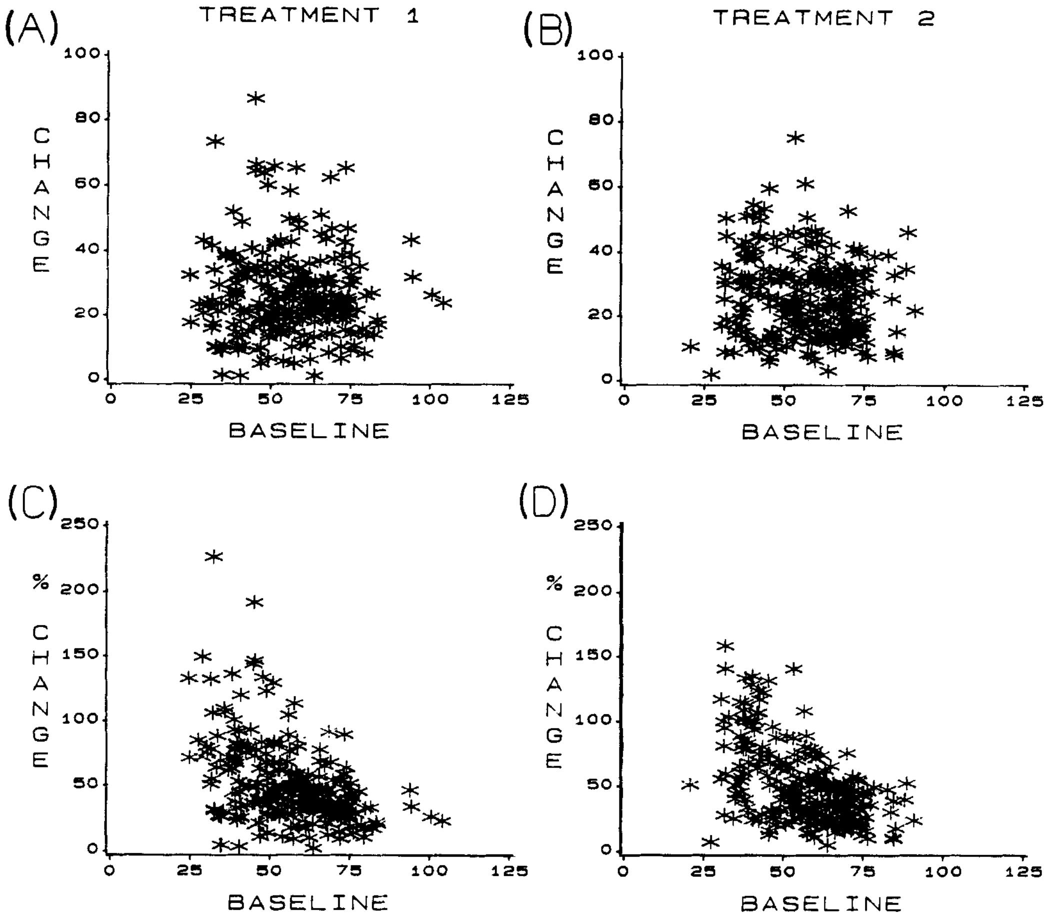

Figure 1. Pulmonary function data: percentage of predicted normal FEV_1. (a), (b) change versus baseline; (c), (d) percentage change versus baseline

Part of the first problem, in a normal theory analysis, is the asymmetry often seen in the distribution of the wrong adjusted response. This one could handle by transformation (for example log or square root) or with a non-parametric analysis. But this is a minor aspect of the problem. More important, extraneous variability is included in the analysis through use of the wrong adjusted response. If, for example, change is independent of baseline, then division of change by baseline makes as much sense as division by a quantity selected from a table of random numbers.

Appendix II shows that if change is independent of baseline, then for any estimator of a treatment group difference based on percentage changes one can construct just as good an estimator which is a function of the changes and the order statistics of the baselines. Thus, without loss of generality when change is independent of baseline, one can ignore the observed relationship between change and baseline. Similarly, for hypothesis testing, when change is independent of baseline, for any test based on the percentage changes, one can construct a (randomized) test with identical size and power which is a function of the changes and the baseline order statistics. Of course, similar statements hold for the analysis of change when percentage change is independent of baseline.

