Paper 211-29

Sample-Size Analysis in Study Planning: Concepts and Issues, with Examples Using PROC POWER and PROC GLMPOWER

Ralph G. O'Brien, Cleveland Clinic Foundation, Cleveland, Ohio
John M. Castelloe, SAS Institute, Cary, North Carolina

ABSTRACT

Ever-improving methods and software, including new tools in the SAS® 9.1, are transforming the practice of sample-size analysis. Statisticians and subject-matter investigators need to understand the concepts and issues involved, but such matters are too often given short shrift in statistics education. We hope this paper helps close that gap.

This presentation covers the most common type of sample-size analysis, classical prospective power analysis, which examines the chance that a given null statistical hypothesis will be rejected ($p \le \alpha$). Essential to this—but rarely taught—is that greater power reduces both positive and negative inference mistake rates (defined herein), which we argue are more relevant to researchers than the classical Type I and II error rates (α and $\beta = 1 -$ power).

In carrying out a strong prospective sample-size analysis, the research team is forced to delineate and critique the rationale undergirding the study and all the components of the research protocol. They must specify tight research questions, the research design, the various measures, and the analysis plan. They must come to agree on and justify (as best they can) some reasonable conjectures for what the "infinite dataset" may be for their study.

Modern software facilitates all this work. These and other concepts and issues are addressed through a case study that uses PROC POWER (SAS/STAT® 9.1).

This tutorial is for statisticians and para-statisticians inexperienced with sample-size analysis, but even those who are quite experienced may discover new ideas and/or new ways of communicating these matters to students and investigators.

Note: Later versions of this paper and all related SAS code are available at www.bio.ri.ccf.org/robrien/OBriCast2004. We intend to add a case study involving an analysis of covariance that uses PROC GLMPOWER and others that use PROC POWER to handle sample-size analyses for confidence intervals and tests of equivalence. The final version will also describe a general method for retrospective power analysis (implemented via a custom SAS macro), which uses existing data to compute lower confidence limits for power.

INTRODUCTION

In their "Perspectives on Large-Scale Cardiovascular Clinical Trials for the New Millennium," Drs. Eric Topol, Robert Califf, and others (1997) provide a fine preamble to our discussions:

> The calculation and justification of sample size is at the crux of the design of a trial. Ideally, clinical trials should have adequate power, ≈90%, to detect a clinically relevant difference between the experimental and control therapies. Unfortunately, the power of clinical trials is frequently influenced by budgetary concerns as well as pure biostatistical principles. Yet an underpowered trial is, by definition, unlikely to demonstrate a difference between the interventions assessed and may ultimately be considered of little or no clinical value. From an ethical standpoint, an underpowered trial may put patients needlessly at risk of a new therapy without being able to come to a clear conclusion.

'Sample Size at Crux of Design'
- power ≈90% to detect relevant effects
- budgetary (and recruiting) constraints
- underpowered study
  - …little or no value
  - …ethical concerns

Let us augment this passage with another key issue in sample-size analysis. All studies are planned based on knowledge gained from previous ones. Richard Feynman, the 1965 Nobel Laureate in Physics and the self-described "curious character," wrote (Feynman 1999),

> Scientific knowledge is a body of statements of varying degrees of uncertainty,
> some mostly unsure,
> some nearly sure,
> none absolutely certain.

