Chapter 10

Sample-Size Analysis for Traditional Hypothesis Testing: Concepts and Issues

Ralph G. O'Brien
John Castelloe

Sample-size analysis continues to be transformed by ever-improving strategies, methods, and software. Using these tools intelligently depends on what the investigators understand about statistical science and what they know and conjecture about the particular research questions driving the study planning. This chapter covers only the most common type of sample-size analysis—power analysis, i.e., studying the chance that a given hypothesis test

From: Dmitrienko A, Chuang-Stein C, D'Agostino R (Ed.), *Pharmaceutical Statistics Using SAS: A Practical Guide*, Cary, NC: SAS Press, 2007, 237-271. This .pdf version was produced and distributed by Ralph O'Brien for his teaching and differs in minor ways from the book version.

Ralph O'Brien, PhD, is Professor, Quantitative Health Sciences, Cleveland Clinic, USA (OBrienRalph@gmail.com). John Castelloe, PhD, is Senior Research Scientist, SAS Institute, USA (John.Castelloe@sas.com). Comments from readers are welcome.

Date of production: 28 January 2007.

will be "statistically significant," $p \leq \alpha$. We focus on the core concepts and issues that the collaborating statistician must master and key investigators must understand.

We begin by reviewing p values and discuss how to conduct sample-size analyses that focus on the *classical* Type I and Type II error rates, α and β. Then we go further to consider two other error rates, the *crucial* Type I error rate, α^*, which is the chance that the null hypothesis is true even though $p \leq \alpha$, and the crucial Type II error rate, β^*, defined as the chance that the null hypothesis is false in some particular way even though $p > \alpha$. We argue that α^* and β^* are just as relevant (if not more so) than α and β. These issues are explored in depth through two examples stemming from a straightforward clinical trial.

10.1 Introduction

In their "Perspectives on Large-Scale Cardiovascular Clinical Trials for the New Millennium," Dr. Eric Topol and colleagues (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.

In addition, it must be stressed that investigators do not plan studies in a vacuum. They design them based on their knowledge and thoughtful conjectures about the subject matter, on results from previous studies, and on sheer speculation. They may already be far along in answering a research question, or they may be only beginning. Richard Feynman, the 1965 Nobel Laureate in Physics and self-described "curious character," stated this somewhat poetically (1999, P. 146):

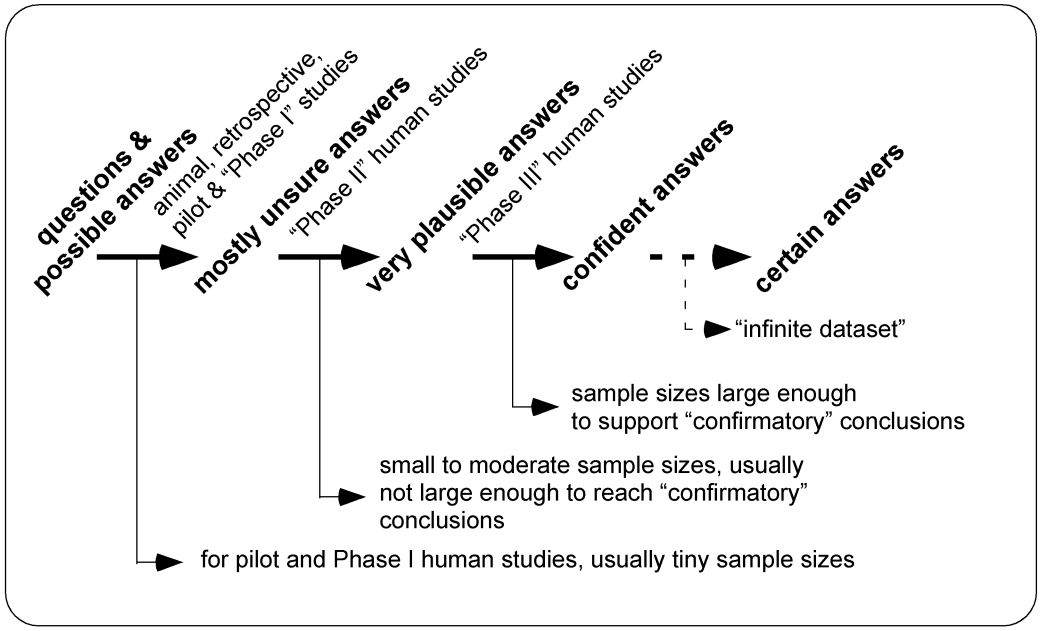
> Scientific knowledge is a body of statements of varying degrees of uncertainty,
> some mostly unsure,
> some nearly sure,
> none absolutely certain.

This reflects what we will call The March of Science, which for clinical research is sketched in Figure 10.1.

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 and more specific hypotheses, not testing them. Accordingly, little or no formal sample-size analyses may be 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 paraphrasing the New York Yankee catcher and populist sage, Yogi Berra, *Don't make the wrong mistake.*

Medical research is still dominated by traditional (frequentist) hypothesis testing and classical power analysis. Here, investigators and reviewers typically ask, What is the chance (inferential power) that some given key p-value will be significant, i.e. less than some specified Type I error rate, α? Thus, one cannot understand inferential power without

Figure 10.1 March of science in clinical research



knowing what p-values are and what they are not. Researchers rely on them to assess whether a given null hypothesis is true, but p-values are random variables, so they can mislead us into making Type I and II errors. The respective *classical* error rates are called α and $\beta = 1 - power$. All of this is reviewed in detail.

This chapter also considers other error rates that relate directly to two crucial questions that researchers should address. First, if a test turns out to be significant, what is the chance that its null hypothesis is actually true (Type I error)? A great many researchers think that this chance is at most α. They might say something like, "We will use $\alpha = 0.05$ as our level for statistical significance, so if we get a significant result, we will be more than 95% confident that the treatments are different with respect to this outcome." Researchers want to be able to make statements like this, but *this particular logic is wrong.* Likewise, if a test turns out to be non-significant, they may ask, "What is the chance that its null hypothesis is actually false (Type II error) to some particular degree?" Many researchers think this is the usual Type II error rate, β. It is not.

So, what is an appropriate way to do this? We describe something we call the *crucial* Type I error rate (here, α^*), which is the chance that the null hypothesis is true even after obtaining significance, $p \leq \alpha$. Similarly, the crucial Type II error rate (β^*) is the chance that the null hypothesis is false in some particular way even though a $p > \alpha$ result has occurred. We argue that α^* and β^* are just as relevant (if not more so) than α and β. We demonstrate how crucial error rates can be guesstimated if investigators are willing to state and justify their current belief about the chance that the null hypothesis is indeed false. Importantly, for a given α level, greater inferential power reduces both crucial error rates.

All these concepts will be developed and illustrated by carrying out a sample-size analysis for a basic two-group trial to compare two treatments for children with severe malaria: usual care only versus giving an adjuvant drug known to reduce high levels of lactic acid. Two planned analyses will be covered. The first compares the groups with respect to a binary outcome, death within the first 10 days. The second compares them on a continuous outcome, the ratio of two amino acids measured in plasma, using baseline values as covariates. The principles covered apply to any traditional statistical test being used to try to reject a null hypothesis, including analyses far more complex than those

discussed here.

While obtaining an appropriate and justifiable sample size is important, going through the analytical process itself may be just as vital in that it forces the research team to work collaboratively with the statistician to delineate and critique the rationale undergirding the study and all the components of the research protocol. The investigators must specify tight research questions, the specific research design, the various measures, and an analysis plan. They must come to agree on and justify reasonable conjectures for what the "infinite dataset" may be for their study. In essence, they must imagine how the entire study will proceed before the first subject is recruited. The "group think" on this can be invaluable.

Our reader audience includes both collaborating statisticians and content investigators. We present almost no mathematical details. While the examples given here involve clinical trials, the principles apply broadly across all of science.

The SAS procedures POWER and GLMPOWER are the primary computational engines, but we only use a small portion of their capabilities. Far more information can be found in the current SAS/STAT User's Guide.

To save space, not all SAS code and output is shown. The complete SAS code and datasets used in this book are available on the book's companion website at `http://support.sas.com/publishing/bbu/companion_site/60622.html`.

10.2 Research Question 1: Does "QCA" Reduce Mortality in Children with Severe Malaria?

According to a report released in 2003 by the World Health Organization, malaria remains one of the world's foremost health problems, killing at least one million people annually, mostly children under five years old in sub-Saharan Africa. Lactic acidosis (toxic levels of lactic acid in the blood) is a frequent complication in severe malaria and is an incremental statistical predictor ("independent risk factor") of death. Moreover, a plausible biological rationale supports the hypothesis that lactic acidosis is a contributing cause of death.

Dr. Peter Stacpoole of the University of Florida has spent decades investigating the safety and efficacy of dichloroacetate (DCA) for treating lactic acidosis in genetic and acquired diseases. In 1997-99, he collaborated with Dr. Sanjeev Krishna of the University of London to lead a team that conducted a small, randomized, double-blind, controlled trial of quinine-only versus quinine+DCA in treating lactic acidosis in Ghanaian children with severe malaria (Agbenyega et al., 2003). 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. 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.

From now on the story is fictionalized. Suppose "quadchloroacetate" (QCA) has the same molecular structure as DCA at the active biological site, and has now been shown in large animal and human studies to be clinically equivalent to DCA in quickly reducing abnormally high blood lactate levels. However, QCA is less expensive to produce (about US$1/dose) and has a longer shelf-life, especially in tropical climates.

"Dr. Sol Capote" heads the malaria research group at "Children's Health International (CHI)," and he and his colleagues are now designing a large clinical trial to be coordinated from "Jamkatnia" in West Africa. Dr. Capote is an experienced investigator, so he knows that substantial thought, effort, and experience must go into developing the sample-size analysis and the rest of the statistical considerations.

The CHI study will use a randomized, double-blind design to compare usual care only (UCO) versus usual care plus a single dose of QCA. After reviewing all previous human studies of both DCA and QCA, the CHI team is convinced that a single dose of QCA is very likely to be safe. Accordingly, after consulting with Jamkatnian health officials and a

bioethicist, they decide that 2/3 of the subjects should get QCA.

10.3 p-Values, α, β and Power

The primary efficacy analysis will yield a p-value that compares the mortality rates of control versus QCA. Smaller p-values indicate greater statistical separation between the two samples, but *how that p-value is determined is an issue that is not critical to understanding the essential concepts in sample-size analysis.* In this case, that p-value may come from one of the many methods to compare two independent proportions, including the likelihood ratio chi-square test, as used here, 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 p is small enough ("significant") and the QCA mortality rates are better, Dr. Capote will report that the study supported the hypothesis that QCA 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.

