0363-5465/96/2424-0863$02.00/0
THE AMERICAN JOURNAL OF SPORTS MEDICINE, Vol. 24, No. 6

Current Concepts

A Statistics Primer

P Values: Probability and Clinical Significance

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

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

In considering the findings of any study, we must question the validity of the methods employed. Take as an example a study of 12-year-old boys evaluating the effect of heading soccer balls and a suspected decrease in cognitive function on standardized tests. The null hypothesis is that there is no difference in cognitive function and the number of times a ball is headed during a season. The research hypothesis is that there is a decrease in cognitive function among players who head the ball on average more than five times per game throughout the season. To test the hypothesis, a study is designed in which players take standardized cognitive function tests before and after the season. Results show that soccer players who headed the ball more than three times on average per game had lower postseason cognitive function scores than those boys who did not head the ball as much.

Is this finding real or are there alternative explanations? Is there *bias* in this study that could account for such a finding? Who is administering the cognitive function tests? Are these well-trained volunteers or are psychologists administering the tests? Under what circumstances is the testing taking place? (On the bench after the game or in a quiet classroom, are players exhausted, well-hydrated, well-fed? Are the circumstances under which the testing is conducted the same for each player in the entire study group?)

Are there any *confounding* factors that might affect the validity of the study conclusions? The researchers have only studied boys, not girls, limiting the generalizability of the study to one sex. Another potential confounder is that only younger players are included because of the age limit of 12 years. Also, there can be great differences in physical maturity among 12-year-old boys. How will this variability affect the study results? Have the investigators considered other explanations for decreased cognitive scoring at the end of the soccer season? Has head trauma, which may have occurred outside the actual soccer game, been tracked? What about the number of times heading has occurred during soccer practice? These questions raise issues that are important in considering the validity of a study finding. Most readers will raise these questions as they peruse journal articles, before making any changes in clinical practice based on the results of one study.

In addition to bias and confounding, the reader must also consider the role that chance plays in a study. Simply put, what is the probability, even when confounding and bias are well controlled in a study, that the findings might simply be due to chance? *Probability* is that branch of mathematics that attempts to quantify uncertainty and randomness. It is a concept we often deal with: Is it likely that the University of Michigan Ice Hockey Team will repeat as national champions? It is virtually certain that taxes must be raised to control the deficit. The odds favor an increase in interest rates next year.

Probability is used to model physical situations and the data are used to support these models. To demonstrate how probability is used in the interpretation of study findings, consider the following example. Suppose we want to see if a die is a fair die or not. We take the die and toss it 10 times and the number of dots that appear is recorded after each toss. (This represents a sample of 10 measurements drawn from the much larger body of tosses, the population, which we could generate if we had the time to sit around and toss this die all day.) Now, suppose that each of the 10 tosses resulted in only 1 dot showing on the upper face of the die. Remember that we want to make an inference concerning the population of tosses and whether the die is balanced. We would likely be somewhat suspicious, having observed 10 tosses resulting in 10 identical measurements and, based on the study, we would

* Address correspondence and reprint requests to Mary Lou V. H. Greenfield, MPH, 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.

likely reject the theory that the die is balanced. If the die is balanced (null hypothesis), then observing 10 identical measurements is improbable. Either we have observed an extremely rare event or the hypothesis is false. Most people would suspect that the die is not balanced and hence reject the null hypothesis as false.[3]

How does this carry over into the research study of soccer heading described earlier? Tests for statistical significance (such as chi-square, Student's *t*, ANOVA) are chosen during the design process of a study. Many investigators will consult with statisticians for the appropriate test to use. 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 is the probability that the null hypothesis is false and indicates the likelihood of obtaining a result by chance alone, assuming there really is no association. As discussed in an earlier article on hypothesis testing (September/October 1996, pages 702 to 703), the *P* value is also the exact *alpha level*, or the prespecified level at which the investigator is willing to risk making a *Type I error* when he or she rejects the null hypothesis.

Looking at the soccer heading study, the null hypothesis is that there is no association between soccer heading and cognitive function test scores. The study investigators suspect that there may be a relationship. Before conducting the study, the investigators specify what statistical tests they will use and how many heading exposures experienced by the boys will suggest a relationship between heading the ball and a decreased cognitive function score. Statistical tests allow the investigator to calculate the exact probability of observing a prespecified number of headers and a decreased cognitive function score. In this study the researchers must specify the average cognitive function score of a 12-year-old boy, the average number of headers per game, the number of headers that they *suspect* might result in a decreased cognitive function score, and how much of a decrease in cognitive function would be considered important.

If the researchers documented that the average number of headers during regular indoor season play is approximately three per player, and those boys who headed more than five times per game had a decrease in cognitive function testing, what is the likelihood that this was just a chance finding and not significant at all?

The *P* value provides the reader of the study with a guide for what the likelihood is that the statistical observation in a study is due to chance alone. In the example of soccer heading, let's suppose that statistical tests for significance have been conducted and the results indicate that there is a relationship between the number of times a soccer ball is headed and a decrease in cognitive function. The *P* value associated with the significance testing is 0.006. This means that the probability of finding a result such as this by chance or random occurrence if there really is no association between excessive heading and reduced cognition is 6 in 1000. The question now becomes, have we observed an extremely rare event or is the null hypothesis (no association between soccer heading and cognitive function) wrong? Most often, the researcher would conclude that the null hypothesis is incorrect and agree that an event that could occur 6 times in 1000 by chance *could* have occurred; but in this case the researcher is likely to reject the null hypothesis in favor of an association between heading and cognitive function. (Just like the die toss example, we reject that the hypothesis that the die is balanced when the study sample results shows that in 10 throws we get 10 identical numbers.)