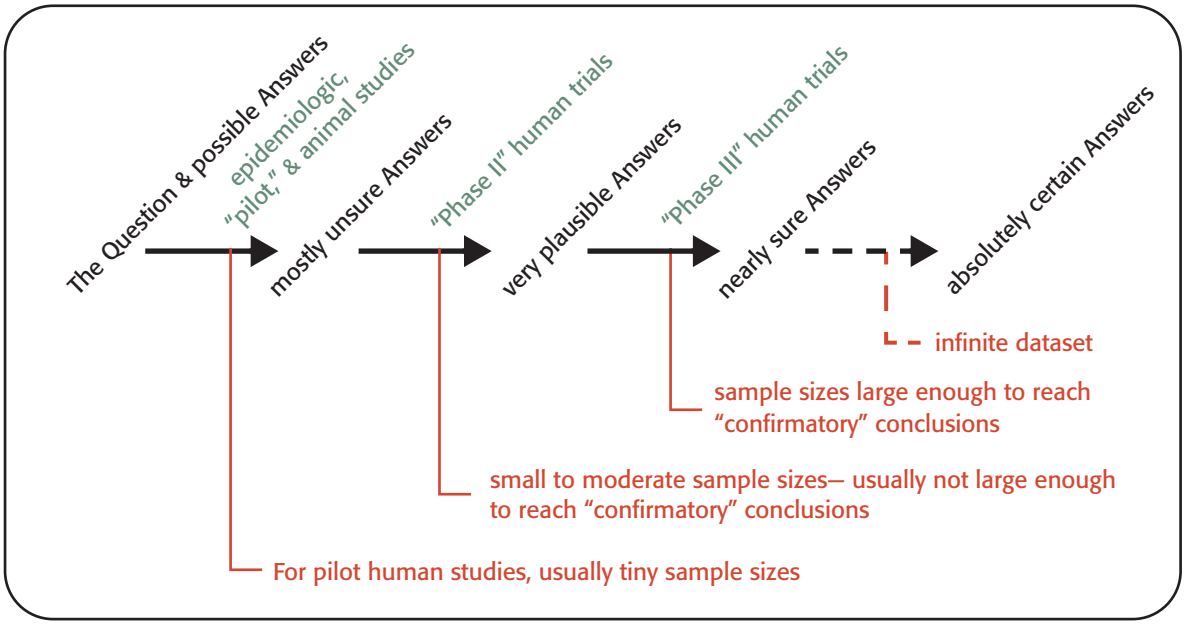

Figure 1: The March of Science in medical research.

This reflects what we call The March of Science, which for medical research is sketched in Figure 1. As we step forward, our sample-size considerations need to reflect what we know. At any point, but especially at the beginning, the curious character inside of us should be free to conduct observational, exploratory, or pilot studies, because as Feynman said, "something wonderful can come from them." Such studies are still 'scientific,' but they are for generating new hypotheses, not testing them. Accordingly, little or no sample-size analysis is usually called for. But to become "nearly sure" about our answers, we typically conduct convincing confirmatory studies under specific protocols. This often requires innovative and sophisticated statistical planning, which is usually heavily scrutinized by all concerned, especially by the reviewers. No protocol is ever perfect, but as the New York Yankee catcher and populist sage once said, "Don't make the wrong mistake."

The March of Science

- Observational, exploratory, or pilot studies may not require sample-size analyses.
- Confirmatory studies usually require sample-size analyses.

This paper

- Reviews the core issues and concepts in sample-size analysis for classical (frequentist) hypothesis testing.
- Defines negative and positive false inference rates, quantities often more relevant to consider than the classical Type I and Type II error rates. Investigators need to understand that greater statistical power reduces *both* types of false inference rates.
- Develops these principles through a fictionalized case study that is tied directly to important medical research. All computations use the SAS System Version 9.1, especially PROC POWER (SAS Institute, 2004a,b).
- Later versions of this paper and all related SAS code are at www.bio.ri.ccf.org/robrien/OBrCast2004 and will eventually also include examples involving
  - analysis of covariance (PROC GLMPOWER)
  - confidence intervals (PROC POWER)
  - equivalency testing (PROC POWER)
  - finding lower confidence limits for power using existing data (custom SAS macro).

HYPOTHESIS TESTING

WILL ADJUNCTIVE THERAPY USING DCA REDUCE MORTALITY IN CHILDREN WITH SEVERE MALARIA? (PART 1)

Peter Stacpoole, MD, PhD, directs the General Clinical Research Center at the University of Florida. He has spent decades investigating the safety and efficacy of dichloroacetate (DCA) for treating lactic acidosis (toxic levels of lactic

acid in the blood) in chronic and acquired diseases. Malaria remains one of the world's foremost health problems. According to a report released in 2003 by the World Health Organization, malaria kills at least one million people annually, mostly children under five in sub-Saharan Africa. Lactic acidosis is a frequent complication in severe malaria. It is also an independent statistical predictor of death, and a plausible biological rationale supports the hypothesis that lactic acidosis is a contributing cause of death. A team led by Dr. Stacpoole conducted a randomized, double-blind controlled trial of quinine only versus quinine+DCA. They concluded that a single infusion of DCA was well-tolerated, did not appear to interfere with quinine, and, as hypothesized, reduced blood lactate levels (Agbenyega, et al., 2003). The sample size of N = 62 + 62 was much too small to support comparing mortality rates. The authors concluded that a large prospective study was warranted.

The Research Question

- Malaria kills hundreds of thousands of children each year.
- Lactic acidosis is a frequent complication in patients with severe malaria.
- Higher lactate levels is an independent statistical predictor of death. Biological rationale suggests a causal link, as well.
- DCA reduces blood lactate levels in many diseases, including malaria.
- Does DCA (in addition to quinine) reduce mortality?

Let us suppose that Dr. Sol Capote (fictitious) heads the tropical disease research program at the World Health Organization (WHO), which has decided to design a large clinical trial that will address this question effectively. Dr. Capote is an experienced investigator, so he knows that substantial thought, effort, and experience need to go into developing the sample-size analysis and the rest of the statistical considerations.

P-VALUES AND POWER

The WHO study will use a randomized, double-blind design to compare quinine-only with quinine+DCA. For reasons discussed below, about 2/3 of the subjects will get quinine+DCA.