10.3.1 Null and Non-Null Distributions of p-Values; Type I and Type II Errors

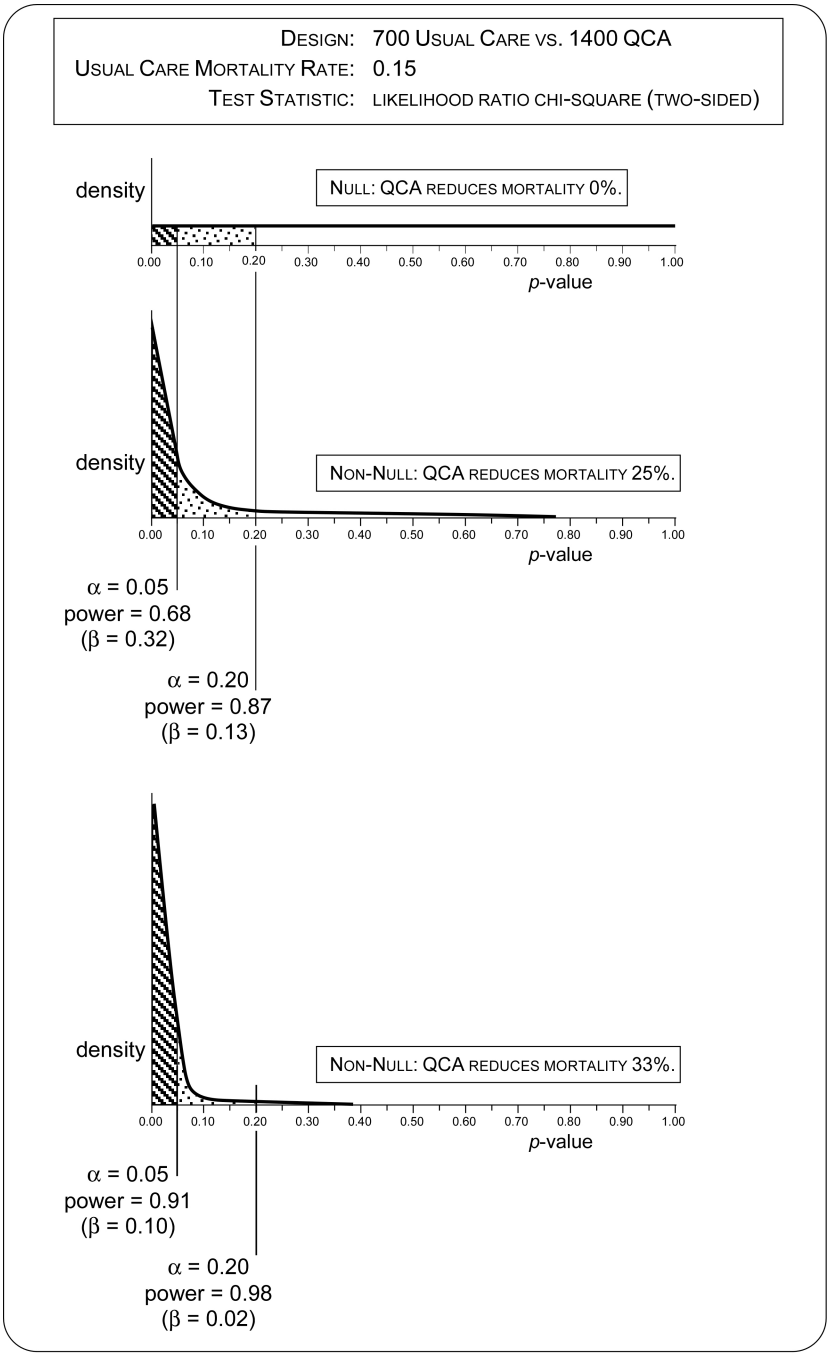
Dr. Capote's quest here is to answer: Does QCA decrease mortality in children with severe malaria? While Mother Nature knows the correct answer, only if we mortals in science were able to gather an infinitely large, perfectly clean dataset could we figure this out ourselves. Rather, we must design a study or, usually, a series of studies, that will yield sample datasets that give us a solid chance of inferring what Mother Nature knows. Unfortunately, Lady Luck builds randomness into those sample datasets, and thus even the best studies can deliver misleading answers.

Please study the top distribution in Figure 10.2. Here, there is no difference between the two groups' mortality in the infinite dataset, so *regardless of the sample size*, all values for $0 < p < 1$ are equally likely. Accordingly, there is a 5% chance that $p \leq 0.05$, or a $100\alpha\%$ chance that $p \leq \alpha$ (in practice, these percentages are rarely exact, because the data are discrete or they fail to perfectly meet the test's underlying mathematical assumptions). If there is no true effect but $p \leq \alpha$ indicates otherwise, this triggers a Type I error, which is why α is called the Type I error rate. α should be chosen after some thought; it should not be automatically set at 0.05.

What if QCA has some true effect, good or bad? Then the non-null (non-central) distribution of the p-value will be skewed towards 0.0, as in the middle and bottom plots of Figure 10.2. The middle one comes from presuming (1) true mortality rates of 0.28 for UCO and 0.21 for QCA, which is a 25% reduction in mortality; (2) 700 patients randomized to UCO versus 1400 to QCA, and (3) the p-value arises from testing whether the two mortality proportions differ (non-directional) using the likelihood ratio chi-square statistic. The bottom plot conforms to presuming that QCA cuts mortality by 33%.

Inferential power is the chance that $p \leq \alpha$ when the null hypothesis is false, which is why α could be called the 'null power.' If there is some true effect, but $p > \alpha$, then a Type II error is triggered. Consider the middle plot, which is based on a 25% reduction in mortality. Using the common Type I error rate, $\alpha = 0.05$, the power is 0.68, so the Type II error rate is $\beta = 0.32$. By tolerating a higher α-level, one can increase power (decrease β). Here, using $\alpha = 0.20$. the power is 0.87, so $\beta = 0.13$. If QCA is more effective (bottom plot, 33% reduction in mortality), then the power rises to 0.91 with $\alpha = 0.05$ and 0.98 with $\alpha = 0.20$. Again, we will never know the true power, because Mother Nature will never tell us the true mortality rates in the two groups, and Lady Luck will always add some natural randomness into our outcome data.

Figure 10.2 Distribution of the p-value under the null hypothesis and two non-null scenarios



10.3.2 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 generally agree, but stress that there should be no standard power threshold that is worshiped blindly as being satisfactory across all situations. Why do so many people routinely perform power analyses using $\alpha = 0.05$ and 80% power ($\beta = 0.20$)? Rarely do they give it any firm thought.

Consider the middle plot in Figure 10.2. We could achieve a much better Type II error rate of 13% if we are willing to accept a substantially greater Type I error rate of 20%.

Investigators should seek to obtain α versus β values that are in line with the consequences of making a Type I error versus a Type II error. In some cases, making a Type II error may be far more costly than making a Type I error. In particular, in the early stages of the March of Science, making a Type I error may only extend the research to another stage. This is undesirable, of course, as are all mistaken inferences in science, but making a Type II error may be far more problematic, because it may halt a line of research that would ultimately be successful. So it might be justified to use $\alpha = 0.20$ (maybe more) in order to reduce β as much as possible. Using such high α values is not standard, so investigators adopting this philosophy must be convincing in their argument.

10.4 A Classical Power Analysis

Dr. Capote and his team plan their trial as follows.

Study Design

This trial follows the small (N = 62 + 62) DCA trial reported by Agbenyega, (2003). It will be double-blind, but instead of a 1:1 allocation, the team would like to consider giving one patient usual care only for every two patients that get QCA, where the QCA is given in a single infusion of 50 mg/kg.

Subjects

Study patients will be less than 13 years old with severe malaria complicated by lactic acidosis. "Untreatable" cases (nearly certain to die) will be excluded. These terms will require operational definitions and the CHI team will 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 Jamkatnia. If needed, they can add more centers in neighboring "Gabrieland" and increase the total size to 2700. Drop-outs should not be a problem in this study, but all studies must consider this and enlarge recruitment plans accordingly.

Primary Efficacy Outcome Measure

Death before Day 10 after beginning therapy. Almost all subjects who survive to Day 10 will have fully recovered. Time to death (i.e., survival analysis) is not a consideration.

Primary Analysis

To keep this story and example relatively simple, we will limit our attention to the basic relative risk that associates treatment group (UCO vs. QCA) with death (no or yes). For example, if 10% died under QCA and 18% died under UCO, then the estimated relative risk would be $0.10/0.18 = 0.55$ in favor of QCA. p-values will be based on the likelihood ratio chi-square statistic for association in a 2×2 contingency table. The group's biostatistician, "Dr. Phynd Gooden," knows that the test of the treatment comparison could be made with greater power through the use of a logistic regression model that includes baseline measurements such as a severity score or lactate levels, etc. (as was done in Holloway et al., 1995). In addition, this study will be completed in a single malaria season, so performing interim analyses is not feasible. These issues are beyond the scope of this chapter.

10.4.1 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 we explain little or nothing, reviewers will question the depth of our thinking and planning, and thus the scientific integrity of our

proposal. Be as thorough as possible and do not apologize for having to make some sound guesstimates. All experienced reviewers have had to do this themselves.

Dr. Stacpoole's $N = 62 + 62$ human study (Agbenyega et al., 2003) had 8 deaths in each group. This yields 95% confidence intervals of [5.7%, 23.9%] for the quinine-only mortality rate (using the "EXACT BINOMIAL" method in PROC FREQ) and [0.40, 2.50] for the DCA relative risk (using the asymptotic "RELRISK" method in PROC FREQ). These wide intervals are of little help in specifying the scenario. However, CHI public health statistics and epidemiologic studies in the literature indicate that about 19% of these patients die within 10 days using quinine only. This figure will likely be lower for a clinical trial, because untreatable cases are being excluded and the general level of care could be much better than is typical. Finally, the Holloway et al. (1995) rat study obtained a DCA relative risk of 0.67 [95% CI: 0.44, 1.02], and the odds ratios adjusting for baseline covariates were somewhat more impressive, e.g., OR= 0.46 (one-sided $p = 0.021$).

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

Topol et al. (1999) wrote 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 we 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 annually, especially because a single infusion of QCA is relatively inexpensive. Should the CHI 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 actually know at this point? What do we think is possible? What scenarios are supportable? Will the reviewers agree with us?

10.4.2 What Allocation Ratio? One-Sided or Two-Sided Test?

Dr. Gooden is aware of the fact that the likelihood ratio chi-square test for two independent proportions can be more powerful when the sample sizes are unbalanced. Her first task is to assess how the planned 1:2 (UCO: QCA) allocation ratio affects the power. As shown in Program 10.1, this is relatively easy to do in PROC POWER. Its syntax is literal enough that we will not explain it, but note particularly the GROUPWEIGHTS statement.

Table 10.1 Effect of the Allocation Ratio, $N_{UCO} : N_{QCA}$, on Power, β, and Sample Size for Two-Sided $\alpha = 0.05$ Assuming 15% Mortality with Usual Care and a Relative Risk of 0.67 in Favor of QCA

		Allocation ratio ($N_{UCO} : N_{QCA}$)			
		1:1	2:3	1:2	1:3
$N_{total} = 2100$	Power	0.930	0.923	0.905	0.855
	β	0.070	0.077	0.095	0.145
	Relative Type II risk ratio	1.00	1.10	1.36	2.07
Power = 0.90	N_{total}	1870	1925	2064	2420
	Relative efficiency	1.00	0.97	0.91	0.77
	Relative inefficiency	1.00	1.03	1.10	1.29

Table 10.1 displays results obtained using Program 10.1 and some simple further computations. For this conjecture of 15% mortality versus $0.67 \times 15\%$ mortality, the most efficient of these four designs is the 1:1 allocation ratio. It has a power of 0.930 or $\beta = 0.070$ with $N_{total} = 2100$ ($\alpha = 0.05$), and to get a 0.90 power requires $N_{total} = 1870$. Compared to the 1:1 design, the 1:2 design has a 36% larger Type II error rate ("relative Type II risk ratio") at $N_{total} = 2100$ and requires 2064 subjects to achieve a 0.90 power. Thus, the 1:2 design has a relative efficiency of $1870/2064 = 0.91$ and requires about 10%

more subjects to achieve 0.90 power (relative inefficiency: 2064/1870 = 1.10). The relative inefficiencies for the 2:3 and 1:3 designs are 1.03 and 1.29 respectively.

