Power and Sample Size: How Many Patients Do I Need?

Gary L. Grunkemeier, PhD, and Ruyun Jin, MD

Providence Health & Services, Portland, Oregon

In our combined consulting experience (42 statistician-years), the two most frequently asked questions are: (1) "Is this result statistically significant?" and (2) "How many patients do I need for my study?" In this expository article, we will attempt to provide an understanding of how the statistical significance of a completed study is determined and how, by adding the notion of power, sample sizes for a future study can be estimated. We start with a test of statistical significance, because that is the number 1 question, and it also introduces some of the concepts needed to answer question number 2. We will emphasize graphical portrayals of these concepts, and relegate statistical details to Appendix 1. Appendix 2 lists some examples of software available for sample size calculation.

Completed Study

There is current interest in postoperative atrial fibrillation (AF), and also current controversy about the role of aprotinin in cardiac surgery. As an illustrative data set, we will use all 1,730 isolated heart valve replacements at Providence St. Vincent Medical Center, Portland, Oregon from 1998 to 2005 in patients without preoperative AF. The numbers of valves implanted with and without the use of aprotinin were about equal (835 and 895, respectively), but the percentage of aprotinin usage increased with time. The incidence of new AF was 21.3% (178 of 835) with aprotinin and 25.9% (232 of 895) without aprotinin, a difference of −4.6%. This result is not risk-adjusted, and it is used for demonstration purposes only. Yet the results agree remarkably with a recent article on thoracic aortic surgery that showed a lower incidence of post-surgical arrhythmia (mostly AF) with the use of aprotinin (20%) compared with the matched controls (25%) [1].

Determining Statistical Significance

The statistical significance of this observed difference is determined as follows. We cannot prove that the true difference is −4.6%. Instead, we set up a straw man, a null hypothesis that there is no difference, and then we saw how improbable the observed difference (−4.6%) would be due to random chance with that scenario. We hoped that the difference was improbable enough to provide sufficient evidence to reject the null hypothesis, thus allowing us to declare the difference statistically significant.

Address correspondence to Dr Jin, 9205 SW Barnes Rd., Suite 33, Portland, OR 97225; e-mail: ruyun.jin@providence.org.

To formalize this exercise, we estimated the distribution of the difference between the two AF percentages under the null hypothesis (shown by the bell-shaped curve in Figure 1) to see how remote the observed difference (point D in Figure 1) would be if the distribution were really centered around zero (the null hypothesis). It is possible that there is no real difference in AF and that the extreme value we discovered was just due to chance. We will determine the probability of a difference at least that extreme arising by chance. The bell-shaped curve in Figure 1 is centered at zero and its distribution is determined from the retrospective data (Appendix 1A). The determination of statistical significance is based on two parameters: (1) the observed difference, and the compactness (standard error) of the distribution of this difference, which is in turn based on (2) its variability (in the case of binomial distribution, a function of the percentages themselves), and (3) the observed sample sizes. The smaller the variability or the larger the sample sizes, the more compact (narrow and tall) was the distribution of the difference. The area under the tails of this distribution corresponding to points that are at least as extreme as the observed difference are shaded. In this figure, they represent about 2.4% of the probability (the entire area under the curve equals 100% probability), 1.2% in each tail (for a two-sided test). By an arbitrary but common convention, if this shaded area is less than 5% (0.05), then the observed difference is declared statistically significant, in this case, with a p value of 0.024, which equals 2.4%.

Prospective Study

The previous exercise used historical data from a retrospective study. Now suppose that we decide to launch a prospective randomized study to test the hypothesis that aprotinin reduces AF. Again the statistical set up begins with a null hypothesis that aprotinin does not affect AF, which again we hope to reject to claim support for our (alternative) hypothesis that aprotinin does indeed reduce the percentage of AF.

Motivation

Before proceeding, we might ask, why bother when we already have the data previously presented (assuming it were risk-adjusted, which it is not)? One reason that evidence from a prospective study is more compelling is that the hypothesis is determined beforehand, and its acceptance will rise or fall based on the data collected. With retrospective studies, there are many data sets that

Ann Thorac Surg 2007;83:1934–9 • 0003-4975/07/$32.00
doi:10.1016/j.athoracsur.2007.01.020

could have been examined and innumerable hypotheses that could have been tested; maybe we looked at many of them and decided to announce only the one that worked out, or that aligned with our preconceived bias. The multiple comparisons required to discover such a significant result would in fact produce an erroneously small *p* value. (Confession: we are in fact guilty of this very deception. During the time period we reviewed, only isolated valve surgeries, not coronary artery bypass grafting nor coronary artery bypass grafting plus valve, achieved a statistically significant difference in AF rates.)

