0363-5465/98/2626-0145$02.00/0
THE AMERICAN JOURNAL OF SPORTS MEDICINE, Vol. 26, No. 1

Current Concepts

A Statistics Primer

Confidence Intervals

Mary Lou V. H. Greenfield,* MPH, MS, John E. Kuhn, MS, MD, and Edward M. Wojtys, MD

From MedSport and the Section of Orthopaedic Surgery, University of Michigan, Ann Arbor, Michigan

The basis for scientific study is the search for truth. Hypothesis testing is a common way of providing evidence to support this search for truth. As discussed in a previous article,[12] study findings such as means or frequencies are associated with critical values from statistical tests along with associated probabilities or *P* values. Recall that tests for statistical significance (such as chi-square, Student's *t*, analysis of variance or ANOVA) are chosen during the design of a study and, regardless of the statistical test chosen, good study design requires that all tests of significance lead to a probability statement or *P* value. "The *P* value provides the reader of the study with a guide for the likelihood that the statistical observation in a study is due to chance alone."[8] By convention, *P* values less than 0.05 are accepted as statistically significant.

This convention of statistical significance set at the 0.05 level, however, presents challenges to the reader in interpreting study findings. First, statistical significance ($P \leq 0.05$) may be present but may not be clinically relevant. Consider a previously presented example of the effects of two treatments on the outcome of length of hospital stay.[8] Suppose the *P* value associated with the study findings is 0.006—a highly statistically significant finding. However, the clinical result is less impressive because the statistical estimate associated with Treatment A results in a length of stay of 6.2 days and the estimate of Treatment B results in a length of stay of 6.8 days. These differences between results are meaningless. Even very small differences such as a few parts per million could be statistically significant if large numbers of subjects compose the sample. However, there may be no practical importance to such a finding.[19]

Conversely, a statistically nonsignificant result does not necessarily imply a clinically unimportant finding.[14] Suppose two different investigators at two different sites implement a study of the effect of a particular exercise program on ACL tears among collegiate athletes. Using $P < 0.05$ as the cutpoint for significance, both investigators should wrongly reject the null hypothesis of no difference 5% of the time. Now, suppose the statistical estimates at the end of both studies are identical: one exercise program (Program A) is associated with a 25% injury rate and the other exercise program (Program B) results in a 50% injury rate. Clearly, the results of these two studies indicate a clinically meaningful finding. However, one investigator has calculated a *P* value of 0.06 at her study site, and the other investigator has calculated a *P* value of 0.04 at her study site. Based on the conventional significance level of 0.05, one investigator will decide that the study finding is not statistically significant, and the other investigator will draw the opposite conclusion.

Recall that the *P* value is the representation of the probability that the investigator incorrectly rejects the null hypothesis. The findings of one investigator indicate that a result as extreme or more extreme than the one that actually occurred (25% versus 50%) could have happened by chance alone (assuming no true difference between the treatment groups) *4% of the time,* and the other investigator's findings indicate that a result as extreme or more extreme than the one that actually occurred could have happened by chance alone *6% of the time.* The first investigator would reject the null hypothesis, and the second investigator would fail to reject the null hypothesis based on a *P* value of 0.05 or less as statistically significant. We do not know what the "truth" is here, but consistent results support that there is a "real" difference. Even with the same clinical finding, in one study the finding is not statistically significant and in the other study it is. How could this happen? It may be that the first

* Address correspondence and reprint requests to Mary Lou V. H. Greenfield, MPH, MS, University of Michigan, Orthopaedic Surgery, TC2914G-0328, 1500 East Medical Center Drive, Ann Arbor, MI 48109.

No author or related institution has received any financial benefit from research in this study.

researcher had a smaller sample size, or more random error, or more variation in her data.

