Evid Based Integrative Med 2004; 1 (4): 225-231
1176-2330/04/0004-0225/$31.00/0

Separation Tests for Early-Phase Complementary and Alternative Medicine Comparative Trials

Mikel Aickin[1,2]

1 Program in Integrative Medicine, University of Arizona, Tucson, Arizona, USA
2 Helfgott Research Institute, National College of Naturopathic Medicine, Portland, Oregon, USA

Abstract

There are some substantial ways in which the current approach to establishing an evidence base for a treatment effect works against complementary and alternative medicine (CAM) research. The standard statistical method in comparative trials is the null hypothesis test, which requires a decision that the 'no treatment effect' hypothesis is confirmed, or it is rejected. While this approach is appropriate for evidence-based medicine, the requirement for a decisive result drives sample sizes upward, or leaves small trials subject to being criticised for being underpowered. Employing the fundamental theory of hypothesis tests, as developed by Neyman and Pearson, it is possible to define a 'separation test' that avoids this problem, by invoking a different inferential rationale. The purpose of a separation test is to find an indication that further research is justified, or that it is not. This change in strategy can considerably lower the required sample size. Since many CAM comparative trials are in early phases of research, where both budgetary and safety considerations argue for a small sample size, separation tests may be especially of interest to CAM researchers. There is, therefore, an opportunity for evidence-based medicine to generate a new kind of study, in which the purpose is to assess whether it is worthwhile to pursue research on an alternative treatment, rather than to determine whether it has been proved effective.

Many areas of biomedicine employ a phased approach to a particular research question. In complementary and alternative medicine (CAM), the issues addressed by a phase I study are the feasibility of both recruitment and the delivery of the intervention and control conditions, and perhaps also some preliminary indication that the intervention might ultimately prove to be successful. Phase II builds on phase I, by attempting solutions to the difficulties discovered in phase I, extending the experience with problems such as adverse effects or toxicities, and again providing some more indication that the intervention might be beneficial. A phase III study follows a successful phase II study and focuses on carrying out a formal trial, with the intent of demonstrating that the intervention provides a definite benefit (or, alternatively, showing that the currently best trial cannot demonstrate a benefit). These phase III studies are of the greatest interest for the purposes of evidence-based medicine. Other areas of biomedicine have various definitions of the phases of research, but all of them benefit from the fact that the appropriate level of analysis and interpretation changes as one moves across the phases.

One of the sharp distinctions between early-phase (phases I and II) trials and later-phase trials lies in the type of evidence that they provide about the potential effectiveness of the intervention. Phase III trials are designed to make a decision, that the intervention has either proved itself worthy of adoption, or not. As a consequence, with these later-phase trials there must be consideration of the statistical power to detect a meaningful clinical effect. This generally means that fairly large sample sizes (often at least 100 per group) are required. The inferential procedure that is universally used in these trials is the null hypothesis test. This approach was advocated by R.A. Fisher in a famous, continuing debate with Jerzy Neyman, which Fisher essentially won.

Null hypothesis testing is also commonly applied in early-phase research. It is not completely clear, however, that this approach is appropriate for phase I and phase II trials. One of the consequences of applying it is that early-phase trials suffer from being regarded as chronically underpowered (in the statistical sense), and therefore they tend to be either discounted or discouraged. More seriously, when a small early-phase trial is negative

(does not demonstrate an intervention effect), this is often taken as evidence against further research on the specific intervention. Because the trial is underpowered, this conclusion is often not warranted; it is called the 'non-informative null' study in epidemiological terminology. But in the absence of any other approach besides null hypothesis testing, it is not clear how this situation might be remedied.

It will be argued here that there is a version of Neyman-Pearson[1] hypothesis testing that is more in the spirit of early-phase trials than is Fisher's null hypothesis testing. This approach is called the separation test. Whereas the result of a null hypothesis test is a rejection or confirmation of the null hypothesis, as a definitive decision, the result of a separation test is either that there is a certain degree of separation between the effects of the control and the intervention, or that there is no such separation. In the former case, there is an indication that it is worthwhile (or not worthwhile, depending on the direction of the results) to pursue research on the intervention, or there is no such indication. A separation test produces an indication of further research effort, rather than a medical or societal decision about the advisability of adopting an intervention. The purpose of this article is to introduce the separation test in the very common situation of comparing the mean outcomes between two groups.