More Parameters Needed

To design the prospective study, we will use the previously mentioned data to provide estimates of (1) the clinically important difference (effect size) that we hope to detect, and (2) the variability of that difference. Other considerations would normally be brought to determine the selection of the effect size [2, 3], but we will use the observed difference (−4.6%), mainly to help demonstrate the relationship between hypothesis testing and sample size estimation. For the prospective study design, we need to designate two more parameters: (1) the probability α (alpha) of a false positive (or type I error), that is, of declaring significance when it does not exist (similar to the *p* value in Fig 1), and (2) the probability β (beta) of a false negative (or type II error), that is of not finding significance when it does exist. For instructive purposes, we will introduce these graphically and one at a time. Given these parameters (plus the fact that the test is two-sided), we will be able solve an equation for the number of patients needed in each arm of the study to satisfy these design requirements.

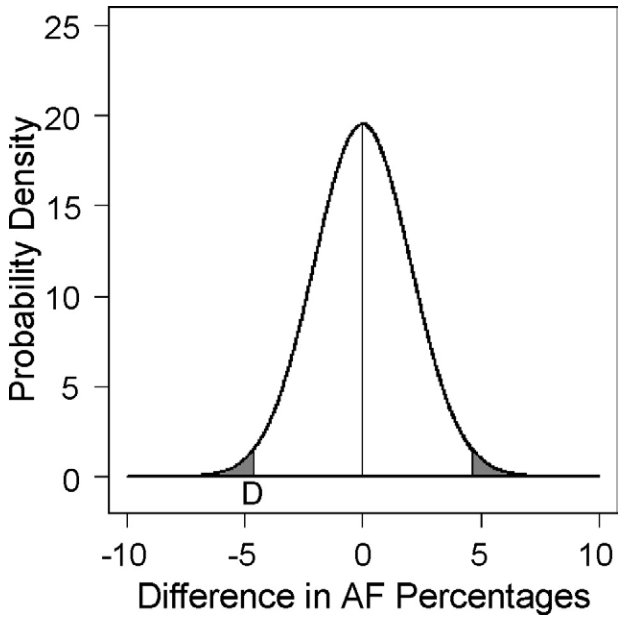

Fig 1. A significance test is performed by comparing the observed difference (D) with the distribution of the difference in atrial fibrillation (AF) percentages under the null hypothesis with a mean of zero. The probability of a value at least as extreme as D, in either direction (for a two-sided test) is 2.4%, 1.2% in each tail (shaded areas). Thus, the p value is 0.024 and the difference is statistically significant.

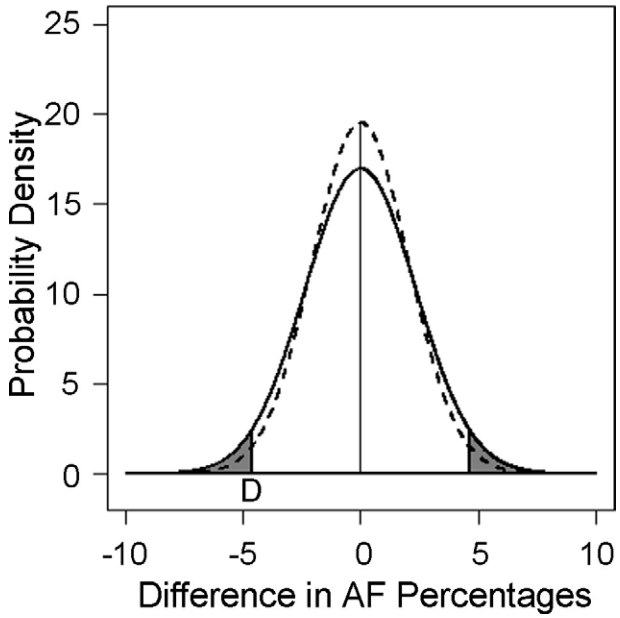

Fig 2. The distribution of the difference in atrial fibrillation (AF) percentages under the null hypothesis that would find the observed difference −4.6% (point D) in AF percentages to be just barely significant (p = 0.05) requires 656 patients in each group. The 5% tail areas (2.5% in each tail) are shaded. For comparison, the dashed curve line shows the curve from Figure 1.