Program 10.1 Compare allocation weights

```
* Powers at Ntotal=2100;
proc power;
    TwoSampleFreq
    GroupWeights =  (1 1) (2 3) (1 2) (1 3)  /* UCO:QCA */
    RefProportion = .15  /* Usual Care Only (UCO) mortality rate*/
    RelativeRisk =  .67   /* QCA mortality vs. UCO mortality */
    alpha = .05
    sides = 1 2
    Ntotal = 2100
    test =  lrchi
    power = .;

* Ntotal values for power = 0.90;
proc power;
    TwoSampleFreq
    GroupWeights =  (1 1) (2 3) (1 2) (1 3)/* UCO:QCA */
    RefProportion = .15  /* Usual Care Only (UCO) mortality rate*/
    RelativeRisk =  .67   /* QCA mortality vs. UCO mortality */
    alpha = .05
    sides = 1 2
    Ntotal = .
    test =  lrchi
    power = .90;
    run;
```

Note that Dr. Gooden uses SIDES=1 2 in Program 10.1 to consider both one-sided and two-sided tests. Investigators and reviewers too often dogmatically call for two-sided tests only because they believe using one-sided tests is not trustworthy. But being good scientists, Dr. Capote's team members think carefully about this issue. Some argue that the scientific question is simply whether QCA is efficacious versus whether it is not efficacious, where "not efficacious" means that QCA has no effect on mortality or it increases mortality. This conforms to the one-sided test. For the design, scenario, and analysis being considered here, the one-sided test requires 1683 subjects versus 2064 for the two-sided test, giving the two-sided test a relative inefficiency of 1.23. At $N = 2100$, the Type II error rate for the one-sided test is $\beta = 0.052$, which is 45% less than the two-sided rate of $\beta = 0.095$. On the other hand, other members argue that it is important to assess whether QCA increases mortality. If it does, then the effective Type II error rate for the one-sided test is 1.00. This logic causes many to never view one-sided tests favorably under any circumstances. After considering these issues with Dr. Gooden, Dr. Capote decides to take the traditional approach and use a two-sided test.

For some endpoints, such as for rare adverse events or in trials involving rare diseases, the argument in favor of performing one-sided tests is often compelling. Suppose there is some fear that a potential new treatment for arthritis relief could increase the risk of gastrointestinal bleeding in some pre-specified at-risk subpopulation, say raising this from an incidence rate in the first 30 days from 8% to 24%, a relative risk of 3.0. A balanced two-arm trial with $N = 450 + 450$ subjects may be well powered for testing efficacy (arthritis relief), but suppose the at-risk group is only 20% of the population being sampled, so that only about $N = 90 + 90$ will be available for this planned sub-group analysis. Using $\alpha = 0.05$, the likelihood ratio test for comparing two independent proportions will provide 0.847 power for the two-sided test and 0.910 power for the

one-sided test. Thus, using a one-sided test cuts the Type II error rate from 0.153 to 0.090, a 41% reduction. Stated differently, using a two-sided test increases β by 70%. However, if this research aim is only concerned with detecting an increase in GI bleeding, why not use the statistical hypothesis—the one-sided version—that conforms to that aim? If using the two-sided test increases the Type II error rate by 70%, why is that more trustworthy?

For completeness, and because it takes so little time to do, Dr. Gooden also uses PROC POWER to find the approximate optimal allocation ratio. After iterating the group weights, she settles on using Program 10.2 to show that while the theoretical optimal is approximately 0.485:0.515, the balanced (0.500:0.500) design has almost the same efficiency.

Program 10.2 Find optimal allocation weights

```
proc power;
    TwoSampleFreq
    GroupWeights =       /* UCO : QCA */
    (.50 .50) (.49 .51) (.485  .515) (.48 .52) (.45 .55) (.33 .66)
    RefProportion = .15  /* Usual Care Only (UCO) mortality rate*/
    RelativeRisk =  .67   /* QCA mortality vs. UCO mortality */
    alpha = .05
    sides = 2
    Ntotal = .
    test =  LRchi /* likelihood ratio chi-square */
    power = .90
    nfractional;
    run;
```

Output from Program 10.2

Index	Weight1	Weight2	Fractional N Total	Actual Power	Ceiling N Total
1	0.500	0.500	1868.510571	0.900	1869
2	0.490	0.510	1867.133078	0.900	1868
3	0.485	0.515	1867.002923	0.900	1868
4	0.480	0.520	1867.245653	0.900	1868
5	0.450	0.550	1876.616633	0.900	1877
6	0.330	0.660	2061.667869	0.900	2062

Should the study use the less efficient 1:2 design? After substantial debate within his team, Dr. Capote decides that the non-statistical attributes of the 1:2 design give it more "practical" power than the 1:1 design. First, nobody has safety concerns about giving a single dose of QCA. Second, Jamkatnian health officials and parents will prefer hearing that 2 out of 3 subjects will be treated with something that could be life saving for some. Third, the extra cost associated with a 10% increase in the sample size is not prohibitive. Given that this study's set-up costs are high and the costs associated with data analysis and reporting are unaffected by the sample size, the total cost will only increase about 3%.

10.4.3 Obtaining and Tabling the Powers

The stage is now set to carry out and report the power analysis. Please examine Program 10.3 together with Output 10.3, which contains the essential part of the results. Version 9.1 of PROC POWER provides plain graphical displays of the results (not shown here), but lacks corresponding table displays. As this chapter was going to press, a general-purpose SAS macro, %powtable, was being developed to help meet this need; see the book's website. Here, Dr. Gooden uses the ODS OUTPUT command and PROC TABULATE to create a basic table. The reader may devise a better method.

Program 10.3 Power analysis for comparing mortality rates

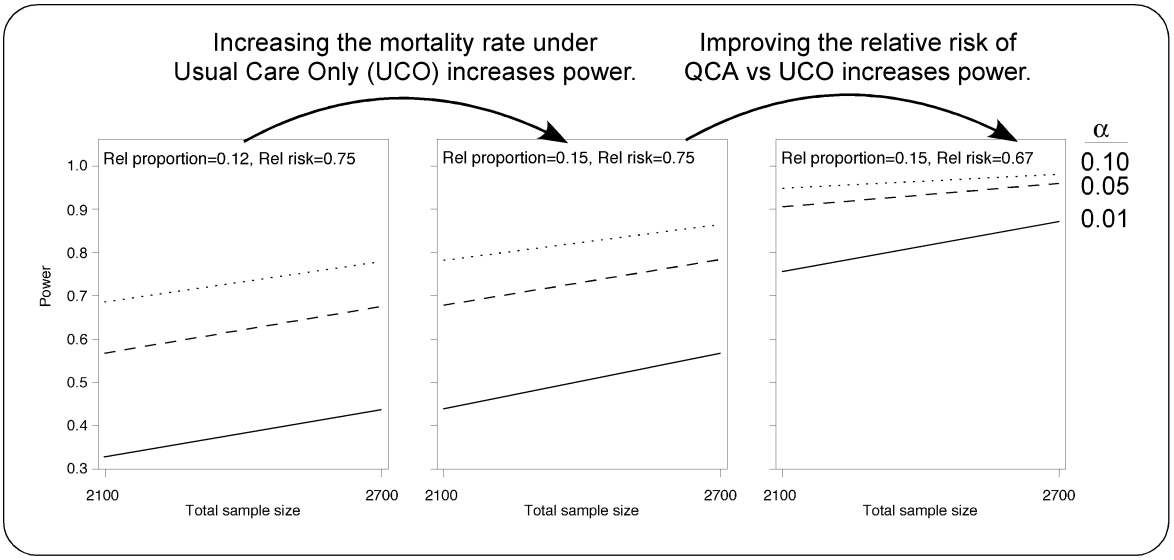
```
options ls=80 nocenter FORMCHAR="|----|+|---+=|-/\<>*";

proc power;
    ODS output output=MortalityPowers;
    TwoSampleFreq
    GroupWeights =  (1 2) /* 1 UCO : 2 QCA*/
    RefProportion = .12 .15 /* UCO mortality rate */
    RelativeRisk =  .75 .67  /* QCA rate vs UCO rate*/
    alpha = .01 .05 .10
    sides = 2
    Ntotal = 2100 2700
    test = LRchi  /* likelihood ratio chi-square */
    power = .;
    plot vary (panel by RefProportion RelativeRisk);

/* Avoid powers of 1.00 in table */
data MortalityPowers;
    set MortalityPowers;
    if power>0.999 then power999=0.999;
      else power999=power;

proc tabulate data=MortalityPowers format=4.3 order=data;
    format Alpha 4.3;
    class RefProportion RelativeRisk alpha NTotal;
    var Power999;
    table
    RefProportion="Usual Care Mortality"
        * RelativeRisk="QCA Relative Risk",
    alpha="Alpha"
        * Ntotal="Total N"
        * Power999=""*mean=" "/rtspace=28;
    run;
```

Output from Program 10.3

		Alpha		
		.010	.050	.100
		Total N	Total N	Total N
		2100	2700	2100	2700	2100	2700
Usual Care Mortality	QCA Relative Risk						
0.12	0.75	.329	.437	.569	.677	.687	.780
	0.67	.622	.757	.823	.905	.893	.948
0.15	0.75	.438	.566	.677	.783	.781	.864
	0.67	.757	.872	.905	.960	.948	.981

Figure 10.3 juxtaposes three plots that were produced by using the same ODS output dataset, MortalityPowers, with a SAS/GRAPH program not given here, but which is available at this book's website. This shows concretely how power increases for larger UCO mortality rates or better (smaller) relative risks for QCA versus UCO.

Figure 10.3 Plots for the mortality analysis showing how changing the reference proportion or the relative risk rate affects power.



Colleagues and reviewers should have little trouble understanding and interpreting the powers displayed as per Output 10.3. If the goal is to have 0.90 power using $\alpha = 0.05$, then $N = 2100$ will only suffice under the most optimistic scenario considered, that is, if the usual care mortality rate is 15% and QCA reduces that risk 33%. $N = 2700$ seems to be required to assure adequate power over most of the conjecture space.

Tables like this are valuable for teaching central concepts in traditional hypothesis testing. One can see with concrete numbers how power is affected by various factors. While we can set N and α, Mother Nature sets the mortality rate for usual care and the relative risk associated with QCA efficacy.

Let us return to the phrase from Topol et al. (1997) that called for clinical trials to have adequate power "to detect a clinically relevant difference between the experimental and control therapies." With respect to our malaria study, most people would agree that even a true 5% reduction in mortality is clinically relevant and it would also be economically justifiable to provide QCA treatment given its low cost and probable safety. But could we detect such a small effect in this study? If QCA reduces mortality from 15% to $0.95 \times 15\% = 14.25\%$, then the proposed design with $N = 900 + 1800$ only has 0.08 power (two-sided $\alpha = 0.05$). In fact, under this 1:2 design and scenario, it will require almost 104,700 patients to provide 0.90 power. This exemplifies why confirmatory trials (Phase III) are usually designed to detect plausible outcome differences that are considerably larger than "clinically relevant." The plausibility of a given scenario is based on biological principles and from data gathered in previous relevant studies of all kinds and qualities. By ordinary human nature, investigators and statisticians tend to be overly optimistic in guesstimating what Mother Nature might have set forth, and this causes our studies to be underpowered. This problem is particularly relevant when new therapies are tested against existing therapies that might be quite effective already. It is often the case that potentially small but important improvements in therapies can only be reliably assessed in very large trials. Biostatisticians are unwelcome and even sometimes disdained when they bring this

news, but they did not make the Fundamental Laws of Chance—they are only charged with policing and adjudicating them.

