BASIC STATISTICS FOR CLINICIANS: 1. HYPOTHESIS TESTING

Gordon Guyatt, *† MD; Roman Jaeschke, *† MD; Nancy Heddle, ‡ MSc; Deborah Cook, *† MD; Harry Shannon, * PhD; Stephen Walter, * PhD

Abstract • Résumé

In the first of a series of four articles the authors explain the statistical concepts of hypothesis testing and *p* values. In many clinical trials investigators test a null hypothesis that there is no difference between a new treatment and a placebo or between two treatments. The result of a single experiment will almost always show some difference between the experimental and the control groups. Is the difference due to chance, or is it large enough to reject the null hypothesis and conclude that there is a true difference in treatment effects? Statistical tests yield a *p* value: the probability that the experiment would show a difference as great or greater than that observed if the null hypothesis were true. By convention, *p* values of less than 0.05 are considered statistically significant, and investigators conclude that there is a real difference. However, the smaller the sample size, the greater the chance of erroneously concluding that the experimental treatment does not differ from the control — in statistical terms, the power of the test may be inadequate. Tests of several outcomes from one set of data may lead to an erroneous conclusion that an outcome is significant if the joint probability of the outcomes is not taken into account. Hypothesis testing has limitations, which will be discussed in the next article in the series.

Dans ce premier article d'une série de quatre, les auteurs expliquent les concepts statistiques que sont les vérifications des hypothèses et les valeurs *p*. Au cours de nombreux essais cliniques, les enquêteurs font l'essai d'une hypothèse nulle selon laquelle il n'y a pas de différence entre un nouveau traitement et un placebo, ou entre deux traitements. Le résultat d'une seule expérience indiquera presque toujours un écart entre les sujets de l'expérience et ceux des groupes témoins. L'écart est-il attribuable au hasard ou est-il assez important pour qu'on rejette l'hypothèse nulle et conclue qu'il y a vraiment un écart entre les effets des traitements? Les tests statistiques donnent une valeur *p* : c'est la probabilité selon laquelle l'expérience démontrera un écart aussi important ou plus important que celui qu'on observerait si l'hypothèse nulle s'avérait. Par convention, les valeurs *p* de moins de 0,05 sont considérées comme importantes sur le plan statistique et les enquêteurs concluent qu'il y a un écart réel. Or, plus l'échantillon est petit, plus grande est la chance de conclure à tort que le traitement expérimental ne diffère pas du traitement témoin — sur le plan statistique, la puissance du test peut être insuffisante. Des essais de plusieurs résultats d'une série de données peuvent inciter à conclure à tort qu'un résultat est important si l'on ne tient pas compte de la probabilité conjointe des résultats. Les vérifications des hypothèses ont leurs limites, sur lesquelles portera le prochain article de la série.

Clinicians are often told that they are supposed to not only read journal articles, but also understand them and make a critical assessment of their validity.[1,2] Clinicians may offer better care if they are able to appraise critically the original literature and apply the results to their practice.[3,4] Criteria for assessing the strength of the methods reported in medical articles can provide clinicians with guidance in recognizing the strengths and weaknesses of clinical research.[5,6] However, such guidelines tend to make only passing reference to statistical methods or interpretation of study conclusions based on statistics.

Some authors have attempted to fill this gap.[7–11] This series has modest goals. We do not intend, for instance, to enable readers to identify or understand the statistical tests used to calculate a *p* value, but we are interested in helping them interpret the *p* values generated by such tests. We

*From the departments of *Clinical Epidemiology and Biostatistics, †Medicine and ‡Pathology, McMaster University, Hamilton, Ont.*

Drs. Guyatt and Cook are the recipients of Career Scientist Awards from the Ontario Ministry of Health. Dr. Cook is a scholar of the St. Joseph's Hospital Foundation, Hamilton, Ont. Dr. Walter is the recipient of a National Health Scientist Award from Health Canada.

***Reprint requests to:** Dr. Gordon Guyatt, Rm. 2C12, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton, ON L8N 3Z5*

This is the first article in a series of four, to appear in the January and February 1995 issues of CMAJ.

