# Statistical Inference by Confidence Intervals: Issues of Interpretation and Utilization

This article examines the role of the confidence interval (CI) in statistical inference and its advantages over conventional hypothesis testing, particularly when data are applied in the context of clinical practice. A CI provides a range of population values with which a sample statistic is consistent at a given level of confidence (usually 95%). Conventional hypothesis testing serves to either reject or retain a null hypothesis. A CI, while also functioning as a hypothesis test, provides additional information on the variability of an observed sample statistic (ie, its precision) and on its probable relationship to the value of this statistic in the population from which the sample was drawn (ie, its accuracy). Thus, the CI focuses attention on the magnitude and the probability of a treatment or other effect. It thereby assists in determining the clinical usefulness and importance of, as well as the statistical significance of, findings. The CI is appropriate for both parametric and nonparametric analyses and for both individual studies and aggregated data in meta-analyses. It is recommended that, when inferential statistical analysis is performed, CIs should accompany point estimates and conventional hypothesis tests wherever possible. [Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys Ther.* 1999;79:186–195.]

**Key Words:** *Confidence intervals, Estimation, Hypothesis testing, Statistical inference.*

*Julius Sim*

*Norma Reid*

Much published physical therapy research that involves statistical inference seems to make exclusive use of hypothesis testing. In this approach, a null hypothesis of no difference (or of no association, according to the nature of the relationship being examined) is posited, and, by means of a statistical test, this hypothesis is either rejected or not rejected. A number of limitations to the hypothesis testing approach have been identified. Among these limitations are that this approach gives little or no indication of the magnitude of statistical relationships, that it reduces statistical inference to a process of binary decision making, and that whether or not statistical significance is achieved may simply be a function of choice of sample size.[1–3]

An alternative approach to statistical inference, using confidence intervals (CIs), assists in addressing some of these limitations. In the medical literature, there has been increasing attention focused on the use of CIs.[4–7] In a discussion of various aspects of statistical inference, Ottenbacher[8] has advocated greater use of CIs in rehabilitation research, and CIs feature prominently in a recent discussion of statistical inference in rehabilitation.[9]

In this article, we examine some of the merits of an approach to statistical inference based on CIs. The conventional approach to hypothesis testing is described, followed by a discussion of the nature and use of CIs. Key strengths of the CI as a means of statistical inference are then considered, in particular, the precision that they attach to statistical estimates and the light they shed on issues of clinical importance. Finally, recommendations are made concerning the use of CIs.

## Conventional Hypothesis Testing

### *Basic Principles*

When an experiment or other form of quantitative study is carried out, it is rarely the case that data are gathered from the entire population of interest. Most commonly, a study sample is selected, and this will be just one of an infinite number of possible samples from the population. At the conclusion of the study, the results are generalized back from the sample to the population.

Certain properties of the study sample can be calculated, such as the variance of all the scores for a given variable, the mean of these scores, or the mean difference in scores between 2 groups within the sample. These values are referred to as *statistics*. They can be calculated for each variable represented in the sample (although different statistics may be appropriate for different variables), and they will normally be slightly different, as different samples are drawn from the population. The corresponding properties of the population are known as *parameters*, and, because there is only one population, these values are fixed. Thus, the mean of a given population is invariant, but the means of a series of samples drawn from that population will typically vary to some degree. A study statistic, such as the mean, is an estimate of the corresponding population parameter. It is an estimate because the true value of the population is almost always unknown.

J Sim, PhD, PT, Professor, Department of Physiotherapy Studies, Keele University, Keele, Staffordshire ST5 5BG, United Kingdom (pta05@keele.ac.uk). Address all correspondence to Dr Sim.

N Reid, DPhil, Professor of Health Sciences, Office of the Vice-Chancellor, University of Plymouth, Plymouth, Devon, United Kingdom.

*This article was submitted April 9, 1998, and was accepted August 23, 1998.*

**Table 1.**
Results From the First Hypothetical Study of the Effectiveness of Treatment for Patients With Fibromyalgia Syndrome (N=50)

| | Visual Analog Scale Scores | | Pain Relief | | Statistical Test | | |
|---|---|---|---|---|---|---|---|
| Subjects[a] | Mean (Pretest) | Mean (Posttest) | Mean | Mean Difference | *t* | *df* | *P* |
| Group 1 | 68.68 | 61.08 | 7.6 | 3.48 | 2.08 | 48 | .043 |
| Group 2 | 66.52 | 55.44 | 11.08 | | | | |

[a] The subjects in the 2 groups received different treatments designed to alleviate pain.