10.5 Beyond α and β: Crucial Type I and Type II Error Rates

Are α and β (or power$= 1 - \beta$) the only good ways to quantify the risk of making Type I and Type II errors? While they may be the classical rates to consider and report, they fail to directly address two fundamental questions:

- If the trial yields traditional statistical significance ($p \leq \alpha$), what is the chance this will be an incorrect inference?
- If the trial does not yield traditional statistical significance ($p > \alpha$), what is the chance this will be an incorrect inference?

To answer these in some reasonable way, we need to go beyond using just α and β.

10.5.1 A Little Quiz: Which Study Provides the Strongest Evidence?

Table 10.2 summarizes outcomes from three possible QCA trials. Which study has the strongest evidence that QCA is effective? Studies #1 and #2 have $N = 150 + 300$ subjects, whereas #3 has $N = 700 + 1400$ subjects. Studies #1 and #3 have identical 0.79 estimates of relative risk, but with $p = 0.36$, Study #1 does not adequately support QCA efficacy. Choosing between Studies #2 and #3 is harder. They have the same p-value, so many people would argue that they have the same inferential support. If so, then #2 is the strongest result, because its relative risk of 0.57 is substantially lower than the relative risk of 0.79 found in Study #3. However, Study #3 has nearly 5 times the sample size, so it has greater power. How should that affect our assessment?

Table 10.2 Which Study Has the Strongest Evidence that QCA Is Effective?

	Deaths/N		Mortality		Relative risk		LR test
Study	UCO	QCA	UCO	QCA	RR	[95% CI]	p-value
#1	21/150	33/300	14.0%	11.0%	0.79	[0.47, 1.31]	0.36
#2	21/150	24/300	14.0%	8.0%	0.57	[0.33, 0.992]	0.05
#3	98/700	154/1400	14.0%	11.0%	0.79	[0.62, 0.995]	0.05

10.5.2 To Answer the Quiz: Compare the Studies' Crucial Error Rates

Suppose that Mother Nature has set the true usual care mortality rate at 0.15 and the QCA relative risk at 0.67, the most powerful scenario we considered above. We have already seen (Figure 10.3, Output 10.3) that with $N = 700 + 1400$ subjects and using $\alpha = 0.05$ (two-sided), the power is 90%. With 150 subjects getting usual care and 300 getting QCA, the power is only about 33%.

Now, in addition, suppose that Dr. Capote and his team are quite optimistic that QCA is effective. This does not mean they have lost their ordinary scientific skepticism and already believe that QCA is effective. Consider another Feynman-ism (1999, P. 200):

> 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 even for the most promising experimental treatments, the clear majority fail to work when tested extensively. In fact, Lee and Zelen (2000) estimated that among 87 trials completed and reported out by the Eastern

Cooperative Oncology Group at Harvard from 1980-1995, only about 30% seem to have been testing therapies that had some clinical efficacy.

Let us suppose that Dr. Capote's team conducted 1000 independent trials looking for significant treatment effects, but Mother Nature had set things up so that 700 effects were actually null. What would we expect to happen if Dr. Capote ran all 1000 trials at average powers of 33%? 90%? Table 10.3 presents some straightforward computations that illustrate what we call the *crucial* Type I and Type II error rates. With 700 null tests, we would expect to get 35 (5%) Type I errors (false positives). From the 300 non-null hypotheses tested with 33% power, we would expect to get 99 true positives. Thus, each "significant" test ($p \leq 0.05$) has an $\alpha^* = 35/134 = 0.26$ chance of being misleading. Note how much larger this is than $\alpha = 0.05$. Some people (including authors of successful statistics books) confuse α and α^*, and hence they also misinterpret what p-values are. A p-value of 0.032 does *not* imply that there is a 0.032 chance that the null hypothesis is true.

Table 10.3 Expected Results for 1000 Tests Run at $\alpha = 0.05$. The true hypothesis is null in 700 tests. For the 300 non-null tests, the average power is 33% or 90%.

	Result of hypothesis test	
	$p \leq 0.05$ ("significant")	$p > 0.05$ ("not significant")
33% average power		
700 true null	5% of 700 = 35	95% of 700 = 665
300 true non-null	33% of 300 = 99	67% of 300 = 201
	Crucial Type I error rate: $\alpha^* = 35/134 =$ **0.26**	Crucial Type II error rate: $\beta^* = 201/866 =$ **0.23**
90% average power		
700 true null	5% of 700 = 35	95% of 700 = 665
300 true non-null	90% of 300 = 270	10% of 300 = 30
	Crucial Type I error rate: $\alpha^* = 35/305 =$ **0.11**	Crucial Type II error rate: $\beta^* = 30/695 =$ **0.04**

The crucial Type II error rate, β^*, is defined similarly. With 33% power, we would expect to get 201 Type II errors (false negatives) to go with 665 true negatives; thus $\beta^* = 210/866 = 0.23$. Note that this is not equal to $\beta = 0.67$.

10.5.3 Greater Power Reduces Both Types of Crucial Error Rates

A key point illustrated in Table 10.3 is that *greater power reduces both types of crucial error rates*. In other words, statistical inferences are generally more trustworthy when the underlying power is greater. Let us return to Table 10.2. Again, which study has the strongest evidence that QCA is effective? Even under our most powerful scenario, a $p \leq 0.05$ result has a 0.26 chance of being misleading when using $N = 150 + 300$, as per Study #2. This falls to 0.11 using $N = 700 + 1400$ (Study #3). Both studies may have yielded $p = 0.05$, but they do not provide the same level of support for inferring that QCA is effective. Study #3 provides the strongest evidence that QCA has *some* degree of efficacy. This concept is poorly understood throughout all of science.

10.5.4 The March of Science and Sample-Size Analysis

Consistent with Lee and Zelen (2000), we think that investigators designing clinical trials are well served by considering α^* and β^*. (Note that Lee and Zelen's definition is reversed from ours in that our α^* and β^* correspond to their β^* and α^*, respectively.) Ioannidis (2005b) used the same logic in arguing "why most published research findings are false."

Wacholder et al. (2004) described the same methodology to more carefully infer whether a genetic variant is really associated with a disease. Their "false positive report probability" is identical to α^*. Also, readers familiar with accuracy statistics for medical tests will see that $1-\alpha$ and β are isomorphic to the specificity and sensitivity of the diagnostic method and $1-\alpha^*$ and $1-\beta^*$ are isomorphic with the positive and negative predictive values.

Formally, let γ be the probability that the null hypothesis is false. We like to think of γ as measuring where the state of knowledge currently is in terms of confirming the non-null hypothesis; in short, its location along its March of Science (Figure 10.1). Thus, for novel research hypotheses, γ will be nearer to 0. For mature hypotheses that stand ripe for solid confirmation with say, a large Phase III trial, γ will be markedly greater than 0. One might regard $\gamma = 0.5$ as scientific equipoise, saying that the hypothesis is halfway along its path to absolute confirmation in that we consider the null and non-null hypothesis as equally viable. Lee and Zelen's calculations put γ around 0.3 for Phase III trials coordinated in the Eastern Cooperative Oncology Group.

Given γ, α and some β set by some particular design, sample size, and non-null scenario, we can apply Bayes' Theorem to get

$$\alpha^* = \text{Prob}[\text{H}_0 \text{ true} \mid p \leq \alpha] = \frac{\alpha(1-\gamma)}{\alpha(1-\gamma)+(1-\beta)\gamma}$$

and

$$\beta^* = \text{Prob}[\text{H}_0 \text{ false} \mid p > \alpha] = \frac{\beta\gamma}{\beta\gamma+(1-\alpha)(1-\gamma)}.$$

To be precise, "H_0 false" really means "H_0 false, as conjectured in some specific manner." For the example illustrated first in Table 10.3, we have $\gamma = 0.30$, $\alpha = 0.05$ and $\beta = 0.67$, thus

$$\alpha^* = \frac{(0.05)(1-0.30)}{(0.05)(1-0.30)+(1-0.67)(0.30)} = 0.261$$

and

$$\beta^* = \frac{(0.67)(0.30)}{(0.67)(0.30)+(1-0.05)(1-0.30)} = 0.232.$$

In Bayesian terminology, $\gamma = 0.3$ is the prior probability that QCA is effective, and $1-\alpha^* = 0.739$ is the posterior probability given that $p \leq \alpha$. However, nothing here involves Bayesian *data analysis* methods, which have much to offer in clinical research, but are not germane to this chapter. Some people are bothered by the subjectivity involved in specifying prior probabilities like γ, but we counter by pointing out that there are many other subjectivities involved in sample-size analysis for study planning, especially the conjectures made in defining the infinite dataset. Indeed, we find that most investigators are comfortable specifying γ, at least with a range of values, and that computing various α^* and β^* values of interest gives them much better insights into the true inferential strength of their proposed (frequentist) analyses.

10.6 Research Question 1, Continued: Crucial Error Rates for Mortality Analysis

In developing the statistical considerations for the QCA/malaria trial, Dr. Gooden understands the value in assessing its crucial Type I and Type II error rates, and she presses her CHI colleagues to complete the exercise faithfully. As mentioned before, they

Program 10.4 Compute crucial error rates for mortality endpoint

```
options ls=80 nocenter FORMCHAR="|----|+|---+=|-/\<>*";

proc power;
  ODS output output=MortalityPowers;
  TwoSampleFreq
    GroupWeights =  (1 2)    /* 1 UCO for every 2 QCA  */
    RefProportion = .12 .15  /* UCO mortality rate     */
    RelativeRisk =  .75 .67  /* QCA rate vs UCO rate   */
    alpha = .01 .05
    sides = 2
    Ntotal = 2700
    test = LRchi  /* likelihood ratio chi-square */
    power = .;
  plot key=OnCurves;
  run;

* Call %CrucialRates macro, given in Appendix B of this chapter;
%CrucialRates(   PriorPNullFalse = .30 .50,
                 Powers = MortalityPowers,
                 CrucialErrRates = MortalityCrucRates   )

proc tabulate data=MortalityCrucRates format=4.3 order=data;
title3 "Crucial Error Rates for QCA Malaria Trial";
  format alpha 4.3;
  class RefProportion RelativeRisk alpha gamma TypeError NTotal;
  var CrucialRate;
  table
    Ntotal="Total N: ",
    RefProportion="Usual Care Mortality"
      * RelativeRisk="QCA Relative Risk",
    alpha="Alpha"
      * gamma="PriorP[Null False]"
      * TypeError="Crucial Error Rate"
      * CrucialRate=""*mean=" "
  / rtspace = 26;
  run;
```

are optimistic that QCA is effective, but to compute crucial error rates, they must now quantify that optimism by setting γ. Initial discussions place γ near 0.75, but the 0.30 value reported by Lee and Zelen (2000) tempers their thinking substantially. They come to settle on $\gamma = 0.50$. Dr. Gooden will also use $\gamma = 0.30$.

Program 10.4 gives the code to handle this. First, a more focused version of Program 10.3 computes the powers. Second, a macro called %CrucialRates (given in Appendix B and available on the book's web site) converts the PROC POWER results into crucial Type I and Type II error rates. Finally, PROC TABULATE organizes these crucial rates effectively.

