CHEST

Official publication of the American College of Chest Physicians

Hypothesis Testing, Study Power, and Sample Size

Bart J. Harvey and Thomas A. Lang

Chest 2010;138;734-737
DOI 10.1378/chest.10-0067

ISSN:0012-3692

CHEST Postgraduate Education Corner

MEDICAL WRITING TIP OF THE MONTH

Hypothesis Testing, Study Power, and Sample Size

Bart J. Harvey, MD, PhD; and Thomas A. Lang, MA

CHEST 2010; 138(3):734–737

Two earlier articles in this Medical Writing Tips series presented several guidelines for reporting important statistical information in scientific articles, including hypothesis testing.[1,2] In this article, we briefly review hypothesis testing to set the stage for discussing two additional statistical procedures used for planning and interpreting scientific studies: estimating sample size and calculating study power.

Hypothesis Testing: *P* Values, α, Type I Error, and CIs

Imagine that a randomized double-blind, placebo-controlled clinical trial, conducted to assess the effectiveness of supplementing standard treatment with a new chemotherapy to reduce lung cancer recurrence, provides the results shown in Table 1. These results indicate that the recurrence rate in the group receiving the new chemotherapy was one-half (30%: six of 20) the rate of the group receiving the placebo (60%: 12 of 20). Although these results appear promising, they can have one or more of three explanations. The first, of course, is that the results are real—that the new chemotherapy does, in fact, lead to a reduction in cancer recurrence. However, there is no way to directly measure or otherwise determine the degree to which this explanation accounts for the observed results. Instead, we must assess this explanation indirectly by addressing each of the two alternative explanations.

The first of these alternative explanations is that the study was flawed, either in its design or in its conduct. For example, if the allocation process was flawed, resulting in a disproportionate number of patients expected to experience recurrence being allocated to the placebo group, this "selection bias" could reasonably explain the apparent effectiveness of the new chemotherapy. To assess the degree to which study errors, biases, and confounding might account for the observed results, we would need to use critical appraisal techniques, such as those described in the Consolidated Standards of Reporting Trials Statement.[3] Such details would include, for example, the source of random numbers, how the allocation schedule was kept secret from those enrolling patients and those assigning patients to the study groups, and the success of blinding patients.

The second alternative explanation is chance—that the observed excess of recurrences in the placebo group is simply the result of a larger proportion of study patients who were ultimately destined to experience a recurrence being randomly allocated to that group. Even with the use of scientifically sound and rigorous randomization processes (ie, all patients have the same chance of being assigned to the treatment or comparison group), the distribution of patients who are ultimately destined to develop recurrence may, by chance, be imbalanced. The probability that such an unbalanced allocation may have occurred can be determined by statistical testing. More specifically, applicable statistical tests answer the question, "If there truly is no difference between the groups (ie, the so-called 'null hypothesis'), what is the probability of observing a result as extreme or more extreme as the one observed, simply by chance alone?"

In our example, if the chemotherapy truly has no effect, then we would expect nine patients in each group to have a recurrence because the 18 (ie, 6 + 12) patients with recurrence would probably be evenly

Manuscript received January 10, 2010; revision accepted March 26, 2010.

Affiliations: From the Dalla Lana School of Public Health, Department of Family and Community Medicine, and the Department of Surgery (Dr Harvey), University of Toronto, Toronto, ON, Canada; and Tom Lang Communications and Training (Mr Lang), Davis, CA.

Correspondence to: Bart J. Harvey, MD, PhD, Dalla Lana School of Public Health, University of Toronto, Room 688, 155 College St, Toronto, ON, M5T 3M7, Canada; e-mail: bart.harvey@utoronto.ca

DOI: 10.1378/chest.10-0067

Table 1—*Example Study Results*

Outcome	Chemotherapy	Placebo	Total
No recurrence	14	8	22
Recurrence	6	12	18
Total	20	20	40

divided between the two study groups. However, the observed results in our hypothetical study differ from this "expected" amount by three patients each. By completing the appropriate statistical test (in this case a χ^2 test) either by hand or by using suitable computer software (such as the web-based calculator at http://www.statpages.org/ctab2x2.html) this probability is found to be 0.112, or 11.2%. That is, assuming that the null hypothesis is true (that the new chemotherapy truly has no effect on the risk of recurrence), a difference as large as the one observed could have occurred simply by chance more than one time in 10 (ie, 11.2%).

