0363-5465/97/2525-0138$02.00/0
THE AMERICAN JOURNAL OF SPORTS MEDICINE, Vol. 25, No. 1


Current Concepts

# A Statistics Primer

## Power Analysis and Sample Size Determination

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*

Consider the following theoretical study. Two groups of female collegiate basketball players are being studied for the incidence of ACL rupture during a regular season of play. Participants are from several universities and all agree to be randomized to either a regular training program or a specially designed training program that emphasizes hamstrings strengthening exercises. Sixty women agree to participate and half are randomized to the experimental group and half are randomized to the control group.

The researchers hypothesize that women who participate in the training program that emphasizes specialized hamstrings strengthening exercises will sustain fewer ACL tears than those who are randomized to the regular training program. The null hypothesis is that there will be no difference in the incidence of ACL tears between the groups. From a review of the literature and from their own clinical experience, the investigators anticipate approximately 50% of the women with a regular training program will suffer an ACL tear. They hypothesize that a reduction by one-half, resulting in only a 25% incidence in ACL injuries in the hamstrings training group, would be a clinically significant finding. The investigators set their significance level at $P = 0.05$. At the end of their study there are fewer ACL injuries in the hamstrings strengthening group, but there is no statistical difference in the number of ACL tears between the two groups. Consequently, the investigators conclude that there is no useful advantage in specialized hamstrings strengthening exercises among female collegiate basketball players in preventing ACL injuries. Accordingly, they do not reject the null hypothesis.

Many would agree with the investigators' conclusion that there is no difference between specialized hamstrings training and regular training among these athletes. That is, while there is a difference between the two groups, it is not statistically significant. But is this really the case? Assuming that the study is well designed and carefully implemented, why would the reader have any reason to doubt or question this conclusion?

Most of us are familiar with setting limits for minimizing a Type I error by selecting a critical $P$ value.[3] To be more specific, in this study of the effect of hamstrings strengthening exercises and ACL tears, the investigators specified before their study that they were willing to accept a 1 in 20 chance ($P = 0.05$) that they would be wrong if they were to conclude that there was a statistically significant reduction in the incidence of ACL tears in the study group compared with the expected incidence of 50% incidence of ACL injuries in the control group. After conducting the statistical analysis, the $P$ value was determined to be 0.09. Because the investigators were unable to demonstrate statistical significance, they concluded that hamstrings training does not make a difference compared with regular training in these athletes. In other words, they fail to reject the null hypothesis of no difference between the two groups.

But what if they are wrong? What if there is a difference that was not captured in their study sample? If the investigators fail to reject the null hypothesis and the null hypothesis is false, this is a Type II error. How does the investigator set the conditions in a study to prevent a Type II error? What is the a priori probability, the power of the study design to detect a statistical difference between two study groups if a difference truly exists?

* 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.

*Statistical power* is the ability to avoid a Type II error.[1] This means that if a Type II error is made in the study just described, the investigators would conclude that hamstrings strengthening exercises make no difference in minimizing ACL tears in collegiate basketball players, when in fact the strengthening exercises make a big difference in preventing injury. A clinical example similar to this Type II error is concluding that a patient does not have Ewing's sarcoma when, in fact, he or she does. The seriousness of the error becomes apparent. In any study, we want to be sure that if we obtain a $P$ value greater than a prespecified value of 0.05, we are not incorrectly accepting a false null hypothesis.

Lieber,[3] in his article, "Statistical Significance and Statistical Power in Hypothesis Testing," discusses the situations in which the statistical power in the experimental design has not been considered. Unfortunately these are all situations many of us have observed. The following examples are taken from his article:

1. A study is conducted in which a small sample size ($N = 3$) is used, statistical analysis is performed, and a $P$ value is obtained that is greater than 0.05. The investigator concludes that the treatment has no effect. Many would protest that the sample size is not large enough to demonstrate a difference even if there is a difference.

2. In another study, the investigator performs an experiment with 10 patients per sample, and obtains a $P$ value of 0.07. The investigator is encouraged to add a few more patients to the sample to achieve statistical significance. In addition to being problematic from a sample size perspective, this example demonstrates bad science: Adding more study patients is inappropriate because it indicates that the study is being driven by the data and not by an hypothesis. In addition, every time the investigator takes a "look" at the data, the chance that a Type I error occurs increases. If an investigator is going to take preliminary "looks" at the data this must be specified before the study is begun and the $P$ value must be adjusted downward to account for it. It is never appropriate to add more patients after the data have been analyzed in the hopes of obtaining statistical significance.

