Chapter 14. Inferential Statistics, Descriptive Statistics and The Analysis Plan

By William P. Coleman, Ph.D.

Provided that the data show that a compound is safe, it is the primary efficacy analysis that determines whether the compound is approved. The secondary analyses support, clarify and confirm, but only one prospectively planned efficacy analysis can be primary in support of an effectiveness claim for a given trial.

Why should this be so? And how can we design the trial and set up the primary efficacy analysis to get the best chance of confirming efficacy? The object of this chapter is to explain, in a simple but concrete way, the theoretical ideas behind the primary efficacy analysis.

Many people treat statistics as mumbo–jumbo, but they don't have to. There are commonsense reasons behind statistical ideas. While one might be impatient to get beyond the concepts and quickly arrive at the bottom line, this hurry is a mistake. Using statistical concepts thoughtfully can mean major (possibly multi–million dollar in a large study) savings in the study design and, much more importantly, can make the difference in turning a clinically beneficial compound into an approved and marketable reality.

YOU AND YOUR STATISTICIAN

I would gratefully like to thank Drs. Andrea De Gaetano, Frank Dorsey, Roberto Fiorentini, Daniel Freedman, Louis Fries, Fred Geisler and Steven Linberg and Devinder Poonian for working with me to clarify these ideas in the course of several projects. I would also like to thank Professor Marvin Zelen, who was directly or indirectly responsible for teaching me much of it to begin with.

My theoretical work on clinical trial design has been generously supported by a grant from the Life Science Division of NASA.

Before starting, a word about your relationship with your statistician. Think of your statistician like you think of your attorney. He or she is there to give you technical support to get what you want. In

---

I would gratefully like to thank Drs. Andrea De Gaetano, Frank Dorsey, Roberto Fiorentini, Daniel Freedman, Louis Fries, Fred Geisler and Steven Linberg and Ms. Devinder Poonian for working with me to clarify these ideas in the course of several projects. I would also like to thank Professor Marvin Zelen, who was directly or indirectly responsible for teaching me much of it to begin with.

My theoretical work on clinical trial design has been generously supported by a grant from the Life Science Division of NASA.

the beginning the two of you have a communication gap. The statistician doesn't know what you want; and, on your side, you don't know all the techniques and services the statistician might have available to help you. You have to iterate back and forth to get it right. *Consult your statistician and do it early.* As this chapter will explain, *prospective* planning is everything: it's what makes the study credible. The better you understand the concepts in this chapter, the better you can explain yourself to the statistician and the better you can understand the replies.

Another way that your statistician is like your attorney is that part of the job is to keep you from getting yourself into trouble. The statistician realizes that you have multiple goals to fulfill and that you have to do so within multiple constraints. While you're concentrating on getting to your main goals, the statistician should be evaluating the technical aspects of your plans to make sure you're not covertly hurting yourself in other ways. Don't be defensive if you get criticism: that's part of the service you're paying for.

Lastly, your statistician is there to represent you. The FDA has a staff of highly competent statisticians, and statistical understanding is also well diffused among their clinical staff. They will usually detect early if you're confused or engaging in wishful thinking. The only solution is to retain the services of someone who can first ensure that you're not thinking wishfully and then can explain your point of view to the FDA. Make sure you have enough dialogue throughout the whole process so that your statistician can give you solid backup.

WHAT PROBABILITY THEORY HAS TO DO WITH IT

Let's consider a real–life example. In 1991, a report was published[1] in the New England Journal of Medicine of a preliminary study of the efficacy of GM–1 ganglioside in acute spinal cord injury. While the results of this study are still awaiting confimation by the multi–center center trial now running, we can use it as an illstration of many issues in clinical trial design.

The primary measure of efficacy was the number of patients able to make an improvement of two or more grades on the Frankel Scale of Motor Function, i.e., to make a major improvement in their ability to live independently.

TABLE 1: GM–1 IN ACUTE SPINAL CORD INJURY.

	improve	not	total
placebo	1	13	14
GM–1	7	7	14
total	8	20	28

The results shown in Table 1 appear to indicate a 50% (7 in 14) success rate for GM–1 compared to only 7.1% for placebo. Is this result true in the general population? Can readers of the article expect similar success with their patients? Or, is the result merely a

feature of the particular 28 patients studied? *The point of inferential statistical testing (as opposed to simple descriptive statistics), indeed its only point, is to provide a framework in which to answer such questions.* Can we extrapolate our result to the general population?

How does probability enter into the analysis of efficacy data like Table 1? The idea is very common sense. Consider the original pool of 28 patients. They differ in their prognosis in a large number of ways, known or unknown: for example, complete vs. incomplete injury, cervical vs. thoracic, extent of associated injuries, and age. Suppose we were to take these 28 patients and divide them arbitrarily into 2 groups of 14, without any difference in treatment: we *merely label* one group "A" and the other group "B" and wait to observe the outcome. Even though both groups are treated the same, it's unlikely that there will be *exactly* the same number of recoveries in each. Chance will likely send more patients with good prognosis into one of the groups, A or B, than the other. This is not a sign that that group is better: it's merely random variation.

This suggests the possibility that the result in Table 1 might also merely be random variation and not an effect of the drug at all. To this suggestion we could reply that, while a small discrepancy might be due to chance, it is very unlikely we would observe a such a large difference unless the drug is effective. All right, well how likely would it be, *exactly*? Can we clarify the question by providing exact quantitation?

Imagine a roster already made up with 14 slots designated for the placebo patients and 14 more designated for GM–1. We are to assign the 28 available patients at random to those 28 slots. Count out all the possible ways of choosing 8 places among the 28 slots for the 8 successful patients. It turns out there are exactly 3,108,105 ways to do this. (The statistician doesn't actually count them. There are short–cut formulas that exactly predict what would happen if you did count. Conceptually, though, one *counts*.) If there were no drug effect, the probability of a dataset like Table 1 is equal to the fraction of these that give 1 of the successful patients to placebo and 7 to GM–1. Out of the total there are 48,048 choices that do so. Therefore the probability of the outcome in Table 1 is 48,048 divided by 3,108,105, or .01546. The other possible outcomes can be analyzed in the same way, with the result shown in Table 2.