The primary analysis will yield a p-value that compares the mortality rates of subjects treated with quinine only versus those treated with quinine+DCA. Smaller p-values indicate greater statistical separation between the two samples, but how that p-value is determined is an issue that is secondary to understanding sample-size analysis. It may come from one of the many ways of comparing two independent proportions, including the classical Pearson chi-square test for 2×2 contingency tables, or it may come from a logistic or hazard modeling that includes co-predictors. Regardless of what test is used to get the p-value, if it is small enough ("significant") and the quinine+DCA mortality rates are better, Dr. Capote will report that the study supported the hypothesis that DCA reduces mortality in children with severe malaria complicated with lactic acidosis. If the p-value is not small enough ("not significant"), then he will report that the data provided insufficient evidence to support the hypothesis.

Concepts and Issues

- Smaller p-values indicate greater statistical effect in the sample.
- How p-values are determined is a secondary issue to understanding sample-size analysis.

Null and non-null distributions of p-values. Our quest here is to figure out Mother Nature. We could do this if she gave us an infinitely large, perfectly clean dataset for the question at hand, so we could calculate the true effect. But we must settle for a random sample. If there is no difference between the two groups' mortality in the infinite dataset, then the p-value obtained from the sample is just as likely to be 0.972 as 0.621 or 0.126 or even 0.001. In short, the null distribution of the p-value is flat between 0.0 and 1.0 (Figure 2, top), and so the chance is 5% that $p \le 0.05$, or $100\alpha\%$ that $p \le \alpha$.

Concepts and Issues

- Mother Nature, infinite datasets, sample.
- Null p-value distributions are flat.
- Non-null p-value distributions are skewed towards 0.0.
- If $p \le \alpha$, the test is "significant." If the true effect is null, this Type I error happens with probability α, which we can set.
- If $p > \alpha$, the test is "not significant.." If the true effect is non-null, this Type II error happens with probability β, which we never know in practice.
- Power = $1 - \beta$.

On the other hand, if DCA has some true effect, good or bad, then the distribution of the p-values will be skewed towards 0.0 (Figure 2, bottom). This is a non-null distribution of the p-values. The statistical power of the test is the chance that $p \le \alpha$. Do you see why α is sometimes called the "null power" of the test? If there is no true effect but $p \le \alpha$ indicates otherwise, this triggers a Type I error, which is why α is called the Type I error rate. If there is some true effect, but $p > \alpha$, then a Type II error is triggered. In Figure 2, consider the common Type I error rate, $\alpha = 0.05$. Under this particular non-null distribution, the power is 0.68, so the Type II error rate is $\beta = 0.32$. We never know the true Type II error rates.

Balancing Type I and II error rates. Recall that Topol, et al. (1997) advocated that the power should be around 90%, which puts the Type II error rate around 10%. We strongly agree.

Why do so many people automatically perform power analyses using $\alpha = 0.05$ and 80% power ($\beta = 0.20$)? We doubt they really think about it!

For the case in Figure 2, we could achieve a much better Type II error rate of 0.13 if we are willing to accept a substantially greater Type I error rate of 20%. That is wise thing to consider if the consequences of a Type II error are greater than for a Type I error. For many studies, making a Type I error may only extend the research to another stage. This is wasteful, of course, but making a Type II error may halt a journey that would ultimately be successful. In the early stages of The March of Science, when the sample sizes are too small to support adequate power at $\alpha = 0.05$, it may be prudent to use higher Type I error rates, say $\alpha = 0.20$ or more. These computations are summarized in Table 1.

