CLINICAL ORTHOPAEDICS AND RELATED RESEARCH
Number 413, pp. 55–62

Part II. Statistical Issues in the Design of Orthopaedic Studies

Statistical Sampling and Hypothesis Testing in Orthopaedic Research

*Joseph Bernstein, MD**; Kevin McGuire, MD*; and Kevin B. Freedman, MD, MSCE**

The purpose of the current article was to review the process of hypothesis testing and statistical sampling and empower readers to critically appraise the literature. When the p value of a study lies above the alpha threshold, the results are said to be not statistically significant. It is possible, however, that real differences do exist, but the study was insufficiently powerful to detect them. In that case, the conclusion that two groups are equivalent is wrong. The probability of this mistake, the Type II error, is given by the beta statistic. The complement of beta, or 1-beta, representing the chance of avoiding a Type II error, is termed the statistical power of the study. We previously examined the statistical power and sample size in all of the studies published in 1997 in the American and British volumes of the Journal of Bone and Joint Surgery, and in Clinical Orthopaedics and Related Research. In the journals examined, only 3% of studies had adequate statistical power to detect a small effect size in this sample. In addition, a study examining only randomized control trials in these journals showed that none of 25 randomized control trials had adequate statistical power to detect a small effect size. However, beta, or power, is less well understood. Because of this, researchers and readers should be aware of the need to address issues of statistical power before a study begins and be cautious of studies that conclude that no difference exists between groups.

Now more than ever, orthopaedic surgeons are asked to critically appraise the evidence on which their treatment decisions are based. This process of evaluating clinical and basic science evidence and incorporating it into practice is called evidence-based medicine. Although some may reasonably argue that evidence-based medicine is but a new name for an old process (that is, that traditional medicine is of course evidence-based), there is no doubt that financial pressures and constraints

From the *Department of Orthopaedic Surgery, and the **Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA.

Dr. Bernstein was supported by an OREF/Zimmer Career development award.

Reprint requests to Joseph Bernstein, MD, 424 Stemmler Hall, University of Pennsylvania, Philadelphia PA 19104–6081. Phone: 215–349-8834; E-mail: orthodoc@post.harvard.edu.

DOI: 10.1097/01.blo.0000079769.06654.8c

on resources have brought the issue of evidence to the fore. At the minimum, explicit consideration of evidence is required. Patients and payers recently have been emboldened to ask clinicians to justify their diagnostic and treatment plans.

In response to this trend toward explicit evidence based medicine, many physicians' organizations, including those in orthopaedic surgery, have attempted to provide guidelines to clinicians. The popularity of the American Academy of Orthopaedic Surgeons' instructional course lectures and its new journal speak to the need for critical reviews. Nonetheless, as the exponents of evidence-based medicine themselves assert, the best way to evaluate the medical evidence is by individual endeavor: broad guidelines cannot tell us as much as evaluations of evidence made for a distinct patient. This is because every patient presents with unique attributes, and the critical judgment of whether a given study in the literature applies to a given patient is best made by the doctor who knows the patient.

Because it is best that individual doctors evaluate the evidence for individual patients, new skills may be demanded of the former: doctors are going to be required to become facile with the critical review of research study; and, perhaps to their dismay, will be required to have some savvy with statistics.

Acquiring a savvy with statistics does not mean that orthopaedic surgeons need to become statisticians, of course. As Greenhalgh[1] correctly states, there is a distinction between knowing how to drive a car and knowing how to build one. In that context, one can say that orthopaedic surgeons will not need to know the intricacies of statistics to the point where they can supplant biostatisticians, but they do need to comprehend what these experts are doing.

The purpose of the current article, therefore, was to review the process of hypothesis testing and statistical sampling to empower readers to critically appraise the literature. Particular emphasis will be given to understanding the meaning of studies in which no significant difference between groups was found, the so-called negative study. This understanding hinges on mastery of one facet of the sampling and testing process: the distinction between disproving something (stating X is false) versus failing to establish that something is true (stating it is not proven that X is true). To understand this distinction, all clinicians must understand how the process of sampling works; and how this process is applied to the scientific method of hypothesis testing.