When a study is carried out, one or more relationships between various statistics will often be examined. For example, a study can examine an association between 2 variables within the sample or a difference in the mean or median values of a variable between 2 (or more) subgroups within the sample. The purpose of a statistical hypothesis test is to determine whether such a relationship is a "real" one (ie, it represents a corresponding relationship in the population) or a "chance" one (ie, it has emerged due to sampling variation and, although accurately reflecting the relationship that exists in the sample, does not necessarily represent a corresponding relationship in the population). The way in which the statistical test accomplishes this is by asking the question: What is the likelihood of finding this relationship in the sample, if, in fact, no such relationship exists in the population from which it was drawn? This assumption of no relationship is referred to as the *null hypothesis*, and the rival assumption (that the relationship does exist within the population) is referred to as the *alternative hypothesis*.[10]

A hypothetical example may serve to illustrate the way in which a statistical hypothesis test is utilized. A sample of 50 patients with fibromyalgia syndrome (FMS) is drawn randomly from a population of such patients and then assigned (again randomly) to receive 1 of 2 treatments designed to alleviate pain. The null hypothesis is that a difference in pain relief will not exist between the 2 groups following treatment. The alternative hypothesis is that such a difference will exist. The chosen outcome variable, pain intensity as measured in millimeters on a 10-cm visual analog scale (VAS), is measured before and after treatment, and a pain relief score is thereby obtained (pretest score minus posttest score). The mean pain relief score can then be calculated for each group.

This pain relief score will almost certainly differ in the 2 groups, but the question is whether such a difference in outcome is attributable to an underlying difference in the treatments received by the groups, rather than simply to the effect of sampling variation or of chance differences in group assignment. If there is a sufficiently high probability that the observed difference in outcome can be attributed to such variation in sampling or group assignment (conventionally, a probability of 5% or above [ie, $P \geq .05$]), then the assumption of no difference between the treatments (ie, the null hypothesis) is retained. That is, the observed difference between groups is considered to be no greater than the difference expected from variation in sampling or group assignment, at the specified level of probability, and the assumption of no underlying difference between groups, therefore, is considered to be plausible. Conversely, if there is a sufficiently low probability that the observed difference in outcome is due to these chance factors (conventionally, a probability of less than 5% [ie, $P < .05$]), then the assumption of no difference between treatments is rejected. That is, it is considered more plausible that the difference in outcome is due to an underlying difference between the groups than that it is due to chance factors such as variation in sampling or group assignment. The null hypothesis is rejected in favor of the alternative hypothesis.

Table 1 shows the results of the hypothetical experiment. A statistical test for differences was used in this study. Because the data concerned are continuous, are approximately normally distributed, and can be argued to lie on an interval scale of measurement, a *t* test for independent measures was performed. The probability associated with the test statistic ($P=.043$) was less than the conventional critical value of .05. This finding provides sufficient grounds for doubting the null hypothesis, which is therefore rejected in favor of the alternative hypothesis. The difference in pain relief between the 2 groups is said to be real.

There are 2 important and related questions that have not yet been answered with respect to this study. The first question relates to the importance of the difference that has been detected. Although a mean difference in pain relief between the 2 treatments has been shown to be real, based on probabilistic statistics, it may be of little practical importance. On a 10-cm VAS, a difference between treatments of 3.48 mm is relatively minimal and could be outweighed by other features of the more successful treatment (eg, it might be more expensive or require more frequent attendance by the patient).

The outcomes of statistical tests need to be considered in the context of the situation to which they relate, and outcomes of clinical research must be subjected to clinical judgment. It is worth noting that the converse of the situation just outlined may also arise. A difference in pain relief may be found not to be real, perhaps due to insufficient sample size or a high degree of variance in subjects' scores (Type 2 error). In such a case, although the observed difference in pain relief is real for this sample, it cannot be assumed to reflect a real difference in the population. In order for such a finding to be applicable to general clinical practice, the observed difference must be shown to be real for the population. This will require a reduction in the risk of a Type 2 error, through an increase in sample size, a more precise method of measurement, or other means of reducing random error.

Another unanswered question relates to the magnitude of this difference in pain relief between groups. This difference is an estimate of the difference that would exist if the full population of patients with FMS were studied. All we know is that the difference found in this particular sample is sufficiently great for it to be attributed to a genuine difference between the treatments rather than to chance variation in sampling or allocation to groups. We do not know how good an estimate it is of the true population difference (ie, the difference we would find had we tested the treatments on the whole population of individuals with FMS). The hypothesis test has told us, on a "yes/no" basis, whether the observed difference is real, but it has not enlightened us as to the true value of this difference in the population. As Abrams and Scragg[11] point out, a probability value conveys no information about the size of the true effect. This is information, however, that we need in order to inform clinical practice.

The CI assists in addressing these questions as to the clinical importance and magnitude of an observed effect and remedies some of the shortcomings of more conventional approaches to hypothesis testing. These points will be considered in detail following an account of interval estimation.