False Positive Probability (α)

We arbitrarily chose 0.05 (5%) for the probability α, also called the size or level of the test. This means that our study must produce a *p* value of 0.05 or less to achieve a significant result. Note that the historical data (Fig 1) with 1,730 (835 plus 895) patients produced a smaller *p* value (0.024) than the prospective study would require. One might wonder how many historical patients would have been needed to (just barely) achieve statistical significance, which is what we would be quite happy to achieve with our prospective study. Using the formulas that produced the *p* value for Figure 1 (Appendix 1A), somewhat in reverse (Appendix 1B), we can determine that only 1,312 patients (656 in each group) would be needed to produce a *p* value of exactly 0.05. The shaded area under the tails of the distribution in Figure 2 contain 5% of the probability, 2.5% in each tail, and just barely contain the observed difference (point D in Figure 2). Note that this distribution is less compact than the distribution in Figure 1 (reproduced as the dashed curve in Fig 2), which had larger sample sizes and was thus taller and narrower.

False Negative Probability (β)

This seemed to be great news. Assuming that the same percentages were obtained in our prospective study, we would need only 656 patients in each arm to (just barely)

achieve statistical significance ($p < 0.05$), that is, fewer patients than in the retrospective study. However this assumes that the aprotinin and non-aprotinin arms of the study produced the same results as the retrospective study, with a difference of exactly −4.6%, which was our alternative hypothesis. Even if the true percentages were exactly as assumed by the alternative hypothesis, it is unlikely that those exact percentages would be observed in a prospective study, because of random variation. The dashed bell-shaped curve in Figure 3 shows the distribution of the differences that might be observed under the alternative hypothesis (true mean = D) by chance, if there were only 656 patients in each group (Appendix 1C). Note that by symmetry there was only a 50:50 chance that the observed difference would be to the left of point D in Figure 3, that is, in the rejection region of the test (the white area under the dashed curve). Thus, even when the alternative hypothesis was true, the probability that this prospective study, with only 656 patients in each arm, would produce a value that was not in the rejection region of the null hypothesis (a false negative, or type II error) was 50% (β = 0.50), as shown by the light gray area under the dashed curve in Figure 3.

Power (1-β)

In Figure 3, the probability (β) of not rejecting the null hypothesis when the alternative hypothesis was true (type II error) is 50% (light gray area). The complement of this probability (1-β), the probability of rejecting the null hypothesis when the alternative hypothesis was true, is called the power of the test. Thus, in Figure 3, the power was only 50% (the white area under the dashed curve). Typically we would not be satisfied to go forward with a study that had such a low chance of finding a significant result. Before beginning our proposed study we wanted some insurance that it would have a high probability of declaring a significant result when the alternative hypothesis was true, so we set the power, again arbitrarily, to 80% (β = 0.20). Of the parameters involved in the study, the only one left to manipulate was the sample size. As we increased the sample size, keeping the effect size, standard errors, and α fixed, we increased the power (by decreasing β). Thus we would pay for this insurance by requiring more subjects in the study.