Null Hypothesis Test

Although null hypothesis testing is certainly very familiar, it may be worth a short recapitulation, to set the stage for the definition of the separation test. In the case of a comparison of means, the customary test statistic is the mean difference: intervention mean minus control mean (we assume for convenience that higher numbers are associated with a benefit). The difference between the sample means is regarded as an estimate of the population, or long-run difference between means. The standard deviation (SD) of the measurements is presumed the same in the control and intervention conditions (there are variations that allow the SD to differ, but we are not interested in pursuing the most general case here). It follows from this that one can estimate the standard deviation of the sampling distribution of the estimate (SDE), that is, the SD of the mean difference, which is described in elementary statistics texts.

Although the acronym SDE appears novel, it is an important concept in the separation test, and so it may be helpful to explain why it is necessary to obviate an historical problem in terminology. In early statistical papers (late 1800s to early 1900s), it became a problem to understand whether a reported 'standard deviation' pertained to the individual measurement or to the mean of a group of measurements. The term 'standard error' (SE) was coined to distinguish the two, by standing for the more cumbersome 'standard deviation of the mean'. Indeed, even in modern sources the SE is defined according to the formula for the SD of the mean. With the advent of computer programs for an increasingly complex variety of statistical models, the term SE has been carried over, even though this conflicts with its most common definition (the formula for the SD of a mean does not generalise to other estimates, such as regression coefficients, for example). This confuses students, and perhaps also some researchers, quite considerably. In order to be clear, here we use SDE to stand for whatever 'standard deviation of the estimate' an appropriate computer program produces (even if it uses the 'SE' terminology). Obviously, this includes cases more general than the difference between means upon which we are concentrating.

Figure 1 provides a graphical look at the elements of a null hypothesis test. The distribution on the left is the sampling distribution of the mean difference. It is centred over zero (the solid circle in the figure), because that is the usual null hypothesis value. Two critical values are established at $\pm 1.96 \times$ SDE. If the observed mean difference falls outside these critical values, then the null hypothesis of 'no intervention effect' is rejected; otherwise, the null hypothesis is confirmed.

Although there is another distribution to the right in figure 1, it actually plays no role in Fisher's approach to hypothesis testing. The $\pm 1.96 \times$ SDE critical values are selected so that the probability of rejecting the null hypothesis when it is true is no more than 0.05, the canonical standard (which was also established by Fisher). That is, although what one wants to demonstrate is an alternative hypothesis (that the intervention is better than the control, for example), it is not necessary in null hypothesis testing to explicitly consider this in the construction of the null hypothesis test.

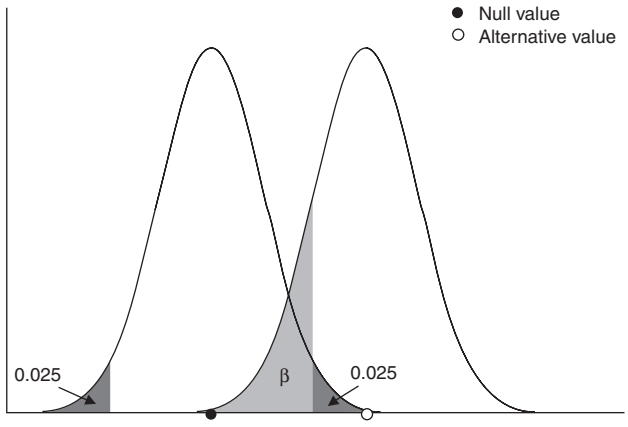

Fig. 1. The usual null hypothesis test rejects the null value when the test statistic is too large or too small (regions denoted 0.025) so that the probability of rejection when it is true is 0.05. If an alternative hypothesis is true, then β is the probability of confirming the null hypothesis (a type II error).

This issue does arise, however, when a study turns out negative (the null hypothesis is confirmed). One would then like to know what is the probability of confirming the null hypothesis when a reasonable alternative, which one would have wanted to detect, happens to be true. This is the role of the right-hand distribution in figure 1. It is the distribution of the mean difference when some clinically meaningful alternative (the open circle in the figure) is true. From the standpoint of design, since the 0.05 standard is rigidly established in biomedical research, it follows inexorably that the probability of confirming the null hypothesis when the alternative is true (β in figure 1) is entirely determined by the sample size.