STATISTICAL SAMPLING AND HYPOTHESIS TESTING

Sampling and testing is a process of making inferences, bound intricately to theories of probability. Hypothesis testing based on sampling does not produce irrefutable facts, but only inferences made with certain degrees of uncertainty. Of course, if that degree of uncertainty is arbitrarily low, one can behave as if it were a fact. As the philosopher David Hume famously noted, our certainty that the sun will rise tomorrow is but an inference drawn from the fact that it has risen every day thus far. All people are comfortable in their day to day life with some elements of uncertainty, whether they know it by that name or not. Therefore, like the trend toward evidence based medicine itself, the novelty here is that tacit assumptions and unspoken ideas are made explicit, but no new assumptions are made.

The reason why uncertainty is introduced when hypotheses are tested on samples is that the process of sampling stands in contrast to the process of complete measuring. A sample is an incomplete measurement. One can have only a certain probability (less than 100%) that this sample observation is the same one that would be derived from a complete measurement.

Sampling Versus Measuring

Consider the following example. Let us say a scientist wishs to know whether the serum calcium in patients with fibrodysplasia ossificans progressiva was higher than in patients with progressive osseous heteroplasia. Fibrodys-

plasia ossificans progressiva and progressive osseous heteroplasia are very rare diseases. Accordingly, it is possible, although difficult perhaps, to find each and every person with these diseases and measure their serum calcium explicitly. This is the process of complete measurement.

It may be found, for instance, that the serum calcium in patients with fibrodysplasia ossificans progressiva is 8.2 mg/dL with a standard deviation of 0.9; and the serum calcium of patients with progressive osseous heteroplasia is 8.1 mg/dL with a standard deviation of 0.8. Therefore, if the question where asked, "is the mean serum calcium of patients with fibrodysplasia ossificans progressiva higher than the serum calcium in patients with progressive osseous heteroplasia?" The answer is yes. All patients with both conditions, as stipulated, were measured and a result was obtained. The standard deviation of this measurement in each group, for the purposes of this question, is irrelevant. Indeed, any other statistical parameter (median, mode, distribution) of these data for this question also is irrelevant.

The magnitude of the difference, at least as the question here was framed, is irrelevant. It may well be that for serum calcium, a clinical difference of 0.1 is trivial, but clinical relevance was not a feature of the question as framed.

The key point from this example is that if a given group was measured and assessed exhaustively, there is no need for statistical testing. Statistical testing is used to answer the general question, How confident can I be that an incomplete sample of a large group I measured represents the entire population of interest from which it was derived?

If the entire population of interest is measured, there is no need for statistical tests. However, most questions in clinical medicine involve populations that cannot be measured completely: only samples of these populations can be assessed. The question then becomes one of whether the sample measured adequately represents the underlying population of interest.

Furthermore, one must note that when it is asked whether the sample adequately represents the underlying population, this question is asked merely from a statistically standpoint. Keep in mind that there are other epidemiologic principles that need to be addressed (and not discussed in the current article). For example, there may be medical facts that help inform a clinician that a given population is not representative of a population. If one were to measure the weight of 10 male patients with a stroke and compare that with the weight of 10 healthy female patients and find that the weight of the healthy group was less, one should not generalize from this sample comparison that patients with strokes are heavier, even though the statistics show a difference. The reason that such an inference would be invalid is that a confounder, gender, was added. As such, the inference is invalid because an epidemiologic principle ("do not add confounders") has been violated, and not because of any statistical rule.