← For prescribing information see page 102

wish to allow readers to understand the conclusions derived from statistical procedures that they find in clinical articles. This series complements our guides to using the medical literature, which focus on study design and application of study results.[12]

Commonly used statistical techniques

We chose to address only the techniques and approaches that clinicians most commonly face. To identify these, we reviewed recent contributions to three major medical journals: original, special and review articles in the *New England Journal of Medicine* (1991; 324: 1–352); diagnosis and treatment, review, and academia articles in the *Annals of Internal Medicine* (1991; 114: 345–834), and original research, current review, and clinical and community studies articles in the *Canadian Medical Association Journal* (1991; 144: 623–1265). Two of us (N.H. and R.J.) independently reviewed 100 articles and noted the statistical techniques used. Discrepancies between the findings of the two reviewers were resolved by consensus.

The results of this review (Table 1) are consistent with those of a similar review.[13] Although a wide variety of statistical techniques were reported, hypothesis tests, confidence intervals, p values and measures of association occurred most frequently. On the basis of this information our series will deal with hypothesis testing, estimation, measures of association, survival analysis, and regression and correlation. Examples will be drawn from the articles surveyed and others.

Hypothesis testing

When we conduct a trial of a new treatment we can assume that there is a true, underlying effect of the treatment that any single experiment can only estimate. Investigators use statistical methods to help understand the true effect from the results of one experiment. For some time the paradigm for statistical inference has been hypothesis testing.

Table 1: Frequency of statistical concepts and techniques in 100 articles published in three medical journals

Concept or technique	No. of articles
p value	66
Confidence interval	43
Hypothesis testing	
Parametric method	36
Nonparametric method	25
Regression or correlation	22
Measure of association	19
Survival analysis	19
Measure of agreement	8

The investigator starts from what is called a "null hypothesis": the hypothesis that the statistical procedure is designed to test and, possibly, disprove. Typically, the null hypothesis is that there is no difference between outcomes as a result of the treatments being compared. In a randomized controlled trial to compare an experimental treatment with a placebo, the null hypothesis can be stated: "The true difference in the effects of the experimental and control treatments on the outcome of interest is zero."

For instance, in a comparison of two vasodilator treatments for patients with heart failure, the proportion of patients treated with enalapril who survived was compared with the proportion of survivors among patients given a combination of hydralazine and nitrates.[14] We start with the assumption that the treatments are equally effective and stick to this position unless the data make it untenable. The null hypothesis in the vasodilator trial could be stated: "The true difference in the proportion surviving between patients treated with enalapril and those treated with hydralazine and nitrates is zero."

In the hypothesis-testing framework we ask Are the observed data consistent with this null hypothesis? The logic behind this approach is the following. Even if the true difference in effect is zero, the results observed will seldom be exactly the same; that is, there will be some difference between outcomes for the experimental and control groups. As the results diverge farther and farther from the finding of no difference, the null hypothesis that there is no difference between treatments becomes less and less credible. If the difference between results in the treatment and control groups becomes large enough, the investigator must abandon belief in the null hypothesis. An explanation of the role of chance helps demonstrate this underlying logic.

The role of chance

Imagine a fair or "unbiased" coin in which the true probability of obtaining heads in any single coin toss is 0.5. If we tossed such a coin 10 times we would be surprised if we saw exactly five heads and five tails. Occasionally, we would get results very divergent from the five-to-five split, such as eight to two, or even nine to one. Very infrequently 10 coin tosses would result in 10 consecutive heads or tails.

Chance is responsible for this variation in results. Games of chance illustrate the way chance operates. On occasion, the roll of two unbiased dice (with an equal probability of rolling any number between one and six) will yield two ones, or two sixes. The dealer in a poker game will, on occasion (and much to the delight of the recipient), dispense a hand consisting of five cards of a single suit. Even less frequently, the five cards will not only belong to a single suit but will also be consecutive.

Chance is not restricted to the world of coin tosses, dice and card games. If a sample of patients is selected from a community, chance may result in unusual distributions of disease in the sample. Chance may be responsible for a sub-

stantial imbalance in the rates of a particular event in two groups of patients given different treatments that are, in fact, equally effective. Statistical inquiry is geared to determining whether unbalanced distributions can be attributed to chance or whether they should be attributed to another cause (treatment effects, for example). As we will demonstrate, the conclusions that may be drawn from statistical inquiry are largely determined by the sample size of the study.