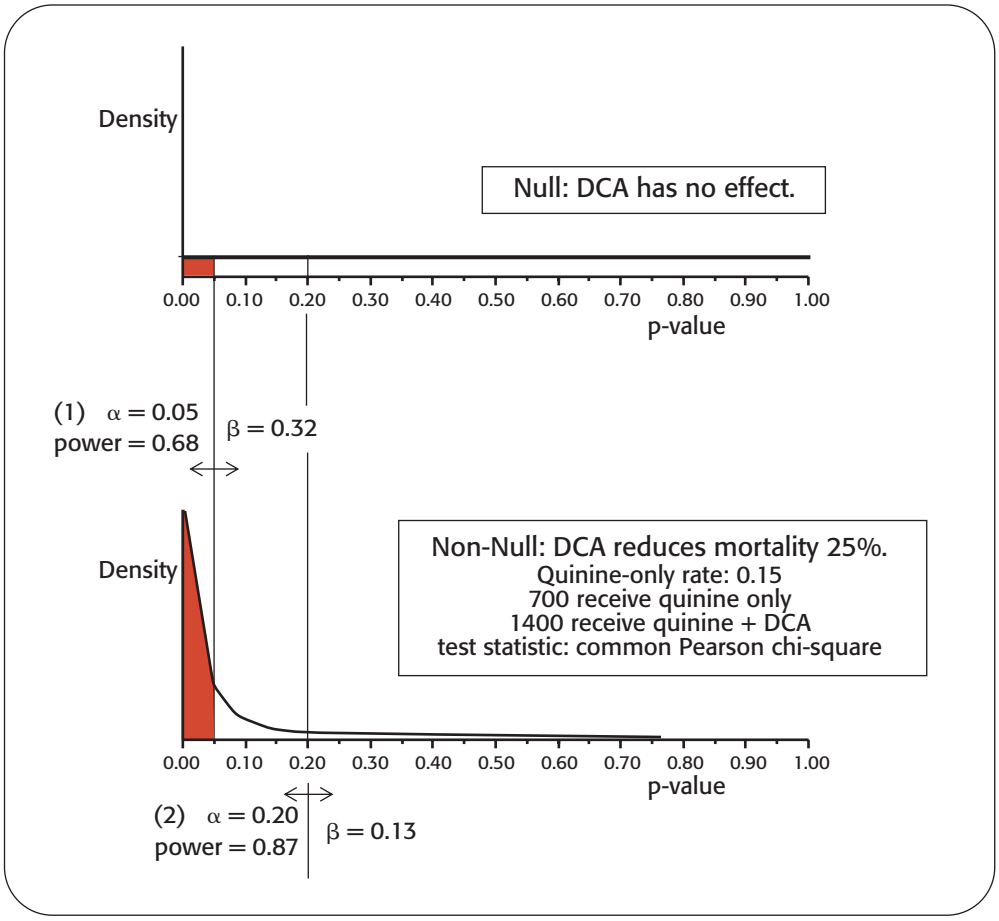

Figure 2: Null and non-null distributions for p-values. Increasing the Type I error rate (α: 0.05 → 0.20) decreases the Type II error rate (β: 0.36 → 0.20).

Concepts and Issues

- Do not automatically set α at 0.05 and/or believe that 80% power is OK.
- As best you can, design a study that reflects the consequences of making Type I and Type II errors.

Table 1: Summary of computations related to Figure 2.

Decision Based on Sample	Truth ("Infinite Dataset")	
	no effect	non-null effect
$p > \alpha$ "insufficient evidence of effect"	correct	Type II error (1) $\beta = 0.32$ (2) $\beta = 0.13$
$p \leq \alpha$ "evidence indicates some effect"	Type I error (1) $\alpha = 0.05$ (2) $\alpha = 0.20$	correct (1) power = 0.68 (2) power = 0.87

P-values are misunderstood, but for a good reason. Is the p-value the estimated probability that the effect being tested is null? For example, does $p = 0.032$ imply that the chance is 96.8% that there is some non-null effect? No.

Rather, this common and important misunderstanding stems not from some picky issue about English, but because the p-value is not what we really need to know:

- If a test turns out to be significant ($p \le \alpha$), what is the chance this will be a Type I error?
- If a test turns out to be non-significant ($p > \alpha$), what is the chance this will be a Type II error?

As Zelen (2003) has also opined, these are the critical questions to ask when designing studies, crafting analysis plans, and choosing sample sizes.

INFERENCE MISTAKE RATES

Consider the three studies and their outcomes given in Table 2.

Table 2: Which study has strong evidence that DCA is effective?

Study	Sample Size		Mortality Rate		DCA Relative Risk [95% CI]	p
	Q-Only	Q+DCA	Q-Only	Q+DCA		
#1	140	280	13.6%	12.1%	0.89 [0.53, 1.51]	.678
#2	140	280	13.6%	7.5%	0.55 [0.31, 0.994]	.046
#3	700	1400	14.0%	11.0%	0.79 [0.62, 0.995]	.046

Because

$$\text{DCA Relative Risk} = \frac{\text{mortality rate for quinine+DCA}}{\text{mortality rate for quinine only}},$$

lower values favor DCA efficacy. Which study gives the best scientific evidence that DCA lowers mortality? Obviously not #1. But deciding between Studies #2 and #3 is not clear cut, because they have the same p-value, so they appear to have the same inferential support. That being the case, most people think #2 is the strongest result, because its relative risk of 0.55 is substantially lower than the relative risk of 0.79 found in Study #3. This is reasonable logic. But Study #3 has 5 times the sample size, so it has greater power. How should that affect your comparison?

Suppose the quinine-only mortality rate is 0.15 and the DCA relative risk is 0.67. With 140 subjects getting quinine-only and 280 getting quinine+DCA, the power using $\alpha = 0.05$ is about 33%. With 700+1400 subjects, the power is 90%. Now, in addition, suppose that Dr. Capote and his team are quite impressed with Dr. Stacpoole's work, so they are as optimistic as they can be that DCA is effective. But does that mean they have lost their ordinary scientific skepticism and already believe that DCA is affective? Again, consider Feynman (1999):

> The thing that's unusual about good scientists is that … they're not so sure of themselves as others usually are. They can live with steady doubt, think "maybe it's so" and act on that, all the time knowing it's only "maybe."