Output 10.4 shows that the most optimistic case considered here presumes that the mortality rate is 0.15 under usual care and it takes QCA to have a prior probability of $\gamma = 0.50$ of being effective, where "effective" is a QCA relative risk of 0.67. If so, then by using $\alpha = 0.05$, the crucial Type I and Type II error rates are $\alpha^* = 0.050$ and $\beta^* = 0.040$, respectively, which seem very good. However, for $\alpha = 0.01$, β^* rises to 0.115. Now consider the most pessimistic case. If $\gamma = 0.30$ and the non-null scenario has a mortality rate of 0.12 under usual care and a QCA relative risk is 0.75, then using $\alpha = 0.05$ gives $\alpha^* = 0.147$ and $\beta^* = 0.127$. The team from Children's Health International decides that they can tolerate these values and thus planning continues around $N = 2700$.

After going through this process, Dr. Capote remarks that if all clinical trial protocols were vetted in this manner, a great many of them would show crucial Type I and Type II error rates that would severely temper any inferences that can be made. This is true.

Output from Program 10.4

Total N: 2700

		Alpha			
		.010		.050	
		PriorP[Null False]		PriorP[Null False]	
		0.3	0.5	0.3	0.5
		Crucial Error Rate	Crucial Error Rate	Crucial Error Rate	Crucial Error Rate
		Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Usual Care Mortality	QCA Relative Risk								
0.12	0.75	.051	.196	.022	.362	.147	.127	.069	.254
	0.67	.030	.095	.013	.197	.114	.041	.052	.091
0.15	0.75	.040	.158	.017	.305	.130	.089	.060	.186
	0.67	.026	.053	.011	.115	.108	.018	.050	.040

10.7 Research Question 2: Does "QCA" Affect "Elysemine:Elysemate" Ratio (EER)?

This section expands Dr. Capote's planning to consider a test that compares the UCO and QCA arms with respect to a continuous outcome, adjusted for baseline covariates. PROC GLMPOWER is used to perform the calculations.

10.7.1 Rationale Behind the Research Question

Now the team turns to investigating potential adverse effects.

A descriptive analysis being completed in Jamkatnia has compared 34 children with severe malaria with 42 healthy children on some 27 measures related to metabolic functioning, including two amino acids, "elysemine" and "elysemate" (both fictitious). Elysemine is synthesized by the body from elysemate, which is abundant in food grains and meat. Phagocytes (a type of white blood cell) need elysemine to fight infection. Low plasma elysemine levels have been shown to be an incremental risk factor for death in critically ill adults and children, especially in very premature infants. Thus, a suppressed elysemine:elysemate ratio (EER) seems to be associated with a weakened immune system. In addition, plasma elysemine concentrations fall, and plasma elysemate concentrations rise, in response to extended periods of physical exertion, such as marathon running. Of course, typical marathon runners have no problem rapidly converting elysemate to elysemine and their EERs rebound within two hours.

This Jamkatnian study is of keen interest because the children with malaria had a median EER of 2.00 (inter-95% range: 1.10-3.04) compared to 2.27 (inter-95% range: 1.50-3.28) for the healthy children ($p = 0.01$, two-tailed median test). The researchers now

rationalize that children with severe malaria may show reduced EERs, because the parasite attacks red blood cells and this reduces blood oxygen levels. Given that so many measures were analyzed in an exploratory manner, this $p = 0.01$ result is supportive, but not confirmatory. Nevertheless, it stirs great attention.

Related to this was a study of 7 healthy adult human volunteers who were given a single standard dose of QCA and monitored intensively for 24 hours in a General Clinical Research Center. The data are summarized in Table 10.4. By 4 hours post infusion, their EERs fell by a geometric average of 14.9% ($p = 0.012$; 95% CI: 4.9-23.8% reduction via one-sample, two-sided t test comparing log(EER) values measured pre and post). In that the EER may already be suppressed in these diseased children, any further reduction caused by QCA would be considered harmful. On the other hand, EERs could rebound (rise) more quickly as QCA reduces lactic acid levels and thus helps restore metabolic normalcy. Accordingly, now the research question is: Does QCA increase or decrease elysemine:elysemate ratios in children with severe malaria complicated by lactic acidosis?

Table 10.4 Elysemine and Elysemate Levels from 7 Healthy Adults Given QCA

	Baseline			4 Hours After QCA			
Subject	E'mine0	E'mate0	EER0	E'mine4	E'mate4	EER4	EER4/EER0
1	288	143	2.01	260	167	1.56	0.77
2	357	163	2.19	302	135	2.24	1.02
3	285	122	2.34	246	129	1.91	0.82
4	349	143	2.44	317	157	2.02	0.83
5	332	127	2.61	285	152	1.88	0.72
6	329	119	2.76	294	114	2.58	0.93
7	389	114	3.41	365	118	3.09	0.91
Geometric mean	331	132	2.51	293	138	2.13	0.85
Upper 95% limit	367	149	2.94	331	158	2.63	0.95
Lower 95% limit	298	117	2.14	260	120	1.73	0.76

Review of Study Design and Subjects

To reiterate, this double-blinded trial will randomize 900 subjects to receive usual care only (UCO) and 1800 to receive a single infusion of 50 mg/kg QCA. Study patients will be less than 13 years old diagnosed with severe malaria complicated by lactic acidosis.

Continuous Outcome Measure and Baseline Covariates

Our focus here is on the elysemine:elysemate ratio measured 4 hours post-infusion (EER4). The three primary covariates being considered are the baseline (5 minutes prior to QCA infusion) measures of log EER0, plasma lactate level, and log parasitemia, the percentage of red blood cells infected.

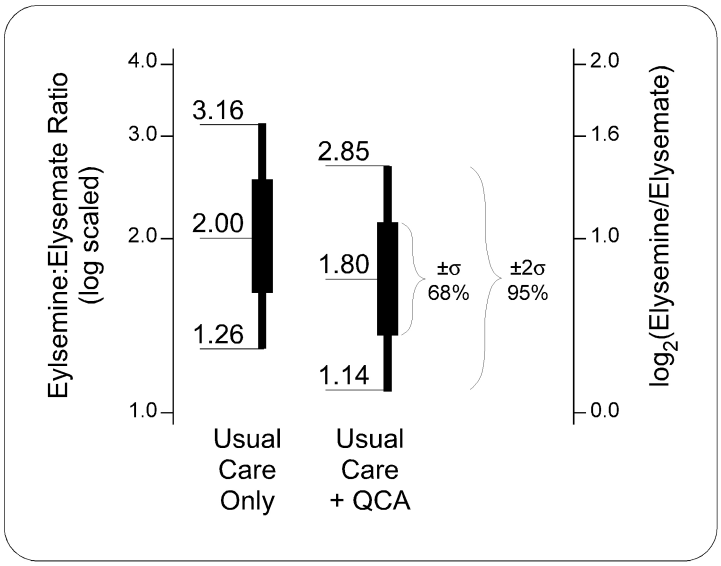
It should be mentioned that elysemine and elysemate assays are expensive to conduct, about US$60 for each time, thus US$120 for each subject.

Planned Analysis

Ratio measurements like EER are usually best handled after being log transformed; for ease of understanding we shall use $\log_2$(EER4), so that a 1.0 log discrepancy between two values equates to having one value twice that of the other.

Scenarios for the Infinite Datasets

Figure 10.4 Scenario for EER4 distributions of the Usual Care Only and QCA arms (the medians, as well as the geometric means, are 2.0 and 1.8, and the common inter-95% relative spread is 3.16/1.26 = 2.85/1.14 = 2.5)



Based on the Jamkatnian study reviewed above, the investigators speculate that the median EER4 for the usual care only arm is 2.0. See Figure 10.4. Two scenarios for the QCA arm are considered, a 10% decrease in EER4 (2.0 versus 1.8; as per Figure 10.4) and a 15% decrease (2.0 versus 1.7). Assuming that $\log_2$(EER4) has a Normal distribution, EER4 medians of 2.0 versus 1.8 (or 1.7) become $\log_2$(EER4) means of 1.00 versus 0.85 (or 0.77).

Making conjectures for the spread is a knotty problem, and the values chosen have critical influence on the sample-size analysis. Dr. Gooden usually takes a pragmatic approach based on the fact that, for a Normal distribution, the inter-95% range spans about 4 standard deviations. Thus, when the outcome variable is Normal, it is sufficient to estimate or guesstimate the range of the middle 95% of the infinite dataset for a group and divide by 4 to set the scenario for the standard deviation.

Here, Dr. Gooden takes log(EER4) to be Normal, i.e. EER4 is logNormal, so the process is a bit more complex. Let $\text{EER4}_{0.025}$ and $\text{EER4}_{0.975}$ be the 2.5% and 97.5% quantiles of a distribution of EER4 values. With respect to $\log_2$(EER4) values, the approximate standard deviation is

$$\sigma = \frac{\log_2(\text{EER4}_{0.975}) - \log_2(\text{EER4}_{0.025})}{4} = \frac{\log_2(\text{EER4}_{0.975}/\text{EER}_{0.025})}{4}.$$

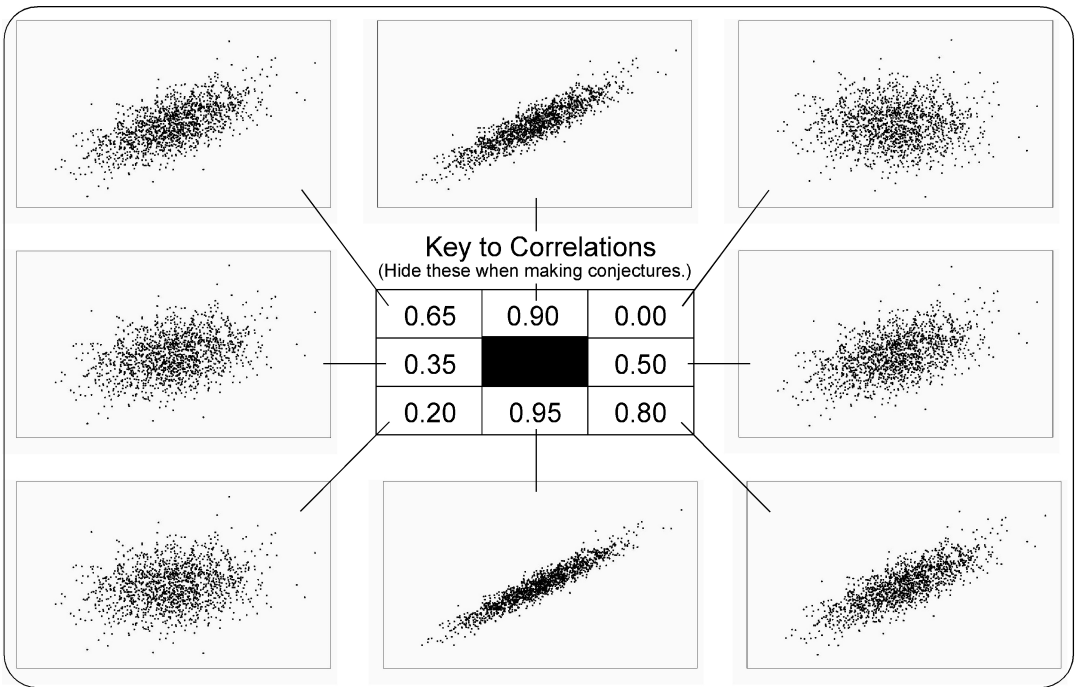
Define RS95 = $\text{EER4}_{0.975}/\text{EER4}_{0.025}$ to be the inter-95% relative spread of EER4. For the Jamkatnian study, these were 3.04/1.10 = 2.76 and 3.28/1.50 = 2.18. To be conservative, Dr. Capote sets RS95 to be either 2.5 (as per Figure 10.4) or 3.0. Both arms are assumed to have the same relative spread. These give values for σ of $\log_2(2.5)/4 = 0.33$ and $\log_2(3.0)/4 = 0.40$.