THE P VALUE

One way that an investigator can go wrong is to conclude that there is a difference in outcomes between a treatment and a control group when, in fact, no such difference exists. In statistical terminology, erroneously concluding that there is a difference is called a Type I error, and the probability of making such an error is designated α. Imagine a situation in which we are uncertain whether a coin is biased. That is, we suspect (but do not know for sure) that a coin toss is more likely to result in heads than tails. We could construct a null hypothesis that the true proportions of heads and tails are equal. That is, the probability of any given toss landing heads is 0.5, and so is the probability of any given toss landing tails. We could test this hypothesis in an experiment in which the coin is tossed a number of times. Statistical analysis of the results would address whether the results observed were consistent with chance.

Let us conduct a thought experiment in which the suspect coin is tossed 10 times, and on all 10 occasions the result is heads. How likely is this result if the coin is unbiased? Most people would conclude that this extreme result is highly unlikely to be explained by chance. They would therefore reject the null hypothesis and conclude that the coin is biased. Statistical methods allow us to be more precise and state just how unlikely it is that the result occurred simply by chance if the null hypothesis is true. The probability of 10 consecutive heads can be found by multiplying the probability of a single head (0.5) by itself 10 times: $0.5 \times 0.5 \times 0.5$ and so on. Therefore, the probability is slightly less than one in 1000. In an article we would likely see this probability expressed as a p value: $p < 0.001$. What is the precise meaning of this p value? If the null hypothesis were true (that is, the coin was unbiased) and we were to repeat the experiment of the 10 coin tosses many times, 10 consecutive heads would be expected to occur by chance less than once in 1000 times. The probability of obtaining either 10 heads or 10 tails is approximately 0.002, or two in 1000.

In the framework of hypothesis testing the experiment would not be over, for we have yet to make a decision. Are we willing to reject the null hypothesis and conclude that the coin is biased? How unlikely would an outcome have to be before we were willing to dismiss the possibility that the coin was unbiased? In other words, what chance of making a Type I error are we willing to accept? This reasoning implies that there is a threshold probability that marks a boundary; on one side of the boundary we are unwilling to reject the null hypothesis, but on the other we conclude that chance is no longer a plausible explanation for the result. To return to the example of 10 consecutive heads, most people would be ready to reject the null hypothesis when the observed results would be expected to occur by chance less than once in 1000 times.

Let us repeat the thought experiment with a new coin. This time we obtain nine tails and one head. Once again, it is unlikely that the result is due to chance alone. This time the p value is 0.02. That is, if the null hypothesis were true and the coin were unbiased, the results observed, or more extreme than those observed, (10 heads or 10 tails, 9 heads and 1 tail or 9 tails and 1 head) would be expected to occur by chance twice in 100 repetitions of the experiment.

Given this result, are we willing to reject the null hypothesis? The decision is arbitrary and a matter of judgement. However, by statistical convention, the boundary or threshold that separates the plausible and the implausible is five times in 100 ($p = 0.05$). This boundary is dignified by long tradition, although other choices of a boundary value could be equally reasonable. The results that fall beyond this boundary (i.e., $p < 0.05$) are considered "statistically significant." Statistical significance, therefore, means that a result is "sufficiently unlikely to be due to chance that we are ready to reject the null hypothesis."

Let us repeat our experiment twice more with a new coin. On the first repetition eight heads and two tails are obtained. The p value associated with such a split tells us that, if the coin were unbiased, a result as extreme as eight to two (or two to eight), or more extreme, would occur by chance 11 times in 100 ($p = 0.11$). This result has crossed the conventional boundary between the plausible and implausible. If we accept the convention, the results are not statistically significant, and the null hypothesis is not rejected.

On our final repetition of the experiment seven tails and three heads are obtained. Experience tells us that such a result, although it is not the most common, would not be unusual even if the coin were unbiased. The p value confirms our intuition: results as extreme as this split would occur under the null hypothesis 34 times in 100 ($p = 0.34$). Again, the null hypothesis is not rejected.

Although medical research is not concerned with determining whether coins are unbiased, the reasoning behind the p values reported in articles is identical. When two treatments are being compared, how likely is it that the observed difference is due to chance alone? If we accept the conventional boundary or threshold ($p < 0.05$), we will reject the null hypothesis and conclude that the treatment has some effect when the answer to this question is that repetitions of the experiment would yield differences as extreme as those we have observed less than 5% of the time.