TABLE 2: WAYS OF CHOOSING 8 AMONG 28 PLACES, 14 LABELED "PLACEBO" AND 14 LABELED "GM–1," FOR 8 SUCCESSFUL PATIENTS (ASSUMING THERE IS NO DRUG EFFECT)

Successes for placebo	Successes for GM-1	number of possible choices	Probability
0	8	3,003	.00097
1	7	48,048	.01546
2	6	273,273	.08792
3	5	728,728	.23446
4	4	1,002,001	.32238
5	3	728,728	.23446
6	2	273,273	.08792
7	1	48,048	.01546
8	0	3,003	.00097
	TOTAL:	3,108,105	1.00000

This table illustrates the discussion so far. The chance of getting *exactly* the same number of successes in each treatment group is .32238, only about 1/3. Still, it is likely to get *approximately* equal numbers; and lopsided results like Table 1 are improbable. I have shaded all the outcomes that are as lopsided as Table 1, or more, in one direction or the other. The combined probability of this shaded area is .00097+.01546+.01546+.00097. This number equals .03285 and is the *P–value* that the published article reports for Table 1.

In our calculation of this result we have assumed that a patient who recovers could equally well belong to GM–1 or to placebo, i.e., that GM–1 has no effect. However, we now find that this assumption would lead us to the conclusion that the actually observed data are improbable, having only a .03285 chance. What are we to do in the face of this paradox? One possibility is simply to accept that the data are improbable. Another possibility, more comfortable for most people, is to give up the assumption that the drug has no effect, instead accepting that it does have an effect. The lower the *P*–value, the more we feel compelled to believe that the apparent drug effect is real. We say, as in this example, that the data show that the difference between drug and placebo is *statistically significant*.

The statistical test that we have used here is *Fisher's Exact Test*. I shall continue to use it as an example since it provides neat, easy to understand illustrations of many ideas. It is the appropriate test when the outcome measure is *binary*, or yes/no. For numerical data, Student's *T*-test is excellent where it is applicable. (However, it is commonly used to the exclusion of all others, even where it is obviously inappropriate: for example in cases where the histogram of the data does not fit the symmetric, bell–shaped curve of the Normal Probability Distribution.). The underlying theoretical ideas that I shall be explaining also apply to the *T*-test and other statistical tests, but the technical details are somewhat more mathematical.

Table 1, by itself, is an example of *descriptive statistics*: it merely describes the particular sample that was actually obtained. However, the inference, prompted by the *P*–value .03285, that the apparent drug effect can be extrapolated to the population as a whole is an example of *statistical inference*. Descriptive statistics can be used freely: they cost nothing more than perseverance by authors and readers, and can be very illuminating. Inference, however, requires thought and care because it involves a risk of error each time. It should be used sparingly, in those instances where it can be justified.

Notice that the probabilities can be calculated exactly, provided that we are able to assume that the patients are assigned to the two treatments with complete randomness: any patient can be in any treatment. If the investigator were to make any form of conscious choice in the selection of the groups, then our assumption that all the possible reassignments are equally likely would be invalid, and the numerical value we obtain at the end, although appearing exact, would be meaningless. It might be close, or far: it's simply a guess with spurious credentials of scientific accuracy. (It is possible to divide the sample into *strata*, to ensure representativeness. If the sample is drawn strictly randomly within the strata then the result can still be valid, although the analysis is more technical.)

How then can we relate this result, describing the particular sample that was actually obtained, to the characteristics of the spinal injury population as a whole? To do this, we have to be able to assume, not only that our sample is *divided* randomly into the two groups, but also that our sample is *obtained* randomly from the population at large. This is not an assumption to be taken lightly, since it is obviously impossible for an investigator to choose subjects from a list of all possible patients. Two sorts of questions arise, and care needs to be taken to ensure explicitly that the answers are satisfactory. First, although the mix of patients seen by a given provider at a given institution may be roughly representative, there are disparities, sometimes large; and also the quality, or merely the conditions, of service by providers can vary in ways that bias the result. Second, the investigator makes choices to include or exclude subgroups within the population: these choices have a direct reflection in the makeup of the target population to which the result can legitimately be extrapolated, and therefore also a direct reflection in the wording of the package insert. We can't say, *on the basis of our data*, that the result applies to females of childbearing age if they were excluded, for whatever good reason, from our sample.

These ideas are summarized in GOLDEN RULE 1: *The valid use of probability theory to calculate the results of statistical tests depends on using data that are sampled randomly from the desired target population. If either (1) your data are not randomly chosen or (2) you cannot clearly identify a target population of which they are representatives, then you should not be using statistical tests.*

Many people feel that statistical procedures are an automatic means that can be applied blindly to "make results scientific." Therefore, they do as many tests as possible (as many as they can get their

software to deliver a number for), so as to look as scientific as possible. Inappropriately applied statistics do not make you appear scientific; they make you appear foolish in the eyes of those who know better.

ADJUDICATION OF PATIENT ELIGIBILITY.

Some time ago, scientists came to notice that, particularly in cancer trials, there were treatments that were so complex or had such side effects that some patients wouldn't complete them, and then often went on to die early. These scientists reasoned that in practice there was little difference whether the patient failed because the treatment was ineffective or because the treatment wasn't followed: a failure is a failure. They developed an approach called *intent–to–treat*. A physician reading our results has a right to the best information about whether starting this treatment, rather than another, will help the patient.

The principle behind this idea is one with which I agree completely. In practice there can be problems. We are trying to estimate a combination of two things: how well a patient is likely to comply, and how well a compliant patient is likely to do. Although it is certainly important to try to estimate both of these, it may only be the second that we have the opportunity to measure with any pretension to scientific method. As for the first, the degree of compliance may be perturbed by the very circumstances of the study. Sometimes when a treatment is experimental people feel that it is a wonder drug and they comply better than if it were routine. Other times patients or physicians feel that an experimental treatment is unproven and won't give it the degree of compliance that they would a treatment known to work, especially if the patient begins to get into trouble. Therefore in some, but not all, studies we are better off confining ourselves to trying to answer the more artificial question of how well a compliant patient is likely to do, but at least give a scientific answer with known characteristics.

Some investigators have turned intent–to–treat into a dogma, that *every* patient that is randomized has to be included in the primary efficacy analysis, no matter what. This no longer makes sense.

The first issue has to do with *eligibility*. For example, if a patient is randomized and later found not to have the disease being studied, then their data should not be included in the primary efficacy analysis. GOLDEN RULE 1 shows that we are attempting to sample randomly from a target population; uncontrolled addition of patients from other populations makes our probability calculations meaningless, biasing the result in ways that cannot be understood or corrected. Despite bitter claims by believers that only an intent–to–treat analysis including such patients can be scientific, the opposite is true.