## The Nature of Confidence Intervals

A sample statistic, such as a sample mean, provides an estimate of a population parameter. It provides a single estimate of the specific value of the parameter on the basis of the observed value of the statistic. As an adjunct to a single estimate, an *interval estimate* can be calculated. This interval estimate specifies a range of values on either side of the sample statistic within which the population parameter can be expected to fall with a chosen level of confidence.[2] To return to the FMS study, the mean pretest VAS score for the sample of patients is 67.6 mm, and the associated 95% CI is 64.15, 71.05. What this CI tells us is that we can be 95% sure that the population mean for this variable lies somewhere between 64.15 and 71.05 mm. It is also the case that values toward the extremes of this interval are rather less likely to represent the population mean than those nearer the center.[2]

The essential meaning of a 95% CI can be expressed as follows. If we were to draw repeated samples from a population and calculate a 95% CI for the mean of each of these samples, the population mean would lie within 95% of these CIs. Thus, in respect of a particular 95% CI, we can be 95% confident that this interval is, of all such possible intervals, an interval that includes the population mean rather than an interval that does not include the population mean. It does not strictly express the probability that the interval in question contains the population mean, as this must be either 0% or 100%. The population mean is either included or not included.[12,13]

The function of a CI, therefore, is essentially an inferential one. A CI is used when examining a characteristic of a sample (in this case, the mean pretest VAS score) in terms of its degree of variability in the corresponding population. If the researcher's concern with sample statistics, however, is a purely descriptive one (ie, if the researcher is interested only in the pretest VAS score of the sample, without reference to the population from which this sample was drawn), conventional measures of dispersion, such as the standard deviation (for a mean) or the semi-interquartile range (for a median), should be used.

### *Width of Confidence Intervals*

For a given level of confidence, the narrower the CI, the greater the precision of the sample mean as an estimate of the population mean. In a narrow interval, the mean has less "room" to vary. There are 3 factors that will influence the width of a CI at a given level of confidence.

First, the width of the CI is related to the variance of the sample scores on which it is calculated. If this sample variance can be reduced (eg, by increasing the reliability of measurements), the CI will be narrower, reflecting the greater precision of the individual measurements. Selecting a sample that is more homogeneous will reduce the variance of scores and thereby increase their precision. This factor, however, is often outside the researcher's influence.[14]

Second, following the principles of sampling theory, sampling precision increases in a curvilinear fashion with increasing sample size. This increase in precision occurs because the variance of a statistic, as expressed by

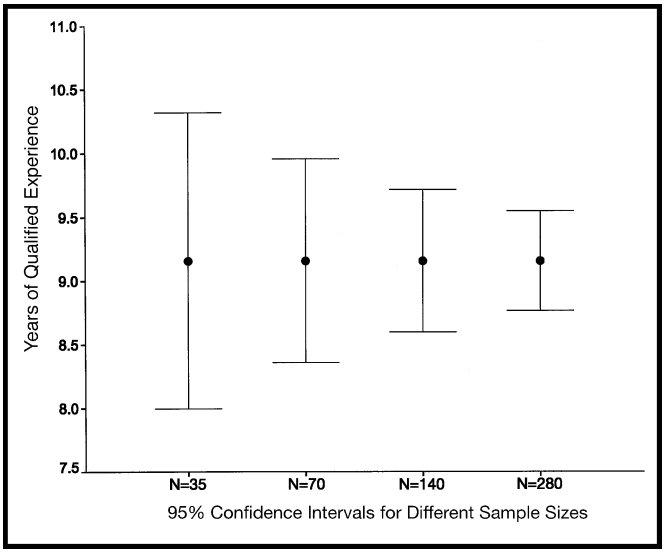


**Figure 1.**
Means and confidence intervals of years of qualified experience for progressively larger samples of physical therapists drawn from a single population.

its standard error, decreases as sample size increases. Figure 1 shows 4 samples of a progressively greater size drawn from a single population of physical therapists and the mean period of postqualification experience for each sample. The mean is precisely the same in each case, but the CI becomes narrower as the sample size increases. As sampling precision is related to the square root of the sample size, doubling the sample size will only decrease the width of the CI by 25%.[15]

Third, the chosen level of confidence will influence the width of the CI. If the investigator wants to be 99% confident of having included the population mean within the interval, this interval would be wider. With a higher level of confidence, the interval needs to be wider in order to support the claim of having included the population parameter at the chosen level of confidence. Conversely, a 90% CI would be narrower than a 95% CI.

It is not the case, however, that, at a given confidence level, a narrow CI is any more (or less) likely than a wider CI to be one that contains the population parameter. The probability of including the parameter is determined by the chosen confidence level, not by the width of the particular CI concerned. If a 95% CI is narrow, this means that only a small range of possible values has to be included in order to be 95% confident that the CI contains the parameter. Correspondingly, a wide CI means that a large range of possible values has to be included in order to be 95% confident that the parameter lies within the CI. The probability of inclusion, however, is the same in both cases. A 95% CI is, by definition, one that is 95% likely to contain the population parameter, irrespective of its width.

The width of a CI is indicative of its *precision* (ie, the degree of random error associated with it), but it does not convey its *accuracy* (ie, whether it includes the population parameter), which is determined by the chosen level of confidence.[9] Choosing a 99% CI rather than a 95% CI will increase the accuracy of the CI (ie, it will have a greater chance of being one of those that includes the population parameter), but will decrease its precision (ie, it will be wider than the corresponding 95% CI).