To summarise, the null hypothesis test postulates that there is no difference between the population mean outcomes and demands stringent evidence before this position is required to be abandoned. No consideration of possible alternative hypotheses is required until one needs to consider the probability of correctly detecting a given, prespecified effect (this would be $1-\beta$ in figure 1).

Separation Test

A general separation test involves not one, but two hypotheses. In the case of mean differences (which we continue here), this introduces two quantities symmetrically placed with respect to 0 (no mean difference), which I denote $-\Delta/2$ and $\Delta/2$. Note that the difference between these is Δ, which I call the 'separation'.

Figure 2 shows a possible situation illustrating the separation test. The left distribution is the sampling distribution of the mean difference estimate, if the long-run mean difference were $-\Delta/2$, favouring the control. If one wanted to take this as the null hypothesis, one would reject this hypothesis if the observed mean difference were more than $1.645 \times$ SDE above $-\Delta/2$, as indicated by C in figure 2 (this is the 0.05-level one-sided hypothesis test). The right distribution in figure 2 is the sampling distribution of the mean difference estimate, if the long-run mean difference were $\Delta/2$, favouring the intervention. If one wanted to take this as the null hypothesis, one would reject this hypothesis if the observed mean difference were less than $1.645 \times$ SDE below $\Delta/2$, as indicated by $-$C in figure 2 (this is the 0.05-level one-sided hypothesis test).

The separation test of figure 2 is used as follows. If the observed mean difference is at least $1.645 \times$ SDE above $-\Delta/2$, we say that there is an indication that the intervention is better than the control. If the observed mean difference is at most $1.645 \times$ SDE below $\Delta/2$, we say that there is an indication that the control is better than the intervention. In the remaining case, we declare no indication.

Remember that an indication points to more research (at least in the case where the indication favours the intervention), but it does not provide a decision in favour of (or against) the intervention in any sense of demonstrated effectiveness.

Although the test embodied in figure 2 is accurate and general, I would like to propose a special case for general use in early-phase testing. To see why, note that in figure 2 the value of Δ had to come from outside the research results. That is, it had to be decided by the investigators as part of designing the trial, and certainly before seeing any results. In the simple separation test that I want to propose, the separation between the control and intervention groups will be determined by the level of precision that is provided by the data themselves.

This is shown in figure 3. It is a specific case of the general separation test. The value of separation, Δ, is not determined by considerations prior to the trial. It is determined by the results of the trial itself. Specifically, Δ is defined as $1.645 \times$ SDE. The value of SDE is determined by the data. Δ represents a measure of how much separation between the control and intervention groups is possible, given the fundamental variability in the observations. The practical advantage of the simple separation test, as indicated in figure 3, is that the two hypothesis values ($-\Delta/2$ and $\Delta/2$) coincide with the critical values for the observed mean difference. That is, if the mean difference exceeds $\Delta/2$ then we reject the hypothesis that the control is at least $\Delta/2$ better than the intervention, in favour of the hypothesis that the intervention is at least $\Delta/2$ better than the control. On the other hand, if the mean difference falls below $-\Delta/2$ then we reject the hypothesis that the intervention

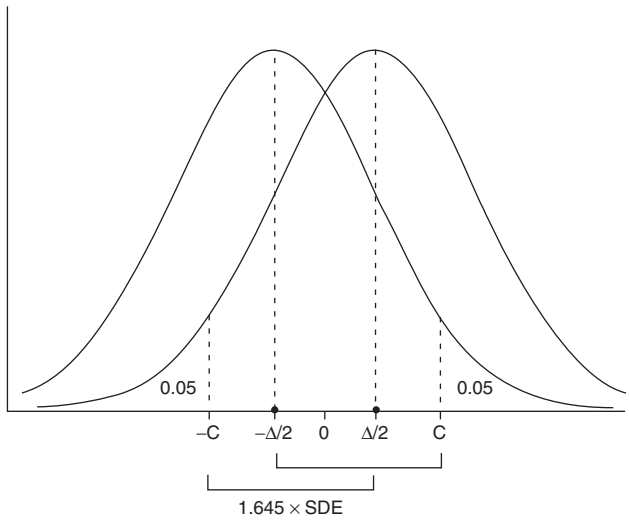

Fig. 2. The separation test postulates two hypotheses, each favouring one of the two treatment groups, and critical values (C and $-$C). The choice of $1.645 \times$ SDE is to guarantee that the type I error with respect to either hypothesis is 0.05. **SDE** = standard deviation of the sampling distribution of the estimate; Δ indicates separation.