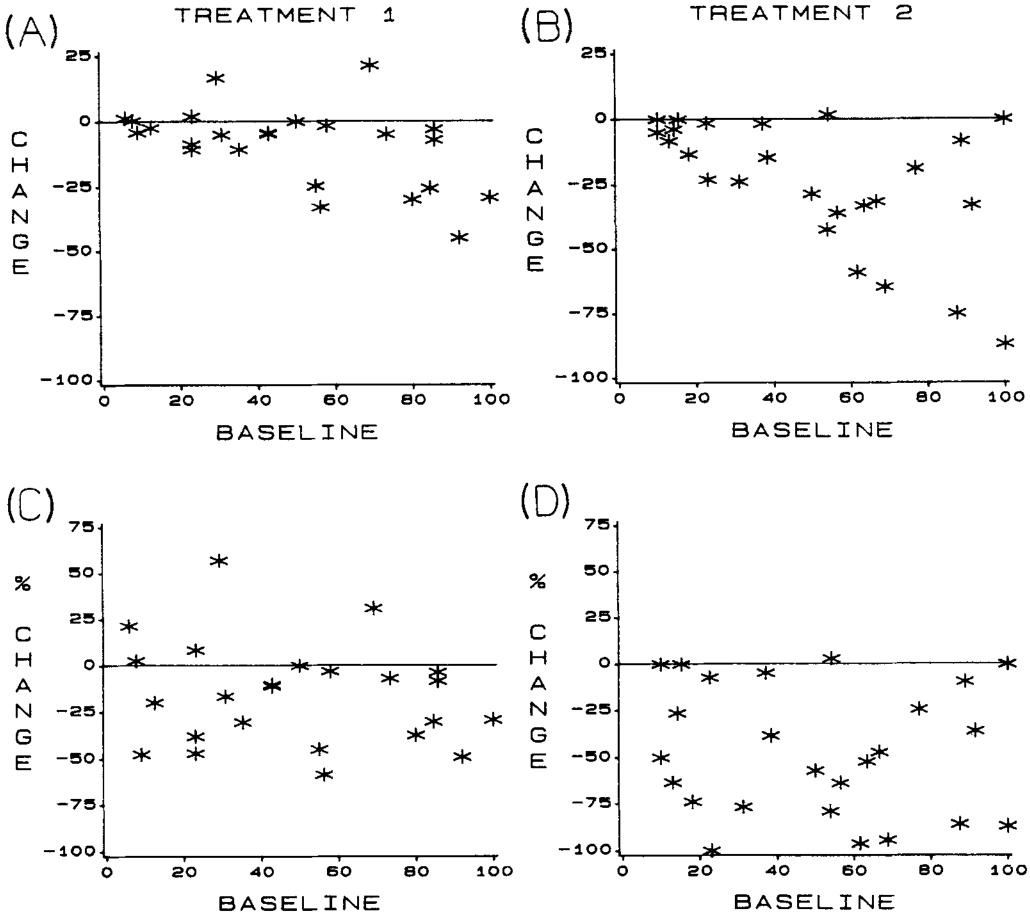

Figure 2. Asthma symptoms data: percentage of days with at least one symptom of asthma. (a), (b) change versus baseline; (c), (d) percentage change versus baseline

To illustrate the loss in power in the analysis of the wrong adjusted response I ran a set of simulations for each example in Section 3. For the pulmonary function data, I selected with replacement two random samples of 25 patients from the treatment 2 group. To one sample I added a constant, delta, to the changes. I calculated the percentage change for each patient and compared the samples with the Wilcoxon rank-sum test and repeated this 500 times. Similarly, for the asthma symptoms data I selected two samples of 25 patients. To one sample I added a constant, delta, to the percentage changes, calculated the change for each patient, and then compared the samples with the Wilcoxon rank-sum test. I repeated this 500 times.

Table I summarizes the results. For the pulmonary function data the power for the analysis of percentage change is substantially below that of change and the opposite holds for the asthma symptoms data. In these examples one can show[11] that the power is approximately $\Phi(\sqrt{(n)}D-1{\cdot}96)$ for some constant D, where Φ is the standard normal distribution function. Use of the simulated power to estimate D for each simulation yields an estimate of the ratio of the sample size needed with the wrong adjusted response relative to that needed with the proper adjusted response for detection of a given difference between treatments with a given

Table I. Simulated power (per cent) of a Wilcoxon rank-sum test at 5 per cent significance level. Each simulated power based on 500 independent repetitions.