The usefulness of a CI depends on the point statistic (eg, the sample mean) on which it is based being an unbiased point estimate. If systematic error is present in a study, the point estimate will lie at some distance from the true value of the parameter. In such a case, a CI based on a large sample will, paradoxically, be more misleading than one based on a small sample.[16] Consider again Figure 1. Imagine that the point estimate of 9.2 obtained is biased and that the true population mean is 8.5. It is evident that, unlike the 2 wider CIs, the narrow CIs, based on the larger samples, actually exclude this value. In the presence of systematic error, the lesser precision afforded by a wide CI actually increases the likelihood of its including the true population value. This example illustrates the fundamental point that increases in sample size will only assist in dealing with random, not systematic, error. Systematic error is usually an issue in study design rather than a function of the statistics used.

### Calculating the Confidence Interval

The 95% CI stated earlier for the mean pretest VAS scores in the FMS study is calculated from the sample mean ($\overline{X}$), the statistic from the $t$ distribution representing a 95% level of confidence at 49 degrees of freedom ($t_{cv}$), and the standard error of the sample mean (SE), according to the following formula:

$$95\%\ \text{CI} = \overline{X} \pm (t_{cv} \times \text{SE})$$
$$= \overline{X} \pm (2.010 \times 1.7183)$$
$$= 67.6 \pm 3.45$$
$$= 64.15, 71.05$$

The terms in the calculation relate to some of the basic concepts considered earlier. The $t$ statistic corresponds to a particular probability level (and thus a confidence level), the degrees of freedom are determined by sample size, and the standard error of the mean represents sample variance.

For a 99% CI, $t_{cv}$ would be 2.683, and the CI will accordingly be wider: 62.10, 72.21. Conversely, for a 90%

CI, $t_{cv}$ would be 1.676, resulting in a narrower CI of 64.72, 70.48.

A 95% CI for a difference in means would be calculated in an analogous manner:

$$95\%\ \text{CI} = (\bar{X}_1 - \bar{X}_2) \pm (t_{cv} \times \text{SE}_{\text{diff}})$$

where $\bar{X}_1$ and $\bar{X}_2$ are the 2 sample means and $\text{SE}_{\text{diff}}$ is the standard error of the difference between these means.

Confidence intervals can also be constructed for sample statistics other than the mean and in relation to samples that do not satisfy the assumptions of parametric statistics.[17–19]

## Advantages of Confidence Intervals

### *Confidence Intervals Attach a Measure of Accuracy to a Sample Statistic*

Determining the accuracy of a point estimate is not possible. A statistic such as a sample mean is just one estimate of the population mean, and, because the population mean is nearly always unknown, it is not possible to know how good the estimate is. Table 2 shows the means and 95% CIs of 10 random samples of 9 cases drawn from a population of 300, and Figure 2 displays the same data graphically. In this case, the population mean is 9.5, but the researcher is unlikely to know this. Any one of these point estimates of the mean, taken on its own, gives no indication of how precise an estimate it is or of how near it lies to the population mean. In contrast, for any of the associated 95% CIs, we can be 95% sure that the population mean lies somewhere between its upper and lower limits. Moreover, the width of the interval gives an indication of the precision of the point estimate. For those samples with smaller variance, the interval is narrower, reflecting greater precision in the point estimate. Because each CI differs from sample to sample, however, it does not follow that any single CI will include the means of 95% of all samples of 9 cases drawn from this population.[20]

### *Confidence Intervals Can Function as Hypothesis Tests*

Confidence intervals have been considered as a means of estimating the value of a parameter. In many cases, however, researchers may want to test a specific hypothesis, as in the example of the FMS study considered earlier. Although the orthodox approach is to conduct a hypothesis test, in this case using the independent-measures *t* test (Tab. 1), CIs can also be used to either reject or retain the null hypothesis, and thereby perform precisely the same function as the orthodox significance test. The appropriate CI for hypothesis testing is determined by subtracting alpha (the criterion probability value for statistical significance) from 1. Thus, a 95% CI would be used to test the null hypothesis at the $P<.05$ level, a 99% CI would be used to test the null hypothesis at the $P<.01$ level, and so forth.[12]

**Table 2.**
Means and 95% Confidence Intervals (CIs) (in Arbitrary Units) for 10 Random Samples (n=9 in Each Sample) Drawn From a Population of 300 Individuals by Type of Hypothetical Data

| Sample No. | Mean | 95% CI |
|---|---|---|
| 1 | 10.57 | 8.87, 12.28 |
| 2 | 8.11 | 5.04, 11.18 |
| 3 | 9.67 | 7.12, 12.22 |
| 4 | 8.88 | 6.00, 11.78 |
| 5 | 10.67 | 8.95, 12.40 |
| 6 | 8.10 | 5.00, 11.19 |
| 7 | 10.11 | 7.63, 12.59 |
| 8 | 9.43 | 7.20, 11.65 |
| 9 | 11.70 | 10.05, 13.34 |
| 10 | 7.71 | 4.60, 10.81 |