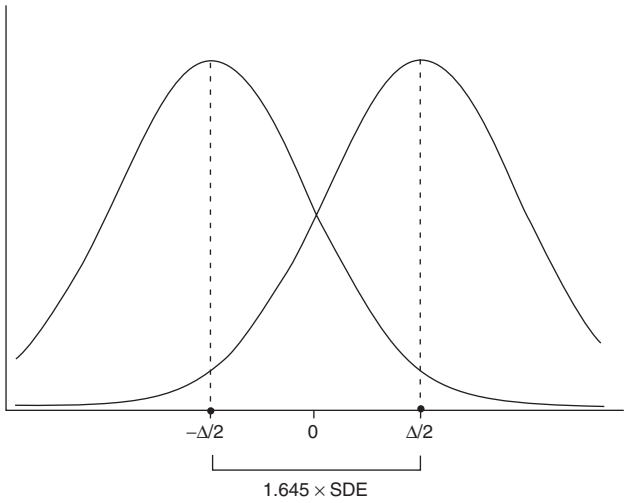

Fig. 3. The simple separation test is the special case of figure 2 in which the critical values (C and −C) coincide with the hypothesis values. **SDE** = standard deviation of the sampling distribution of the estimate; Δ indicates separation.

is at least Δ/2 better than the control, in favour of the hypothesis that the control is at least Δ/2 better than the intervention. If neither of these rejections occurs, we say that we have no indication in favour of intervention or control.

Example 1: Magnet Therapy for Fibromyalgia

The data in table I come from a study to test two kinds of magnetic bed pads (A with constant negative field, B with varying field) against sham pads (S) and against usual care (U), in the treatment of fibromyalgia.[2] The outcome shown in the table is for the Fibromyalgia Impact Questionnaire, measuring functional status. Negative values in the 'Change' column favour the first group in the 'Contrast' column.

The data in the 'Change' column are reproduced from the publication and were said to be adjusted for baseline values.[2] If the confidence intervals (in parentheses in table I) were to be used for significance testing, one would confirm the null hypothesis of no treatment difference on any row where zero fell into the interval. (The use of 0.99 confidence intervals, rather than the customary 0.95 or 0.90 intervals, was due to a multiple testing adjustment, using the Bonferroni procedure.) The conventional conclusion would be, therefore, that no differences had been demonstrated between groups.

The use of a stringent significance test (with an additional multiple testing penalty added on) can be questioned. It is clear from the article that this was an early-phase study, with small treatment groups (A, n = 37; B, n = 30; S, n = 27; U, n = 17) and a large amount of variability in the outcome measure, and therefore a separation test would be warranted.[2] From the information given in the article, one can compute the SDEs of the treatment effects, and from this the separation and critical values for the simple separation test, as in table I. The tests are carried out by comparing the changes (second column) with the negatives of the values in the right-most column. The conclusions are that one would confirm an indication of A as being better than any of the other groups, but that there were no further indications.

In order to interpret these results, one must know that the SD of the functional status score itself was about 13. The separation in the five tests was a bit more than one-half SD, a moderate value.[3] In other words, the hypotheses explicit in the simple separation test specified a moderate amount of difference between treatment groups. Since small separation is better, this study did not attempt to test for the kinds of effects that one might build into the design of a larger trial. On the other hand, by not posing a large separation, it promised to contribute to the evidence favouring further testing. And the conclusion is that A seems worthy of future testing in a larger, more definitive trial. Since A is favoured over B, the suggestion is that the larger trial not include a B group, but rather test A against either S or U. This interpretation seems a more sensible summary of the data in table I than simply saying that no statistically significant differences were observed.