Dr. Capote's team understands that for even the most promising experimental treatments, the vast majority fail to work when tested extensively. They feel that if Dr. Capote could present to Mother Nature 1000 ideas similar to this one, she would tell him that perhaps 700 are worthless (null). What would you expect to happen if Dr. Capote ran all 1000 trials at average powers of 33% or 90%?

The breakdown of expected results appears in Table 3. If 700 of the studies are null in their infinite datasets, then 35 false positive tests (5% of 700) are expected. But with only 33% power there will be only 100 true positives. Thus, when a study is significant at $p \le 0.05$, the chance is 35/135 = 0.26 that it is a Type I error (positive inference mistake). Likewise, if $p > 0.05$, the chance is 200/865 = 0.23 that it is a Type II error (negative inference mistake). On the other hand, if the power is 90%, these inference mistake rates drop to 0.11 and 0.04. Because the purpose of statistical hypothesis testing is to get the right answer, Study #3 has stronger results than those for #2. Too few people understand that greater power gives you more assurance in making both positive and negative inferences.

Table 3: Inference mistake rates for α = 5%, powers of 33% and 90%, and under a belief that there is still a 70% chance there is no true effect.

	Outcome of Study	
N = 140 + 280	$p \le 0.05$ "significant"	$p > 0.05$ "not significant"
700 studies no true effect	5% of 700 = 35	95% of 700 = 665
300 studies average power 33%	33% of 300 = 100	67% of 300 = 200
	Positive inference mistake rate: 35/135 = 0.26	Negative inference mistake rate: 200/865 = 0.23
N = 700 + 1400		
700 studies with no true effect	5% of 700 = 35	95% of 700 = 665
300 studies average power 90%	90% of 300 = 270	10% of 300 = 30
	Positive inference mistake rate: 35/305 = 0.11	Negative inference mistake rate: 30/695 = 0.04

WILL ADJUNCTIVE THERAPY USING DCA REDUCE MORTALITY IN CHILDREN WITH SEVERE MALARIA? (PART 2)

Here is a synopsis and commentary on how the statistical planning might proceed for Dr. Capote's trial.

Position along The March of Science. There is now convincing evidence that deaths from malaria are related to lactic acidosis and that DCA is generally effective in lowering high lactate levels in the blood (Stacpoole, Nagaraja, Hutson, 2003). In a study using a rat model for malaria and lactic acidosis, Holloway, Knox, Bajaj, et al., (1995) found and that DCA increased survival when venous lactate concentrations are 5-8.9 mmol/liter (odds ratio > 2.2, P < 0.021). The animals treated with DCA had a 33% reduction in mortality at 50 hours (relative risk for DCA of 0.67). The evidence at hand supports a large-scale human trial, but Dr. Capote maintains a healthy scientific skepticism. When asked what he thinks about the hypothesis that DCA has some true positive effectiveness in children with severe malaria and lactic acidosis, he states that there is still a 50% chance that DCA is not effective. He could never be more optimistic until a major trial is completed and efficacy is soundly confirmed.

Study Design. This will be a greatly enlarged version of the trial reported by Agbenyega, Planche, Bedu-Addo, et al. (2003): a randomized, double-blind trial comparing quinine-only versus quinine+DCA, where the DCA is given in a single infusion of 50 mg/kg. The group's biostatistician, Dr. Anna Tholus, (fictitious) is aware of the fact that the Pearson chi-square test for two independent proportions has slightly more power if the sample sizes are a little unbalanced (O'Brien and Muller, 1993). Based on the cost of DCA and its excellent safety record at this dose, the WHO team is willing randomize more subjects to quinine+DCA in order to put this potentially beneficial therapy into more children on study. West African public health officials would prefer that 2/3 of the subjects get DCA, even if more subjects would need to be studied in total.

Subjects. Dr. Capote and his colleagues will study "children" with "severe malaria" who have "lactic acidosis." All of these terms will require operational definitions. The team must formulate the other inclusion/exclusion criteria and state them clearly in the protocol. They think it is feasible to study up to 2100 subjects in a single malaria season using just centers in their tropical disease research network. If needed, they can add more centers and increase this to 2700.

Primary outcome measure. Satisfactory recovery within 10 days of beginning therapy. Almost all subjects who do not recover satisfactorily within this time will have died. Time to recovery or death (i.e., survival analysis) is not a consideration.

Primary analysis. To keep this case study as simple as possible, we will limit our attention to the basic relative risk that associates treatment (quinine-only vs. quinine+DCA) with satisfactory recovery by 10 days (no or yes). For example, if 90% recovered under quinine+DCA (risk = 0.10) and 82% recovered under quinine-only (risk = 0.18), then the estimated relative risk would be 0.10/0.18 = 0.55 in favor of DCA. Two-sided p-values will be based on the ordinary Pearson chi-square statistic for association in a 2×2 contingency table. The astute reader will see that this analysis could be sharpened through the use of a logistic regression model that includes baseline measurements on lactate levels, etc. (as was done in Holloway, et al., 1995). The study will be completed in a single malaria season, so performing interim analyses is not feasible. These issues are beyond the scope of this paper.