Next are the related issues of *compliance* and *premature termination*. Sometimes, as noted above, a patient fails to comply with the treatment regimen as defined by the protocol because the treatment is complicated to administer or because it has undesirable side effects. The data of such

patients should be included in the analysis just as if treatment had been followed. However. it also occurs that a patient fails to comply solely for technical reasons of the study. For example, in the GM–1 study, one patient's drug was lost during a room change, and it was then impossible to resupply without breaking the blind; during ordinary clinical practice there would be no problem since drug could simply be ordered from the pharmacy. Similarly, a patient who is prematurely terminated or dies, for medical reasons, should be included; but a patient who dies on study in an automobile accident (or as a result of assassination by the KGB) should be excluded. Thus, a distinction has to be made between *failure to complete treatment* and *failure to complete the study*, and patients should be included or excluded accordingly.

There is also the issue that for a patient lost to follow up there simply may be no rational way to assign or guess a numerical value for the efficacy measure. (Assuming an arbitrary worst case is not rational.)

It is now common to present a "completers" analysis (including every patient who completes teatment and is followed up) and an "intent–to treat" analysis (including every patient randomized). Neither of these is precisely what we want. If our object is to inform a treating physician about the likely prospects for the patient, compliant or not, we ought rather to present a primary analysis that incudes every eligible patient who complies with treatment and can be assigned a reasonable value of the efficacy measure, and then a second analysis that also includes eligible patients who did not comply.

Administratively, there should be several committees conducting the study. They do not need to be large; and for simplicity and economy, they may share members as far as possible and their meetings can be informal, even by telephone. *What is important is that their respective functions be carried out and documented.* The *Coordinating Committee* should be in overall charge. Its members should be blinded to the treatment code until the final dataset is prepared and the planned analyses are performed. There should be an *Extramural Monitoring Committee* that advises the Coordinating Committee on patient safety and on science. As the name indicates, at least a majority of the members should neither be employees of the sponsoring company nor consultants paid to design the trial in question. They should have access to any data they feel necessary, including the treatment code, whenever they wish. There should be an *Adjudication Committee* charged with making *final* decisions about the matters discussed in this section, after reading the patient's file, ensuring the quality of all data relating to efficacy and safety analyses. For trials with simple, well–understood endpoints they might review the files only of those patients with whom an irregular situation is known to exist. For many other trials they should review every patient's file; the cost per patient is small in order to insure the integrity of data that is very expensive to collect. This committee must perforce remain blinded to the treatment code to avoid bias in

its decisions. They should decide whether each patient is *eligible*, *technically ineligible* (has a minor, but not disabling, deviation from protocol), or *truly ineligible*. They should decide whether a patient is *compliant*, *failed to complete treatment*, or *failed to complete study*. Based on these considerations, they should decide which patients are to be included in the primary efficacy analysis, and the values of any missing or questionable variables. When they feel that the data set is accurate enough to allow the efficacy analysis to proceed, they should vote to freeze the data. After this vote any changes in the data should be suspected of biasing the result and therefore should be avoided. They should keep regular minutes of their decisions for each patient and of any general policy decisions.

GOLDEN RULE 2: *Your object is to be able to inform a treating physician reading your results about the potential for the therapy to help the patient. Design your target population thoughtfully. In accordance with* GOLDEN RULE 1, *include in your primary efficacy analysis <u>all</u> and <u>only</u> the patients that fit the target population. Otherwise, the P-value cannot be legitimately computed by the laws of probability and the result cannot be scientific, despite the vigorous claims of advocates of "intent–to–treat" analyses that obligatorily include every randomized patient. Final decisions about inclusion and exclusion should be made by a blinded Adjudication Committee.*

In thinking about testing compounds for clinical use, many people quickly fall into the paradigm of regarding the compound as a kind of candidate, that once it has proven its worth, might be admitted to a club and given the privilege of joining the other members. This attractive thinking is made more plausible by the fact that the compound is usually sponsored by people who stand to make money from approval and who are vigorously pressing its case. Nonetheless, this way of looking at the matter is misconceived.

This attitude that we have a privilege to confer upon the drug, if and when we find it worthy, stands in the way of clarifying our thinking about what patients to include in the primary efficacy analysis. The object is not to pass judgment on the drug, either to help or to hurt it, but to help the patient.

Further, from this point of view, it's important to keep ineffective drugs off the market, but it's also important to get effective drugs on the market. It is to this matter that we now turn.

STATISTICAL POWER: HOW TO DETECT A REAL TREATMENT EFFECT

The calculations in Table 2 assume that there isn't really a treatment effect of the drug. We come now to the question that's more interesting to most readers: "What if there *is* a treatment effect?"

The third column of Table 3 reproduces the probabilities shown in the right column of Table 2. The other columns show what the probabilities would be if the odds for a patient given GM–1 to recover were higher than the odds for a patient given placebo, making it more probable that more of the 8 successes are in the GM–1 column. (The *odds* for recovery are the probability for recovery divided by the probability for no recovery. For example, if the probability for recovery on placebo is .1, or 1/10, then the odds for recovery on placebo are 1 to 9. If, then, the odds for recovery on GM–1 were twice as great, they would be 2 to 9 and the corresponding probability of recovery would be 1/11.)

Suppose that we have made it our rule that if the observed test statistic is in the shaded area, the *critical region*, we will reject the *null hypothesis* that there is no treatment effect. In particular, if it is in the part of the critical region that is lower in Table 3, we will conclude that the drug is better. For a drug 1.5 times as good as placebo the chance of this occurring, and thus of our detecting the treatment effect, are low: only .0462+.0043 = .0505. For a better drug our chances are better. Still, even a wonder drug with 10 times better odds gives us a chance of detecting that is only .422+.264 = .686, slightly better than 2 in 3.

Why is the chance of detection so poor in this example? The sample size is very low. In fact, reflecting on Table 3 suggests that, for Table 1 to have been observed, either the effect of GM–1 must be very large or else the investigators must have had the luck, good or bad, to draw an unusual sample.

Assuming that your statistician arranges other matters competently, two factors determine your probability of detecting a treatment effect, the *statistical power* of your test. One factor is the size of the

TABLE 3: RECOMPUTATION OF PROBABILITIES IN TABLE 2, WITH INCREASING VALUES OF ODDS FOR PATIENT RECOVERY ON GM–1 COMPARED TO ODDS FOR RECOVERY ON PLACEBO

Successes for placebo	Successes for GM-1	Probability, if GM-1 odds equal to placebo odds	Probability, if GM-1 odds 1.5 times placebo odds	Probability, if GM-1 odds 2 times placebo odds	Probability, if GM-1 odds 4 times placebo odds	Probability, if GM-1 odds 10 times placebo odds
8	0	.000966183	.000169026	.000042426	.000000951	.000000003
7	1	.015458932	.004056631	.001357639	.000060873	.000000422
6	2	.087922677	.034608137	.015443141	.001384851	.000024020
5	3	.234460472	.138432546	.082363421	.014771739	.000640526
4	4	.322383149	.285517127	.226499407	.081244565	.008807233
3	5	.234460472	.311473230	.329453682	.236347824	.064052606
2	6	.087922677	.175203692	.247090262	.354521736	.240197272
1	7	.015458932	.046207567	.086888883	.249333968	.422324873
0	8	.000966183	.004331959	.010861110	.062333492	.263953046
	Totals:	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000

treatment effect. This is not under your control. The other factor is the sample size. This is very much under your control, within your constraints of ethics, budget, and ability to recruit subjects.