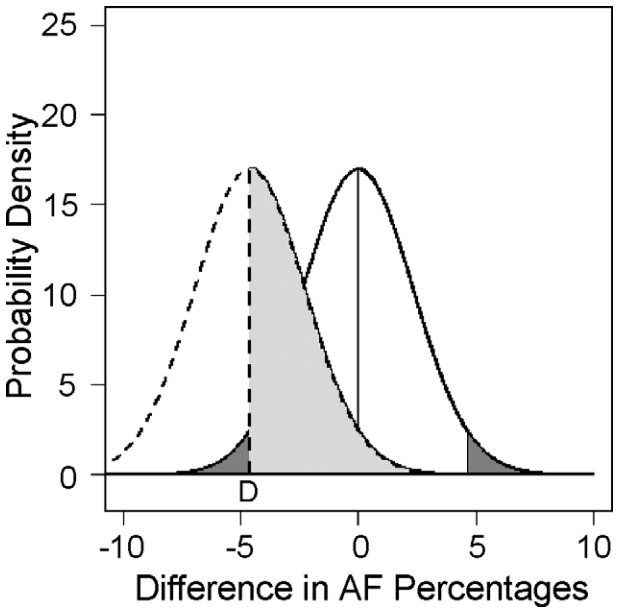

Fig 3. Distributions of the difference in atrial fibrillation (AF) percentages under the null (solid curve) and alternative (dashed curve) hypotheses, with just 656 patients in each group. The probability (β) of a type II error is 50% (light gray area), which corresponds to a power (1 − β) of only 50%. The probability of false positive (α) is 5%, 2.5% in each tail of the solid curve (dark gray area). (D = true mean.)

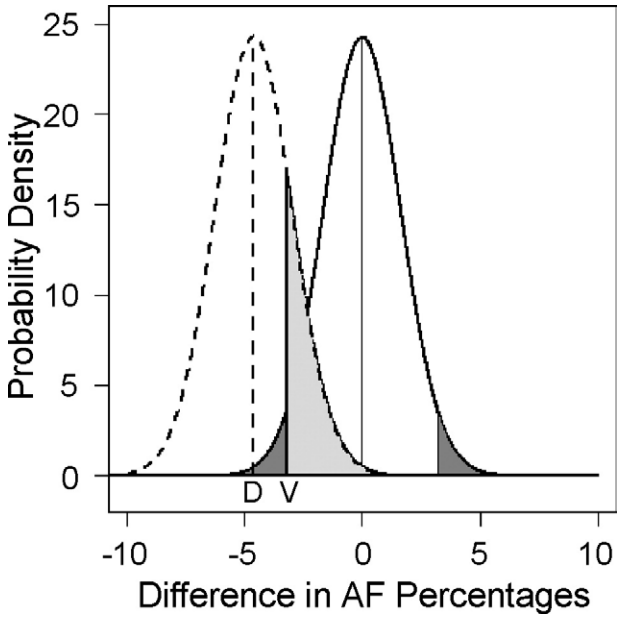

Fig 4. Distributions of the difference in atrial fibrillation (AF) percentages under the null (solid curve) and alternative (dashed curve) hypotheses, with the sample sizes in each group increased to 1,335 to be able to decrease the probability (β) of a type II error to 20% (light gray area) for a power (1 − β) of 80%. The probability of false positive (α) is 5%, 2.5% in each tail of the solid curve (dark gray area). The critical value V is used in the derivation of the sample size formula (Appendix 1D). (D = effect size.)

Sample Size

As the sample size increased, the distributions in Figure 3 would become more concentrated around their mean values (ie, taller and narrower) and the light gray area would become smaller. When we solved the sample size equation with the AF percentages observed in the retrospective study and α = 0.05, we found that 1,335 patients in each arm of the study would just satisfy the requirements (Appendix 1D). The resulting distributions are shown in Figure 4. The difference that would just reach statistical significance with these sample sizes is shown by point V (the critical value) in Figure 4. Note that the effect size (point D in Fig 4) was now farther out in the tail of the null distribution. The relationship between power

and sample size is shown in Figure 5 for 4 levels of α. On the curve for α = 0.05, the light gray circle at 50% power corresponds to the situation depicted in Figure 3, with 656 patients in each arm of the study, and the dark gray circle corresponds to 80% power and 1,335 patients, shown in Figure 4. If we repeated the significance test in the first section (Fig 1) with 1,335 patients, we would get a much more significant difference ($p = 0.005$). In fact, that would be the same as setting the power to 50% so that the result could be seen from the white circle in Figure 5.

Post-Script: "Post-Hoc" or "Observed" Power

Suppose we had completed the study and it did not result in a significant difference. Suppose the observed difference was at point O in Figure 6, to the right of the critical value (point V), so that the null hypothesis was not rejected ($p > 0.05$). Facing such a negative study, it is tempting to compute the "observed" or "post hoc" power (Appendix 1E) to conclude that the difference was not significant because the study was underpowered, implying that the difference would have been significant if a more appropriately powerful study had been done. However, this is a misuse of power and should be avoided [2, 4–7].