By convention, *P* values of 0.05 are most often accepted as statistically significant. It must be understood that a *P* value of 0.05 means that if we were to reject the null hypothesis, we may be wrong 1 in 20 times due to chance alone! The decision to use a *P* value of 0.05 should be specified before the study begins and should reflect the investigator's "evaluation of the impact that could be caused if an incorrect conclusion were reported."[1]

Remember that a *P* value does not tell the whole story, but it often gets the most attention in a study. Other important factors must be considered in interpreting *P* values and judging whether findings are statistically significant. First, the word *significant* needs to be weighed cautiously in evaluating scientific papers. *Statistical significance* is not the same as *clinical significance*. Prudence needs to be exercised in both the exposition of study findings as well as in the reading and interpretation of these findings. For example, a study looked at the length of hospital stay for two groups of patients with different total knee arthroplasty devices. Using appropriate statistical tests, there was a highly significant difference ($P = 0.006$) between the two groups of patients regarding the average length of stay for each group after surgery. $P = 0.006$ would most likely be considered a highly *statistically* significant finding. However, the *clinical* finding is less impressive. The first group of patients' average hospital stay was 6.2 days and the second group of patients' average hospital stay was 6.8 days. Common sense indicates that there is no meaningful difference in terms of clinical practice in this *highly significant finding*. "The importance of clinical significance cannot be overstated and should be a driving force in planning clinical studies."[2] The same can be said for the interpretation of study results.

Second, sample size may affect statistical significance. Although this will be discussed in greater detail in a future article on sample size and power analysis, it is important to consider the number of subjects being studied. Generally speaking, the larger the sample size, the more reliable the conclusions are regarding the statistical tests for significance. That is, the role of chance can be decreased by increasing the sample size. (As with the role of the die, as skeptical as we are that the die is balanced with 10 tosses producing identical face values of 1, we are more convinced when 500 tosses produce the same face value on the die.) In the soccer heading study, it seems almost intuitive that findings from a group of 500 boys will be much more convincing than the same findings from a group of 5 boys. The one exception is when the event of interest is extremely rare. For example, the incidence of slipped capital femoral epiphysis is approximately 7 in 100,000. Even with a study group of 10,000, it is possible that no cases will be found because the condition is rare.

Third, when *multiple comparisons* are made in a study, conventional P values of 0.05 are *not* acceptable. The more comparisons performed within a study, the chance that a statistically significant result will be due to chance alone increases. In the study of soccer heading and cognitive function, the relationship between height and weight and scholastic records in addition to cognitive function and heading exposure could be examined; three new variables (height, weight, scholastic record) have been included introducing the possibility of finding something statistically significant just because more tests have been performed on more variables. As an illustration, consider the interpretation of blood chemistry panels done as a routine screen on an annual physical examination for a patient. Forty different tests results are reported and one, chloride level, is high. Is this an important finding? Is there anything in the care of this patient that should be changed as a result of this one finding when every other test result is normal in a healthy patient?

Because of the effects of chance alone when multiple comparisons are made, the P value should be adjusted downward. This should be discussed in the "Materials and Methods" section of a paper. (Some statistical software packages automatically adjust for multiple comparisons when doing tests involving multiple comparisons, but this should be specified in the paper.) In papers in which the P value has not been adjusted for multiple comparisons, and a multitude of different tests were employed, an easy rule of thumb for making a quick adjustment to an acceptable P value (and hence a significant finding) is to divide the conventional $P = 0.05$ by the total number of comparisons. For example, a study with seven comparisons would be 0.05 divided by 7 to give a P value 0.007. This would be the P value associated with statistical significance in this study.

Fourth, as already discussed, we must rule out that the results are due to bias or confounding. In addition, the *randomness* of the sample must be examined. If we are studying the relationship between soccer heading and cognitive function, and we only include 12-year-old boys, the study findings cannot be used to characterize professional soccer players, no matter how significant the findings. If we invite all 12-year-old boys in a particular league to participate in this study, are the results generalizable to other leagues? In addition, if we ask every 12-year-old boy in a particular league to participate in this study, but only half of the boys eligible for the study agree to participate, is there something different between the group that declined and the group that participated? Perhaps the boys who participated had suffered more head injuries and headaches and their parents wanted them to be tested. This would affect the study result. Response rates or participation rates should always be reported in studies because these rates may change both the meaning of study findings as well as the generalizability and applicability of the study results to other populations.

Fifth, *exact* P values should always be used in reporting study findings. Reporting a P value <0.05 is not as helpful to the reader as reporting the exact value of the finding. Reporting exact values allows the reader to evaluate the extent to which the data presented agree or disagree with the null hypothesis. "In particular, it enables each reader to choose his or her own personal value α and then decide whether or not the data lead to the rejection of the null hypothesis."[3] A P value <0.05 can be 0.049 or it can be 0.001. In either case, the reader should be given the exact information.

Sixth, the largest P value that will lead to the declaration of a statistically significant result should be specified by the investigators *before* conducting the experiment, along with the null hypothesis and the research hypothesis. The reader (as well as the investigator) should avoid drawing the conclusion that because there is an abundance of data, there must be something clinically significant.

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 article.

REFERENCES

1. Janssen HF: Experimental design and data evaluation in orthopaedic research. *J Orthop Res 4(4):* 504–509, 1986
2. Lindgren BR, Wielinski CL, Finkelstein SM, et al: Contrasting clinical and statistical significance within the research setting. *Pediatr Pulmonol 16:* 336–340, 1993
3. Mendenhall W, Beaver RJ: *Introduction to Probability and Statistics.* Boston, PWS-Kent Publishing Co, 1991