In the randomized trial mentioned earlier, treatment with enalapril was compared with treatment by a combination of hydralazine and nitrates in 804 male patients with heart failure. This trial illustrates hypothesis testing when

there is a dichotomous (Yes–No) outcome, in this case, life or death.[14] During the follow-up period, which ranged from 6 months to 5.7 years, 132 (33%) of the 403 patients assigned to the enalapril group died, as did 153 (38%) of the 401 assigned to the hydralazine and nitrates group. Application of a statistical test that compares proportions (the χ^2 test) shows that if there were actually no difference in mortality between the two groups, differences as large as or larger than those actually observed would be expected 11 times in 100 ($\chi^2 = 0.11$). We use the hypothesis-testing framework and the conventional cut-off point of 0.05, and we conclude that we cannot reject the null hypothesis — the difference observed is compatible with chance.

Risk of a false-negative result

A clinician might comment on the results of the comparison of enalapril with hydralazine and nitrates as follows: "Although I accept the 0.05 threshold and therefore agree that we cannot reject the null hypothesis, I still suspect that treatment with enalapril results in a lower mortality rate than treatment with the combination of hydralazine and nitrates. The experiment leaves me in a state of uncertainty." This clinician recognizes a second type of error that an investigator can make: falsely concluding that an effective treatment is useless. A Type II error occurs when we erroneously fail to reject the null hypothesis (and, therefore, we dismiss a useful treatment).

In the comparison of treatment with enalapril and with hydralazine and nitrates, the possibility of erroneously concluding that there is no difference between the treatments looms large. The investigators found that 5% fewer patients receiving enalapril died than those receiving the alternative vasodilator regimen. If the true difference in mortality really were 5%, we would readily conclude that patients benefit from enalapril. Despite this result, however, we were unable to reject the null hypothesis.

Why were the investigators unable to conclude that enalapril is superior to hydralazine and nitrates despite having observed an important difference between the mortality rates? The study did not enrol enough patients for the investigators to be confident that the difference they observed was real. The likelihood of missing an important difference (and making a Type II error) decreases as the sample gets larger. When there is a high risk of making a Type II error, we say the study has inadequate power. The larger the sample, the lower the risk of Type II error and the greater the power. Although 804 patients were recruited by the investigators conducting the vasodilator trial, for dichotomous outcomes such as life or death very large samples are often required to detect small differences in the effects of treatment. For example, the trials that established the optimal treatment of acute myocardial infarction with acetylsalicylic acid and thrombolytic agents recruited thousands of patients to ensure adequate power.

When a trial fails to reject the null hypothesis ($p > 0.05$) the investigators may have missed a true treatment effect, and we should consider whether the power of the trial was adequate. In such "negative" studies, the stronger the trend in favour of the experimental treatment, the more likely the trial missed a true treatment effect.[15] We will explain more about deciding whether a trial had adequate power in the next article in this series.

Some studies are designed to determine not whether a new treatment is better than the current one but whether a treatment that is less expensive, easier to administer or less toxic yields the same treatment effect as standard therapy. In such studies (often called "equivalence studies"[16]) recruitment of an adequate sample to ensure that small but important treatment effects will not be missed is even more important. If the sample size in an equivalence study is inadequate, the investigator risks concluding that the treatments are equivalent when, in fact, patients given standard therapy derive important benefits in comparison with those given the easier, cheaper or less toxic alternative.

Continuous measures of outcome

All of our examples so far have used outcomes such as Yes or No, heads or tails, or dying or not dying, that can be expressed as proportions. Often, investigators compare the effects of two or more treatments using numeric or ordinal variables such as spirometric measurement, cardiac output, creatinine clearance or score on a quality-of-life questionnaire. These outcomes are continuous: a large number of values are possible.

For example, in the study of enalapril versus hydralazine and nitrates in the treatment of heart failure the investigators compared the effect of the two regimens on exercise capacity (a continuous variable). In contrast with the effect on mortality, which showed better results with enalapril treatment, exercise capacity improved with hydralazine and nitrates but not with enalapril. The investigators compared the change in exercise capacity from baseline to 6 months in the two treatment groups with the use of a statistical test for continuous variables (Student's *t*-test). Exercise capacity in the group receiving hydralazine and nitrates improved more than it did in the other group, and the difference between the two groups was unlikely to have occurred by chance ($p = 0.02$). *P* values for Students' *t*-test and others like it are obtained from standard tables.