When evaluated at the observed difference, the power will always be less than 50% (with rare exceptions). This

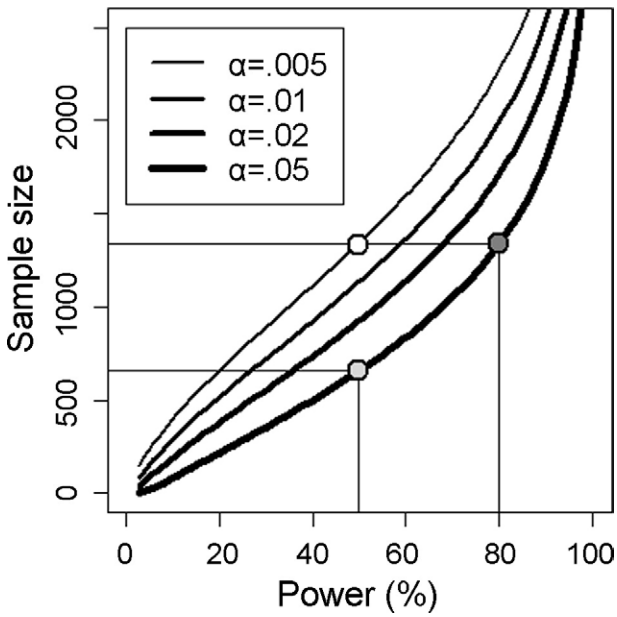

Fig 5. Relationship between power and sample size for the atrial fibrillation percentages observed in the retrospective study and 4 levels of α. For α = 0.05, the light gray circle at 50% power corresponds to Figure 3, with n = 656, and the dark gray symbol at 80% power corresponds to Figure 4, with n = 1,335. If the statistical test is performed after the study is completed with n = 1,335, then the difference is much more significant (p = 0.005). This is the same as a test with α = 0.005 and a ("post hoc") power of 50%, as shown by the white circle.

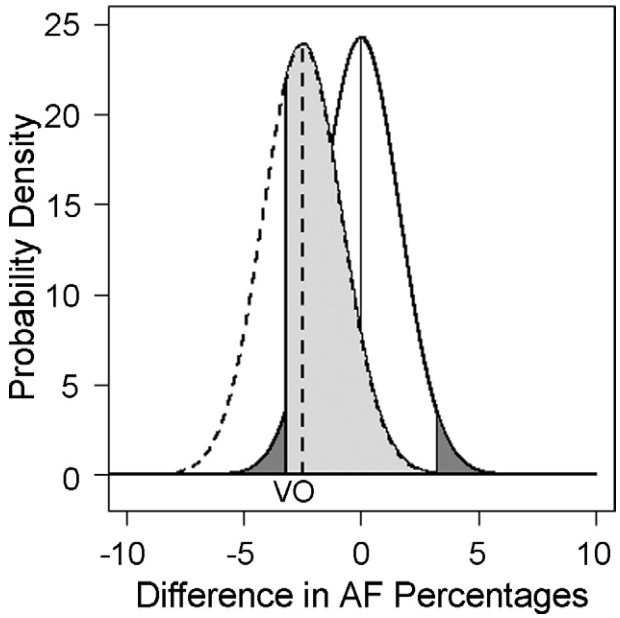

Fig 6. Example of a "post-hoc" power calculation, when the observed difference (O) is not significant. The null hypothesis, centered at 0, is shown as the solid curve, with the extreme 2.5% in each tail shown in dark gray. A power calculation is done with the alternative hypothesis (dashed curve) centered at O, to imply that the study was underpowered. Because the light gray area under the distribution of the mean of the alternative hypothesis (β) will always be greater than 50%, the post-hoc power (1-β) will always be less than 50%. (AF = atrial fibrillation; V = critical value.)

can be seen from Figure 5, in which the alternative hypothesis (dashed curve) was now centered at the observed value O. The probability of a type II error (1-power) was larger than 50% (light gray area), and you could see that it would still exceed 50% as the observed value (O) approached the critical value (C), and become even larger as the observed difference moved toward zero. Post-hoc power does not add information beyond that of the p value, because it is derived by just solving the sample size equation for power after plugging in the observed values from the study. For these reasons, several authors have roundly criticized its use: "... this ill-advised exercise answers an already answered question" [4]; "... if your car made it to the top of the hill, it was powerful enough; if it didn't, it was not powerful enough" [2]; and, "... like trying to convince someone that buying a lottery ticket was foolish (the before-experiment perspective) after they hit a lottery jackpot (the after-experiment perspective)" [5].