Example 2: Massage for Surgical Anxiety

The data in table II come from an early-phase trial of massage to reduce anxiety and other factors important for cardiac catheterisation patients.[4] The control group contained 35 patients, and the massage group contained 43. The effects on the first three outcome measures in table II were measured as changes from baseline to just before surgery, since the experiment was designed to address presurgical conditions. The effects on the last three were measured as changes from baseline to postsurgery.

The data tables in the source article did not report SDs.[4] They did present p-values, however, from which it is possible to deduce

Table I. Comparison of changes in functional status among four magnetic bed-pad treatment groups with fibromyalgia[2]

Contrast	Change[a] (99% CI)	SDE	Δ	Δ/2
A vs S	−7.3 (−17.6, 3.0)	4.44	7.30	3.65
A vs U	−7.3 (−19.5, 4.9)	5.26	8.65	4.33
A vs B	−3.9 (−14.1, 6.3)	4.39	7.24	3.62
B vs S	−3.4 (−14.0, 7.3)	4.59	7.55	3.78
B vs U	−3.4 (−15.9, 9.0)	5.37	8.83	4.42

a Change in Fibromyalgia Impact Questionnaire.

A = constant negative field; **B** = varying field; **S** = sham pad; **SDE** = standard deviation of the sampling distribution of the estimate; **U** = usual care; Δ indicates separation.

Table II. Changes (post- minus pre-surgery) for six outcome measures in the investigation of the effects of massage among cardiac catheterisation patients[4]

Outcome	Control	Massage	Δ/2	Mean difference	Indication
Anxiety	−6.80	−16.20	4.43	−9.40	Lower
Pain	−2.30	−4.90	3.11	−2.60	
Drug use	5.70	4.70	0.72	−1.00	Lower
MABP	−7.20	−7.60	1.51	−0.40	
Heart rate	−3.30	−3.50	2.02	−0.20	
Resp. rate	−0.90	−1.00	0.78	−0.10	

MABP = mean arterial blood pressure; **resp.** = respiratory; Δ indicates separation.

the SDE (standard deviation of the mean difference in this case). The lowest of the reported p-values was 0.081, indicating that there were no conventionally significant differences. As shown in table II, there were two indications by the separation test that massage might be followed up to determine whether it lowered anxiety or drug use. Treating this study as a phase III trial would conclude that no effects were demonstrated, whereas the separation test shows that two effects were worthy of follow-up studies, and importantly, one of these was anxiety, which was the primary outcome measure in the study.

Example 3: Massage for Consequences of Bone Marrow Transplantation

The data in table III come from a study of the potential effects of massage or therapeutic touch on the complications and perceived benefits of therapy among bone marrow transplant patients.[5] This early-phase study randomised 25 women to control, 24 to massage therapy and 28 to therapeutic touch.

The means and SDs were reported for each group separately, so that it was possible to obtain the SDE for mean differences, and this was the basis for the data in table III. The separation test indicates that further investigation is warranted for improvement in pain, central nervous system (CNS) complications and gastrointestinal complications for massage therapy, and for CNS complications, pulmonary complications, gastrointestinal complications and circulation complications for therapeutic touch. There was, however, also an indication that therapeutic touch might be associated with increased skin complications. Among the perceived benefits (affective and comfort), there were indications that both

Table III. Mean outcomes for complications and perceived benefits of massage and therapeutic touch versus control among bone marrow transplant patients[5]

Outcome	Control	Massage	TT	Massage vs control			TT vs control		
				Δ/2	mean difference	indication	Δ/2	mean difference	indication
Complication									
Pain	1.71	1.40	1.86	0.23	−0.31	Lower	0.19	0.15	
Performance	2.18	2.11	2.25	0.14	−0.07		0.13	0.07	
Food intake	2.36	2.39	2.54	0.13	0.03		0.13	0.18	Higher
CNS	1.61	0.64	1.31	0.26	−0.97	Lower	0.24	−0.30	Lower
Pulmonary	1.49	1.52	1.22	0.30	0.03		0.27	−0.27	Lower
Cardiac	0.68	0.56	0.75	0.27	−0.12		0.27	0.07	
Hepatic	0.74	0.87	0.85	0.28	0.13		0.24	0.11	
Gastrointestinal	2.21	1.93	2.01	0.15	−0.28	Lower	0.16	−0.20	Lower
Genitourinary	1.04	1.16	0.96	0.31	0.12		0.28	−0.08	
Skin	1.17	1.30	1.44	0.27	0.13		0.24	0.27	Higher
Circulation	1.60	1.44	1.15	0.27	−0.16		0.27	−0.45	Lower
Mean	1.53	1.42	1.49	0.15	−0.11		0.12	−0.04	
Perceived benefit									
Affective	14.85	19.00	17.15	1.61	4.15	Higher	1.84	2.30	Higher
Comfort	13.00	21.07	18.46	1.29	8.07	Higher	1.59	5.46	Higher
Total	27.42	40.07	35.62	2.87	12.65	Higher	3.38	8.20	Higher