Shortly, I shall talk about how to plan your sample size. However, first we present another commercial from our sponsor, the laws of probability. Before we can talk realistically about how much things cost, we have to make sure we understand the implications of our prices.

IF IT'S IMPROBABLE, *CAN* IT HAPPEN? IF IT'S PROBABLE, *MUST* IT HAPPEN?

We found above that datasets as lopsided as Table 1 have only about a 1 in 30 chance of occurring merely as a result of random variability, if there is no drug effect. Does this mean that they *can't* occur? They're effectively impossible? Can we conclude definitely that the apparent drug effect *must* be real?

Conversely, if we set out upon a trial with a .686, or 2 in 3, chance of detecting a treatment effect, can we rest assured that our expensive investment in conducting the trial will indeed vindicate our even more expensive investment in development of the compound?

The answer to these and to any questions involving probabilistic events is a resounding "Maybe!"

If the morning weather announces a 20% chance of rain, many of us think, "Twenty per cent. That's low. It won't rain. Let's leave the umbrella home." However, if we stood back from the particular needs of today and thought about the matter with any clarity we would conclude that this behavior will get us wet on 1 out of any 5 such days. The laws of probability don't *protect* us. On the contrary, although they guarantee that 80% of such days will be dry, they also equally guarantee that 20% will be wet.

The fallacy in our thinking is the inference from the fact that the probability is low to the conclusion that it won't rain. The only rational way to think about it is in terms of the costs and benefits of a plan for long–term behavior: is it worth carrying the umbrella 5 times to avoid rain once?

There is perhaps another, subtler, fallacy in our thinking. When chances are low, say 1 in 20, we think it can't happen on the first try, that it wouldn't happen until sometime in the middle, or the end, of a series of trials. Therefore, if we're only going to do it once, or only a few times, we're sure to escape. This thinking is 100% incorrect. If a probabilistic event is known to happen exactly once in a series of 20 actual trials, then it can happen with equal probability anywhere within the series, early or late or even first.

We don't usually do more than a very small number of trials for any one compound, so it's more useful to think of these ideas as applying across the development programs for several compounds. Even

if we're only ever going to test one compound once, we still have to take the probabilities seriously. Think about it this way. The chance of getting four straight heads in four flips of a coin is 1 in 16, about .0625. This might seem remote, but if all the students in a classroom of 30 try for it the chances are excellent (about .856) that *somebody* will succeed. I consult for many clients doing clinical trials and some of them are going to be unlucky. That somebody could be you as well as any of the others.

You have to think of the human, financial and other costs involved in taking risks as real costs and budget for them, according to costs and benefits, soberly. In a way they are even costlier because they are uneven and unpredictable. You can take a 1 in 10 risk ten times and pay nothing on nine of them; but the remaining time (possibly the first) you will pay for all ten at once. This kind of irregularity is what makes me buy life insurance: it's highly improbable that I'll die before the kids get through college and the house is paid off, but if it did happen the costs would otherwise be too devastating.

GOLDEN RULE 3: *If the probability of an event is* n *out of* m, *then confidently expect* <u>*both*</u> *that it will happen about* n *out of* m *times* <u>*and*</u> *that it will not happen the other* m – n *times. Budget ahead for losses either way.*

There are really 4 possible outcomes depending on whether there really is or is not a treatment effect and on whether the sample we happen to pick leads us to believe or to disbelieve in such an effect.

	We are led to believe that there *is* a treatment effect	We are led to believe that there *isn't* a treatment effect
There really *is* a treatment effect	(win)	"Type II Error"
There really *isn't* a treatment effect	"Type I Error"	(win)

The FDA is charged with ensuring that both types of error are avoided. However, there are enough applicants trying to chisel on Type I that the FDA tends to overemphasize it, although they do often require applicants also to calculate their statistical power, or probability of avoiding Type II.

This tendency for an adversarial relationship between the applicant and the FDA, centering on Type I Error, muddies the water in a number of ways that are unfortunate. First, many applicants do have a native sense of responsibility that would lead them to see that having ineffective compounds on the market is bad business: aside from the very poor ethical and legal implications, it makes it much more difficult subsequently to develop or get FDA approval or community acceptance for compounds that do work. Everybody has a legitimate interest in keeping bad drugs off the market.

This feeling that the objective is to beat the FDA at the Type I game also tends to blind applicants to the importance of Type II. Since, as I shall explain shortly, getting enough statistical power depends on having a large enough sample, and since sample size costs money and inconvenience, many

