Alpha errors, beta errors and negative trials

DESMOND LEDDIN, MB, MRCPI, FRCPC

ABSTRACT: Reports of negative trials are increasing in number as standard therapy for many gastrointestinal diseases is refined. The validity of a negative report depends on the number of patients in the trial, the alpha and beta error and the difference in efficacy which the trial is able to detect. The relationship between these parameters is discussed and a formula given for the calculation of trial size. All reports of negative trials should include not only the number of patients involved and the level of significance of the results but also the beta error and the detectable difference in efficacy of the treatments. **Can J Gastroenterol 1988;2(4):147-50**

Key Words: *Alpha error, Beta error, Negative trials, Trial size*

THE EMPHASIS IN THE INTRODUCTION of many new medications for the treatment of gastrointestinal disorders is not that the new drug is therapeutically superior to the old but that it is as effective as the old. If the newer medication has an advantage over standard treatment, such as a more acceptable dosage regimen or fewer side effects, then presumably it is to be preferred.

As a consequence of this we are presented with trials that appear to show no difference between treatments. So called 'negative' trials have peculiar methodological problems which it is important to appreciate. The most important limitation of these trials is that it is virtually impossible to ensure comparable efficacy.

Negative trials may, of course, be very useful but they may also be detrimental if they result in the acceptance of newer medications which are not as effective as standard treatment or if they delay the introduction of useful therapy.

The purpose of this article is to explore the statistical background to 'negative' trials in its simplest form and to present the methodology for calculation of trial size.

Division of Gastroenterology, Faculty of Medicine, Memorial University of Newfoundland
Correspondence and reprints: Dr Desmond Leddin, Division of Gastroenterology, Faculty of Medicine, Memorial University of Newfoundland, St John's, Newfoundland A1B 3V6
Received for publication May 1, 1988. Accepted June 10, 1988

NORMAL DISTRIBUTION

The normal distribution is really a probability distribution. It is symmetric and bell-shaped (Figure 1). This type of distribution is common for factors which show variability and are continuous. The distribution is described by its mean and the deviation of values from the mean, ie, the standard deviation. The area under the curve is 100% and corresponds to a probability of 1. Roughly 95% of values lie within the area defined by the mean ± 2 standard deviations from the mean. The area α in each tail of Figure 1 is 2.5% of the total area and corresponds, therefore, to a probability of 0.025. The probability of the next measured value falling into either of the shaded areas is then 0.05.

Populations and samples: It is necessary to grasp the difference between populations and samples before understanding the concept of error since error is a

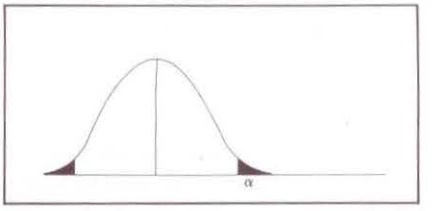

Figure 1) *The normal distribution. The shaded area is alpha*

theoretical concept based on population not on sample.

The aspartate aminotransferase (AST) of healthy Canadian males is a population of values. We will never, for economic and practical reasons, measure the AST of every Canadian male. What we can do is measure the AST of a sample of Canadian males and use this to make statements about the AST of the population of all Canadian males. This sample may or may not reflect the population accurately as sampling is open to various forms of systematic and random bias. It is this extrapolation backwards from sample to population that introduces the problem of error.

Alpha error: Assume that the population distribution of AST values is known for every healthy Canadian male and every healthy British male. In reality they are identical as illustrated in Figure 2.

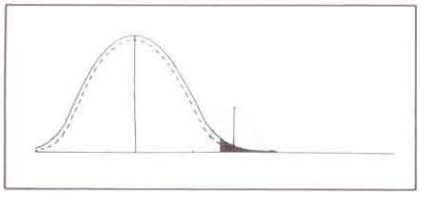

Figure 2) *Alpha error. The distribution of AST values for the population of Canadian (—) and British (---) males and a sample of British males (····)*

For practical reasons the population values will never be known and we are forced to resort to sampling if we wish to compare the two populations. It is quite possible that the mean of a sample (which should accurately reflect the mean of the population) of British males will fall sufficiently far from the mean of the population of Canadian males that we would conclude on the basis of the sample that it is likely that the AST values of Canadian population and British population are different. In fact they are not and this is an alpha error. There is no difference between the populations but our sample has mislead us into believing that there is. Alpha error occurs when a difference between the populations studied is claimed but no actual difference exists.