References

1. Sedrakyan A, Wu A, Sedrakyan G, Diener-West M, Tranquilli M, Elefteriades J. Aprotinin use in thoracic aortic surgery: safety and outcomes. J Thorac Cardiovasc Surg 2006;132:909–17.
2. Lenth RV. Some practical guidelines for effective sample size determination. The American Statistician 2001;55:187–93.
3. Fleiss JL, Levin B, Paik MC. Chapter 4. Determining sample sizes needed to detect a difference between two proportions.

Statistical methods for rates and proportions. Hoboken, NJ: Wiley-Interscience, 2003:64–85.
4. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365:1348–53.
5. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994;121:200–6.
6. Levine M, Ensom MH. Post hoc power analysis: an idea whose time has passed? Pharmacotherapy 2001;21:405–9.
7. Hoenig JM, Heisey DM. The Abuse of power: the pervasive fallacy of power calculations in data analysis. The American Statistician 2001;55:19–24.
8. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 1981;2:93–113.
9. Machin D, Campbell MJ. Chapter 3. Comparing two binomial proportions. statistical tables for the design of clinical trials. Oxford UK: Blackwell Scientific, 1987:10–34.
10. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 1980;36:343–6.
11. Dupont WD, Plummer WD Jr. Power and sample size calculations. A review and computer program. Control Clin Trials 1990;11:116–28.

Appendix 1

Formulas Used to Draw the Figures and Produce the Statistics

Note that event percentages are used for presentation purposes in the text and figures, but proportions are used for the calculations in this Appendix. Percentage = 100∗proportion, where the asterisk symbol (∗) is used to indicate multiplication.

A. (Fig 1). Given the observed percentages and sample sizes, determine the *p* value. With large sample sizes such as this, the binomial distribution for the null (no difference) hypothesis can be approximated by a normal distribution with mean zero, and standard error (SEn) derived from the binomial distribution. We define some parameters and formulas based on them to produce the *p* value for Figure 1. The *p* value is the probability of a difference at least as extreme as that observed, if there is really no difference in atrial fibrillation (AF) proportions between the two groups. A continuity correction is sometimes used, which makes the test slightly more conservative (larger *p* value).
   A1. n1 and p1 are the number of patients and proportion of AF without aprotinin.
   A2. n2 and p2 are the number of patients and proportion of AF with aprotinin.
   A3. D = p2–p1 is the difference in the proportion of AF attributed to aprotinin.
   A4. p = (p1 + p2)/2 is the average proportion of AF (a weighted average is sometimes used).
   A5. q = 1 − p is the overall proportion without AF.
   A6. SEn = $\sqrt{}$(p∗q/n1 + p∗q/n2), where $\sqrt{}$(x) indicates the square root of x.
   A7. Zn = D/SEn = −2.25, the value for a standard (mean = 0, SE = 1) normal distribution (Z).
   A8. *p* value = 0.024 is the (two-sided) probability that Z would be as extreme as Zn.

B. (Fig 2). Given the observed proportions, determine the (equal) sample sizes needed for each group to produce a *p* value of 0.05 (without regard to power).
   B1. n = n1 = n2 is the required sample size for each group.
   B2. SEn = $\sqrt{}$(2∗p∗q/n), the same as equation A6.
   B3. α = 0.05.
   B4. Zα = 1.96 is the (two-sided) value of Z corresponding to α.
   B5. D/SEn = 1.96.
   B6. SEn^2 = 2∗p∗q/n = (D/1.96)^2, by using B5 and B2.
   B7. n = 2∗p∗q∗(1.96/D)^2 = 655.3, by solving B6 for n.

C. (Fig 3). The distribution based on the alternative hypothesis is superimposed onto Figure 2. The mean is at D and the SE corresponding to the alternative hypothesis (SEa) has a slightly different formula (than B2):
   C1. SEa = $\sqrt{}$(p1∗q1/n + p2∗q2/n).

