Statistical Significance and Clinical Relevance

The Importance of Power in Clinical Trials in Dermatology

Sachin S. Bhardwaj, MD; Fabian Camacho, MS; Amy Derrow, MS; Alan B. Fleischer, Jr, MD; Steven R. Feldman, MD, PhD

When evaluating the validity of a study, the reader must consider both the clinical and statistical significance of the findings. A study that claims clinical relevance may lack sufficient statistical significance to make a meaningful statement. Conversely, a study that shows a statistically significant difference in 2 treatment options may lack practicality. The concept of power of a clinical trial refers to the probability of detecting a difference between study groups when a true difference exists. We will discuss statistical power by examining studies too small to identify important differences, studies so large as to identify differences that are not clinically significant, difficult-to-design studies without very large patient populations, and those studies with both adequate power and clinically relevant findings. Dermatologists should not focus on small $P$ values alone to decide whether a treatment is clinically useful; it is essential to consider the magnitude of treatment differences and the power of the study.

Arch Dermatol. 2004;140:1520-1523

Statistical analysis in clinical research is used to show that the findings are not likely due to chance. However, it is easy to misinterpret the results of statistical tests. Often, the language of statistics obscures the findings of clinical trials. For example, a small study that claims clinical relevance may lack sufficient statistical power to justify its conclusions. Conversely, authors of a study may speak of the statistical significance of a treatment effect that has little, if any, clinical utility. Therefore, when evaluating the validity of a study presented in the dermatologic literature, the reader must consider both the clinical and statistical significance of the findings.

The literature already offers physicians descriptions of the different terms used in statistics.1-5 The purpose of the present article is to provide dermatologists with a conceptual understanding of 1 statistical concept essential to clinical research: power. The power of a statistical study is the probability of detecting a difference when one exists. Rather than further explaining power on a mathematical basis, we will examine the importance of power using examples from the dermatologic literature.

Understanding the direct relationship between sample size and power is critical to interpreting the conclusions that can be drawn from a study.¹ The failure to detect a clinically important difference between 2 groups can occur as the result of inadequate sample size; that is, inadequate power.³ This occurrence is more likely in studies involving rare events, but it can also be a hindrance to studies involving more common events. As the power of a statistical study increases, the study’s ability to detect progressively smaller differences increases. The concept of a particular study having too much power must also be considered. Studies with very large sample sizes may detect statistically significant differences that are clinically irrelevant.

We have arbitrarily selected from the literature examples of clinical studies that illustrate the importance of power in...
terpreting the clinical relevance of clinical trial results. Each example will be briefly described and followed by a discussion of the study’s power and the resultant effects on the author’s conclusions.

EXAMPLES FROM THE LITERATURE

Studies With Inadequate Power To Detect Meaningful Differences

In a 1989 article comparing the atrophogenic potential of mometasone furoate ointment and hydrocortisone ointment in the treatment of psoriasis, 51 patients with psoriasis vulgaris were treated simultaneously with each medication on opposite bilateral lesions (102 target treatment sites). Each patient underwent 6 weeks of treatment, and a 3 × 3-cm target area was examined and inspected for 6 signs of cutaneous atrophy. After 6 weeks of treatment, 2 of the 51 sites treated with mometasone and 1 of the 51 treated with hydrocortisone showed evidence of cutaneous atrophy. The authors concluded that the 2 therapies demonstrated comparable atrophogenic potential, while mometasone was more efficacious than hydrocortisone in the treatment of psoriasis. The authors further suggest that “a dissociation of potency from increased risks of side effects including dermal atrophy has been achieved with the mometasone molecule.”

Was the study large enough (that is, did it have enough power) to draw these conclusions? Mometasone treatment sites did improve from the baseline scores more than the sites treated with hydrocortisone (P < .001). After 1 week, mean improvement percentage in the mometasone-treated lesions was 45% compared with the 32% mean improvement percentage for hydrocortisone (P < .01); after 2 weeks, 3 weeks, and 4 weeks, the mean improvement percentages for mometasone were 54%, 55%, and 60%, respectively, while the values for hydrocortisone were 39%, 37%, and 38%, respectively (P < .01, P < .001, and P < .001, respectively). Atrophy was seen with hydrocortisone 2.0% of the time and with mometasone 4.0% of the time. If these numbers are assumed to represent the true atrophogenic potential of each agent, mometasone would have 2 times the atrophogenic potential of hydrocortisone.

However, to have a 50% probability to show that this difference was not due to chance (P < .05), the study would have required at least 580 subjects. As designed, using 51 subjects, the mometasone study had only a 2% probability of detecting a difference in atrophy, if one existed. This study did not show that mometasone and hydrocortisone have similar atrophogenic potential, nor did it show a dissociation of potency from safety. Indeed, the design of this trial was such that efficacy differences between the 2 treatments were detected, but differences in the rate of adverse events (which occur uncommonly), even ones that are clinically significant, were not apparent.