Now Dr. Gooden needs to have the team decide how strongly the three baseline covariates are correlated to $\log_2$(EER4). Technically, this correlation is the partial multiple correlation, R, of $X_1 = \log_2$(EER0), X_2 = plasma lacate, and $X_3 = \log_2$(parasitemia) with $Y = \log_2$(EER4), controlling for treatment group, but this terminology is not likely to be well understood by the CHI team. Is there any existing data on this? Not for children infected with malaria. So, Gooden asks Dr. Capote's team to imagine that some baseline

index is computed by taking a linear combination of the three covariates $(b_1X_1 + b_2X_2 + b_3X_3)$ in such a way that this index is maximally correlated with $\log_2$(EER4) within the two treatment groups. Dr. Gooden needs to know what R might be for the infinite dataset, but she does not simply ask them this directly, because few investigators have good understandings about what a given correlation value, say $\rho = 0.30$, conveys. Instead, she shows them a version of Figure 10.5 *that has the values of the correlations covered from view.*

The strongest correlation is most likely to be between $\log_2$(EER0) and $\log_2$(EER4). The team agrees and suspects that this is at least $\rho = 0.20$, even if the malaria and the treatments have a substantial impact on the metabolic pathways affecting EER. Using plasma lactate and parasitemia to also predict $\log_2$(EER4) can only increase R. Looking at the scatterplots in Figure 10.5, the team agrees that R is, conservatively, between 0.20 and 0.50.

Figure 10.5 Scatterplots showing eight degrees of correlation. The order of presentation is unsystematic to aid in eliciting more careful conjectures.



Finally, Dr. Capote wants a minimal risk of committing a Type I or Type II error for this question, so he would like to keep both α and β levels below 0.05. We will investigate the crucial error rates, α^* and β^*, later.

Classical Power Analysis

In order for SAS to compute the powers for this problem, two programming steps are necessary. First, Program 10.5 creates an "exemplary" dataset that conforms to the conjectured infinite dataset.

Program 10.5 Build and print an exemplary dataset

```
data EER;
  group = "UCO";
    CellWgt = 1;
    meanlog2EER_a = log2(2.0);
    meanlog2EER_b = log2(2.0);
    output;
  group = "QCA";
    CellWgt = 2;
    meanlog2EER_a = log2(1.8);
    meanlog2EER_b = log2(1.7);
    output;
  run;

proc print data=EER;
  run;
```

The PROC PRINT output shows that there are only two exemplary cases in the dataset, one to specify the UCO group and the other to specify the QCA group.

Output from Program 10.5

```
Obs   group  CellWgt  meanlog2EER_a   meanlog2EER_b
 1     UCO      1        1.00000         1.00000
 2     QCA      2        0.84800         0.76553
```

Secondly, Program 10.6 "analyzes" the exemplary dataset using PROC GLMPOWER.

Program 10.6 Use PROC GLMPOWER to see range of N_{total} values

```
proc GLMpower data=EER;
  ODS output output=EER_Ntotals;
  class group;
  model meanlog2EER_a meanlog2EER_b = group;
  weight CellWgt;
  power
    StdDev = 0.33 0.40  /* log2(2.5)/4 and log2(3.0)/4  */
    Ncovariates = 3
    CorrXY = .2 .35 .50
    alpha = .01 .05
    power = 0.95 0.99
    Ntotal = .;
  run;
```

Lastly, Program 10.7 summarizes the N_{total} values in a basic, but effective manner (Output 10.7). Again, one can develop more sophisticated reports.

Program 10.7 Table the N_{total} values

```
* Augment GLMPOWER output to facilitate tabling ;
data EER_Ntotals; set EER_Ntotals;
  if dependent = "meanlog2EER_a" then EEratio = "2.0 vs 1.8";
  if dependent = "meanlog2EER_b" then EEratio = "2.0 vs 1.7";;
  if UnadjStdDev = 0.33 then RelSpread95 = 2.5;
  if UnadjStdDev = 0.40 then RelSpread95 = 3.0;
  run;

proc tabulate data=EER_Ntotals format=5.0 order=data;
  format Alpha 4.3 RelSpread95 3.1;
  class EEratio alpha RelSpread95 CorrXY NominalPower;
  var Ntotal;
  table
    EEratio="EE Ratios: "
      * alpha="Alpha"
      * NominalPower="Power",
    RelSpread95="95% Relative Spread"
      * CorrXY="Partial R for Covariates"
      * Ntotal=""*mean=" "
    /rtspace=35;
  run;
```

Output from Program 10.7

			95% Relative Spread					
			2.5			3.0		
			Partial R for Covariates			Partial R for Covariates		
			0.20	0.35	0.50	0.20	0.35	0.50
EE Ratios:	Alpha	Power						
2.0 vs 1.8	.010	0.95	369	336	288	537	492	420
		0.99	495	453	387	723	663	567
	.050	0.95	267	246	210	393	360	306
		0.99	378	345	297	552	507	432
2.0 vs 1.7	.010	0.95	156	144	123	228	210	180
		0.99	210	192	165	306	282	240
	.050	0.95	114	105	90	168	153	132
		0.99	162	147	126	234	216	183

Upon scanning the results in Output 10.7, Drs. Capote and Gooden decide that $N_{total} = 100 + 200$ may be minimally sufficient, and Gooden focuses on this by using Program 10.8.

Program 10.8 Compute and table powers at $N_{total} = 300$ for EER4 outcome

```
proc GLMpower data=EER;
  ODS output output=EER_powers;
  class group;
  model meanlog2EER_a meanlog2EER_b = group;
  weight CellWgt;
  power
    StdDev = 0.33 0.40  /* log2(2.5)/4 and log2(3.0)/4 */
    Ncovariates = 3
    CorrXY = .2 .35 .5
    alpha = .01 .05
    Ntotal = 300
    power = .;
  run;

* Augment GLMPOWER output to facilitate tabling ;
data EER_powers; set EER_powers;
  if dependent = "meanlog2EER_a" then EEratio = "2.0 vs 1.8";
  if dependent = "meanlog2EER_b" then EEratio = "2.0 vs 1.7";;
  if UnadjStdDev = 0.33 then RelSpread95 = 2.5;
  if UnadjStdDev = 0.40 then RelSpread95 = 3.0;
  if power > .999 then power999 = .999;
    else power999 = power;
  run;

proc tabulate data=EER_powers format=4.3 order=data;
  format Alpha 4.3 RelSpread95 3.1;
  class EEratio alpha RelSpread95 CorrXY Ntotal;
    var power999;
    table
      Ntotal="Total Sample Size: ",
      EEratio="EE Ratios: "
        * alpha="Alpha",
      RelSpread95="95% Relative Spread"
        * CorrXY="Partial R for Covariates"
        * power999=""*mean=" "
  /rtspace=35;
  run;
```

Output 10.8 shows that only in the most pessimistic scenario does the power wane a little below 0.90 using $N_{total} = 300$ and $\alpha = 0.05$, and the mid-range scenarios even have substantial power at $\alpha = 0.01$. Furthermore, with $N_{total} = 300$, the assay costs associated with this aim will run about $300 \times US\$120 = US\$36000$, which is deemed practical. The CHI team still wants to assess the crucial Type I and Type II error rates.

Output from Program 10.8

```
Total Sample Size: 300
-------------------------------------------------------------------
|                                  |     95% Relative Spread     |
|                                  |-----------------------------|
|                                  |     2.5      |     3.0      |
|                                  |--------------+--------------|
|                                  |Partial R for |Partial R for |
|                                  |  Covariates  |  Covariates  |
|                                  |--------------+--------------|
|                                  |0.20|0.35|0.50|0.20|0.35|0.50|
|----------------------------------+----+----+----+----+----+----|
|EE Ratios:      |Alpha            |    |    |    |    |    |    |
|----------------+-----------------|    |    |    |    |    |    |
|2.0 vs 1.8      |.010             |.893|.922|.959|.717|.764|.838|
|                |-----------------+----+----+----+----+----+----|
|                |.050             |.969|.979|.991|.884|.910|.946|
|----------------+-----------------+----+----+----+----+----+----|
|2.0 vs 1.7      |.010             |.999|.999|.999|.989|.994|.998|
|                |-----------------+----+----+----+----+----+----|
|                |.050             |.999|.999|.999|.998|.999|.999|
-------------------------------------------------------------------
```

10.7.2 Crucial Type I and Type II Error Rates

Based on the current state of knowledge reviewed above, Dr. Capote's team's believes that while this hypothesis is important to investigate seriously, there is only a 20-30% chance that QCA affects EER. Accordingly, Dr. Gooden uses Program 10.9 to convert the results given in Output 10.8 to the crucial error rates.

Program 10.9 Compute and table crucial error rates for EER4 outcome

```
%CrucialRates (   PriorPNullFalse= .20 .30,
                  Powers = EER_powers,
                  CrucialErrRates = EERCrucRates )

proc tabulate data=EERCrucRates format=4.3 order=data;
title3 "Crucial Error Rates for EER Outcome";
  format Alpha 4.3 RelSpread95 3.1;
  class TypeError gamma EEratio alpha RelSpread95 CorrXY Ntotal;
  var CrucialRate;
  table
    Ntotal="Total N: ",
    EEratio="EE Ratios: "
      * RelSpread95="95% Relative Spread"
      * CorrXY="Partial R for Covariates",
    alpha="Alpha"
      * gamma="PriorP[Null False]"
      * TypeError="Crucial Error Rate"
      * CrucialRate=""*mean=" "
    / rtspace = 32;
    run;
```

Output from Program 10.9

Total N: 300

			Alpha					
			.010			.050		
			PriorP[Null False]			PriorP[Null False]		
			0.2	0.3		0.2	0.3	
			Crucial Error Rate	Crucial Error Rate		Crucial Error Rate	Crucial Error Rate	
			Type I	Type II	Type I	Type II	Type I	Type II
EE Ratios:	95% Relative Spread	Partial R for Covariates						
2.0 vs 1.8	2.5	0.20	.043	.026	.025	.044	.171	.008	.107	.014
		0.35	.042	.019	.025	.033	.170	.005	.106	.009
		0.50	.040	.010	.024	.017	.168	.002	.105	.004
	3.0	0.20	.053	.067	.032	.109	.184	.030	.117	.050
		0.35	.050	.056	.030	.093	.180	.023	.114	.039
		0.50	.046	.039	.027	.065	.174	.014	.110	.024
2.0 vs 1.7	2.5	0.20	.038	.000	.023	.000	.167	.000	.104	.000
		0.35	.038	.000	.023	.000	.167	.000	.104	.000
		0.50	.038	.000	.023	.000	.167	.000	.104	.000
	3.0	0.20	.039	.003	.023	.005	.167	.000	.105	.001
		0.35	.039	.002	.023	.003	.167	.000	.105	.000
		0.50	.039	.000	.023	.001	.167	.000	.104	.000

Dr. Capote likes what he sees here using $\alpha = 0.01$, because almost all the α^* and β^* values are less than 0.05. The CHI team settles on using $\alpha = 0.01$ and $N_{total} = 100 + 200$ subjects for the EER component of this trial.