Beta error: Again, assume that the populations are known. This time in reality there is a difference between the populations. The distribution of values of the two populations overlap somewhat as in Figure 3. We obtain a sample of British males as before. By chance, the mean of the sample now falls sufficently close to the mean of the Canadian males for us to say that there is no difference between the AST values of Canadians and those of the British. The populations, however, really are different. We have extrapolated back from our sample of British males to conclude that there is no difference between the populations when there actually is a difference. This is a beta error. Beta error occurs when no difference between the populations being studied is claimed but a difference actually exists. This type of error is especially important in 'negative' trials, ie, trials in which no difference is claimed.

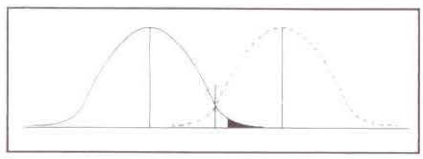

Figure 3) *Beta error. — Population Canadian males; --- Population British males; ···· sample British males*

Relationship of alpha and beta: In Figure 4, the two populations are different, but overlap. The alpha level of population A (the solid shading) defines an area of population B (the hatched shading). The hatched area corresponds to the risk of a beta error.

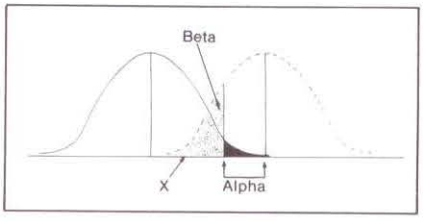

Figure 4) *The relationship of alpha and beta error. — Population A; --- Population B*

This is more easily seen by way of an example. If we obtained a sample of population B whose mean fell at point X we would conclude that population A and population B were identical. We would conclude this because the mean of the sample (which is being used to estimate the mean of population B) falls sufficiently close to the mean of population A for us to say that there is no statistical difference between populations A and B. Since these populations are actually different this is a beta error.

As might be expected from Figure 4 there is a mathematical relationship between alpha and beta. This has been calculated for various levels of alpha and beta and tables are available (Geigy). The function, f(alpha,beta), is shown in Table 1. If, for example, one wishes to set the alpha level at 0.05 and beta at 0.1 then the value of f(alpha,beta) is 10.5. This value is used in the calculation of trial size as will be seen.

Choosing alpha and beta levels: In practice, the risk of alpha error is usually taken as 5%. In other words, there is a probability of 0.05 that an alpha error may occur. There will be a 5% chance that we will detect a difference when no difference actually exists.

Beta error is usually set at 0.1 or 0.2. There are good theoretical grounds for choosing this level. These relate to the excessive size of samples required at levels more stringent than 0.1. At beta levels greater than 0.2 the risk of a false negative result is generally considered to be unacceptable. A beta level of 0.2 means that there is a 20% chance that we will miss a difference even if a difference actually exists.

TABLE 1
Relationship of alpha and beta error

Alpha	Beta	f(alpha,beta)
0.05	0.05	13.0
0.05	0.1	10.5
0.05	0.2	7.9
0.05	0.5	3.8
0.01	0.05	17.8
0.01	0.1	14.9
0.01	0.2	11.7
0.01	0.5	6.6

DETERMINANTS OF TRIAL SIZE

Effect of increasing the number of subjects (n): Intuitively we can recognize that increasing the sample size will increase the likelihood of the sample accurately reflecting the population which we wish to study. Increasing the number of patients in a trial also tightens the standard error or spread of the sample and hence will decrease the risk of error. In addition, increasing trial size increases the power of many statistical tests. Small trials are particularly prone to beta error since the spread of the sam-