Baseline differences

Authors of articles often state that hypothesis tests have been "adjusted" for baseline differences in the groups studied. Random assignment, in which chance alone dictates to which group a patient is allocated, generally produces comparable groups. However, if the investigator is unlucky, factors that determine outcome might be unequally distributed between the two groups. For example, in a trial to compare two treatments, let us say that it is known that older pa-

tients have a poorer outcome. After random assignment, the investigator discovers that a larger proportion of the older patients are assigned to one of the two treatments. This age imbalance could threaten the validity of an analysis that does not take age into account. So the investigator performs an adjustment in the statistical test to yield a p value corrected for differences in the age distribution of the two groups. In this example, readers are presented with the probability that would have been generated if the age distribution in the two groups had been the same. In general, adjustments can be made for several variables at once, and the p value can be interpreted in the regular way.

Multiple tests

University students have long been popular subjects for experiments. In keeping with this tradition, we have chosen medical students as the subjects for our next thought experiment.

Picture a medical school in which an introductory course on medical statistics is taught by two instructors, one of whom is more popular than the other. The dean of the medical school has no substitute for the less popular faculty member. She has a particular passion for fairness and decides that she will deal with the situation by assigning the 200 first-year medical students to one instructor or the other by random assignment, in which each student has an equal chance (0.5) of being allocated to one of the two instructors.

The instructors decide to use this decision to illustrate some important principles of medical statistics. They therefore ask Do any characteristics of the two groups of students differ beyond a level that could be explained by chance? The characteristics they choose are sex, eye colour, height, grade-point average in the previous year of university, socioeconomic status and favourite type of music. The instructors formulate null hypotheses for each of their tests. For instance, the null hypothesis associated with sex distribution is as follows: the students are drawn from the same group of people; therefore, the true proportion of women in the two groups is identical. Since the investigators know in advance that the null hypothesis in each case is true, any time the hypothesis is rejected represents a false-positive result.

The instructors survey their students to determine their status on each of the six variables of interest. For five of these variables they find that the distributions are similar in the two groups, and the p values associated with statistical tests of the differences between groups are all greater than 0.10. They find that for eye colour, however, 25 of 100 students in one group have blue eyes and 38 of 100 in the other group have blue eyes. A statistical analysis reveals that if the null hypothesis were true (which it is) then such a difference in the proportion of people with blue eyes in the two groups would occur slightly less than five times in 100 repetitions of the experiment. If the investigators used the conventional boundary the null hypothesis would be rejected.

How likely is it that, in six independent hypothesis tests on two similar groups of students, at least one test would have crossed the threshold of 0.05 by chance alone? ("Independent" means that the result of a test of one hypothesis does not, in any way, depend on the results of tests of any of the other hypotheses.) This probability is calculated as follows: the probability that we would not cross the 0.5 threshold in testing a single hypothesis is 0.95; in testing two hypotheses the probability that neither one would cross the threshold is 0.95 multiplied by 0.95 (the square of 0.95); in testing six hypotheses, the probability that not a single one would cross the 0.5 threshold is 0.95 to the sixth power, or 0.74. Therefore, when six independent hypotheses are tested the probability that at least one result is statistically significant is 0.265 or approximately 1 in 4, not 1 in 20. If we wish to maintain our overall boundary for statistical significance at 0.05, we have to divide the threshold p value by six, so that each of the six tests uses a boundary value of $p = 0.008$. That is, you would reject the null hypothesis that none of the characteristics differed significantly only if any one of the differences was significant at $p < 0.008$.

There are two messages here. First, rare findings happen on occasion by chance. Even with a single test, a finding with a p value of 0.01 will happen 1% of the time. Second, we should beware of multiple hypothesis testing, because it may yield misleading results. Examples of this phenomenon abound in the clinical literature. Pocock, Hughes and Lee,[2] in a survey of 45 trials from three leading medical journals, found that the median number of endpoints was 6, and most results were tested for statistical significance. A specific example of the dangers of using multiple endpoints is found in a randomized trial of the effect of rehabilitation after myocardial infarction on quality of life.[17] The investigators randomly assigned patients to standard care, an exercise program or a counselling program and obtained patient reports on work, leisure, sexual activity, satisfaction with outcome, compliance with advice, quality of leisure and work, psychiatric symptoms, cardiac symptoms and general health. For almost all of these variables, there was no difference between the three groups. However, the patients were more satisfied with exercise than with the other two regimens, the families in the counselling group tried to protect the patients less than those in the other groups, and work hours and frequency of sexual activity were greater at 18 months' follow-up in the counselling group than in the other groups. Does this mean that the exercise and counselling programs should be implemented because of the small number of outcomes in their favour, or that they should be rejected because most of the outcomes showed no difference? The authors concluded that their results did not support the effectiveness of either exercise or counselling programs in improving quality of life. However, a program advocate might argue that, even if only a few of

the results favoured such programs, they are worth while. Hence, the use of multiple variables opens the door to controversy.