CNS = central nervous system; **TT** = therapeutic touch; Δ indicates separation.

therapies were worth following up, and similarly with regard to the total score.

The source article did not carry out individual group comparisons. Instead, it reported p-values from F-tests of the equality of means across the three groups. This approach answers the question of whether it is reasonable to believe that all three group means are the same, but it does little to elucidate which group means might be different. Thus, among the complications, only CNS was identified as showing statistical significance without any more informative inferential statements. Among perceived benefits, the affective component was not identified as being significant, but the comfort component was, as was the total perceived benefit score.

The separation test gives an indication that both components of the perceived benefit are worth following up. It also exhibits a richer pattern of potential effects among complications than is provided by the omnibus F-test, thereby providing leads for further study. Since it is possible from the source data to compute that the effect size (Δ divided by the SD of the underlying measurement) was about 0.47 for the complications, and about 0.65 for the perceived benefits, it is sensible to conclude that the indicated effects are in the medium range.[3]

Sample Size Issue

It is worth emphasising that the Δ used in the separation test is not the detectable effect, as specified in conventional designs. Table IV shows, for a range of sample sizes (numbers of individuals per treatment group), what the corresponding separation Δs are, and what the effects detectable at 95% power are. Both are reported here as effect sizes, so that the units are expressed in terms of the underlying SD. Clearly the detectable effects are considerably larger than the separation Δs. This reflects the fact that detectable effects are stated in terms of definitive decisions, and all of the sample sizes in table IV are inadequate to detect small effect sizes (in the sense of Cohen,[3] p. 26). Since the separation test provides only an indication, the corresponding separation is smaller for each sample size, because less is being delivered (an indication instead of a decision). The important point here is that if small early-phase studies are to be judged by the criteria of phase III trials, they will automatically be underpowered. By reframing the purpose of a phase I/II trial as an indication rather than a decision, the corresponding separation values that can be achieved with small sample sizes are seen to be reasonable.

Table IV. Comparison of the separation (Δ) and the conventional effect detectable with 95% power. Sample sizes adequate for the separation test would not be adequate for conventional tests at the same power

No./group	Δ	Detectable effect
10	0.74	1.75
20	0.52	1.24
30	0.42	1.01
40	0.37	0.88
50	0.33	0.78
60	0.30	0.72
70	0.28	0.66
80	0.26	0.62
90	0.25	0.58
100	0.23	0.55

As a footnote, it should be noted that I have used the normal-distribution critical values of 1.645 and 1.96 throughout. This was done for convenience of exposition. In small samples, one is advised instead to use critical values from the t-distribution tables. These values are always larger than the normal values and converge to them as the sample size increases.

Conclusions

The separation test provides an interval around a null hypothesis value (no treatment difference) that is determined by the data and provides the basis for an indication for further research. Although it superficially resembles the 'equivalence test',[6] it has a different form and purpose. The equivalence test is designed to reject the hypothesis that the benefit (or harm) from a treatment exceeds a certain value, which must be determined before the trial. By rejecting such a hypothesis, the test leads to the decision that the treatments are essentially the same, within the prespecified difference. It is thus identical to null hypothesis testing, in that it intends to reach a decision rather than an indication, and detectable differences must be decided before seeing the data, rather than being computed from the data, as in the separation test.