A recent study compared the efficacy of once-daily vs twice-daily application of betamethasone valerate in a foam vehicle (Luxiq; Connetics Corporation, Palo Alto, Calif) for the treatment of scalp psoriasis. The trial included 79 patients randomized to treatment either once daily or twice daily for 4 weeks. Patients were evaluated at 0 and 4 weeks by a blinded physician grader who graded the scalp for signs and symptoms of psoriasis. There was a statistically significant decrease in erythema and plaque thickness with both once-daily dosing and twice-daily dosing. The magnitude of this improvement demonstrated clinical relevance (although the lack of a placebo group might limit one’s confidence in the finding). The authors concluded: “Although both once-daily and twice-daily application showed significant improvements, the difference between them was not statistically significant.” This does not suggest that once-daily dosing was as effective as twice-daily dosing, only that this study did not show a difference in the 2 treatment schedules. While the sample size was adequate to determine that both treatments were efficacious, a larger sample size or longer duration of follow-up in this trial was needed to demonstrate a meaningful difference in efficacy between once- and twice-daily dosing.

Studies With Excessive Power That Detect Differences That Are Not Clinically Meaningful

In a study of treatment of herpes labialis with penciclovir cream, 2209 patients were enrolled in a double-blind, placebo-controlled trial to compare the safety and efficacy of topical penciclovir in the treatment of recurrent cold sores. The trial’s main outcome variable was lesion healing, although time to loss of lesion pain and time to cessation of viral shedding were also measured. There was a statistically significant decrease in healing time as well as a shorter time to loss of pain and viral shedding in penciclovir-treated patients than among patients who applied the vehicle control. Healing of lesions in the treatment group occurred in a median of 4.8 days vs 5.5 days in the placebo group. This result was statistically significant (P < .001). The authors also stated that pain (median duration, 3.5 vs 4.1 days; hazard ratio, 1.22; P < .001) and viral shedding (median duration, 3 days vs 3 days [sic]; hazard ratio, 1.35; P = .003) resolved significantly more quickly. The hazard ratios for pain and viral shedding for those patients who used penciclovir cream were greater than 1, which indicated a greater risk for developing the mentioned effects.

However, the results of this study, while statistically significant, lack much clinical relevance. The power of this clinical trial was adequate to detect a difference of as little as 15% between the efficacies of the active drug and the placebo. The observed differences of 12% to 15% in the efficacy of penciclovir and placebo were of this order. This study, by using a very large sample size, detected a difference so small that it is probably not of much clinical benefit to patients. This is an instance of a study using too high a power, thus allowing the detection of a very slight difference.

Studies such as these, while apparently well designed and executed, make it imperative that the reader determine what he or she considers to be of clinical value. The reader should not focus on small P values alone to make decisions about whether a treatment is clinically useful; it is essential to consider the magnitude of the observed differences between the 2 treatment groups. The reader of such a study should also consider whether an
appropriate outcome measure was used. Looking at small differences in the time to clearing is not likely to be very relevant clinically. Another approach would be to choose a clinically relevant measure of success and compare the success rates between the drug and placebo groups.

The use of topical minoxidil for the treatment of early male pattern baldness is another example of a study with statistical significance but limited clinical utility. A total of 126 men with similar degrees of early male pattern baldness (no greater than a type VI male pattern alopecia classification) were treated with either 2% or 3% topical minoxidil or with vehicle for 4 months. Evaluation of efficacy was based on total hair counts as well as on patient's subjective overall cosmetic assessment. Results of the multivariate study indicated that there was a statistically significant greater increase in total hair count in the 3% topical minoxidil group than in the placebo group (P = .04). This study appears “overpowered” in that a statistically significant difference exists without a clinically meaningful degree of improvement. The hair count difference did not appear to be clinically significant in that there was no difference in subjective cosmetic assessment between the treatment groups.

IMPORTANT CLINICAL QUESTIONS THAT CANNOT BE ANSWERED IN SMALL, RANDOMIZED, CONTROLLED CLINICAL TRIALS

There are new diagnostic technologies constantly surfacing in dermatology, each hoping to deliver results superior to those of the current tools. However, as exciting as each new discovery may be, it is essential to temper our optimism by remembering that some of these new tests might never be adequately evaluated for statistical significance. The incidences of the disorders they seek to detect are sometimes so rare that it would be almost impossible to find a sample population large enough to attain statistically significant results.

For example, dermoscopy is reported to deliver more sensitivity and specificity than clinical examination alone in evaluating a patient for melanoma and the need to perform a biopsy. However, to show that dermoscopy is more sensitive than clinical examination in detecting melanomas is problematic. The sensitivity of a dermatologist in detecting a melanoma is very high. Assuming that dermatologists have a 96% sensitivity in identifying melanomas, and dermoscopy 99%, to achieve an 80% probability of showing a difference we would need about 335 subjects with melanoma. Assuming that 1 in 10 persons with suspect nevi canvassed for study enrollment actually had melanoma, more than 3000 subjects would need to be canvassed. Such a study size would be difficult (though not impossible) to achieve, and obviously even more subjects would be needed if fewer than 1 in 10 had a true melanoma. While there may be good reasons to use dermoscopy, careful clinical examination and a low threshold for biopsy are already very good screening tools for melanoma, and it will be quite difficult to show that dermoscopy is better.