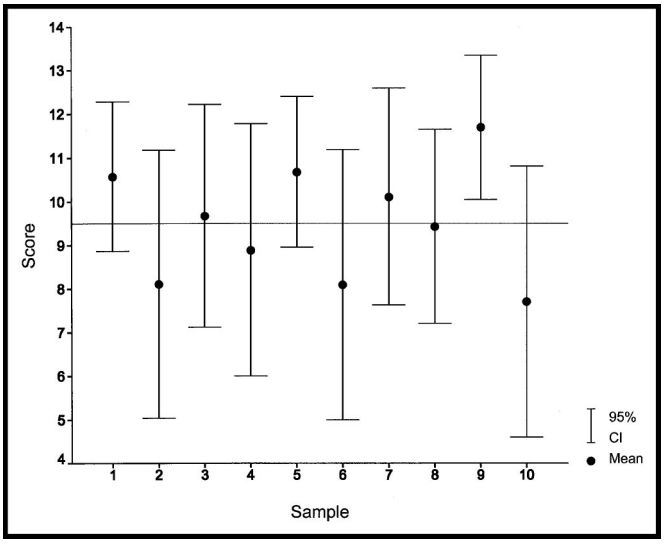


**Figure 2.**
Graphic representation of the data presented in Table 2. The solid horizontal reference line represents the population mean of 9.5. The 95% confidence interval (CI) for the ninth sample fails to include the population mean; it is as likely as not that this will be the case for 1 in 10 samples. This CI is actually the narrowest of the 10 CIs, illustrating the fact that the precision expressed by a CI is independent of its accuracy.

The way in which the null hypothesis is tested by means of a CI is by determining whether the null value (ie, the value specified in the null hypothesis) lies within the CI. If the null value lies within the CI, we cannot exclude it as being the population parameter at the chosen level of confidence. In contrast, if the null value lies outside the CI, we can exclude the null value from the possible values of the population parameter at this level of confidence. For example, Table 3 shows a 95% CI for the mean difference between the pain relief scores for the 2 groups in the FMS experiment, in addition to the results of the *t* test reported previously. The null value is that of

**Table 3.**
Results From the First Hypothetical Fibromyalgia Study, Showing 95% Confidence Interval (CI) for Mean Difference in Pain Relief

| | Pain Relief | | | Statistical Test | | |
|---|---|---|---|---|---|---|
| Subjects[a] | Mean | Mean Difference | 95% CI for Mean Difference | *t* | *df* | *P* |
| Group 1 | 7.60 | 3.48 | 0.11, 6.85 | 2.08 | 48 | .043 |
| Group 2 | 11.08 | | | | | |

[a] The subjects in the 2 groups received different treatments designed to alleviate pain.

**Table 4.**
Results From the Second Hypothetical Fibromyalgia Study, Using a Sample of 100 Patients

| | Pain Relief | | | Statistical Test | | |
|---|---|---|---|---|---|---|
| Subjects[a] | Mean | Mean Difference | 95% Confidence Interval for Mean Difference | *t* | *df* | *P* |
| Group 1 | 7.60 | 3.48 | 1.21, 5.75 | 3.05 | 98 | .003 |
| Group 2 | 11.08 | | | | | |

[a] The subjects in the 2 groups received different treatments designed to alleviate pain.

zero difference, and the 95% CI does not include zero. That is, the researcher can be 95% confident that the difference that would be found in the population of patients with FMS would be greater than zero, which is equivalent to rejecting the null hypothesis of no difference at the $P<.05$ level by means of a *t* test.

The 95% CI gives additional information not afforded by the outcome of the *t* test. This advantage of the CI can be illustrated by considering the results of a second experiment, this time carried out with a larger sample of 100 patients with FMS (Tab. 4). The effect size (ie, the mean difference in pain relief) is the same as in the study using 50 patients, but the probability value is much smaller. The same effect size returns different probability values, depending on the size of the sample. Thus, the probability value is not a direct function of the effect size alone, but is instead a function of both the effect size and the sample size, just as the weight of a cylindrical object is a function of both its height and its diameter. It follows that knowledge of the probability value returned by a hypothesis test does not, on its own, provide an indication of the likely effect size in the population, just as the reported weight of a cylinder does not, on its own, indicate its height. The CI, in contrast, provides this information.

The 95% CI has a further advantage over conventional hypothesis testing because it is equivalent to a hypothesis test for not just one population value but a range of possible population values, in respect to the sample on which it has been calculated.[21] To illustrate this point, imagine that a third study was done on a sample of patients with FMS (n=50), using a different pair of clinical interventions. The *t* test performed rejected the null hypothesis that the population difference between treatments is zero (Tab. 5). If the null value, however, lies outside the 95% CI, the corresponding null hypothesis will be rejected at the $P<.05$ level. Thus, in this example, the 95% CI serves to reject not only a null hypothesis based on a difference of zero, but also a null hypothesis, in respect to this particular sample, based on any population difference outside the limits of the CI. Based on the evidence of this study, the researcher not only can be 95% confident that there is a nonzero difference between treatments, but can also be 95% confident that this difference is not less than 2.63 or greater than 11.69. The *t* test, in contrast, only serves to exclude a difference of zero and does not allow inferences to be made about other possible values of the population parameter.