Example	Delta	Adjusted response Change	 Percentage change
Pulmonary function data	3	14·4	11·0
	6	39·0	34·2
	9	71·2	55·2
	12	89·6	76·8
Asthma symptoms data	10	16·4	19·4
	20	38·4	48·2
	30	54·4	80·0
	40	79·6	91·8

power. This ratio is approximately 140 per cent for the pulmonary function data and 145 per cent for the asthma symptoms data. Thus, one pays a substantial price for analysis of the inappropriate adjusted response.

We can encounter the second problem at several levels of comparison: treatment groups in the same study, different centres in a multicentre study, patient subgroups, studies in an NDA overview, or results from published studies. In each case the problem is the same: with the wrong adjusted response, differences in baselines between the groups influence the comparison of summary statistics of the adjusted response.

For example, Table II contains summary statistics of the pulmonary function data with patients grouped by steroid dependence status. Steroid dependent patients, not surprisingly, tend to have lower baselines. Comparison of the median percentage changes between the subgroups for each treatment might suggest that steroid dependent patients respond better than those steroid independent. Inspection of the summary statistics for change, however, indicates that this trend in the percentage changes results from baseline imbalance.

One could perform a statistical analysis of percentage change in this example which adjusted for baseline, although this would have to account for the curvilinear regression and inhomogeneity of variance evident in Figures 1(c) and (d). This, however, is a much more difficult means to obtain the answer apparent from the analysis of change.

For the third problem, consider the pulmonary function data and the reporting of treatment 2 results with use of a median and an interquartile range. For percentage change the results are 45 (28, 64) and for change the results are 25 (16, 34). One could apply either of these sets of statistics to predict results with a new patient. Once, however, one knows the baseline for this patient (say it is 30 per cent of predicted normal) the first set of statistics becomes less meaningful. Because percentage change tends to increase with decreasing baseline, there is a greater than 50 per cent chance that this patient will have an increase of greater than 45 per cent of baseline. On the basis of these summary statistics, we cannot really specify how much greater. The summary statistics for change, however, apply to all patients, regardless of baseline. This patient could expect to increase 25 per cent of predicted normal. Further, the statistics for change can be used to predict the percentage change for this patient as 25/30 = 83 per cent of baseline, with a 75 per cent chance of increasing greater than 16/30 = 53 per cent of baseline, and a 25 per cent chance of increasing greater than 34/30 = 113 per cent of baseline.

Table II. Median (interquartile range) for two patient subgroups for the pulmonary function data

			Variable		
Treatment	Steroid dependent	*N*	Baseline	Change	Percentage change
1	No	141	60·1 (49·1–68·9)	24·9 (18·5–32·7)	39·8 (28·6–59·2)
	Yes	107	52·6 (40·8–69·3)	24·1 (18·0–33·2)	48·2 (32·1–67·9)
2	No	143	59·8 (48·7–68·7)	25·2 (16·0–34·0)	43·0 (28·4–58.6)
	Yes	104	55·4 (42·4–67·8)	25·3 (16·9–34·0)	45·7 (28·7–70·3)

5. DISCUSSION

Published reports of clinical trials generally do not discuss the choice for analysis of change or percentage change, nor do they generally present plots such as those in Figures 1 and 2 that permit the reader to determine the variable more appropriate for analysis. It is therefore difficult to criticize authors' choices. It is, however, possible to find both change[12,13] and percentage change[4,12] used in a therapeutic area, namely, pulmonary function in the study of beta agonists in asthmatics.

From my own experience as a statistician in the pharmaceutical industry, the choice between change and percentage change often rests on historical precedent and/or intuitive feelings regarding which adjustment is more 'clinically meaningful'. For example, support for the use of percentage change is a feeling that a given change is 'more important' to a patient with a low than a high baseline. We should abandon such reasoning. Since it does not take account of the relationship between the adjusted response and baseline it can potentially lead to the difficulties illustrated in the previous section.