Luus et al.[14] note in a similar example "that the interpretation of the results has now been shifted away from the actual difference to that of probabilities." The cutpoint or significance level of 0.05 is arbitrary, and may ignore important, clinically meaningful findings. Many investigators and readers will disregard potentially useful clinical findings without further analysis simply because study findings are based on P values greater than 0.05. This is because the reader frequently interprets the P value to reflect clinical significance when in fact it only represents statistical significance.[2] The problem here is that we are in danger of throwing out good work based on arbitrary cutpoints set as a standard of statistical significance.[15]

In addition to throwing out good work based on P values, another danger with hypothesis testing is misinterpretation of findings associated with P values. It is not unusual that we see a P value reported that is nonsignificant ($P > 0.05$). Freiman et al.,[5] in a review of 71 studies, demonstrated that, in most cases, the absence of significance in studies has been interpreted by investigators as meaning that a treatment was not effective. A P value that is greater than 0.05 simply means that there is a lack of evidence to reject the null hypothesis. Saying that there is no evidence that two treatments are different does *not* mean that the investigators have established that the two treatments are the same, or equivalent. Put another way, "No proof of a difference is not equivalent to proof of no difference."[6]

Too much emphasis on P values may lead to a lack of critical thinking about research findings and incorrect conclusions.[5,9,16] How can investigators and readers alike avoid inappropriately branding studies as "negative" or "positive," avoid misinterpretation of P values, avoid "throwing out good work," and avoid placing importance on clinically meaningless findings? Rather than solely using probabilities to determine whether results are medically important, many researchers suggest that clinicians complement good clinical judgment with information gained through the use of confidence intervals.

In most clinical studies the researcher is interested in estimating some value, e.g., a mean or proportion, in the study sample to make inferences about the population from which the sample was drawn. The true value for the mean or proportion in the population remains unknown. (Actually, if a value is from a population, it is, by definition, "true." Therefore, when we refer to the "population value" we are referring to the "true value.") For example, in the ACL study described earlier the investigators were interested in the proportion of collegiate athletes with ACL injuries after two different exercise programs. Even in a random sample where everyone in the population of interest has an equal chance of being assigned to either of the treatment groups, every sample that an investigator draws will be slightly different because humans vary greatly. Thus, when the investigators draw two random samples to study their exercise programs, they expect to get slightly different results because of the variability in the larger population. The study result is an estimate, and the uncertainty associated with this estimate can be determined by a confidence interval. Confidence intervals help the reader to avoid misinterpreting nonsignificant results ($P \geq 0.05$) because they demonstrate whether the study finding is consistent with clinically useful true effects.[1]

Where P values in hypothesis testing may lead the reader to reject or fail to reject a null hypothesis (because hypothesis testing is solely concerned with the presence or absence of effect), confidence intervals capture the point estimates within intervals, providing a context within which the reader can assess whether a result is strong or weak, definitive or not.[10] Typically, we want to know not just whether a treatment has an impact or not, but also *how much* impact.[2] A confidence interval provides a range of values that is likely to capture the population value. The range will be larger (wider) for studies with smaller samples, emphasizing unreliability. This is particularly important with negative studies ($P > 0.05$), because the confidence interval presents a range within which the true value may lie. A 95% confidence interval tells us that if this procedure were to be repeated 100 times, the interval that is generated would capture the population value 95 times. A P value cannot tell the reader anything about the magnitude of a difference between two treatments in a study, nor can the P value tell the reader about the direction of the difference between two treatments[7]; however, assuming appropriate study design and research methods, a confidence interval can. This makes confidence intervals particularly attractive to the clinician because the confidence interval contains more information than the P value.

Confidence intervals require some estimate of truth (mean or proportion), and the sampling variability (standard error). In addition, some level of assurance or confidence is specified. Readers are most often familiar with seeing 95% confidence intervals reported. For any given estimated mean or proportion, the 95% confidence interval is the range between two estimated values: from approximately *minus* 2 times (actually 1.96 for Normally distributed data) the standard error of the estimated statistic to approximately *plus* 2 times the standard error.[19] Thus, the estimate *minus* 1.96 × SE represents the lower bound of the confidence interval and the estimate *plus* 1.96 × SE represents the upper bound of the confidence interval.† Expressed as a formula,

$$95\%\ \text{CI} = (\text{estimated mean or proportion} - [1.96 \times \text{SE}])\ \text{to}\ (\text{estimated mean or proportion} + [1.96 \times \text{SE}])$$

How do we interpret this? The 95% confidence interval is *not* interpreted to mean that the probability is 95% that the interval calculated contains the population value, e.g., a mean or proportion. Because repeated samples of the same size drawn from the same population would each