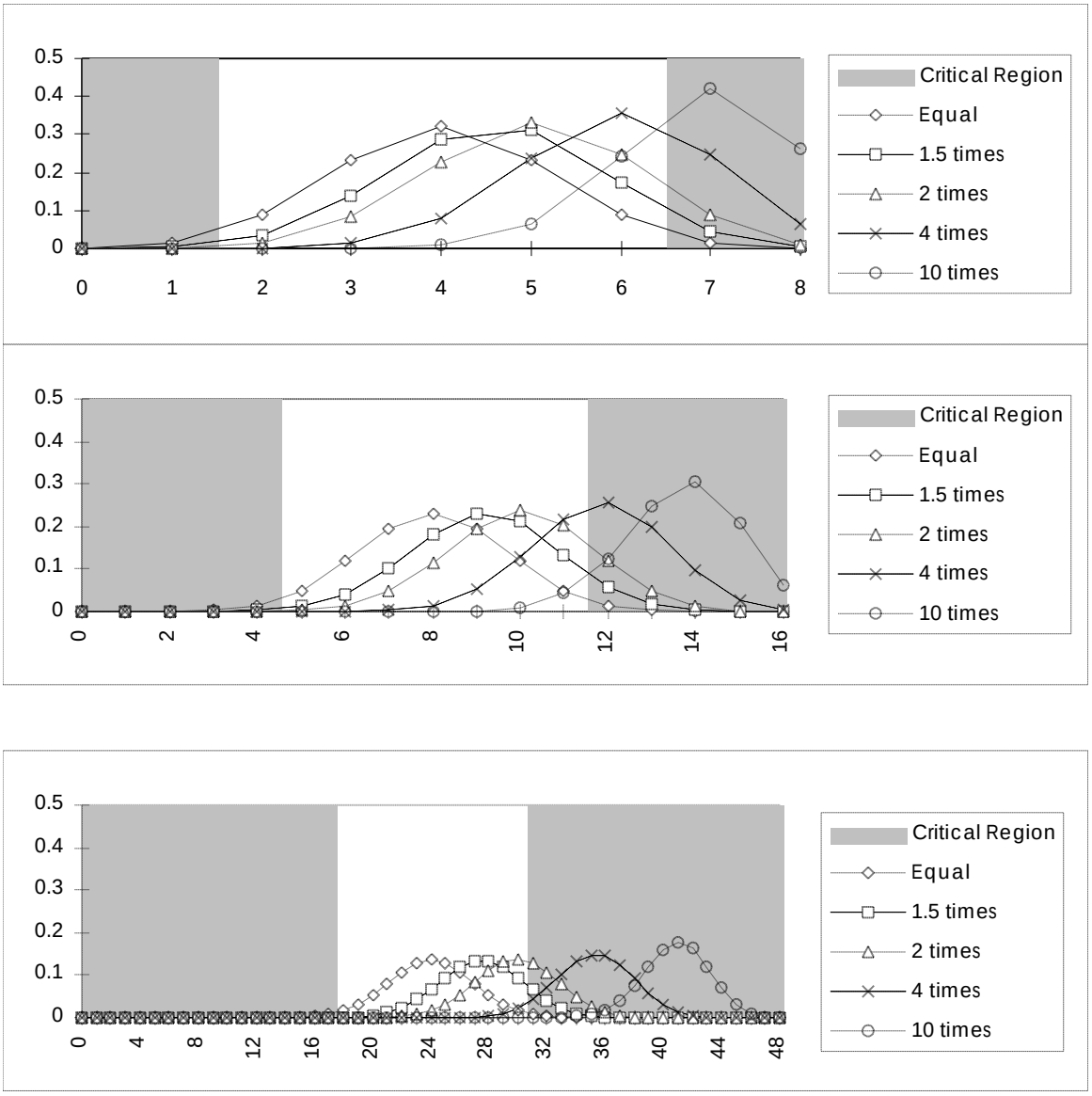
applicants view statistical power as an expensive theoretical nicety that makes the statisticians happy. This is far from true.

GOLDEN RULE 4: *The only way to get a compound approved is to avoid Type II error (failing to detect a treatment effect when there is one). If you don't spend enough extra to get adequate statistical power, then the still rather large amount of money that you are investing in your trial is buying you nothing but the likelihood of no approval. If you're really dumb enough to gamble your compound away, then at least don't pay for a clinical trial to do so: see a fortune teller instead; they're cheaper.*

SAMPLE SIZE AND STATISTICAL POWER.

Table 3 showed how the size of the treatment effect determines the probability that you will be able to detect it. The other factor determining statistical power is sample size. The top of Figure 1 is a graphical representation of Table 3. We see the "equal" curve symmetrical in the center, with the gray critical region covering its right and left ends. The curves for higher effect sizes are skewed progressively more to the right, and therefore put progressively more probability in the right part of the gray critical region. The amount of probability each curve has in this right part represents our chance of detecting an effect.

FIGURE 1: PROBABILITIES WITH 14 IN EACH GROUP, WITH 28 IN EACH GROUP, AND WITH 84 IN EACH GROUP.



In the middle part of Figure 1, I have revised the computations, assuming double the number of patients (28 per group for a total of 56) and double the total number of successes (16) to split between the groups. (There is no reason in advance when starting a new experiment to expect any particular total number of successes; you only know this number afterwards. I have merely picked 16 to keep things proportional so as to illustrate the probabilities neatly.) The effect of the larger sample size in the middle is that each of the curves is less spread out, more bunched together horizontally, than its counterpart on the top. The result is to separate the curves, and therefore to make it easier to detect a treatment effect. Although the visual change in the shape of the "equal" curve seems subtle, it's pronounced enough to increase markedly the relative width of the gray critical region. This, combined with the similarly increased narrowness of, for example, the "10 times" curve, puts much more of their probability of the "10 times" curve within the critical region, increasing our chance of detection.

In the lower part of Figure 1, this process has gone much farther. Here I have assumed 84 patients in each group, with 48 successes to divide between them. Now the "10 times" curve lies just about entirely within the right part of the gray critical region, the "4 times" curve is nearly so, and even the "2 times" curve gives almost half its probability to the critical region.

Thus we see that for any effect size, increasing the sample size narrows the curves — it sharpens the focus —, increasing the probability of detecting the effect, eventually making detection almost certain.

However, the figure also shows that this relationship is slower than many people think or would like. To have a chance of detecting a doubling of the recovery rate, a large decrease in human suffering in a disabling disease like spinal cord injury, requires a very large sample. (Still, we shouldn't be surprised. We are all familiar with television ratings and with opinion polling, assessing the relative performance of two programs or of either of two candidates against a third. We know the sample sizes used and the margin of error that the polls report. For a treatment with yes/no outcome, our position is analogous.) If a trial fails to reach $P < .05$ many people believe that the treatment effect, if any, must be small. Figure 1 shows that such beliefs are often totally unfounded, unless the investigators have supplied an explicit calculation of their statistical power. Type II error abounds.

GOLDEN RULE 5: *(a) There's nothing sacred about .05 except for the fact that it's traditional. It's a measure of risk. How much risk can you afford? (b) If you don't get P < .05, or whatever, it doesn't mean that the compound isn't effective, unless you have also controlled for Type II Error. It means that you have failed to find a treatment effect. How hard did you look?*

HOW TO PLAN FOR AN ADEQUATE SAMPLE SIZE.

How large a sample will be enough? First, we have to decide how much power, how much protection against Type II error, we need. As GOLDEN RULE 3 suggests, this should be weighed in terms of costs and benefits. In this respect, GOLDEN RULE 4 suggests that the protection against Type II error usually ought to be comparable to the protection against Type I, or about 95%. There's simply too much to be lost: you can lose your entire investment in developing the compound so far, and society can lose the benefit of a useful treatment. In my opinion, the common practice of testing with power less than about 90% is often negligent, from your point of view, and reckless, from society's.

If we know how much power we want, then the thing we need to know, in order to calculate the sample size, is the effect size. Unfortunately, this is precisely the thing that we don't know and, in fact, are trying to estimate. Indeed, my experience is that the literature often contains depressingly little upon which to base even an educated guess. Sometimes we do have pilot data. For example, the data of Table 1 were considered as pilot data to design a larger, multi–center clinical trial.