What if neither change nor percentage change is nearly independent of baseline? There are two possibilities worth mentioning. First, if the response is independent of baseline then no adjustment is necessary and one should analyse the response itself. It is good practice to examine plots of the response versus baseline in addition to those for change and percentage change. Second, when there is a linear regression of response on baseline with variation about the regression independent of baseline, but with the slope of the regression not near zero or one, then one should consider analysis of covariance. For other possibilities, the data analyst should use his experience and common sense.

6. CONCLUSION

The recommendation discussed here is easily applied. In those cases where the choice is clear, the use of the appropriate adjusted response yields a more sensitive and meaningful analysis. I recommend that authors explicitly address their rationale for choice of adjusted response and indicate the degree of dependence of the adjusted response on baseline. If the adjusted response is nearly independent of baseline the reader will know that the author used the most sensitive scale for treatment comparisons and that the results are more broadly applicable when compared with other studies and when used to predict individual patient response.

APPENDIX I

Let (Y_{ij}, X_{ij}) denote a response/baseline pair for patient j in group i, $i=1, 2, j=1, \ldots, n_i$. Consider two models for the data:

$$Y_{ij} - X_{ij} \sim \mathrm{N}(\mu_i, \sigma^2) \tag{1}$$

$$Y_{ij} \sim \mathrm{N}(X_{ij}\beta_i, X_{ij}^2\sigma^2). \tag{2}$$

A least squares analysis of (1) corresponds to a t-test on the changes, $C_{ij} = Y_{ij} - X_{ij}$. A (weighted) least squares analysis of (2) corresponds to a t-test on the percentage change $P_{ij} = 100(Y_{ij} - X_{ij})/X_{ij}$. A reasonable way to choose between these two models is to calculate the ratio of the maximum likelihoods. It is straightforward to show that the ratio of the maximum likelihood under (1) to that under (2) is $R^{n/2}$, where

$$R = (\bar{X}_g)^2 \sum_{ij} [(P_{ij} - \bar{P}_i)/100]^2 \Big/ \sum_{ij} (C_{ij} - \bar{C}_i)^2$$

with $\bar{X}_g$ the geometric mean of the baselines and $\bar{P}_i$ and $\bar{C}_i$ the arithmetic means of the percentage changes and the changes in group i, respectively. Choose the change if R is greater than one; otherwise, use percentage change.

For the pulmonary function data of Section 3, $R = 1{\cdot}72$; for the asthma symptoms data, $R = 0{\cdot}36$. The likelihood ratios lead to the same choices of adjusted response as those in Section 3.

While R was derived from normal theory models, limited simulations suggest that it works well even with substantial positive skewness in the distribution of the proper adjusted response. Also, the proper choice is made, with use of R, more frequently the larger the sample size, the larger the coefficient of variation in the baseline, and the less skewness in the distribution of the proper adjusted response.

Finally, it is important that R not be relied on solely. Plots may indicate that both adjusted responses depend substantially on baseline, while R will always lead to a choice of one of the two adjusted responses.

APPENDIX II

Let C_{ij} and X_{ij} denote the change from baseline and the baseline for patient j in group i, $i=1, 2$, $j=1, \ldots, n$. Assume that the X_{ij}s are identically distributed. Further, assume that the $4n$ random variables are mutually independent so that the change is the appropriate variable for analysis. For any estimator $f(X_{11}, \ldots, X_{2n}, C_{11}, \ldots, C_{2n})$ define

$$g(X_{(1)}, \ldots, X_{(2n)}, C_{11}, \ldots, C_{2n}) = \frac{1}{(2n)!} \sum f(X_{k(1)}, \ldots, X_{i(2n)}, C_{11}, \ldots, C_{2n}), \tag{3}$$

where $k(1), \ldots, k(2n)$ is a permutation of the symbols $11, \ldots, 1n, 21, \ldots, 2n$; the summation is over all $(2n)!$ such permutations; and $X_{(1)}, \ldots, X_{(2n)}$ denotes the order statistics of baseline. Because of the mutual independence and the identical distribution of the X_{ij}s, the joint distribution of $X_{11}, \ldots, X_{2n}, C_{11}, \ldots, C_{2n}$ is the same as that of $X_{k(1)}, \ldots, X_{k(2n)}, C_{11}, \ldots, C_{2n}$ so that $Eg = Ef$. Let E_{p} and $\mathrm{Var}_{\mathrm{p}}$ denote the expectation and variance over the permutations of $X_{11}, \ldots, X_{2n}$ for a given set of baseline order statistics and C_{ij}s. Then $\mathrm{Var} f = \mathrm{Var}(E_{\mathrm{p}} f) + E(\mathrm{Var}_{\mathrm{p}} f)$. Since $g = E_{\mathrm{p}} f$ it follows that $\mathrm{Var} f \geqslant \mathrm{Var}\, g$.