3. A scientist performs an experiment with a small sample size comparing a "new" technique and a "standard" technique. Based on a high $P$ value ($P = 0.64$), the scientist concludes that there is no significant difference between the two treatments and therefore the *new* treatment should be used.

All of these examples have one thing in common: They fail to consider the statistical power in the design phase of their studies. Often a study finding of "no statistical difference" is related to low power rather than an actual lack of difference between study groups. It is imperative that calculating power and sample size for a study is accomplished during the design phase of the study. (While there are numerous formulas and techniques for calculating power and sample size, these formulas and techniques are beyond the scope of this paper.) The reader of medical literature should find a power analysis and sample size estimation in the "Materials and Methods" section of a paper or presentation. It is particularly important that the reader look for this discussion when the study results are negative, e.g., not statistically significant. It is also important to note that a finding of no statistical significance does not mean that the two groups are equivalent in their outcomes. No difference between groups in study outcomes does not mean that the groups or the treatments are equivalent.

It is generally accepted that the power of study should be at least 0.8 or 80%. This means that 80% of the time the investigator will be correct when accepting the conclusion that there is no difference between treatment groups. Put another way, the investigator is stating that when he or she concludes no statistical difference between two treatment groups with a power of 0.8, that 8 of 10 times he or she will be correct (and 2 of 10 times he or she will be incorrect!). Of course the power of a study will vary with the critical nature of committing a Type II error. For example, low power in a study of a new drug therapy for treatment of Ewing's sarcoma could mean the difference between life and death for future patients; i.e., if the power in such a study was set too low, the investigators may fail to detect an important difference between two treatments, one of which might be life-saving to patients, concluding the treatments were not significantly different. (For this reason, most pharmaceutical studies of drug therapies set the power of the study quite high at 0.95 or 0.99.)

Although there are many things that can affect power and sample size in a study, there are two that are important to emphasize here.

1. The power of a study increases with an increase in the difference the investigator is trying to detect. For example, in the hamstrings strengthening study discussed previously, if the investigators had conducted a power analysis during the study design phase they would have discovered that in order to detect a difference in ACL injuries of 25% in the study group and 50% in the control group, they would have needed 66 women in each group. Their conclusion (no difference between the outcomes) is suspect because their power is low. Put another way, they cannot be sure that there is no difference in outcome because they may not have had enough women in each group to detect a difference, if one really existed.

To carry this example further, if the investigators believed strongly that the exercise treatment was going to have a better effect, say a 50% incidence compared with a 15% incidence, only 33 women in each group would be required.

2. Another "power" truism is that as the sample size increases, the power increases. The more people the investigator has enrolled in the study, the more likely he or she is to capture a treatment difference, if one truly exists. In the ACL study, if the investigators had 1000 women who were interested in participating in the study, one might think that a power analysis is unnecessary because intuitively this seems like enough women to detect a difference in treatment outcomes. However, why expose more women than absolutely necessary to the study

protocol? If one treatment is potentially harmful, the investigators would want to minimize this effect by limiting enrollment to the lowest number possible. At the very least, the power analysis and sample size determination would save money because it enables the investigators to enroll the smallest number of subjects possible to demonstrate an outcome.

In conclusion, the reader should be looking for a power analysis and sample size determination in the "Materials and Methods" section of a paper or presentation, regardless of the study's conclusion. In addition, in the case of negative results (i.e., no statistically significant difference) there should be a discussion of this conclusion in terms of the power in the study in the "Discussion" section of the paper.

## 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. Hirsch RP, Riegelman RK: *Statistical First Aid: Interpretation of Health Care Data*. Boston, Blackwell Scientific Publications, 1992, p 54
2. Lachin JM: Introduction to sample size determination and power analysis for clinical trials. *Controlled Clin Trials 2(2):* 93–113, 1981
3. Lieber RL: Statistical significance and statistical power in hypothesis testing. *J Orthop Res 8:* 304–309, 1990