If a series of CIs are calculated on a given data set, the risk of a Type I error (ie, the risk of rejecting the null hypothesis when it is true) will rise accordingly, just as in the case of multiple tests of significance.[2] This is a particular problem when unplanned, *post hoc* comparisons are made. In such circumstances, it is often appropriate to adjust the level of confidence for the CI to maintain the same risk of a Type I error, in a manner analogous to the Bonferroni procedure for adjusting probability values.[10] Thus, if 5 CIs were to be calculated on a single data set for the purpose of testing 5 null hypotheses, the confidence level could appropriately be adjusted from 95% to 99% for each CI, and the overall level of confidence will thereby be maintained at 95%.

**Table 5.**
Results From the Third Hypothetical Fibromyalgia Study, Showing a 95% Confidence Interval (CI) for Mean Difference in Pain Relief

| | Pain Relief | | | Statistical Test | | |
|---|---|---|---|---|---|---|
| Subjects[a] | Mean | Mean Difference | 95% CI for Mean Difference | *t* | *df* | *P* |
| Group 1 | 6.52 | 7.16 | 2.63, 11.69 | 3.18 | 48 | .003 |
| Group 2 | 13.68 | | | | | |

[a] The subjects in the 2 groups received different treatments designed to alleviate pain.

When parametric analysis involving the testing of multiple hypotheses is being performed, there are techniques that take account of the number of such comparisons more efficiently than manual adjustment of probability values (eg, the various multiple range tests associated with analysis-of-variance procedures).[13] In such situations, it is probably most advisable initially to conduct the process of hypothesis testing by a technique of this sort and then report CIs for the pair-wise comparisons found to be both statistically significant and clinically important.

### *Confidence Intervals Are Informative on Questions of Clinical Importance*

The CI can provide valuable information when trying to determine the clinical importance of the outcome of a trial. In the third FMS study (Tab. 5), the mean difference in pain relief of just over 7 mm on the VAS was sufficient to reject the null hypothesis of no difference with an independent *t* test. Despite the statistical significance (a high probability that the result was real) attained by this outcome, the lower limit of the 95% CI reveals that the true value of the difference between treatments could be as low as 2.63 mm. The CI for the pain-relief scores is wider in this FMS study than in the previous FMS study (Tab. 4) due to the greater variance of these scores. Although a mean difference of 7.16 mm on a VAS is arguably likely to be clinically important and is the best estimate of the corresponding population parameter for this study sample, a difference of 2.63 mm, which is arguably not likely to be clinically important, is also compatible with rejection of the null hypothesis at the $P<.05$ level. Accordingly, 2 conclusions can be drawn from these results:

1. On the basis of this study, the researcher can be more than 95% confident that there is a difference in the effectiveness of the treatment for the population of patients with FMS between the 2 interventions tested. This conclusion can be inferred from either the *t*-test results or the 95% CI.

2. Despite having rejected the null hypothesis, the researcher cannot be 95% confident that the value of this difference in the population of patients with FMS is clinically important, and this conclusion emerges from the 95% CI alone.

We believe that it is equally important for CIs to be reported when the null hypothesis is not rejected. Gore pointed out that a nonsignificant statistical test is "a statement that the trial results are consistent with there being no difference between treatments, and is not at all the same as saying that there is actually no difference."[15(p660)]

A 95% CI of the differences in scores that includes zero, and thereby causes the null hypothesis to be retained, may nonetheless contain differences, in either direction, that could be clinically meaningful and that may represent the "true" population value. Such information is potentially important to the clinician but is not revealed by inspection of the probability value alone. The Physicians' Health Study[23] provides a case in point. In this randomized, controlled trial investigating the effects of aspirin and a placebo in the prophylaxis of stroke (N=22,071), the odds ratio (ie, the ratio of the likelihood of death from stroke in the group that received aspirin to the likelihood of death in the group that received a placebo) was 3.0.[23] This 3-fold difference, however, was not statistically significant ($P=.16$). This finding is confirmed by the fact that the 95% CI for the odds ratio (0.75, 11.98) includes 1, which is the null value for an odds ratio. Inspection of the upper limit of the CI for the odds ratio reveals that an almost 12-fold difference in fatal stroke between the 2 groups cannot be excluded as the population parameter with 95% confidence. Thus, there is considerable lack of precision in the odds ratio for this study, which is due to the low incidence of strokes among the subjects (6 in the group that received aspirin, 2 in the group that received a placebo). Consequently, despite the nonsignificant finding, we would hesitate to conclude that aspirin has no effect on stroke mortality.