As an example, if $P_{ij} = 100 C_{ij}/X_{ij}$ and $f = \bar{P}_1 - \bar{P}_2$ then

$$g = 100\left(\frac{1}{2n} \sum \frac{1}{X_{ij}}\right)(\bar{C}_1 - \bar{C}_2).$$

Both f and g estimate the difference in mean percentage changes but g has the smaller mean squared error.

A similar argument applies in the hypothesis testing problem (Reference 14, Section 3.4). Let $\phi(X, C)$ be a test which is a function of the baselines and changes and takes values in the interval [0, 1], denoting the probability of rejection of the null hypothesis. As in (3), average ϕ over the permutations of the baseline to obtain the randomized test ψ. As above, $E\phi = E\psi$ and the tests have the same operating characteristics.

ACKNOWLEDGEMENTS

I thank the editor and the reviewers for their helpful comments.

REFERENCES

1. Captopril Multicenter Research Group. 'A placebo-controlled trial of Captopril in refractory chronic congestive heart failure', *Journal of the American College of Cardiology*, **2**, 755–763 (1983).
2. Wahl, J., Singh, B. N. and Thoden, W. R. 'Comparative hypotensive effects of acebutolol and hydrochlorothiazide in patients with mild to moderate essential hypertension: a double-blind multicenter evaluation', *American Heart Journal*, **111**, 353–362 (1986).
3. Hopkins, R., Bird, H. A., Jones, H., Hill, J., Surrall, K. E., Astbury, C., Miller, A. and Wright, V. 'A double-blind controlled trial of etretinate (Tigason) and ibuprofen in psoriatic arthritis', *Annals of the Rheumatic Diseases*, **44**, 189–193 (1985).
4. Kemp, J. P., Chervinsky, P., Orgel, H. A., Meltzer, E. O., Noyes, J. H. and Mingo, T. H. 'Concomitant bitolterol mesylate aerosol and theophylline for asthma therapy, with 24 hour electrocardiographic monitoring', *Journal of Allergy and Clinical Immunology*, **73**, 32–43 (1984).
5. Amsterdam, J. D., Kaplan, M., Potter, L., Bloom, L. and Rickels, K. 'Adinazolam, a new triazolobenzodiazepine, and imipramine in the treatment of major depressive disorder', *Psychopharmacology*, **88**, 484–488 (1986).
6. Lester, R. S., Schachter, G. D. and Light, M. J. 'Isotretinoin and tetracycline in the management of severe nodulocystic acne', *International Journal of Dermatology*, **24**, 252–257 (1985).
7. Egger, M. J., Coleman, M. L., Ward, J. R., Reading, J. C. and Williams, H. J. 'Uses and abuses of analysis of covariance in clinical trials', *Controlled Clinical Trials*, **6**, 12–24 (1985).
8. Laird, N. 'Further comparative analyses of pretest-posttest research designs', *American Statistician*, **37**, 329–330 (1983).
9. Samuels, M. L. 'Use of analysis of covariance in clinical trials: a clarification', *Controlled Clinical Trials*, **7**, 325–329 (1986).
10. Murphy, E. A. *Biostatistics in Medicine*, Johns Hopkins University Press, Baltimore, 1982.
11. Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, Oakland, California, 1975.
12. Busse, W. W., Smith, A. and Bush, R. K. 'The use of a single daily theophylline dose and metered-dose albuterol in asthma treatment', *Journal of Allergy and Clinical Immunology*, **78**, 577–582 (1986).
13. Pedersen, S. 'The importance of a pause between the inhalation of two puffs of terbutaline from a pressurized aerosol with a tube spacer', *Journal of Allergy and Clinical Immunology*, **77**, 505–509 (1986).
14. Lehmann, E. L. *Testing Statistical Hypotheses*, Wiley, New York, 1959.