In the absence of additional information suggesting that a sample is not representative, the best metric to assess whether a sample is representative of the group is the size of the sample. Sample size is important because of the statistical law of large numbers. Simply stated, the law of large numbers asserts that statistical truths are more apparent as the number of observations is increased. Imagine you wanted to determine whether a given coin was balanced evenly. You know that a balanced coin when flipped should be heads approximately ½ the time. You then flip this coin three times and all three times heads were returned. Can you conclude then that this is an imbalanced coin? Perhaps not: even a truly balanced coin will return three heads in one of eight such trials. However, the likelihood of a truly balanced coin returning 1000 heads in 1000 flips is essentially nil. Should that result be returned, one would be considerably more justified concluding that this was indeed an imbalanced coin.

Hypothesis Testing

The fundamental method by which biologic facts are established is the controlled experiment. Strictly speaking, an experiment refers

to a situation in which the examiner manipulates the environment, applying one action to one group and a second action to another group. In clinical medicine, many inferences are made from retrospective observations, not experiments; the examiner was not the one who decided which action was applied to each group. Nonetheless, the process of inference from a controlled observation is similar to that of a controlled experiment.

In a controlled experiment, there are two or more groups, and after all interventions have been made and all data are collected, the question is asked, is there a difference between groups? It is important to note that the question is not asked, Is there no difference between the groups? Hypothesis testing specifically is asking, is there a difference between groups?

The scientific method proceeds as follows: The researcher offers a conjecture, such as prophylactic antibiotics decreases the rate of postoperative infections. He or she then offers a null hypothesis, a statement that is contrary to the conjecture and may be true if no differences are seen between experimental groups. The research then does his or her experiment. Based on the data of the experiment, he may be able to refute the null hypothesis, by showing that there are differences between groups. This, then, would support the original conjecture.

When the results of the study are considered in the context of the actual (unknown) but gold standard truth, one of four possible outcomes is possible (Fig 1). The first possible outcome to consider, shown in Cell 1, is when in the study the infection rates are different, and by the gold standard they truly are different. In this case, the null hypothesis is correctly refuted. The other outcome in which the correct inference is drawn is shown in Cell 4. This represents the outcome when the study shows that the rates of infection are not different, and in point of fact they are truly not different. Here, the null hypothesis is not refuted: the correct response.

Two incorrect inferences might be made; these are shown in Cells 2 and 3. Cell 2 represents the outcome in which the infection rates are found to be different, but by the gold standard truth criterion they actually are not different. Wrongly concluding that the groups are different when they are not truly different is called a Type I error. Cell 3 represents the outcome in which the infection rates are found to be not different, but in reality, the groups are different. Failing to detect true differences is called a Type II error.

Because one almost never knows the truth, all results have to be stated in terms of probability. Researchers reporting data should attach a label to their results that quantifies the chances of Type I and Type II errors. Standard statistical

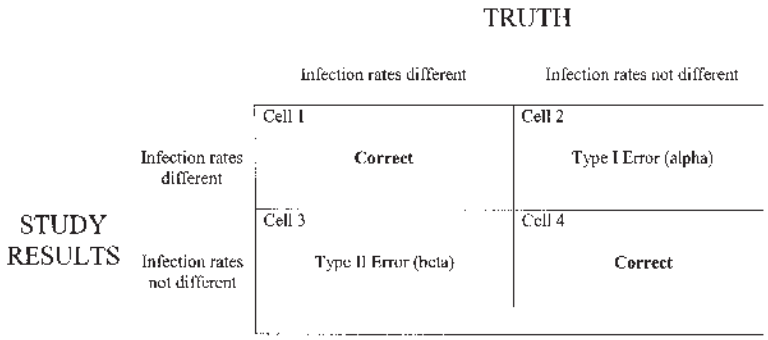

Fig.1. This chart shows the possible outcomes of hypothesis testing. In this hypothetical study, the rates of infection after administering prophylactic antibiotics versus administering no prophylactic antibiotics are compared.

tests can provide measurements for both, but traditionally greater emphasis has been given to limiting the likelihood of a Type I error. Indeed, most journals require reporting the p value, which quantifies the likelihood of Type I error. The more confusing situation arises when the researcher may fail to refute the null hypothesis. In that case, it may be that the original conjecture was wrong, and there really are no differences between groups. However, it also is possible that the groups are different, but that for some (perhaps inexplicable) reason, the researcher just failed to show the difference (i.e., refute the null hypothesis).