Analogies between statistical inference and the reasoning of the legal system are often drawn. In our situation, the null hypothesis test corresponds to a criminal proceeding, in which the standard for a 'guilty' verdict is 'proof beyond a reasonable doubt'. Because a decision is being made which is sometimes irrevocable, definitive evidence is required. Separation tests, on the other hand, correspond to civil suits, in which the requirement to find for one side is 'a preponderance of the credible evidence'. It is not necessary for either side to prove its case, but only for one side to make a stronger demonstration than the other. In a civil case, the jurors may well leave the box feeling that neither side had demonstrated very much, but the judge required them to make a decision nonetheless, because that is the purpose of a civil court. Likewise, the separation test may not be in a good position to assert very much, to the extent that the separation Δ is large, but it will do a

better job than the courts because it will allow no decision when the evidence supports no decision. The jurors can also wind up feeling that on balance the evidence favours one side to a sufficient degree, and when the separation Δ is small this will often be the conclusion of the separation test.

With this view in mind, it seems that the null hypothesis significance test is overused in biomedical research. This probably arises from a general desire to have a single, simple, well understood method of doing inference that will fit all situations. It may also reflect the fact that most scientists find little of interest in the technical or philosophical intricacies of inference *per se*. They do not want to debate whether their results are real, but just to assert them in a conventional way. Nevertheless, one has the feeling in many publications that the scientific issues underlying the specific research project were not sufficiently advanced to warrant a significance test, and in any case the sample size was too small, or the results too variable, to permit a definitive conclusion.

Although CAM is certainly not the only area of biomedical research where early-phase designs are appropriate, it is perhaps distinguished because of the large number of CAM approaches that need to be carefully tested for the first time. In some cases this is simply because no prior research has been done, and in other cases it is because the quality of the prior research is felt to be in need of improvement. In clinical naturopathy, a large number of approaches are in daily use, many of which could be incorporated into early-phase research trials. It is worthwhile for these trials to be funded, if for no other reason than the fact that the therapies are actually being used. In the competition for funding, however, later-phase studies have a built-in advantage, because they fit better into the conventional null hypothesis testing framework. This means that, in effect, larger trials have a 'survival advantage' in terms of attracting funding, because they appear to be on a more substantial inferential footing. It is not appropriate, however, to propose a larger-than-necessary phase I trial simply to attract funding. The size of the trial should be roughly proportional to how much of the science of the modality is understood, and how much previous scientific evidence is available. Whether a therapeutic approach with little research history is worthy of further study should be based on a small early-phase trial. The separation test provides a firm inferential base for the appropriate early-phase trial, based on the metaphor of an indication as opposed to a decision.

Traditionally, evidence-based medicine has been concerned with the logical and scientific form of the evidence about a treatment, and the strength of the statistical results that would lead to adopting the treatment. This frequently involves cumulation of evidence over trials and is ultimately prescriptive. It has not been seen to be in the realm of evidence-based medicine to deal with issues concerning which kinds of therapies or specific therapies are worthy of further research. And yet this is an important part of the overall purview of evidence-based medicine, since some proportion of therapies abandoned owing to early negative results will have been discarded in error, principally because of the inadequate sample sizes (or inadequate inferential method) prevalent in small studies. This leads to the possibility of a new type of evidence-based medicine research, a study of the 'promise' of a newly researched therapy, with the potential conclusion that further research is warranted, rather than simply saying that effectiveness has not been demonstrated. Such research does not appear to be possible within the confines of the classical null hypothesis test, but the separation test seems to resolve this problem.

Acknowledgements

The authors have provided no information on sources of funding or on conflicts of interest directly relevant to the content of this article.

References

1. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A 1933; 236: 333-80
2. Alfano A, Taylor AG, Foresman PA, et al. Static magnetic fields for treatment of fibromyalgia: a randomized controlled trial. J Altern Complement Med 2001; 7 (1): 53-64
3. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdate (NJ): Lawrence Erlbaum, 1987
4. Okvat HA, Oz MC, Ting W, et al. Massage therapy for patients undergoing cardiac catheterization. Altern Ther Health Med 2002; 8 (3): 68-75
5. Smith MC, Reeder F, Daniel L, et al. Outcomes of touch therapies during bone marrow transplant. Altern Ther Health Med 2003; 9 (1): 40-9
6. Wellek S. Testing statistical hypotheses of equivalence. London: Chapman & Hall, 2002

Correspondence and offprints: Dr *Mikel Aickin*, Program in Integrative Medicine, University of Arizona, PO Box 245153, Tucson, AZ 85724-5153, USA.
E-mail: maickin@earthlink.net