There are several statistical strategies for dealing with multiple hypothesis testing of the same data. We have illustrated one of these in a previous example: dividing the p value by the number of tests. We can also specify, before the study is undertaken, a single primary outcome on which the main conclusions will hinge. A third approach is to derive a global test statistic that combines the multiple outcomes in a single measure. Full discussion of these strategies for dealing with multiple outcomes is beyond the scope of this article but is available elsewhere.[18]

Limitations of hypothesis testing

Some readers may, at this point, have questions that leave them uneasy. Why use a single cut-off point when the choice of such a point is arbitrary? Why make the question of whether a treatment is effective a dichotomy (a Yes–No decision) when it may be more appropriate to view it as a continuum (from Very unlikely to be effective to Almost certain to be effective)?

We are extremely sympathetic to such readers; they are on the right track. We will deal further with the limitations of hypothesis testing in the next article, which will present an alternative approach to testing for the presence of a treatment effect and to estimating a range of plausible values of such an effect.

Conclusion

We avoided listing the statistical procedures used to test the null hypotheses in the studies we have cited; we do not expect readers to recognize the many methods available or to question whether the appropriate test has been chosen. Rather, we have provided a guide to interpreting p values and a warning about their interpretation when multiple outcome measures are examined. We have alluded to the limitations of hypothesis testing and the resulting p values. In the next article, which will deal with confidence intervals, we will describe complementary techniques to address some of these deficiencies.

References

1. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre: How to read clinical journals: I. Why to read them and how to start reading them critically. *Can Med Assoc J* 1981; 124: 555–558

2. Pocock SJ, Hughes MD, Lee RJ: Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med* 1987; 317: 426–432

3. Evidence-Based Medicine Working Group: Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* 1992; 268: 2420–2425

4. Guyatt GH, Rennie D: Users' guides to reading the medical literature. [editorial] *JAMA* 1993; 270: 2096–2097

5. Sackett DL, Haynes RB, Guyatt GH et al: *Clinical Epidemiology, a Basic Science for Clinical Medicine*, Little, Brown and Company, Boston, 1991

6. Wasson JH, Sox HC, Neff RK et al: Clinical prediction rules. Applications and methodological standards. *N Engl J Med* 1985; 313: 793–799

7. Clegg F: Introduction to statistics. I: Descriptive statistics. *Br J Hosp Med* 1987; 37: 356–357

8. O'Brien PC, Shampo MA: Statistics series. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. *Mayo Clin Proc* 1988; 63: 813–815

9. Altman DG, Gore SM, Gardner MJ et al: Statistical guidelines for contributors to medical journals. *BMJ* 1983; 286: 1489–1493

10. Gardner MJ, Altman DG: Estimating with confidence. *BMJ* 1988; 296: 1210–1211

11. Gardner MJ, Altman DG: *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*, British Medical Journal, London, England, 1989

12. Oxman AD, Sackett DL, Guyatt GH for the Evidence-Based Medicine Working Group: A users' guide to the medical literature. Why and how to get started. *JAMA* 1993; 270: 2093–2095

13. Emerson JD, Colditz GA: Use of statistical analysis in the *New England Journal of Medicine*. *N Engl J Med* 1983; 309: 709–713

14. Cohn JN, Johnson G, Ziesche S et al: A comparison of enalapril with hydralazine–isosorbide dinitrate in the treatment of chronic congestive heart failure. *N Engl J Med* 1991; 325: 303–310

15. Detsky AS, Sackett DL: When was a "negative" trial big enough? How many patients you needed depends on what you found. *Arch Intern Med* 1985; 145: 709–715

16. Kirshner B: Methodological standards for assessing therapeutic equivalence. *J Clin Epidemiol* 1991; 44: 839–849

17. Mayou R, MacMahon D, Sleight P et al: Early rehabilitation after myocardial infarction. *Lancet* 1981; 2: 1399–1401

18. Pocock SJ, Geller NL, Tsiatis AA: The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; 43: 487–498

For prescribing information see page 115 →