Measuring the risk of a Type II error requires a calculation of a different sort; one that rarely is done. Therefore, the reader must actively weed out this error on his or her own. This is where some grasp of statistics is helpful. They key concept to master is the difference between measuring and sampling.

Integrating Sampling and Hypothesis Testing

Consider another example: a researcher wants to determine whether the serum calcium of patients with osteoporosis is different from the serum calcium in patients with asymptomatic Paget's disease. Here, the diseases in question are extremely prevalent and it essentially would be impossible to measure all patients. Therefore, the research would proceed by sampling. The researcher could measure the serum calcium in 50 patients with each diagnosis. He or she may find that the serum calcium is 8.2 mg/dL in patients with osteoporosis and 8.1 mg/dL in patients with Paget's disease. Does this prove that the serum calcium is different in patients with these two diseases? That answer is found through statistical testing.

According to the scientific method outlined previously, the researcher sets up a null hypothesis. This null hypothesis is that the serum calcium in patients with osteoporosis is not different from the serum calcium in patients with Paget's disease. The researcher then attempts to reject this null hypothesis; that is, to show that there is a difference in serum calcium levels between the two groups. Importantly, the researcher does not set out to prove the null hypothesis. Because samples are used, one now enters the realm of probability. One cannot say that the null hypothesis is necessarily not false; rather, one must state that it is probable only to a given degree that the null hypothesis is true. When this given degree of probability is below a certain threshold, one deems the null to be false.

The criterion threshold of probability is the alpha level. That is, when p is below alpha, one deems the null hypothesis to be false, although that judgment may be wrong. In most clinical studies, alpha is set to 0.05. This alpha level allows a possibility of error (Type I error), but one accepts that risk because it is low. If p is less than alpha = 0.05, there is less than a 5% chance that a Type I error has occurred.

Because the researchers never really know the gold standard truth, the p value will report how likely it is that an error of inference was made. And as one might see, setting alpha too low, in hopes of minimizing Type I error, introduces a greater risk of Type II error: nothing is free. After all, if alpha is very low, there may be cases in which the null hypothesis is not rejected, although it would have been with a more liberal (higher) alpha. In that instance, the groups would have been deemed distinct and no Type II error would occur.

STATISTICAL TESTING

The probability that the null hypothesis is true, given the study results, is derived from statistical tests. In the case of continuous data such as serum calcium measurements in two groups of patients, a Students' t test is used. If the data were not continuous but binary, (dead versus alive, for instance) outcomes differences between groups are measured as ratios and a different test, the chi square test would be used. Other situations call for yet other tests. The decision of which test to use is beyond the current review but is discussed cogently elsewhere.[1]

When evaluating continuous data, the researcher enters each measurement from the

patients in the two groups into a mathematical formula to provide a t statistic. Commercial computer programs are available for this, but calculating the t statistic requires no higher math skills beyond multiplication and addition. This statistic then can be compared with published tables to derive a probability. This probability answers the question, What is the chance, given these data, that two samples from the same underlying population would present with different sample means as observed?"

Clinicians need not know the exact mechanics of that process but two features are worth considering. First, the probability of making an erroneous conclusion goes down with increasing the sample size. This, again, is simply the effect of the law of large numbers: 50 measurements have more truth telling power than five measurements, but less than 500 such measurements. Second, one notes that if that mean value of 8.2 were derived by 50 individual measurements that ranged from 8.1 to 8.3 (that is with low variability) one can be more certain that this 8.2 mean value represents the underlying population compared with the situation where the values ranged from 7.2 to 9.2 (higher variability).