Scenario for the infinite dataset. A prospective sample-size analysis requires the investigators to characterize the hypothetical infinite dataset for their study. Too often, sample-size analysis reports fail to explain the rationale undergirding the conjectures. If you explain little or nothing, reviewers will question the depth of your thinking and planning, and thus the scientific integrity on your proposal. Be as thorough as possible and do not apologize for having to make some sound, educated guesses. All good reviewers have done this themselves.

Concepts and Issues
- Explain the rationale undergirding your scenarios for the infinite data.
- How do you define a "clinically relevant effect?"
- Clinically relevant effects can often be too small statistically to use when powering studies.

Dr. Stacpoole's N = 62 + 62 study (Agbenyega, et al., 2003) had 8 deaths in each group, giving 95% confidence intervals of [5.7%, 23.9%] for the quinine-only mortality rate and [0.40, 2.50] for the DCA relative risk. These results are of little help in specifying the scenario. However, WHO public health statistics and epidemiologic studies in the literature indicate that about 19% of these patients do not recover by 10 days using quinine only. This figure will likely be lower for a clinical trial, because the general level of care will be better than average. Finally, the rat study had a DCA relative risk of 0.67 [95% CI: 0.44, 1.02].

Given this information, the research team conjectures that the quinine-only mortality rate is 12-15%. They agree that if DCA is effective, it is reasonable to conjecture that it will cut mortality 25-33% (relative risk of 0.67-0.75).

Topol, et al. (1999) spoke about needing sufficient power to detect "a clinically relevant difference between the experimental and control therapies." Some authors speak of designing studies to detect the smallest effect that is clinically relevant. How do you define such things? Everyone would agree that mortality reductions of 25-33% are clinically relevant. What about 15%? Even a 5% reduction in mortality would be considered very clinically relevant in a disease that kills so many people, especially because a single infusion of DCA is relatively inexpensive. Should the WHO team feel they must power this study to detect a 5% reduction in mortality? As we shall see, this is infeasible. It is usually best to ask: What do we know actually at this point? What do we think is possible? What scenario is supportable? Will the reviewers agree?

Using PROC POWER. Dr. Tholus first uses PROC POWER to discover that giving DCA to 3/5 of the patients is a near optimal allocation in terms of power. Because giving DCA to 2/3 of the patients is still favored by officials in West Africa and little power will be sacrificed, the WHO team agrees that this minor loss statistical efficiency will be offset by gains in political efficiencies at the study sites.

What about UnifyPow.sas?
- Users of UnifyPow.sas (O'Brien, 1998; www.bio.ri.ccf.org/UnifyPow) will notice similarities in syntax with PROC POWER, but POWER is *not* 'UnifyPow ported to SAS.' POWER is being developed anew by SAS Institute, and thus receives more rigorous quality control than UnifyPow ever could have had.
- In time, both authors hope that PROC POWER incorporates all of the good functionality in UnifyPow, so that RO'B can retire from freeware development and support.

The PROC POWER syntax to perform a traditional power analysis for this problem is given in Input 1a. Input 1b gives a macro that creates a custom RTF-formatted table that is suitable for inclusion into your word processor. After some minor editing (here, using Adobe® FrameMaker®), we produced Table 4In practice, it is often worthwhile for the statistician to bring a portable computer to planning meetings in order to run various scenarios "live." These "elicitation" sessions can be exciting and fun, and they help to fuse everyone around a common plan. In no way, however, should anyone be playing "statistical games" just to find some scenario that mathematically fits some pre-set sample size you want to justify. Good scientists will strive to get close to some reasonable and defensible scenarios, recognizing all the time

Concepts and Issues
- Avoid perfunctory effort and statistical gamesmanship in developing scenarios.
- The process of sample-size analysis improves all aspects of the study design and increases the proposal's chances for success.
- Content investigators make final decisions on design and sample-size choice.

that there are no perfect answers in doing this. Most importantly, going through this process forces the key personnel (including the statistician) to collaborate on a high scientific level, all before the first subject is studied.

Table 4: Power for Testing Quinine-only with Quinine+DCA

2/3 will get DCA		Alpha .01		Alpha .05	
Quinine-only Risk	DCA Relative Risk	Total N 2100	2700	Total N 2100	2700
0.12	0.75	.345	.452	.576	.681
	0.67	.634	.762	.822	.902
0.15	0.75	.452	.577	.681	.784
	0.67	.762	.871	.902	.957

.