† Likewise, levels of assurance associated with 90% and 99% would be expressed, respectively, as: a 90% confidence interval would extend from −1 times the standard error to approximately +1 times the standard error; a 99% confidence interval would extend from −3 times the standard error to approximately +3 times the standard error.

result in different sample means and standard deviations, confidence intervals constructed for each sample will be slightly different. A 95% confidence interval of a value tells us that, if repeated samples of a given size are drawn from the population, 95% of the interval estimates will include the population value. (Although multiple random samples are seldom drawn, the interpretation of the confidence interval always considers this large number of hypothetical samples, each of the same size.[3]) Thus, we are asserting that the population value is likely to fall within the interval we have established with 95% confidence. Put another way, the population value will be contained somewhere within the upper and lower bounds in 95 of 100 confidence intervals constructed from random samples drawn from the same population; 5 of 100 such intervals will not enclose the population value. Note that the actual value in the population remains unknown. It is important to emphasize that statements about confidence intervals are not valid with respect to any particular value such as a sample mean or sample proportion, and thus are not probability statements in classic statistical interpretation; i.e., confidence intervals are statements about belief in the statistical process and do not have probability implications.

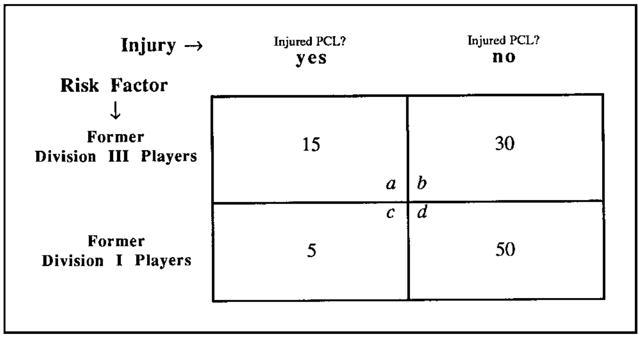
Several examples will serve as illustrations for how confidence intervals might be useful in clinical research findings. Scott K. Powers[17] studied the finish times of nine competitive runners in a 10 kilometer (10K) race. The mean time to finish the race was 33.79 minutes with a standard deviation of 0.518 minutes and a standard error of 0.173 minutes. The 95% confidence interval was (33.45,34.13). Assuming that this is a random sample of the running times of all competitive runners' 10K races, this confidence interval can be interpreted as follows: if we were to draw 100 random samples of 9 runners in a 10K race, and construct a 95% confidence interval from each sample, we expect 95 of these confidence intervals to contain the mean *population* time for a 10K race. We expect that 5 of 100 intervals constructed in this manner would completely miss the *population* mean (perhaps by a lot, perhaps by a little). The sample mean, the standard deviation, the standard error, and hence the confidence interval generated for each of these studies would differ. We would not know if any one of these confidence intervals (e.g., the one from the study we actually did) captured the population mean or not. We do know, though, that the process of calculating a confidence interval from such a sample will result in an interval that captures the population mean 95% of the time. (Notice the narrow width of this particular confidence interval. It indicates that even with a sample as small as nine, the variability in running times is quite small among 10K competitive racers.)

Consider another example borrowed from Riegelman and Hirsch.[18] Suppose that study results show a mean ± standard error for cholesterol in a sample to be equal to 150 ± 15 mg/dl. The 95% confidence interval for this study is equal to 150 ± 30 dg/ml. We estimate that the true population mean for cholesterol would lie somewhere between 120 and 180 mg/dl. This 95% confidence interval may or may not capture the population mean, but we have information that our estimate is not very precise.

The last two examples were illustrations using *means.* What about *proportions?* Consider this hypothetical example: suppose an investigator is interested in the National Football League (NFL) draft picks and the incidence of injured posterior cruciate ligaments (PCLs) among players in Division I colleges versus those who played in Divsion III colleges (Fig. 1).

An odds ratio (discussed in a previous "Current Concepts" article[11]) can be calculated for this study. Among NFL players who played in Division III colleges, the odds of sustaining a PCL injury is 5 times greater than if a player was in a Division I college. The standard error associated with this odds ratio is 0.602. The 95% confidence interval is (3.82,6.18). The lower bound for the 95% confidence interval is 3.82 and the upper bound is 6.18. This indicates that this study's estimate for the odds of sustaining a PCL injury given that the player *did not* play Division I ball is 5 times that of the player who *did* play Division I college ball. We estimate that the population odds ratio could lie somewhere between 3.82 and 6.18. If the confidence interval had captured the value of 1.0 then we would interpret this as the odds of being injured after having played Division III college football are less than the odds for those who played Division I college football, indicating no difference in effect.