If the statistical test returns a t statistic below the given threshold one will reject the null hypothesis and conclude, in this case, that the serum calcium does indeed differ in the two diseases. The interesting question, for the purposes of this review especially, is what it means when the p value is above the threshold. Does that mean it was proven that the two groups are not different; or does it mean only that the researcher simply failed to prove that they are different?

TYPE II ERROR

As noted, when the p value of a study lies above the alpha threshold, the results are said to be not statistically significant. Does one take it that when no significant difference is found that the two groups are therefore the same? This was not the initial question. The initial question was, is there a difference between groups? Our answer was no differences were found.

A study can produce results that are not statistically significant for two reasons. The simple reason is that real differences do not exist; the two samples emanate from the same underlying population. In that case, one would be correct to infer that the lack of a difference proves similarity. But there is an alternative possibility: that real differences do exist, but the study was insufficiently powerful to detect them. In that case, the conclusion that two groups are equivalent is wrong. The probability of this mistake, the Type II error, is given by the beta statistic. The complement of beta, or 1-beta, represents the chance of avoiding a Type II error, is termed the statistical power of the study.

In practical terms, the likelihood that the reader will be susceptible to a Type I error is low: the p value is almost always published. Because most readers know that a Type I error can occur if the p value is above 0.05, the likelihood of a Type I error in reporting study results usually is limited to 5% or less. A Type II error, however, is more apt to occur, because the issues of statistical power have not been publicized to the same extent. It also is more likely because many clinical trials simply do not have large enough samples.

LIMITING TYPE II ERROR

Type II error occurs only in those cases where the p value is above the alpha threshold, usually $p > 0.05$. Accordingly, anything that tends to decrease the p value (for a given underlying population) will decrease the risk of Type II error. This risk can be measured explicitly, however, with the beta statistic. Beta quantitates the risk of committing the Type II error in the same way alpha quantitates the risk of committing a Type I error. The generally accepted value for beta is that it should be less than 0.2; that is, there should be a less than 20% chance that there actually is a difference between two groups although the study does not show it. Because 0.2 is the generally ac-

cepted value for beta, the minimum power (1-beta) by convention is 80%. Yet the statistical power of a result rarely is reported and generally is less familiar to researchers and readers.

Because statistical power rarely is reported and sample sizes in studies frequently are inadequate, Type II error is likely, even in studies reported in prominent orthopaedic journals.[2–4] For example, the statistical power and sample size in all of the studies published in 1997 in the American and British volumes of the Journal of Bone and Joint Surgery, and Clinical Orthopaedics and Related Research were examined.[1,2] In those journals, only 3% of studies showed adequate statistical power to detect a small effect size in this sample.[2] In addition, a study examining only randomized control trials in these journals showed that none of 25 randomized control trials had adequate statistical power to detect a small effect size.[1] In each study examined, no significant difference was observed between groups; however, the groups were not proven to be the same, and the likelihood of a Type II error may have been unacceptably high.

Power is determined primarily by the sample size, but is influenced by two other important variables: the magnitude of the effect of interest and variability of the data. The variability of the data contributes to power in that if the data are bunched tightly about the mean, it takes fewer measurements to determine that mean. In fact, if the variation of the mean of a given group is 0, only one measurement is needed to determine that mean.

Beyond variability, the second parameter affecting power is the magnitude of effect of interest. The effect of interest represents difference between the groups that is clinically important. This is a parameter that is simply asserted, not derived mathematically. (The researcher might be called on to defend his or her chosen effect of interest, but this is an argument of clinical or aesthetic judgment, not calculation.) For example, a researcher may articulate that only those differences in serum calcium concentration greater than 0.5 mg/dL are clinically important and therefore worthy of detection. Therefore, if the study showed a difference smaller than this which was not significant statistically, the researcher would not be disappointed: after all, this difference was deemed a priori to be not clinically important. That is, the consequences of a Type II error are inconsequential, on clinical grounds.