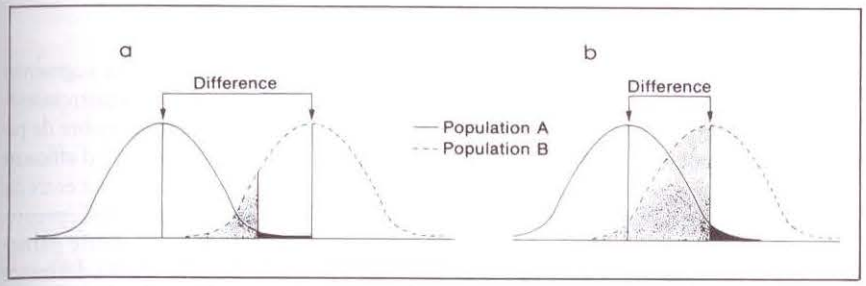

Figure 5) *The effect of decreasing the difference between studied groups on the magnitude of beta error*

ple allows greater possibility of overlap.

However, there are constraints on the number of patients studied such as cost, availability of patients, accrual rate and physician time. It is necessary to calculate a practical trial size which will fulfil the theoretical requirements but still provide the information wanted.

Effect of alpha and beta error: Clearly, the number of patients required will have to take alpha and beta levels into account. This is not difficult since there is a defined relationship between these variables which can be expressed mathematically as f(alpha,beta).

Effect of treatment difference: In Figure 5a the difference between population A and B is huge. The means are widely separated and the overlap is small. The risk of beta error is correspondingly small. If we push the two populations closer, as in Figure 5b, it can be seen that the risk of beta error increases dramatically. Pushing the samples closer is, in effect, decreasing the difference between A and B and corresponds to the trial situation of trying to detect small differences between treatments. The smaller the difference between treatments that we wish to detect, the greater the risk of overlap and hence, of beta error. This is a major problem in trials designed to show no difference and a frequent flaw in apparently negative trials.

To show no difference whatsoever means that populations A and B are superimposed. In this situation, infinitely large numbers of patients would be required and in practice this is never achieved. Some limitation is imposed. One never shows that two groups are identical but simply that it is unlikely that they are different.

Effect of outcome: Finally, one could reason that the endpoint under study will have a place in the estimation of trial size. If, for example, relapses in ulcerative colitis were exceedingly rare then clearly the number of patients required in a trial would be affected.

CALCULATION OF SIZE FOR NEGATIVE TRIALS

Any formula for calculation of trial size must incorporate alpha error, beta error the difference between treatments that is to be detected and the frequency of the studied event.

There are many different mathematical methods for calculating trial size. The formula to be used depends on whether the outcome is qualitative or quantitative and on the design of the trial. The following formula was designed for calculation of a qualitative outcome and specifically for negative trials (Makuch and Simon).

$$n=\frac{2p\times(100-p)\times f(a,b)}{d^2}$$

Where n = number of patients required on each treatment; p = percentage of successes that will occur on standard treatment; d = acceptable difference in efficacy between the old and new treatments; f(a,b) = function of alpha and beta.

For example, consider comparing the efficacy of sulfasalazine and 5-ASA in maintaining remission in ulcerative colitis. Ninety percent of patients with ulcerative colitis treated with 4 g of sulfasalazine daily remain in remission for six months. This is the percentage success rate, p. We wish to test whether 5-ASA is as effective as sulfasalazine in maintaining remission over a period of six months. Let us set alpha at 0.05 and beta at 0.1, ie, we will accept a 5% chance of a false positive or alpha error and a 10% chance of a false negative or beta error. From Table 1 the function f(alpha,beta) equals 10.5. Let us also accept that 5-ASA will still be a useful treatment if it is 10% less effective than sulfasalazine then d = 10. Putting these values into the equation:

$$n=\frac{2\times90\times(10)\times10.5}{10^2}=\frac{18900}{100}$$

giving 189 patients on each treatment or 380 in the study.