In another example, Kujala et al.[13] studied lumbar mobility and back pain in adolescent athletes and controls. They reported, among other findings, the risk of developing future low back pain given that a boy was in the lowest third of subjects with maximal lumbar flexion at baseline. The relative risk calculated was 2.5. This indicates that boys who were in the lowest third of maximal lumbar flexion had a 2.5 times greater risk of developing future low back pain compared with those boys in the highest third at baseline. In this study, the relative risk is the point estimate. The 95% confidence interval is (0.8,8.0). The confidence interval gives the reader much more information than the point estimate, 2.5, alone. We estimate that the true risk of developing future low back pains remains unknown in the population, but we estimate that it lies between 0.8 times and 8.0 times the risk for this group compared with the other group. Note that the lower



Injury →	Injured PCL? yes	Injured PCL? no
Risk Factor ↓ Former Division III Players	15 a	30 b
Former Division I Players	c 5	d 50

Figure 1. Hypothetical risk factors and PCL injuries among college football players.

bound (0.8) is less than 1.0. A relative risk of 1.0 indicates equivalent risk for both groups, and a relative risk less than 1.0 indicates a protective effect. Finally, the upper bound is an eightfold risk. The 95% confidence interval encloses a risk less than 1.0 and as high as 8.0, providing more information for the reader to use in drawing his or her own conclusion than the study's point estimate of 2.5 alone. Because confidence intervals provide so much more information in these examples than the point estimates alone, it is very important that readers look for confidence intervals for all studies reporting odds ratios, relative risks, as well as estimates of the population values, such as means and proportions.

So far we have not discussed the relationship between confidence intervals and hypothesis testing. An instructional course lecture by Ebramzadeh et al.[4] from the American Academy of Orthopaedic Surgeons provides an excellent example of why we should question study conclusions that are based solely on *P* values. Consider these authors' hypothetical example. An investigator studies the effect of a new drug on infectious disease using a prospective, randomized design. He compares the outcomes in two groups treated with Drug A (the experimental drug) and Drug B (the conventional drug). At the end of the study there is a 30% higher rate of recovery in the experimental group. However, the corresponding *P* value is 0.11. The investigator concludes from this *P* value that there is no statistical difference between the two drugs and he recommends the conventional Drug B be used in future patients with this infectious disease. He publishes his finding and recommendations, and other centers stop further use and study of Drug A based on his study.

Ebramzadeh et al. point out that the interpretation and decision-making process represented in this example is quite common and represents a serious misuse of statistics. That is, if the experimental drug can save lives, even without statistical significance ($P = 0.11$), it should not be so readily dismissed. A 95% confidence interval reveals that a 30% recovery rate is associated with limits of -15% to 75%. It is quite possible that the authors in this hypothetical example are rejecting the potential for large, important clinical improvement with the experimental drug. Instead, these authors should understand that their *P* value indicates that there is only a slightly greater than a 1 in 10 ($P = 0.11$) chance that the 30% difference in survival rate occurred by accidental selection (chance); thus, the experimental drug ought to be used while futher studies are conducted. The value in the confidence interval here is that it provides a range of values with which to make an assessment of clinical significance or importance. The *P* value gives an idea of the probability of the study finding (point estimate) being due to chance, but the study finding should not be rejected outright based on the *P* value alone.

In this discussion we have limited our examples of confidence intervals to means and proportions. Confidence intervals, however, can be estimated for many statistics such as medians, regression slopes, survival rates, hazard ratios, and so forth.

There are several caveats for interpreting confidence intervals. Confidence intervals do not control for errors in study design or poor or improper selection of subjects. If biases exist or if random sampling was not used, the estimation errors may actually be greater than the width of the confidence interval might lead us to believe.[3] Thus, confidence intervals represent the smallest estimate for the real error.

The width of the confidence interval reflects the precision of the estimate. In cases where the interval is wide, it may be that the sample size is small, or that the variability of the data is great, or both. Increasing the sample size frequently will narrow the width of the interval, resulting in greater precision of the point estimate under study. Studies in which confidence intervals are wide (in terms of clinical relevance) may be because these studies have insufficient sample size or imprecise point estimates.