Inputs 1c and 1d give SAS macro code that combines the alpha rates and power probabilities given in Table 4 with Dr. Capote's current subjective "null prior" probability of 0.50 that DCA is not truly effective. The result is Table 5, which displays the positive and negative inference mistake rates, which were computed as illustrated in Table 3. Dr. Tholus respects her colleagues scientific integrity and their ability to understand Tables 4 and 5, so she leaves the decision making mostly up to them. They are excited to see that all the positive inference mistake rates are below 10%, even for N = 2100. However, they are concerned that even with N = 2700, a nonsignificant study at $\alpha = 0.05$ might well be a Type II error (negative inference mistakes rates of 4-25%).

Table 5: Inference Mistake Rates for DCA Malaria Trial

Null Prior Probability: 0.50

		Alpha .010				Alpha .050			
		Total N 2100		Total N 2700		Total N 2100		Total N 2700	
Quinine-only Risk	DCA Relative Risk	Outcome Sig	Not	Outcome Sig	Not	Outcome Sig	Not	Outcome Sig	Not
0.12	0.75	.028	.398	.022	.356	.080	.308	.068	.252
	0.67	.016	.270	.013	.194	.057	.158	.053	.093
0.15	0.75	.022	.356	.017	.299	.068	.251	.060	.185
	0.67	.013	.194	.011	.115	.053	.093	.050	.043

```
proc power;
 ODS output output=PowerResults;
 TwoSampleFreq
  GroupWeights =  (1 2) /* 1 qnine only : 2
qnine+DCA */
  RefProportion /* qnine only */= .12 .15
  RelativeRisk /* qnine+DCA */ =  .75 .67
  alpha = .01 .05
  sides = 2
  Ntotal = 2100 2700
  test = pchi /*Pearson's common chi-square*/
  power = .;
 plot key=OnCurves; run;
```

Input 1a. PROC POWER for DCA Study.

What sample size would you choose? Should other analyses be run? The code for this case study is at www.bio.ri.ccf.org/robrien/OBrCast2004, so you could quickly download it and run other examples yourself.

If relative risk is 0.95. What if the WHO team presumes that DCA is clinically effective, but the relative risk may only be 0.95? Using a quinine-only risk rate of 0.15, $\alpha = 0.05$, 90% power, and a 1:2 randomization weighting, PROC POWER finds that the total sample size needs to be almost 105,000. For 80% power and a 2:3 (nearer to optimal) weighting, the study still requires over 73,000 subjects. This demonstrates why studies are rarely designed to find effects that are at the lower range of "clinically relevant."

CONCLUSION

Looking back through this paper, you will see only one equation, a simple ratio defining relative risk. In addition, the SAS code was straightforward. Yet we had an extensive discussion about matters that rely heavily on statistical theory and numerical analysis. Modern software often allows us to all but ignore the underlying mathematics of statistical methods in order to concentrate on the subject-matter at hand—in order to do statistical science. Adapting another thought by Richard Feynman, working with the equations might help some people gain understanding, but until you understand the concepts and issues of what you using those equations to do (here, for sample-size analysis) and figure out how they relate to a particular subject-matter problem, then you are just working with some equations.

We first adopt new technologies in order to be more efficient or to improve what we are producing. But as those technologies mature, they drive fundamental changes in the way we approach and solve problems. Just as modern computing has transformed the ways we do data management and analysis, so is it finally transforming how we do statistical planning. And just as data analyses have become evermore sophisticated, so will sample-size analyses. Better computing tools will allow us to better customize our sample-size analyses to fit our proposed designs and our conjectures for the underlying infinite datasets. Doing this earnestly, honestly and creatively leads to better study designs and thus better research outcomes. Accordingly, when reviewers see a strong sample-size analysis section in a research proposal, they are much more likely to conclude that the study plan has been well developed.

References

Agbenyega T, Planche T, Bedu-Addo G, Ansong D, Owusu-Ofori A, Bhattaram VA, Nagaraja NV, Shroads AL, Henderson GN, Hutson AD, Derendorf H, Krishna S, Stacpoole PW (2003), "Population Kinetics, Efficacy, and Safety of Dichloroacetate for Lactic Acidosis Due to Severe Malaria in Children," *Journal of Clinical Pharmacology,* 43, 386-396.

Feynman RP (1999), *The Pleasure of Finding Things Out,* Cambridge MA: Perseus Books.

Holloway PA, Knox K, Bajaj N, Chapman D, White NJ, O'Brien R, Stacpoole PW, Krishna S (1995), "Plasmodium Berghei Infection: Dichloroacetate Improves Survival in Rats with Lactic Acidosis," *Experimental Parasitology,* 80, 624-632.

O'Brien RG (1998), "A Tour of UnifyPow: A SAS Module/Macro for Sample-Size Analysis," in Proceedings of the SAS Users Group International (SUGI) Conference, SAS Institute (Cary, NC), pp. 1346-1355.

O'Brien RG, Muller KE (1993), "Unified Power Analysis for T-Tests through Multivariate Hypotheses," ed. Edwards L, New York: Marcel Dekker, pp. 297-344.