What statistics from the pilot data are important? A simple approach is to use the observed values. For example, Table 1 indicates a recovery rate of 7.1% for placebo and 50% for GM–1. We would estimate Δ, the difference between treatments, to be 50% – 7.1% = 42.9%.

There are tables and software packages that allow us to use these numbers to estimate the required sample size. In order to do so, we would have to decide how much protection we need against Type I Error. This number, called α, denotes how low the *P*–value will have to be in order for us to conclude that there is a difference between the compound and placebo. Many scientists customarily set $\alpha = .05$. For our example, in which we want to test an assumed recovery rate of 50% versus one of 7% at $\alpha = .05$, the tables or software will tell us that we need a total sample of about 54 patients, or 27 per group.

However, we might also give some thought to the fact that these estimates, 50% and 7%, are based on a sample of only 28 patients, so there might be a lot of random variability in them. The true values could be higher or lower than we observed in the pilot data. One way of quantifying this variability is to use a *confidence interval*, to see the whole range of probable values, rather than a *point estimate*, which just indicates the most likely number within that range. There is a standard procedure that gives an exact one–sided 95% confidence interval for the recovery rate. This procedure would indicate that the recovery rate of patients given placebo could be anything up to 29.7%, a lot higher than we might previously have thought. Similarly, we would find that the recovery rate for patients with GM–1 could be as low as 26.4%. (These results do not necessarily indicate that GM–1 could be worse than placebo: the probability for both of these extremes to happen at the same time is vanishingly small. If we put our data together to compute a 95% confidence interval for the difference Δ, we find that

approximately it might be anything between 6.9% and 79.5%.) Altogether, the estimates underlying the calculation in the preceding paragraph may be too optimistic.

There is another, completely different way of thinking about the effect size: we do not try to figure out what the effect size is likely to be, but rather what effect size would be the minimum that would make the compound clinically worth while and commercially viable. The philosophy here is "Why speculate about what the effect size could be? Let's figure out what would work for us." This requires a value judgment based on the facts of the individual case. In the example, the prognosis for spontaneous recovery after spinal injury is poor: case registries suggest perhaps about 7%. Therefore an increase of $\Delta = 7\%$ would mean a doubling of the recovery rate. For such a devastating disease, this would be quite an improvement.

In this particular example, both methods, using pilot data to estimate the range of possible effect sizes and using judgment to estimate the minimum acceptable benefit, lead comfortingly to approximately the same conclusion: in the worst case, plan for drug to raise the recovery rate from 7% to at least 14%.

In other studies, the answers delivered by these two methods might not be as consistent. A good portion of the range of likely effect sizes might lie *below* the minimum acceptable. In this case the prospects for the clinical trial look dim, and one wants to weigh carefully whether a trial is worth the expense and the exposure of patients to risk. The opposite would be that there might be a good likelihood that the effect size lies well *above* the minimum. Here one might consider a sequential design, discussed below.

GOLDEN RULE 6: *To plan the sample size, use your pilot data, if available, to develop a guess about the whole <u>range</u> within which the effect size might possibly lie. <u>Also</u>, make a judgment as to what the minimum effect size is that would make further testing and development of the compound clinically and commercially worth while. Buy enough protection, enough sample, to cover yourself for the minimum effect size and also across a reasonable segment of the likely range.*

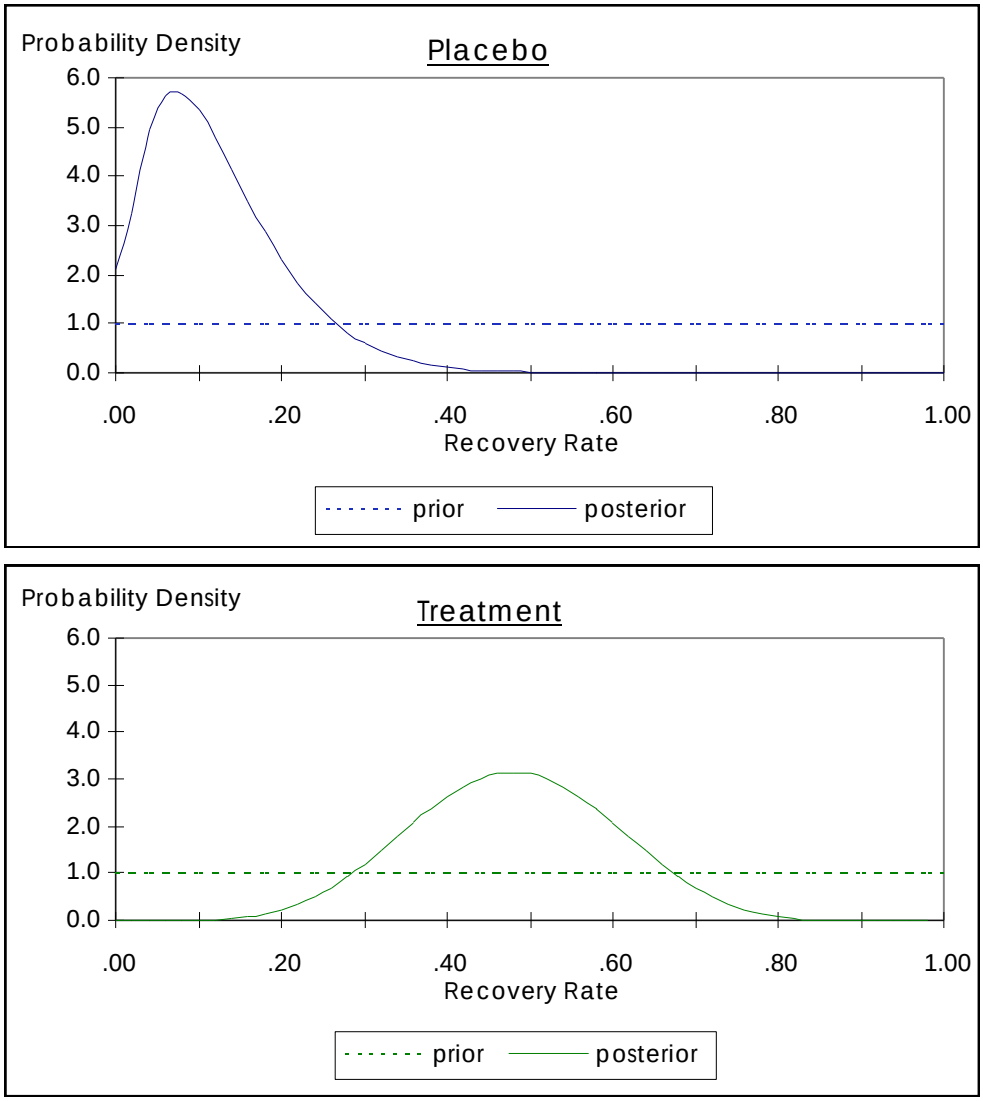
Our current thinking about the GM–1 example, leads to a very large sample size. For 7% versus 14%, the tables tell us that we would need a total sample of 1042, or 521 per group if we want to get 95% power at $P = .05$. Even for 10% versus 20%, we would need 345 per group. This seems so much larger than what we compute using the simple estimates. Is such a large sample really necessary? We can check our ideas by using a more probing analysis that illuminates the issues further. This is a *Bayesian analysis*.