Finally, it must be emphasized that the 95% confidence interval is not interpreted to mean that the probability is 95% that the interval calculated contains the population mean or proportion. Because repeated samples of the same size drawn from the same population would each result in different sample means and standard deviations, the confidence intervals constructed for each sample will be slightly different. Under repeated sampling from the same population, 95% of the confidence intervals constructed will include the real (but unknown) population value. Five percent of the confidence intervals will not.

In conclusion, readers evaluating clinical studies want to know whether the treatment under study has any effect or not, and what the size of this effect might be.[2] The *P* value provides the reader with the likelihood that point estimate is due to chance. Confidence intervals, on the other hand, give the reader an idea about the magnitude of the effect of a study, the direction of the effect, and the differences between treatments. This is considerably more information than that provided by the *P* value alone.

ACKNOWLEDGMENT

The authors thank Dr. M. A. Schork from the University of Michigan, School of Public Health, Department of Biostatistics, for his review of this manuscript.

REFERENCES

1. Altman DG, Gardner MJ: Confidence intervals for research findings. *Br J Obstet Gynaecol 99:* 90-91, 1992
2. Borenstein M: The case for confidence intervals in controlled clinical trials. *Control Clin Trials 15:* 411-428, 1994
3. Braitman LE: Confidence intervals extract clinically useful information from data. *Ann Intern Med 108:* 296-298, 1988
4. Ebramzadeh E, McKellop H, Dorey F, et al: Challenging the validity of conclusions based on P-values alone: A critique of contemporary clinical research design and methods. *Instr Course Lect 43:* 587-600, 1994
5. Freiman JA, Chalmers TC, Smith H Jr, et al: The importance of beta, the type II error and sample size in the design and interpretation of the randomized controlled trial: Survey of 71 "negative" trials. *N Engl J Med 299:* 690-694, 1978
6. Gallagher EJ: No proof of a difference is not equivalent to proof of no difference. *J Emerg Med 12:* 525-527, 1994
7. Gardner MJ, Altman DG: Confidence—and clinical importance—in research findings. *Br J Psychiatry 156:* 472-474, 1990

8. Greenfield ML, Kuhn JE, Wojtys EM: A statistics primer. *P* values: Probability and clinical signficance. *Am J Sports Med 24:* 863-865, 1996
9. Grimes DA: The case for confidence intervals. *Obstet Gynecol 80:* 865-866, 1992
10. Guyatte G, Jaeschke R, Heddle N, et al: Basic statistics for clinicians: 2. Interpreting study results: Confidence intervals. *Can Med Assoc J 152:* 169-173, 1995
11. Kuhn JE, Greenfield ML, Wojtys EM: A statistics primer. Prevalence, incidence, relative risks, and odds ratios: Some epidemiologic concepts in the sports medicine literature. *Am J Sports Med 25:* 414-416, 1997
12. Kuhn JE, Greenfield ML, Wojtys EM: A statistics primer. Hypothesis testing. *Am J Sports Med 25:* 702-703, 1996
13. Kujala UM, Taimela S, Oksanen A, et al: Lumbar mobility and low back pain during adolescence: A longitudinal three-year followup study in athletes and controls. *Am J Sports Med 25:* 363-368, 1997
14. Luus HG, Müller FO, Meyer BH: Statistical significance versus clinical relevance: Part II. The use and interpretation of confidence intervals. *S Afr Med J 76:* 626-629, 1989
15. Murray GD: Statistical aspects of research methodology. *Br J Surg 78:* 777-781, 1991
16. Poole C: Beyond the confidence interval. *Am J Public Health 77:* 195-199, 1987
17. Powers SK, Dodd S, Deason R, et al: Ventilatory threshold, running economy and distance running performance of trained athletes. *Res Q Exerc Sport 54:* 179-182, 1983
18. Riegelman RK, Hirsch RP: *Studying a Study and Testing a Test: How to Read the Health Science Literature.* Third edition. Boston, Little, Brown and Company, 1996, p 274
19. Selvin S, White MC: Research methods: Description and reporting of statistical methods. *Am J Infect Control 21:* 210-215, 1993