The relation between magnitude of the effect of interest and the statistical power is a more subtle concept. In general, it requires a larger sample size (and with it, greater confidence that the calculated mean represents the true group mean) to detect smaller effects of interest. That is because small differences are more likely to be the product of random variation and not true differences.

The power of a study also could be augmented by raising the alpha threshold. Recall that power decreases the chance of a Type II error. A Type II error can be committed, of course, only when the results are found to be not statistically different. If alpha is raised, say to 0.10, with all else the same, it is of course more likely that the results will be statistically significant (by that new standard): all those trials which produce p values from 0.05 to 0.10 now are deemed significant, whereas they were not before that change in threshold. Needless to say, the alpha level of 0.05 is embedded deeply in convention and not really subject to change, but if one were willing to accept this relaxed threshold, the power is indeed higher.

Calculating power (or the sample size needed to attain a study of desired power) are tasks whose description are beyond the scope of the current review. In general, trials attempting to detect an effect of even moderate magnitude (assuming moderate variability and alpha of 0.05) require approximately 75 subjects per group. This is an extremely rough estimate, but it suggests that most trials that hope to show an effect smaller than that which would be obvious to the naked eye require a fairly large cohort. But increasing sample size indefinitely is not good either: it means that, at the least, resources were wasted; and it may mean that some patients were unnecessarily subjected to one treatment that was inferior.

Precise sample size calculations is therefore a requirement of all researchers. Many funding agencies will not consider funding a trial unless sample size calculation was done.

Readers should be alert to studies that do not calculate the beta or power of the study. In studies without such consideration, it is better to assume that no differences found means that the conclusions merely were unproven, rather than inferring that the groups are the same. This caveat is germane especially to those consulting older studies in the literature, written at a time when editors were less attuned to the issue of power and Type II error.

Many if not most studies in the orthopaedic literature are undersized and underpowered. Therefore, readers must take into account the possibility of a Type II error when reading studies in which there was no difference between two groups.

The goal of most studies is to show a difference between groups. Even an underpowered study may succeed at that, but if the study is underpowered and no differences were found, inferences cannot be made. If a study has as its goal to show a lack of difference between groups (similarity), strong consideration must be given to the necessary sample size, and the magnitude of effect of interest. If, to use the previous example, a 1% difference in infection for prophylactic antibiotics is considered important, than a study should be designed with adequate sample size to detect this difference of 1%. In that case, negative results (no statistical difference seen) would be helpful. The reader could be confident that had there been a difference of 1%, the study would have found it (at least n% of the time, where n is the power of the study). Although the groups are not proven to be the same with perfect certainty, an impossible goal, the reader is confident that the likelihood of a clinically important Type II error is low enough to avoid concern.

Statistics are used to give researchers and readers a probability assessment whether the groups examined represent the larger population of interest, within the arbitrary defined limits of probability defined by alpha and beta. Alpha is a generally well-known entity among clinicians. However, beta (or its complement, the power of the study), is less well understood. Because of this, researchers and readers should be aware of the need to address issues of statistical power before a study begins and be cautious of studies that conclude that no difference exists between groups.

References

1. Greenhalgh T: How to Read a Paper. London. BMJ Books 2001.
2. Freedman KB, Back S, Bernstein J: Sample size and statistical power in randomized control trials in orthopaedics. J Bone Joint Surg 83B:397–402, 2001.
3. Freedman KB, Bernstein J: Current concepts review: Sample size and statistical power in clinical orthopaedic research. J Bone Joint Surg 81A:1454–1460, 1999.
4. Lochner H, Bhandari M, Tornetta P: Type II error rates (beta errors) in randomized trials in orthopaedic trauma. J Bone Joint Surg 83A:1650–1655, 2001.