Similarly, consider the issue of collagen propeptide blood tests in the detection of liver disease in patients treated with methotrexate. The standard of care in dermatology has been to recommend a liver biopsy after cumulative consumption of about 1 to 1.5 g of methotrexate. Compared with biopsies, propeptide blood tests are safer and noninvasive; however, designing a study to demonstrate equal or better sensitivity presents a logistical problem. The complication of cirrhosis in patients taking methotrexate is a relatively rare occurrence. One would have to prospectively observe a large number of patients to have enough to compare the 2 tests. If we assume that 5% of methotrexate-treated patients develop cirrhosis, that liver biopsy is 90% sensitive, and that collagen propeptide is 95% sensitive, we would need about 310 patients with cirrhosis, or approximately 6000 patients undergoing treatment with methotrexate to have approximately an 80% chance of finding a difference between the 2 tests. Researchers would face difficult logistic and financial requirements to gather enough study participants within a practical amount of time for such a study. Proposals to replace liver biopsy with blood propeptide levels as a means to monitor hepatic toxic effects should be viewed with appropriate caution.

A STUDY WITH ADEQUATE POWER TO DETECT CLINICALLY SIGNIFICANT DIFFERENCES

A 1998 article about the use of tacrolimus ointment for the treatment of atopic dermatitis in children is an example of a trial that is both statistically significant and clinically relevant. The goal of the study was to determine the safety and efficacy of tacrolimus ointment in pediatric patients with moderate to severe atopic dermatitis. Children were treated with 1 of 3 concentrations of tacrolimus ointment (0.03%, 0.1%, or 0.3%) or with vehicle twice daily for up to 22 days. The mean percentage improvement for each of the 3 treatment groups (72%, 77%, and 83%, respectively) was significantly greater than that of the vehicle group (26%), and no serious systemic adverse reactions were noted. The median percentage reduction in pruritus was also significantly better in the treatment groups than in the vehicle group (74%, 89%, and 51%, respectively).

The statistical methods used in this trial were sound, with a sample size of at least 43 in each of the study’s 4 arms. The authors assumed an effective rate in the vehicle group and the lowest-concentration tacrolimus group to be 50% and 80%, respectively. Given this assumption, a sample size of 40 patients per group was necessary to have an 80% chance to detect a statistically significant difference. The marked differences in mean percentage improvement and in pruritus in the treatment group show clear benefit to the patient.

CONFIDENCE INTERVALS

Examination of confidence intervals (CIs) provides helpful information not provided by P values alone. For example, consider the following CIs for the ratio of efficacy of a drug and a placebo. The 95% CI 0.9 to 1.1 includes 1, indicating that no statistically significant difference was found between the drug and the placebo. A CI of 1.001–1.002, indicates that a statistically significant difference was found but that the magnitude of this...
difference was so small that it would not likely be clinically significant. A CI of 3 to 10 would indicate both a statistically and a clinically meaningful difference. Finally, a CI of 0.8 to 10 indicates that no statistically significant difference was observed but that the power of the study was not sufficient to rule out a rather large difference between drug and placebo. Misleading interpretations of study findings are common in the dermatology literature. Attention to CIs and study power is helpful for avoiding such misinterpretations.

**DETERMINING APPROPRIATE POWER**

When clinical trials are designed, a subject population of the appropriate size should be chosen. Often, this is based on preliminary studies that provide an estimate of the expected effect size. For example, if preliminary studies show a new psoriasis treatment to be successful in 50% of drug-treated patients and placebo to be successful in 10% of placebo-treated patients, then a sample of 18 subjects per group provides 80% power (an 80% chance of showing a statistically significant difference) to detect a difference with $P < .05$ (a difference large enough that it would occur by chance alone $<5\%$ of the time).

Notice that this hypothetical study is powered to detect success. It is not necessarily powered to detect statistically significant differences from other outcomes. While this design offers sufficient power for the efficacy outcome (successful treatment), it would not have the power to show statistically significant differences in adverse events that occur uncommonly. It would be a mistake to conclude, just because the study was sufficiently powered for efficacy, that we can draw strong conclusions about safety. Indeed, studies may claim to show a treatment is safe and effective, but such studies often have proven only efficacy.

**CONCLUSIONS**

The above examples serve to demonstrate the importance of thoroughly examining the methods as well as the results of clinical trials reported in the literature. An assessment of study power is essential in determining both the statistical significance and clinical relevance of any study and has serious implications for any conclusions that can be drawn. Consequences of an inappropriate sample size can be dangerous in either extreme. An excessively large sample may show statistical significance even when there is no clinical practicality; an incorrectly small sample will fail to demonstrate important clinically significant differences. When dermatologists evaluate studies reporting significant differences, they should ask whether these differences are both statistically and clinically meaningful.

**Accepted for Publication:** October 6, 2004.

**Correspondence:** Steven R. Feldman, MD, PhD, Department of Dermatology, Wake Forest University School of Medicine, Medical Center Blvd, Winston-Salem, NC 27157-1071 (sfeldman@wfubmc.edu).

**Acknowledgment:** We thank Margueritte Cox for her help with analyses and review of this article.

**REFERENCES**