(The Bayes procedure can be found in intermediate level statistics books and will be well known to your statistician. A note, to say what the technical jargon is: in the following paragraphs we are assuming that the data in the rows of Table 1 have independent binomial distributions whose parameters

have a conjugate family of Beta Distributions with a Uniform prior. The results shown in Figures 2 and 3 and in Table 4 follow by direct numerical computation from these assumptions. Table 5 then follows from sample size tables or software, as discussed previously.)

In both halves of Figure 2, the horizontal axis represents the possible values of the patient recovery rate and the vertical axis represents the relative probability density that the value is the true one.

FIGURE 2: PRIOR AND POSTERIOR PROBABILITIES FOR TABLE 1 DATA.



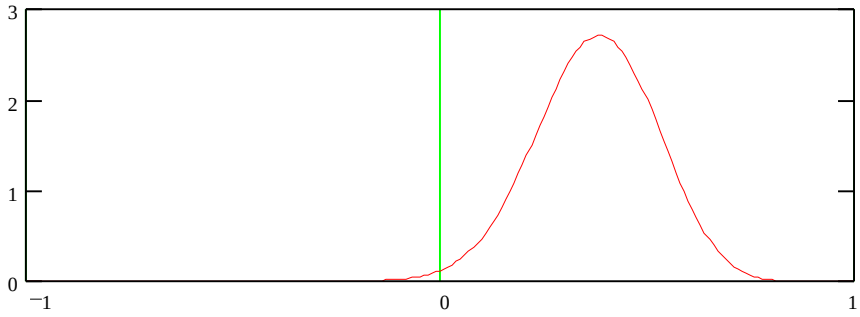
The straight dashed line indicates the state of our ideas before observing any data, our *prior probability density*: the recovery rate could be any value from 0 to 1 with equal probability. The curved line in the top half of the figure is our *posterior probability density* for the placebo group; it shows how our ideas

have changed after observing 1 recovery out of 14 patients assigned to placebo. Most of the probability is that the recovery rate is below 20%, although there is some chance that it is between 20 and 30%, a small chance that it is between 30 and 40% and even a tiny chance that is above 40%. Note that these numbers agree substantially with the 1–sided 95% confidence interval noted above; however they give a clearer idea of relative probabilities within the range. This analysis shows that possibilities other than the naive point estimate 1/14 = 7.1% are real.

Similarly the curved line in the bottom of Figure 2 shows the posterior probabilities for patients given GM–1. We see that, although the most probable value of the recovery rate is the point estimate 50%, considerable divergence from that value is perfectly possible; even values above 75% or below 25% cannot be completely ruled out.

Figure 3 combines the two parts of Figure 2 into an estimate of the *posterior probability* of Δ; it shows the relative likelihood of various values of the difference Δ in recovery rates after observing the

FIGURE 3: POSTERIOR PROBABILITY DENSITY OF Δ



data of Table 1. The possible values of Δ range from −1 (if placebo is completely effective and GM–1 is completely ineffective) through 0 (if they are equally effective) to 1 (if placebo is completely ineffective and GM–1 is completely effective). The posterior probability that placebo is more effective than GM–1 equals the area under the part of the curve that is to the left of the vertical line at Δ = 0. This area is small, equaling .0071. The bulk of the probability is rather broadly spread out between 0 and about .7.

The next question is how to use these estimates to design the sample size in a new trial. Table 4 shows numerical values derived from the posterior probabilities in both halves of Figure 2, in an effort to

TABLE 4: JOINT PROBABILITIES FOR VARIOUS RECOVERY RATES ON PLACEBO AND ON GM–1

		Placebo									
		0–.1	.1–.2	.2–.3	.3–.4	.4–.5	.5–.6	.6–.7	.7–.8	.8–.9	.9–1
Drug	Prob	.464	.382	.132	.030	.005	.000	.000	.000	.000	.000
0–.1	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.1–.2	.004	.002	.002	.001	.000	.000	.000	.000	.000	.000	.000
.2–.3	.046	.021	.018	.006	.001	.000	.000	.000	.000	.000	.000
.3–.4	.163	.076	.062	.022	.005	.001	.000	.000	.000	.000	.000
.4–.5	.286	.133	.109	.038	.009	.001	.000	.000	.000	.000	.000
.5–.6	.286	.133	.109	.038	.009	.001	.000	.000	.000	.000	.000
.6–.7	.163	.076	.062	.022	.005	.001	.000	.000	.000	.000	.000
.7–.8	.046	.021	.018	.006	.001	.000	.000	.000	.000	.000	.000
.8–.9	.004	.002	.002	.001	.000	.000	.000	.000	.000	.000	.000
.9–1	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

show the probability of various combinations. For example, from Figure 2 the posterior probability that the recovery rate on placebo is in the range 0 – .1 is .464 (from the top margin of the first column); the posterior probability that the recovery rate on GM–1 is in the range .2 – .3 is .046 (from the left margin of the third row); and therefore the probability of the combination if these events is the product of the two numbers, namely .021 (from the body of the table; first column, third row).

The entries that are shaded in Table 4 are all those that individually have a probability of at least .01. This choice of cutoff point yields a group of entries that jointly has a probability of .958. Stated another way, we can be approximately 96% sure that the true effect of drug and of placebo will lie within the shaded area. It is therefore a reasonably safe bet that the realistic possibilities that we have to plan for are those in the shaded region.

How large a sample will we need in this shaded region? Table 5 lists the required sample sizes corresponding to the midpoints of the shaded entries in Table 4, and therefore approximately

TABLE 5: SAMPLE SIZE (COMBINED PLACEBO AND DRUG)
CORRESPONDING TO THE MIDPOINTS OF THE SHADED ENTRIES IN TABLE 4;
STATISTICAL POWER EQUAL TO .95 FOR $P = .05$.