## The Role of Confidence Intervals in Meta-analysis

Recently, there has been a call for a more rigorous and systematic approach to the review of the existing research literature in a given area.[24] In place of the narrative literature review, the systematic review should be the approach of choice.[25] The process of meta-analysis also plays a key role in systematic reviewing.[26–28] *Meta-analysis* has been defined as "the statistical analysis of results from a large number of individual research studies so as to integrate their findings."[29(p390)] The strengths of CIs become particularly evident in meta-analysis.[30]

In order for a meta-analysis to be performed, a common measure of effect size must be extracted from, or retrospectively calculated for, each study included in the analysis. The odds ratio is an appropriate measure of effect size for studies that examine the relative incidence of a dichotomous outcome, as opposed to differences in an outcome variable measured on a continuous scale.[31] The odds ratio is the ratio of the odds (likelihood) of achieving a certain outcome under one treatment condition to the odds of achieving that outcome under another treatment condition, with an odds ratio of 1.0 denoting no difference. To illustrate, in a systematic review of randomized, controlled trials of intensive versus conventional therapy for stroke, Langhorne et al[32] found an overall odds ratio for death or deterioration of 0.54. This odds ratio means that the likelihood of death or deterioration during intensive therapy is 54% that of conventional therapy.

Studies will vary in the contributions they make to the total odds ratio, based on sample sizes and other factors that may or may not control random error. Displaying the CI for each study on what is known as a forest plot illustrates clearly the relative merits of the separate studies. Those studies that are based on larger samples have correspondingly narrower CIs, and the CI for the total odds ratio is the narrowest, as this CI is based on the aggregated sample.

Figure 3 shows a forest plot for hypothetical studies of biofeedback versus control treatment for habitual shoulder dislocation, in which the dichotomous outcome measure was whether a recurrence of dislocation occurred within the 8 weeks following treatment. The figure also shows the total odds ratio, which is calculated by statistical analysis of the aggregated data from the individual studies (odds ratios can be calculated retrospectively for studies that used other outcome measures, such as risk ratios). A ratio of less than 1.0 indicates that biofeedback is associated with a lower likelihood of recurrence than the control treatment. Note that, although the odds ratio from study "c" approximates the

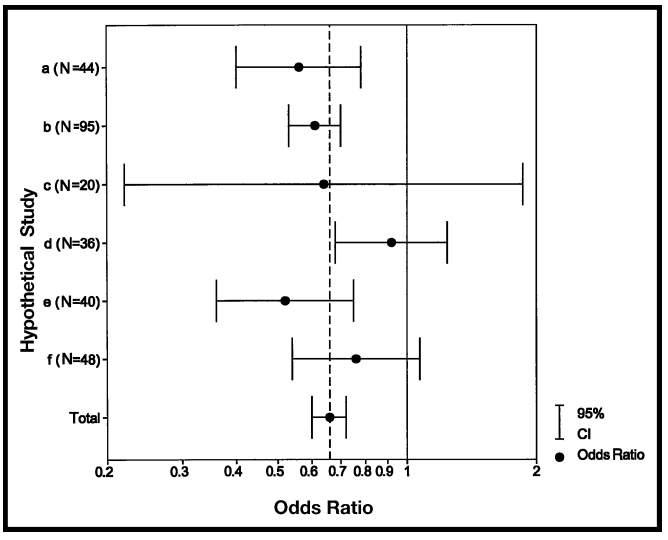


**Figure 3.**
A forest plot of 6 hypothetical studies (a-f) of biofeedback for habitual shoulder dislocation, showing odds ratios for the individual studies and the total odds ratio for the aggregated data. The solid reference line denotes an odds ratio of 1.0 (no difference), and the broken reference line indicates the total odds ratio. The horizontal scale is logarithmic.

total odds ratio very closely, the width of the associated 95% CI shows that it would be very difficult to draw a meaningful inference from the results of this study alone. The narrow width of the CI for the total odds ratio reflects the precision that results from aggregating data, and the fact that it excludes 1.0 indicates that the ratio is statistically significant at the $P<.05$ level. Thus, CIs assist considerably in the interpretation of the results of meta-analytic studies.

## Conclusion

Interval estimation is a valuable form of statistical inference that we believe has certain advantages over conventional hypothesis testing based on tests of significance. We contend that CIs also lend themselves readily to graphic portrayal and, therefore, are a useful means of "eyeballing" relationships in a set of data.[33] Based on our review of the physical therapy literature, however, we believe that CIs are underutilized. We, therefore, make the following recommendations:

- A CI should be included whenever a sample statistic such as a mean (or difference in means) is presented as an estimate of the corresponding population parameter (the standard deviation of the mean should be presented if no inference to the population is intended).
- Confidence intervals should be provided in addition to (or even instead of) the results of hypothesis tests, with the level of confidence for the CI matched to the level of statistical significance for the hypothesis test (eg, 95% CI for $P<.05$, 99% CI for $P<.01$).