It is important to note that *P* values such as this one should almost always be based on a two-tailed hypothesis test. Two-tailed tests are used when differences can potentially occur in either direction; for our example, that recurrence can be either higher or lower in the chemotherapy than in the placebo group. Two-tailed tests give the probability for a difference in either direction and so should be used when the direction of a difference is unknown,[1] whereas one-tailed tests should only be used in those infrequent instances when the direction of the result is known in advance. When one-tailed tests are used, however, they should be clearly identified as such and their use justified.[1]

So, how low does this probability (ie, *P* value) need to be for researchers to consider an alternative explanation for the observed results—that is, that the observed results reflect a true effect, or in the case of our example, that the new chemotherapy really does reduce the risk of experiencing a cancer recurrence? This cutoff value is called the "α level" and should always be selected by the researchers before the study. Whereas an α level of 0.05 (ie, < one chance in 20) is traditionally chosen, depending on the study, researchers may also select other α levels (eg, 0.01, 0.001, or even 0.1).

The α level depends largely on how comfortable researchers are with the balance between wrongly concluding that there is an effect and not detecting a real effect (discussed in the "Statistical Power, β, Type II Error, and CIs" section). For instance, in our example, the researchers might have selected a lower α level (eg, 0.01) so that they could further reduce the risk of wrongly concluding that the chemotherapy is effective, particularly in light of its potential side effects, costs, and limited availability. However, even if the *P* value is as low as 0.01, chance might still be responsible for the observed results (in this case, it is a low probability of one chance in 100). As such, the probability of chance being responsible for observed results never goes to zero because it is always possible for chance to be responsible for the observed results. However, the key question is: How probable is it that chance is responsible for the observed results? Wrongly concluding that there is a difference (or effect) when one truly does not exist is called a "Type I" error—with the *P* value indicating the probability of this error occurring (ie, that the observed results were not the result of a real effect but actually occurred by chance).

The most common statistical reporting error in the literature is confusing statistical significance—a small *P* value—with clinical importance. One characteristic of hypothesis testing is that small and biologically trivial differences can be statistically significant if there are a large number of subjects, and that biologically important differences can be missed if the number of subjects is small. So, although *P* values should be incorporated into the interpretation of study results, biologic plausibility and the clinical importance of the observed result should also be considered.

As such, researchers are advised (and even required by some publications) to calculate and report a CI to indicate the precision of the estimated effect size. A CI is the range of values within which the real difference would be expected to occur with a specified probability (ie, 95% of the time for a 95% CI). By completing the applicable statistical procedure by hand or by using suitable computer software (such as the calculator at http://department.obg.cuhk.edu.hk/researchsupport/Independent_2x2_table.asp), the 95% CI in our example can be determined to be 0.23 to 1.07, indicating that the "true" difference will occur in this range in 95 of 100 similar studies (ie, with a probability of 95%).

When the *P* Value Is Not Statistically Significant

In our example, the *P* value was sufficiently large (ie, 0.112, or 11.2%) that we would accept chance as a reasonable explanation for the observed results. So, should we now conclude that the new chemotherapy is ineffective at reducing the risk of cancer recurrence? As a first step, let us reconsider the apparent effect of the new chemotherapy. As we noted previously, the results of the study indicate that receiving the chemotherapy reduced the rate of recurrence by half (30% vs 60%). So, although the observed results were not statistically significant (ie, the *P* value was not less than the chosen α level), they do suggest a clinically important effect.

Analogous to the previous discussion, an examination of this statistically nonsignificant result requires

the consideration of three possible explanations. The first, of course, is that the observed results are real—that the new chemotherapy is, in fact, not effective at reducing the risk of lung cancer recurrence. Again, however, there is no way to directly measure or otherwise determine whether this explains the statistically nonsignificant result. Instead, this determination is made indirectly by assessing each of two alternative explanations.

The first of these alternative explanations is that the study was flawed, resulting in a real difference between the groups not being detected. For example, if a larger number of persons in the chemotherapy group were lost to follow-up, reducing the number of cancer recurrences being identified in this group, this "measurement bias" could reasonably account for the apparent ineffectiveness of the new chemotherapy. Again, to assess the degree to which study errors, biases, and confounding might account for the apparent lack of effect, we would need to know and critically assess the specific details about the research design and activities, such as those described in the Consolidated Standards of Reporting Trials Statement.[3] In this case, a "best-case/worst-case analysis" (ie, sensitivity analysis)[4] could be done to estimate the potential impact the differences in the rate of patients lost to follow-up might have had on the observed results.

The second alternative explanation is once again chance. But instead of asking, as we did previously, what the probability is that the observed results arose by chance, we ask what the probability is that the study failed to detect a result of a given size if such an effect truly existed. This probability is called Type II or β error and is related to "study power," which is the probability of a study detecting a difference of a given size, if one truly exists. Study power is equal to the complement of β error, $1 - \beta$.