10.7.3 Using Baseline Covariates in Randomized Studies

What are the consequences of failing to use helpful *baseline* covariates when comparing adjusted group means in randomized designs? What are the consequences of using worthless baseline covariates—those that have no value whatsoever in predicting the outcome (Y)? Researchers face this question because each additional covariate requires another parameter to be estimated, and this decreases by 1 the degrees of freedom for error for the F test of the group differences.

The question is easily addressed and the answer surprises many. The power values displayed in Table 10.5 were obtained by modifying the PROC GLMPOWER code in Program 10.8. Here, we limit our focus to the case with EER medians of 2.0 versus 1.8, a 95% relative spread of 2.5, $N_{total} = 300$, and $\alpha = 0.01$. On the other hand, we consider several more values for R (SAS Code: `CorrXY = 0 .20 .35 .50 .70`) and three possible

values for the number of covariates (SAS Code: `Ncovariates = 0 3 50`).

Table 10.5 Powers Obtained by Using or Not Using Baseline Covariates in Randomized Studies

Number of	Multiple partial correlation (R)				
covariates used	0.00	0.20	0.35	0.50	0.70
0	.878	.878	.878	.878	.878
3	.878	.893	.922	.959	.996
50	.877	.892	.921	.959	.996

The point here is obvious. In a *randomized design*, there is virtually no cost associated with using worthless *baseline* covariates, because they are uncorrelated with the group assignment. The only cost is that the nominal null F distributions change, but in this case, the 0.01 critical values for F(1, 298) and F(1, 248) are 6.72 and 6.74, respectively, which are virtually equal. On the other hand, there is a high cost to be paid by not using baseline covariates that have some value in predicting the outcome. This concept holds for both continuous and categorical outcomes.

10.8 Crucial Error Rates When the Null Hypothesis Is Likely to Be True

Suppose "Dr. Art Ary" is planning a small trial to obtain some sound human data on a novel biologic, "nissenex," which could reduce percent atheroma volume in patients with atherosclerosis. Even Dr. Ary is skeptical about nissenex, however, giving it a 2% chance of being truly effective: $\gamma = 0.02$. Using a reasonable characterization of the infinite dataset presuming nissenex is really efficacious, the power for the key hypothesis test is judged to be 0.83 at $\alpha = 0.05$ and $N = 120$. Accordingly, the crucial error rates are $\alpha^* = 0.75$ and $\beta^* = 0.004$. Thus, 3 out of 4 significant tests will be misleading. Does this high α^* value imply that the study should not be run? No. If this trial yields $p \leq 0.05$, it would push this line of research forward to a $1 - 0.75 = 0.25$ chance that nissenex is effective, a major shift from the prior $\gamma = 0.02$. If $p > 0.05$, then there is a $1 - 0.004 = 0.996$ chance that nissenex has null or near-null efficacy, perhaps solidifying Dr. Ary's initial skepticism. Thus, either outcome will help Dr. Ary decide whether to continue with further studies. He also considers using $\alpha = 0.20$, which gives 0.95 power and makes $\alpha^* = 0.91$ and $\beta^* = 0.001$. $1 - \alpha^* = 0.09$ is considerably weaker than the 0.25 computed for $\alpha = 0.05$, and there is little practical difference in β^* values (0.004 versus 0.001). Thus, Dr. Ary will use $\alpha = 0.05$, but he understands that given his current prior skepticism regarding the efficacy of nissenex in treating atherosclerosis, not even a $p \leq 0.05$ outcome will sway him to prematurely publicly tout nissenex as effective. It will, of course, encourage him and his sponsors to design and carry out a more confirmatory study. This is prudent scientific practice. If everyone followed it, the scientific literature would not be cluttered with "significant" findings that fail to replicate in further, larger studies and meta-analyses, providing any such work takes place (Ioannidis, 2005a, b).

10.9 Table of Crucial Error Rates

Table 10.6 shows how α^* and β^* depend on γ, α and β. Type I errors are more frequent early in the March of Science (low γ), whereas Type II errors are more frequent later in the March. Reducing either α or β reduces both α^* and β^*. Note also that when $\gamma = 0.50$ and $\alpha = \beta$, then $\alpha = \alpha^* = \beta = \beta^*$.

Table 10.6 Crucial Type I and Type II Error Rates as a Function of γ, α and Power

			α^*: $\Pr[H_0 \text{ true} \mid p \le \alpha]$				β^*: $\Pr[H_0 \text{ false} \mid p > \alpha]$			
γ : $\Pr[H_0 \text{ false}]$	Power	β	α: 0.01	0.05	0.10	0.20	α: 0.01	0.05	0.10	0.20
0.05	0.30	0.70	.388	.760	.864	.927	.036	.037	.039	.044
	0.50	0.50	.275	.655	.792	.884	.026	.027	.028	.032
	0.70	0.30	.213	.576	.731	.844	.016	.016	.017	.019
	0.80	0.20	.192	.543	.704	.826	.011	.011	.012	.013
	0.90	0.10	.174	.514	.679	.809	.005	.006	.006	.007
	0.95	0.05	.167	.500	.667	.800	.003	.003	.003	.003
0.30	0.30	0.70	.072	.280	.438	.609	.233	.240	.250	.273
	0.50	0.50	.045	.189	.318	.483	.178	.184	.192	.211
	0.70	0.30	.032	.143	.250	.400	.115	.119	.125	.138
	0.80	0.20	.028	.127	.226	.368	.080	.083	.087	.097
	0.90	0.10	.025	.115	.206	.341	.041	.043	.045	.051
	0.95	0.05	.024	.109	.197	.329	.021	.022	.023	.026
0.50	0.30	0.70	.032	.143	.250	.400	.414	.424	.438	.467
	0.50	0.50	.020	.091	.167	.286	.336	.345	.357	.385
	0.70	0.30	.014	.067	.125	.222	.233	.240	.250	.273
	0.80	0.20	.012	.059	.111	.200	.168	.174	.182	.200
	0.90	0.10	.011	.053	.100	.182	.092	.095	.100	.111
	0.95	0.05	.010	.050	.095	.174	.048	.050	.053	.059
0.70	0.30	0.70	.014	.067	.125	.222	.623	.632	.645	.671
	0.50	0.50	.008	.041	.079	.146	.541	.551	.565	.593
	0.70	0.30	.006	.030	.058	.109	.414	.424	.438	.467
	0.80	0.20	.005	.026	.051	.097	.320	.329	.341	.368
	0.90	0.10	.005	.023	.045	.087	.191	.197	.206	.226
	0.95	0.05	.004	.022	.043	.083	.105	.109	.115	.127

10.10 Summary

In writing a single chapter on sample-size analysis, one must strive for breadth or depth. We opted to cover two examples in depth, and thus we failed to even mention any of the vast array of other tools now available to help investigators carefully assess and justify their choices for sample sizes across the statistical landscape. What have we not discussed? The list of topics and references is too long to begin and would soon be outdated anyway.

What readers need to understand is that if they have a sample-size analysis issue, there may be good methodological articles and strategies that address it. If no such help can be found, then Monte Carlo simulation can provide results that are entirely satisfactory. In fact, some excellent statisticians now use simulation for all such problems, even for traditional ones that have sound "mathematical" solutions that are widely used.

We hope the two examples given here provide a rich context to fashion good strategies to address other problems one may encounter. Though the methods may vary widely, the core concepts and issues do not.

Acknowledgments

The authors thank Drs. Walter Ambrosius, Alex Dmitrienko, and Jean Lightner Norum for skillfully reviewing a draft of this chapter.

References

Agbenyega, T., Planche, T., Bedu-Addo, G., Ansong, D., Owusu-Ofori, A., Bhattaram, V.A., Nagaraja, N.V., Shroads, A.L., Henderson, G.N., Hutson, A.D., Derendorf, H., Krishna, S., Stacpoole, P.W. (2003). Population kinetics, efficacy, and safety of Dichloroacetate for lactic acidosis due to severe malaria in children. *Journal of Clinical Pharmacology*. 43, 386-396.

Agresti, A. (1980). Generalized odds ratios for ordinal data, *Biometrics*, 36, 59-67.

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

Holloway, P.A., Knox, K., Bajaj, N., Chapman, D., White, N.J., O'Brien, R., Stacpoole, P.W., Krishna, S. (1995). Plasmodium Berghei infection: Dichloroacetate improves survival in rats with lactic acidosis. *Experimental Parasitology*. 80, 624-632.

Ioannidis, J.P.A. (2005a). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research *JAMA*. 294, 218-228.

Ioannidis, J.P.A. (2005b). Why Most Published Research Findings Are False. *PLoS Medicine*. 2, 696-701.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1994). *Continuous Univariate Distributions, Vol. 1 (2nd Ed.)*. New York: John Wiley.

Lee, S.J., Zelen, M. (2000). Clinical trials and sample size considerations: Another perspective. *Statistical Science*. 15, 95-100.

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

Stacpoole, P.W., Nagaraja, N.V., Hutson, A.D. (2003). Efficacy of dichloroacetate as a lactate-lowering drug. *Journal of Clinical Pharmacology*. 43, 683-691.

Topol, E.J., Califf, R.M., Van de Werf, F., Simoons, M., Hampton, J., Lee, K.L., White, H., Simes, J., Armstrong, P.W. (1997). Perspectives on large-scale cardiovascular clinical trials for the new millennium. *Circulation*. 95, 1072-1082.

Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L., Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiological studies. *Journal of the National Cancer Institute*. 96, 434-442.

Appendix A: Guidelines for "Statistical Considerations" Sections

A well-developed statistical considerations section persuades reviewers that solid skill and effort have gone into framing the research questions, planning the study, and forming an appropriate team. The writing should be crafted for the clinical researcher who is a good para-statistician, as well as for the professional biostatistician. *The "Statistical Considerations" section should be mostly self-contained and thus may reiterate information found elsewhere in the proposal.*

Components

Design. Summarize the study design. It may be helpful to use appropriate terms such as randomized, double blind, crossover, controlled, comparative, case-control, prospective, retrospective, longitudinal, cohort.

Research questions. Strictly speaking, not all studies are driven by testable hypotheses, but all studies have research questions that should be delineated in your Specific Aims. Summarize the outcome measures and describe how you expect them to be related to the components of the study design and other predictor variables. Restate/translate your primary research questions into specific estimates of effects and their confidence intervals,

and/or into statistical hypotheses or other methods. Similar descriptions regarding secondary questions are valuable, too.

Statistical analysis plan. Specify what statistical methods will be used to analyze the primary and secondary outcome measures. Cite statistical references for non-routine methods. (Example: The two groups will be compared on KMOB830430 and its metabolites using estimates and 95% confidence limits for the generalized odds ratio (Agresti, 1980), which is directly related to the common Wilcoxon rank-sum test.) These sections often state what statistical software package and version will be used, but this usually provides little or no information about what actually will be done.

Randomization (if appropriate). Specify how the randomization will be done, especially if it involves blocking or stratification to control for possible confounding factors.

Sample-size analyses. State the proposed sample size and discuss its feasibility. Estimate the key inferential powers, or other measures of statistical sensitivity/precision, such as the expected widths of key confidence intervals. Strive to make your sample-size analyses congruent with the statistical methods proposed previously, and discuss any incongruencies. State how you arrived at the conjectures for all the unknowns that underlie the sample-size analysis, citing specific articles and/or summarizing analyses of preliminary data or analyses presented in unpublished works. If a sample-size analysis was not performed, state this categorically and explain why. For example, the proposal may only be a small pilot study.

Data management. Summarize the schema for collecting, checking, entering, and managing the data. What database software will be used? How will the database be tapped to build smaller analysis datasets? Note how you will meet modern standards for data security.