These figures show that a very large number of patients are required to show comparable efficacy. In fact, if we would accept only a 5% difference in efficacy then the numbers are 756 in each group or 1512 in all. Clearly, to show absolutely no difference in efficacy is impossible. On the other hand, if a 30% difference in efficacy was acceptable then the number required is 21 on each treatment or 42 in all.

Similarly if we are prepared to accept a 20% chance of a false negative result, as opposed to a 10% risk, then the numbers become 142 in each treatment. It is still a trial of imposing size.

What does not significant mean? From this discussion it should be apparent that the frequently made statement 'there is no difference between the groups as $P > 0.05$' does not have much meaning without reference to the beta error and to the size of difference that was being sought. The P value refers to the risk of a false positive result. Since 'negative' trials do not have a positive result the P value per se is not helpful.

It may indeed be the case that there is no difference between the groups but the failure to detect a difference may also be because of a large beta error and a small (but clinically important) difference between groups. We would not accept a report of a difference between treatments if information was not given on the results of significance testing. Why should we accept reports of no difference between treatments unless the corresponding beta error is given?

TRIAL DESIGN

Trial design represents a balance between the statistical requirements (which tend to increase patient numbers) and clinical practicality (which tends to minimize patient numbers). When designing

trials we cannot change the level of alpha error. The beta error may be varied somewhat but is not a critical determinant of size. The response to standard treatment is a biological fact. The only parameter which is variable and a major determinate of size is the difference between treatments. This difference has a major effect on trial size since the number of patients required is approximately inversely proportional to the square of the difference (the smaller the difference the greater the number of patients required). Consciously in designed trials, or unconsciously in poorly designed trials, this is the parameter that is altered.

There is nothing wrong with a small trial designed to detect a difference of 30% in efficacy as long as this is recognized by the author and reported to the reader. A statement that there is no difference between treatments without reference to the difference which it would have been possible to detect may be misleading.

In the example of sulfasalazine and 5-ASA, the question could be asked as to whether 5-ASA would be an acceptable treatment if it was 10% less effective than sulfasalazine. That is a matter of judgement. A decrease in efficacy might well be acceptable for those patients intolerant of sulfasalazine but not acceptable for patients who tolerate the drug well. The wisest course is to answer these questions in conjunction with a statistician before starting the trial.

Small trials which claim to show no difference between treatments should be viewed with caution if information is not given on the size of the beta error and the detectable difference between treatments. All trials should report both the alpha and beta level incorporated into the trial. This is especially important for trials for which no difference is being claimed. It will then be for the reader to decide if the trial is truly 'negative'.

Les erreurs alpha et beta et les épreuves négatives

Le nombre de mises au banc d'essai et les rapports négatifs continuent à augmenter au fur et à mesure que les thérapies standard des nombreuses maladies gastrointestinales se perfectionnent. La validité d'un rapport négatif dépend du nombre de patients sur qui porte l'analyse, des erreurs alpha et beta et de la différence d'efficacité que l'étude parvient à détecter. La discussion porte sur la relation qui existe entre ces paramètres et l'on propose une formule destinée à calculer l'ampleur de l'épreuve. Tous les rapports négatifs devraient inclure non seulement le nombre de participants et la portée significative des résultats mais encore l'erreur beta et la différence décelable dans l'efficacité des traitements.

ACKNOWLEDGEMENTS: My thanks to Sylvia Ficken of Medical Audio-Visual Services, Health Sciences Center for the illustrations.

RECOMMENDED READING

Geigy. Scientific Tables, 7th edn. Basel: Geigy, 1970.

Pocock SJ. Clinical Trials – A Practical Approach. John Wiley and Sons Ltd, 1983.

Altman DG. Statistics and ethics in medical research: Misuse of statistics is unethical. Br Med J 1980;281:1182-4.

Altman DG. Statistics and ethics in medical research: How large a sample? Br Med J 1980;281:1336-8.

Freiman JA, Chalmers TC, Smith H, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomised control trials. N Engl J Med 1978;290:690-4.

Gore SM. Assessing clinical trials – trial size. Br Med J 1981;282:1687-9.

Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep 1978;62:1037-40.