D. (Figs 4, 5). The sample size formula can be derived by considering the critical value (V) that precisely separates the two shaded regions in Figure 4. This point can be defined in two ways:
   D1. V = 0 - Zα∗SEn is the relationship derived from the null hypothesis.
   D2. V = D + Zβ∗SEa, is the relationship from the alternative hypothesis.
   D3. D = −Zα∗SEn - Zβ∗SEa, by combining equations D1 and D2.
   D4. D = −Zα∗$\sqrt{}$(2∗p∗q/n) − Zβ∗$\sqrt{}$((p1∗q1 + p2∗q2)/n), by using B2 and C1 in D3.
   D5. n = (Zα∗$\sqrt{}$(2∗p∗q)/D + Zβ∗$\sqrt{}$(p1∗q1 + p2∗q2)/D)^2 = 1334.6, solving D4 for n. Various simplifications of this formula are sometimes used, a continuity correction can be used to make it slightly more conservative, and unequal allocation in the two arms can be specified [3, 8–10].

E. (Fig 6). This is an example of the questionable practice of calculating "power" after a negative study is completed. We assumed that the mean AF of patients with and without aprotinin was 25.9% and 23.4%, respectively, so the observed difference (O) was 2.5% (= 0.025). Equation D4 can be solved for the "observed" power, which will always be less than 50%, because β (the light gray area) will be greater than 50% (with very rare exceptions).

Appendix 2

Resources for Sample Size Calculations

Several tools are available for performing sample size calculations, including (1) web-based applications, (2) free downloadable special programs, (3) commercial special programs, (4) a free downloadable general statistical program, and (5) commercial statistical packages. Here is an example of each:

1. Java Applets for Power and Sample Size (http://www.stat.uiowa.edu/~rlenth/Power/), by Russell V. Lenth, Department of Statistics and Actuarial Science at the University of Iowa, is a user-friendly web application using a clever slider interface that can be overridden for more precision [2].
2. PS: Power and Sample Size Calculation (http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize), by William D. Dupont and Walton D. Plummer, Department of Biostatistics at Vanderbilt University, is a versatile, free downloadable program for Windows [11].

3. PASS: Power Analysis and Sample Size (NCSS, Kaysville, UT) (http://www.ncss.com/pass.html) is a commercial software package.
4. R (http://www.r-project.org/) is a free, open source statistical computing and graphics program for Windows or Mac OS, with advanced capabilities and a very active user community that contributes routines for specialized calculations. One of the contributed packages is *Hmisc* (http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/Hmisc), a treasure chest of analytical capabilities, written by Frank E. Harrell, Jr, Department of Biostatistics at Vanderbilt University. The function *bsamsize* computes the sample sizes for comparing binomial proportions.
5. Stata (StataCorp LP, College Station, TX) is a commercial statistical program for Windows and Mac OS with advanced capabilities and a user community that contributes routines for specialized calculations. The function *sampsi* computes the sample sizes for comparing binomial proportions.

Online Discussion Forum

Each month, we select an article from the *The Annals of Thoracic Surgery* for discussion within the Surgeon's Forum of the CTSNet Discussion Forum Section. The articles chosen rotate among the six dilemma topics covered under the Surgeon's Forum, which include: General Thoracic Surgery, Adult Cardiac Surgery, Pediatric Cardiac Surgery, Cardiac Transplantation, Lung Transplantation, and Aortic and Vascular Surgery.

Once the article selected for discussion is published in the online version of *The Annals*, we will post a notice on the CTSNet home page (**http://www.ctsnet.org**) with a **FREE LINK** to the full-text article. Readers wishing to comment can post their own commentary in the discussion forum for that article, which will be informally moderated by The Annals Internet Editor. We encourage all surgeons to participate in this interesting exchange and to avail themselves of the other valuable features of the CTSNet Discussion Forum and Web site.

For June, the article chosen for discussion under the Pediatric Cardiac Dilemma Section of the Discussion forum is:

Unidirectional Monovalve Homologous Aortic Patch for Repair of Ventricular Septal Defect With Pulmonary Hypertension

Bo Zhang, MD, Shuming Wu, MD, Jiali Liang, MD, Guangfu Zhang, MD, Guanhua Jiang, MD, Min Zhou, MS, and Xiangling Li, MD

Tom R. Karl, MD
The Annals *Internet Editor*
UCSF Children's Hospital
Pediatric Cardiac Surgical Unit
505 Parnassus Ave, Room S-549
San Francisco, CA 94143-0118
Phone: (415) 476-3501
Fax: (212) 202-3622
e-mail: karlt@surgery.ucsf.edu