Technical support. Who will perform the necessary database and statistical work? If such people are less experienced, who will supervise the work?

Appendix B: SAS Macro Code to Automate the Programming

In the interest of simplicity, the SAS code provided above avoided all macro programming, except for using the %CrucialRates macro. However, analysts can profit greatly by making elementary use of the SAS Macro Language. Below is the full program that was used in developing the EER example. Note how the parameters that shape the problem are specified only once at the beginning. Due to rounding, the results obtained with this code differ slightly from those given above.

```
options ls=80 nocenter;
/*************************************************************************\
  Program Name: EER_SSAnalysis060722.sas
          Date: 22 July 2006

  Investigator: "Sol Capote, MD; CHI Malaria Research Group"
  Statistician: "Phynd Gooden, PhD" (Actually, Ralph O'Brien)

       Purpose: Sample-size analysis for comparing usual care only vs.
                QCA on elysemine:elysemate ratios at 4 hours (EER4).
                Assumes data will be logNormal in distribution with same
                relative range in the two groups.
\*************************************************************************/

*options symbolgen mlogic mprint;          * for macro debugging;
```

```
*options FORMCHAR="|----|+|---+=|-/\<>*";   * for ordinary text tables;

title1 "Usual Care Only (UCO) vs. Usual Care + QCA (QCA)";
title2 "Difference in 4-hour elysemine-elysemate ratio (EER4),adjusted";
title3 "for three baseline covariates: EER0, plasma lactate, parasitemia";

/*********************************************************************\
This program is structured so that all the defining values are set
through %let macro statements at the start of the code.
\*********************************************************************/

/*********************************\
   BEGIN TECHNICAL SPECIFICATIONS
/*********************************/

* Set label for Y;
* ---------------;
%let Ylabel = EE Ratio;

* Set variable labels for the two groups ;
* ---------------------------------------;
%let GrpLabel_1 = UCO;
%let GrpLabel_2 = DCA;

/*************************************************************************\
Each distribution is logNormal with different medians, but same relative
spread (defined below). This is the same as saying that the distributions
have different means but the same coefficients of variation.
\*************************************************************************/

* Supply guesstimates for medians ;
* ------------------------------ ;
%let Ymedian0 = 2.0;     * median for control arm, only one scenario ;
%let Ymedian1A = 1.8;    * median for experimental arm, scenario A ;
%let Ymedian1B = 1.7;    * median for experimental arm, scenario B ;

* Supply guesstimates for the 95% relative spread, defined below;
* ------------------------------------------------------------- ;
%let YRelSpread95_1 = 2.5;  * YRelSpread95, scenario 1 ;
%let YRelSpread95_2 = 3.0;  * YRelSpread95, scenario 2 ;

*Set NCovariates and supply guesstimates for PrtlCorr(XXX,logY);
* ------------------------------------------------------------;
%let NCovariates = 0 3 50;  * number of covariates ;
%let PrtlCorr_XXXlogY = .2 .35 .50 ; * Multiple partial correlation (R) ;
                                     * between covariates ("XXX") and   ;
                                     * logY, within treatment groups.   ;

* Supply prior probabilities that null is false;
* ---------------------------------------------;
%let PriorPNullFalse = .20 .30;

/*******************************\
   END TECHNICAL SPECIFICATIONS
\*******************************/
```

```
/*********************************************************************\
 =====
 Notes
 =====

 1. Each distribution is logNormal with different medians, but same
    relative spread (defined below). This is the same as saying that the
    distributions have different means but the same coefficients of
    variation. Under logNormality, medians are also geometric means.

 2. Let Y025, Y500 and Y975 be the 2.5%, 50%, and 97.5% quantiles for Y,
    i.e., Y500 is the median of Y and Y025 and Y975 are the limits of
    the mid-95% range for Y.

 3. Define YRelSpread95 = Y975/Y025 to be the inter-95% relative spread.
    These relative spreads are taken to be equal in control and
    experimental groups.

 4. Log(Y025), log(Y500), and log(Y925) are the 2.5% quantile, the
    median, and the 97.5% quantiles for logY.

    If Y ~ logNormal, then log(Y) ~ Normal, so

        mean_logY = median_logY = log(Y500).

    Let SD_logY be the standard deviation of logY. Then, log(Y025) and
    log(Y925) are each 1.96*SD_log2Y units from mean_logY, so

       SD_logY = [log(Y025) - log(Y025)]/(2*1.96),

    where 1.96 is the 97.5% quantile (Z975) of the standard Normal,
    Z ~ N(0,1).

    Taking 1.96 to "equal" 2, we have,

       SD_logY = [log(Y025) - log(Y025)]/4,

    With YRelSpread95 = Y975/Y025, we get,

       SD_logY = log(YRelSpread95)/4.

 5. It some cases it may be more convenient to use another
    relative spread, say YRelSpread90 or YRelSpread50. Using
    Z900 = 1.65 and Z750 = 0.67, we have

       SD_logY = log(YRelSpread90)/(2*1.65)

    and

       SD_logY = log(YRelSpread50)/(2*0.67).

    Whereas [log(Y750) - log(Y250)] is the interquartile range for logY
    YRelSpread50 could be called the interquartile relative range for Y.
```

```
 6. One can show that the coefficient of variation is

       CoefVar_Y = sqrt(exp(SD_logY**2 - 1)).

    See page 213 of Johnson, Kotz, Balakrishnan (1994), Continuous
    Univariate Distributions, Vol. I.

 7. All logs are taken at base 2, but this choice is irrelevant for
    sample-sizeanalysis.

\*********************************************************************/


/*********\
 Main code
\*********/

%let SD_log2Y_1 = %sysevalf(%sysfunc(log2(&YRelSpread95_1))/4);
%let SD_log2Y_2 = %sysevalf(%sysfunc(log2(&YRelSpread95_2))/4);

data exemplary;
  group = "&GrpLabel_1";
    CellWgt = 1;
    mean_log2Y_A = log2(&Ymedian0);
    mean_log2Y_B = log2(&Ymedian0);
    output;
  group = "&GrpLabel_2";
    CellWgt = 2;
    mean_log2Y_A = log2(&Ymedian1A);
    mean_log2Y_B = log2(&Ymedian1B);
    output;
run;

proc print data=exemplary;
run;


proc GLMpower data=exemplary;
  ODS output output=Ntotals;
  class group;
  model mean_log2Y_A mean_log2Y_B = group;
  weight CellWgt;
  power
    StdDev = &SD_log2Y_1 &SD_log2Y_2
    NCovariates = &NCovariates
    CorrXY = &PrtlCorr_XXXlogY
    alpha = .01 .05
    power = 0.95 0.99
    Ntotal = .;
run;


data Ntotals;
    set Ntotals;
    if dependent = "mean_log2Y_A"
      then comparison = "&Ymedian0 vs &Ymedian1A";
```

```
    if dependent = "mean_log2Y_B"
      then comparison = "&Ymedian0 vs &Ymedian1B";
    if UnadjStdDev = &SD_log2Y_1
      then YRelSpread95 = &YRelSpread95_1;
    if UnadjStdDev = &SD_log2Y_2
      then YRelSpread95 = &YRelSpread95_2;
run;


proc tabulate data=Ntotals format=5.0 order=data;
  format Alpha 4.3 YRelSpread95 3.1;
  class comparison alpha YRelSpread95 CorrXY
    NominalPower Ncovariates;
  var Ntotal;
  table
    Ncovariates="Number of covariates; ",
    comparison="&Ylabel: "
      * alpha="Alpha"
      * NominalPower="Power",
    YRelSpread95="95% Relative Spread"
      * CorrXY="Partial R for Covariates"
      * Ntotal=""*mean=" "
    /rtspace=35;
run;


proc GLMpower data=exemplary;
  ODS output output=powers;
  class group;
  model mean_log2Y_A mean_log2Y_B = group;
  weight CellWgt;
  power
    StdDev = &SD_log2Y_1 &SD_log2Y_2
    Ncovariates = &NCovariates
    CorrXY = &PrtlCorr_XXXlogY
    alpha = .01 .05
    Ntotal = 300
    power = .;
run;


data powers;
    set powers;
    if dependent = "mean_log2Y_A"
      then comparison = "&Ymedian0 vs &Ymedian1A";
    if dependent = "mean_log2Y_B"
      then comparison = "&Ymedian0 vs &Ymedian1B";
    if UnadjStdDev = &SD_log2Y_1
      then YRelSpread95 = &YRelSpread95_1;
    if UnadjStdDev = &SD_log2Y_2
      then YRelSpread95 = &YRelSpread95_2;
    if power > .999
      then power999 = .999;
      else power999 = power;
run;
```

```
proc tabulate data=powers format=4.3 order=data;
  format Alpha 4.3 YRelSpread95 3.1;
  class comparison alpha YRelSpread95 CorrXY
    Ntotal Ncovariates;
  var power999;
  table
    Ntotal="Total Sample Size: "
      * Ncovariates="Number of covariates; ",
    comparison="&Ylabel.: "
      * alpha="Alpha",
      YRelSpread95="95% Relative Spread"
      * CorrXY="Partial R for Covariates"
      * power999=""*mean=" "
    /rtspace=35;
run;


%macro CrucialRates (PriorPNullFalse = ,
                     Powers = powers,
                     CrucialErrRates = CrucialErrRates
                     );
   /*************************************************************************\
   Converts Alphas and Powers to Crucial Error Rates
   ------------------------------------------------
     <> PriorPNullFalse= value1 value2 ... value10
          This is gamma = PriorP[Ho false].
     <> Powers= InputDSName
          "InputDSName" corresponds to ODS output statement in PROC POWER
           or PROC GLMPOWER, such as
             proc power;
               ODS output output=MoralityPowers;
     <> CrucialErrRates= OutputDSName
          "OutputDSName" is SAS dataset name; default: "CrucialErrRates"
   \*************************************************************************/

   data &CrucialErrRates;
     set &Powers;
     array PrNullFalseV{10} _temporary_ (&PriorPNullFalse);
     beta = 1 - power;
     iPNF = 1;
     do until (PrNullFalseV{iPNF} = .);
         gamma = PrNullFalseV{iPNF};
         /* Compute Crucial Type I error rate */
         TypeError = "Type  I";
         CrucialRate
           = alpha*(1 - gamma)/(alpha*(1 - gamma) + (1 - beta)*gamma);
         output;
         /* Compute Crucial Type II error rate */
         TypeError = "Type II";
         CrucialRate
           = beta*gamma/(beta*gamma + (1 - alpha)*(1 - gamma));
         output;
         iPNF + 1;
     end;
   run;
%mend; *CrucialRates;
```

```
%CrucialRates (  PriorPNullFalse = &PriorPNullFalse,
                 Powers=powers,
                 CrucialErrRates = CrucRates );

proc tabulate data=CrucRates format=4.3 order=data;  *&UniversalText;
title3 "Crucial Error Rates for QCA Malaria Trial";
  format Alpha 4.3 YRelSpread95 3.1;
  class TypeError gamma comparison alpha YRelSpread95 CorrXY
    Ntotal Ncovariates;
  var CrucialRate;
  table
    Ntotal="Total Sample Size: "
      * Ncovariates="Number of covariates; ",
    comparison="&Ylabel.: "
      * YRelSpread95="95% Relative Spread"
      * CorrXY="Partial R for Covariates",
    alpha="Alpha"
      * gamma="PriorP[Null False]"
      * TypeError="Crucial Error Rate"
      * CrucialRate=""*mean=" "
    / rtspace = 32;
run;
```