- Advantage should be taken of the information provided by CIs to assess the clinical importance of study findings.
- If multiple CIs are calculated, the level of confidence should be adjusted to maintain the desired risk of a Type 1 error.
- When systematic reviews are conducted, CIs from individual studies should be reported (or calculated if missing from the original study), and these CIs should be displayed in any meta-analysis performed.

## References


**1** Oakes M. *Statistical Inference: A Commentary for the Social and Behavioural Sciences.* Chichester, England: John Wiley & Sons Ltd; 1986.

**2** Gardner MJ, Altman DG. Estimation rather than hypothesis testing: confidence intervals rather than *p* values. In: Gardner MJ, Altman DG, eds. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London, England: British Medical Association; 1989:6–19.

**3** Chow SL. *Statistical Significance: Rationale, Validity, and Utility.* London, England: Sage Publications Ltd; 1996.

**4** Rothman K. A show of confidence. *N Engl J Med.* 1978;299: 1362–1363.

**5** Langman MJS. Towards estimation and confidence intervals. *BMJ.* 1986;292:716.

**6** Bulpitt CJ. Confidence intervals. *Lancet.* 1987;1:494–497.

**7** Gardner MJ, Altman DG, eds. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London, England: British Medical Association; 1989.

**8** Ottenbacher KJ. Why rehabilitation research does not work (as well as we think it should). *Arch Phys Med Rehabil.* 1995;76:123–129.

**9** Bryant TN, Machin D. Statistical methods. In: Wilson BA, McLellan DL, eds. *Rehabilitation Studies Handbook.* Cambridge, England: Cambridge University Press; 1997:189–204.

**10** Bland M. *An Introduction to Medical Statistics.* 2nd ed. Oxford, England: Oxford University Press; 1995.

**11** Abrams KR, Scragg AM. Quantitative methods in nursing research. *J Adv Nurs.* 1996;23:1008–1015.

**12** Huck SW, Cormier WH. *Reading Statistics and Research.* 2nd ed. New York, NY: Harper Collins; 1996.

**13** Howell DC. *Statistical Methods for Psychology.* 4th ed. Belmont, Calif: Duxbury Press; 1997.

**14** Hurlburt RT. *Comprehending Behavioral Statistics.* Pacific Grove, Calif: Brooks/Cole Publishing; 1994.

**15** Gore S. Statistics in question: assessing methods—confidence intervals. *BMJ.* 1981;283:660–662.

**16** Brennan P, Croft P. Interpreting the results of observational research: chance is not such a fine thing. *BMJ.* 1994;309:727–730.

**17** Neave HR, Worthington PL. *Distribution-Free Tests.* London, England: Routledge & Kegan Paul Ltd; 1988.

**18** Campbell MJ, Gardner MJ. Calculating confidence intervals for some non-parametric analyses. In: Gardner MJ, Altman DG, eds. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London, England: British Medical Association; 1989:71–79.

**19** Cliff N. *Ordinal Methods for Behavioral Data Analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc; 1996.

**20** Colquhoun D. *Lectures on Biostatistics: An Introduction to Statistics With Applications in Biology and Medicine.* Oxford, England: Clarendon Press; 1971.

**21** Armitage P, Berry G. *Statistical Methods in Medical Research.* 3rd ed. Oxford, England: Blackwell Scientific Publications Ltd; 1994.

**22** Hennekens CH, Eberlein KA. A randomized controlled trial of aspirin and β-carotene among US physicians. *Prev Med.* 1985;14: 165–168.

**23** Steering Committee of the Physicians' Health Study Research Group. Preliminary report: findings from the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med.* 1988;318:262–264.

**24** Gray JAM. *Evidence-Based Healthcare: How to Make Health Policy and Management Decisions.* New York, NY: Churchill Livingstone Inc; 1997.

**25** Mulrow CD. The medical review article: state of the science. *Ann Intern Med.* 1987;106:485–488.

**26** Light RJ. Accumulating evidence from independent studies: what we can win and what we can lose. *Stat Med.* 1987;6:221–231.

**27** Mulrow CD. Rationale for systematic reviews. In: Chalmers I, Altman DG, eds. *Systematic Reviews.* London, England: British Medical Association; 1995:1–8.

**28** Egger M, Smith GD. Meta-analysis: potentials and promise. *BMJ.* 1997;315:1371–1374.

**29** Wood P. Meta-analysis. In: Breakwell GM, Hammond S, Fife-Schaw C, eds. *Research Methods in Psychology.* London, England: Sage Publications Ltd; 1995:386–399.

**30** Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ.* 1997;315:1533–1537.

**31** Farmer R, Miller D. *Lecture Notes on Epidemiology and Public Health Medicine.* Oxford, England: Blackwell Scientific Publications Ltd; 1991.

**32** Langhorne P, Wagenaar R, Partridge C. Physiotherapy after stroke: more is better? *Physiother Res Int.* 1996;1:75–88.

**33** Browne RH. On visual assessment of the significance of a mean difference. *Biometrics.* 1979;35:657–665.