SAS Institute Inc. (2004), *Getting Started with the Power and Sample Size Application,* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2004), SAS/STAT® User's Guide, Version 9.1, Cary, NC: SAS Institute Inc.

Stacpoole PW, Nagaraja NV, Hutson AD (2003), "Efficacy of Dichloroacetate as a Lactate-Lowering Drug," *Journal of Clinical Pharmacology,* 43, 683-691.

Topol EJ, Califf RM, Van de Werf F, Simoons M, Hampton J, Lee KL, White H, Simes J, Armstrong PW (1997), "Perspectives on Large-Scale Cardiovascular Clinical Trials for the New Millennium," *Circulation,* 95, 1072-1082.

Zelen M (2003), "The Training of Biostatistical Scientists," *Statistics in Medicine,* 22, 3427-3430.

Acknowledgments

Ralph O'Brien's work was supported in part by Public Health Service National Center for Research Resources grant No. M01-RR018390 to the General Clinical Research Center at the Cleveland Clinic Foundation. Separately, Dr. O'Brien discloses that he serves as a contracted external consultant to The SAS Institute.

Contact information and relevant websites

The sample-size analysis tools in The SAS System and our writings on these matters are all shaped by input from others. Please send us your comments and questions.

• Ralph G. O'Brien, PhD

Department of Biostatistics and Epidemiology / WB4
Cleveland Clinic Foundation
Cleveland, OH 44195
robrien@bio.ri.ccf.org
http://www.bio.ri.ccf.org/UnifyPow

- John M. Castelloe, PhD
  SAS Institute Inc.,
  SAS Campus Drive
  Cary, NC 27513
  John.Castelloe@sas.com
  http://www.sas.com/statistics


```
%macro TablePower_RR (PowerResults=PowerResults);
*=============================================;
* Tables powers for relative risk problem.     ;
* PowerResults corresponds to ODS statement    ;
* in PROC POWER:                               ;
*    ODS OUTPUT OUTPUT=PowerResults            ;
*=============================================;

data temp; set &PowerResults;
    if Power > .999 then Power = .999;
    run;

proc tabulate data=temp format=4.3 order=data;
title2 "Power Probabilities";
  format Alpha 4.3;
  class RefProportion RelativeRisk Alpha
        sides NTotal;
    var Power;
    table RefProportion="Reference Risk"
        * RelativeRisk="Relative Risk",
      Alpha="Alpha"
        * Ntotal="Total N"
        * Power="Power"*mean=" "
          /rtspace=23;
    run;
title2;
%mend; *TablePower_RR;

ods rtf file='C:\miscSASoutput\SUGI04_Case1Power.rtf';
%TablePower_RR
ods rtf close;
```

Input 1b. TablePower_RR macro to get table of powers for relative risk problem.

```
%macro IMRate (NullPrior=, PowerResults=PowerResults,
   IMRateResults=IMRateResults);
*================================================;
* Converts Powers to Inference Mistake Rates.    ;
* NullPrior = Prob[Ho is true] (subjective).     ;
* PowerResults corresponds to ODS statement in   ;
* PROC POWER:                                    ;
*    ODS output output=PowerResults              ;
* IMRateResults is output dataset.               ;
*================================================;

data &IMRateResults; set &PowerResults;
    NullPrior = &NullPrior;
    result = "Sig";
    IMRate = NullPrior*alpha/
            (NullPrior*alpha + (1 - NullPrior)*power);
    output;
    result = "Not";
    IMRate = (1 - NullPrior)*(1-Power)/
            ((1 - NullPrior)*(1-Power)
                   + NullPrior*(1 - alpha));
    output;
    run;
%mend; *IMRate;

%IMRate (NullPrior=.50)
```

Input 1c. TablePower_RR macro to get table of powers for relative risk problem.

```
%macro TableIMRate_RR (IMRateData);
*============================================================;
* Tables Inference Mistake Rates                            ;
* IMRateData is dataset created by calling                  ;
*    %IMRate(___, ___, IMRateData)                          ;
*============================================================;

proc tabulate data=&IMRateData format=4.3 order=data;
title2 "Inference Mistake Rates (IM Rate)";
  format Alpha 4.3;
  class RefProportion RelativeRisk Alpha
        sides NullPrior result NTotal;
    var IMRate;
    table NullPrior="Null Prior Probability: ",
      RefProportion="Reference Risk"
        * RelativeRisk="Relative Risk",
      Alpha="Alpha"
        * Ntotal="Total N"
        * result="Outcome"
        * IMRate="IM Rate"*mean=" "
          /rtspace=23;
    run;
title2;
%mend; *TableIMRate_RR;


ods rtf file='C:\miscSASoutput\SUGI04_Case1IMRates.rtf';
%TableIMRate_RR (IMRateResults)
ods rtf close;
```

Input 1d. TableIMRate_RR macro to get table of inference mistake rates for relative risk problem.