Statistical Power, β, Type II Error, and CIs

To determine the statistical power of our hypothetical study (ie, the probability that it would detect an effect of a certain size), we must identify and specify three characteristics of the study: (1) the minimum difference between the groups that would be considered clinically important to detect, (2) the sample size of each study group, and (3) the α level used to declare a statistically significant result (ie, to conclude there is a difference). With this information, we can calculate study power (using a tool such as the web-based calculator for comparing two proportions available at http://www.stat.uiowa.edu/~rlenth/Power/index.html).

This calculation gives a result of 35%, meaning that the power of our hypothetical study to detect a 50% reduction in recurrence using an α level of 0.05 and a sample of 40 was about one in three. So, what does statistical power of a study tell us? First, let us consider how high this probability would need to be for us to conclude that the study had sufficient statistical power to detect a reasonable or desirable difference if one truly existed. Traditionally, researchers consider a study power of 80% to be the minimum, although higher values are commonly used (eg, 90% or 95%). As such, this study's statistical power of 35% helps to explain why this observed twofold difference in recurrence rates was not statistically "detected." Not enough patients were studied to have a reasonable chance of detecting it. Similarly, the wide and heterogeneous (ie, imprecise) and includes the null value of 1.0 CI presented earlier (ie, 0.23-1.07) also clearly indicates that the study lacked sufficient statistical power. So, how could such a situation have been avoided?

Estimating the Sample Size for a Study

To avoid conducting a study only to find that it has insufficient power, researchers are strongly encouraged—even expected—to specify the desired study power and to calculate the sample size needed to achieve this power when planning the study. In our example, how many patients would have been required to provide sufficient study power? To estimate this number, the applicable statistical calculation would be completed using a suitable statistical procedure (such as the web-based sample size calculator available at http://www.statpages.org/proppowr.html). In our example, to have an 80% chance of detecting a 50% difference in the frequency of cancer recurrence at an α level of 0.05, the power calculation

Table 2—*Variables Included in Statistical Calculations for a Comparison of Two Percentages Using a χ^2 Test and the Effect of Each Variable on the Desired Sample Size for Each Group*

Variable	Group 1, %	Group 2, %	α	Power ($1 - \beta$)	Sample Size (No.)
Two-tailed test	60	30	0.05	0.8	48
One-tailed test	60	30	0.05	0.8	39
↓ Difference	60	**40**	0.05	0.8	107
↓ α	60	30	**0.01**	0.8	69
↑ Power	60	30	0.05	**0.9**	62

Values in bold have been varied from the first line to show how changes in each variable affect sample size (determined, with continuity correction, using http://www.statpages.org/proppowr.html).

indicates that each of the two groups should have at least 48 patients. As you would expect, this sample size will change if any of the specifications are changed (Table 2). For example, to detect a smaller difference in recurrence rates (eg, 60% vs 40%), at least 107 patients would be needed in each study group. Similarly, the use of an α level of 0.01 would require each study group to have at least 69 patients. Further, to achieve a statistical power of 90%, each group would require at least 62 patients.

By applying these principles and procedures, researchers are able to better ensure that their studies have an adequate number of patients and therefore sufficient statistical power to detect clinically important differences if they truly exist. In fact, if a proper sample size calculation is completed during the planning of a study and if the required number of subjects is recruited into the study, a calculation of power after the completion of the study should never be needed. As such, post hoc calculations of study power should ideally never be necessary. Although our hypothetical study compares the differences between two percentages, the same principles and analogous procedures are available for other study situations, such as correlation coefficients, regression slopes, and testing the differences in means and rates.

Summary

This article, building on earlier articles in this Medical Writing Tips series,[1,2] reviews the principles and practices of hypothesis testing and, further, discusses the ability, if necessary, to assess the statistical power of a study with statistically nonsignificant results. Finally, the article also describes how sufficiently powered studies can be best assured through the completion of an appropriate sample size calculation while the study is being designed.

Acknowledgments

Financial/nonfinancial disclosures: The authors have reported to *CHEST* the following conflicts of interest: Dr Harvey receives royalties from the statistics book, *Statistics for Medical Writers and Editors*. Dr Lang receives royalties from the statistics book, *How to Report Statistics in Medicine*.

Other contributions: We thank Professors Paul Corey and David Streiner for their valuable comments, suggestions, and corrections. All referenced web-based calculators were located at and accessed from http://www.statpages.org.

References

1. Lang T. Documenting research in scientific articles: guidelines for authors: 2. Reporting hypothesis tests. *Chest*. 2007; 131(1):317-319.
2. Lang T. Documenting research in scientific articles: guidelines for authors: reporting research designs and activities. *Chest*. 2006;130(4):1263-1268.
3. Altman DG, Schulz KF, Moher D, et al; CONSORT GROUP (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001;134(8):663-694.
4. Fletcher RH, Fletcher SW. *Clinical Epidemiology: The Essentials*. 4th ed. Philadelphia, PA: Lippincott, Williams & Wilkins, 2005:118,121-122.