Drug \ Placebo	0-.1	.1-.2	.2-.3	.3-.4	.4-.5	.5-.6	.6-.7	.7-.8	.8-.9	.9-1
0-.1										
.1-.2										
.2-.3	174	860								
.3-.4	94	252	1120							
.4-.5	58	122	304							
.5-.6	40	72	140							
.6-.7	28	46	78							
.7-.8	18	30								
.8-.9										
.9-1										

representative of their needs. Although sizes are low in the lower left of the table, they rise extremely steeply as we approach the upper right of the shaded area (the main diagonal of the table as a whole). To guard against all the possibilities in the shaded area we would need to take the worst case: 1120 patients altogether. The interpretation would be that this many patients would give us at least 95% power in at least 95% of the most likely possibilities. This conclusion is perhaps a bit too extreme and we might back off to our previous estimate of a few hundred per group.

Some management executives might find this conclusion outrageously expensive and argue either that the true value of Δ "shouldn't be too far from" the observed value .43 or that there's no need to set β, the probability of Type II error, as low as .05. Unfortunately, both these arguments are wishful thinking. Considering the small total sample size in the pilot study of only 28, we can expect that our estimates are subject to a lot of random variability. Indeed, this variability is objectively quantified in Table 4. Next, although we'd be happy to see again a treatment effect as good as that observed in the pilot study, and chances are that we shall; still there's also a realistic chance that it's not that high. Thinking about it more soberly, an increase from 10% on placebo to 20% on drug would represent a considerable relief in human suffering to spinal cord injury victims, and therefore would also represent a considerable commercial possibility for the manufacturer. Take this value as, not the most likely possibility, but a real possibility that could materialize. If so, the statistical power has the simple direct meaning that it is the probability that we shall be lucky enough to draw a sample that will exemplify the

treatment effect and therefore get approval from the FDA. If I were a stockholder, I wouldn't think 95% was too high.

The prospect of a clinical trial with 345 patients per group may sound expensive, but it sounds cheap compared to the potential of throwing away the development cost of the drug, the cost of the pilot study, the cost of the new study, and the profit potential if the drug were approved.

All right then, but how can such expenses legitimately be kept down? One partial solution is to run larger pilot studies. This would have the effect of narrowing the curves in Figure 2 and Figure 3, reducing the uncertainty: the "worst case" estimates would be closer to the "typical" estimates and you wouldn't have to pay so much for insurance. Of course, larger pilot studies are more expensive in themselves. A second possibility is to reduce the variation within groups by using a crossover design. Another possibility is to focus on subgroups of your population that are more homogeneous. Still another possibility is to consider a sequential design. I shall discuss these options further below. In general, collaborate closely with an inventive and empathetic statistician.

CROSSOVER DESIGNS.

In a *parallel group design* we have two separate groups, one for placebo and one for the compound. This means, roughly, that each patient is compared to the *average* of the other group. An alternative possibility is to start some patients out with placebo and then cross them over to the compound (the P→C arm), and start some patients out with the compound and then cross them over to placebo (the C→P arm). The good point of this idea is that each patient's outcome on the compound is compared to *their own* value on placebo, reducing the variance within groups and decreasing the sample size, perhaps considerably.

As attractive as this idea is, sometimes there are drawbacks. First, preliminary data are sparse in the literature, making sample size estimates more of a guess. Second, sometimes they simply are inapplicable: for example, in studies of acute trauma, or of terminal illness. Another difficulty is the possibility of *carry over* of effect: when the patient leaves one treatment, there are still effects that carryover into the observation period of the other. These effects may either enhance or interfere with the second treatment. The reason for having both the P→C arm and the C→P arm is to help avoid biasing the *direction* of the treatment difference: both get a chance to go first. However, depending on whether the two treatments are cooperative or inhibitory, there may be a dilution or an enhancement of the *size* of the treatment difference, changing the apparent *P*–value and making interpretation of results problematic. (The 2 arms have to be understood as 2 subgroups and thus subject to the considerations expressed below.)

Crossover studies can be an excellent idea, but the number of trials in which they can be used is limited and they need to be designed in close collaboration between experienced clinical and statistical personnel. GOLDEN RULE 7: *Use a crossover design to reduce variability and sample size, if the chance presents itself, and if you're sure you know what you're doing.*

SEQUENTIAL DESIGNS: HOW TO (PARTLY) HAVE YOUR CAKE AND EAT IT TOO.

The analysis in the section before last left us in a frustrating position. We apparently have to buy a far larger sample than we probably need, merely in order to have insurance against the small but realistic chance that Δ might be about .1. Not only does this enormously increase the cost of the trial to us, but it delays publication of the results and the wide availability of the drug to the population of spinal injury victims. It also delays the time until the drug becomes profitable.

Couldn't we start out planning for a large trial, but also take an interim look along the way? If things go as they are likely to, we can declare an early win and get out; if not, we just keep on until the end. The answer is that we can indeed do this, but we need to plan ahead and make appropriate corrections.

Some of the issues in designing a sequential clinical trial can be illustrated by the following anecdote. Two friends, *X* and *Y,* have coffee together. They agree to toss a coin to see who pays. *X* tosses, *Y* calls "Heads," and the toss comes up Tails. *Y* says, "Well, let's try for best 2 out of 3." *X* refuses and tells *Y* to pay up. As many of us intuitively feel, *X* is right. We can analyze why exactly. The possible outcomes for 3 tosses are listed in Table 6.

TABLE 6: THE 8 POSSIBLE OUTCOMES FOR 3 COIN TOSSES.

Outcome	Toss 1	Toss 2	Toss 3	Total # Heads	Probability	Result
a:	Heads	Heads	Heads	3	1/8	*Y* wins on 1st toss
b:	Heads	Heads	Tails	2	1/8	*Y* wins on 1st toss
c:	Heads	Tails	Heads	2	1/8	*Y* wins on 1st toss
d:	Heads	Tails	Tails	1	1/8	*Y* wins on 1st toss
e:	Tails	Heads	Heads	2	1/8	*Y* wins on 3rd toss
f:	Tails	Heads	Tails	1	1/8	*X* wins
g:	Tails	Tails	Heads	1	1/8	*X* wins
h:	Tails	Tails	Tails	0	1/8	*X* wins

First, although it is a fair bet for *Y* to win if Heads comes up in the first toss (happens in 4 outcomes out of 8: *a, b, c,* and *d*) and it is also a fair bet for *Y* to win if Heads comes up at least twice in three tosses (happens in 4 outcomes out of 8: *a, b, c,* and *e*), it is *not* a fair bet if *Y* can win in *either* of these (happens in 5 outcomes out of 8: *a, b, c, d* and *e*). Similarly, if a clinical trial had an interim analysis with a win if *P* is less than $\alpha = .05$ and then also had a *final analysis* with a win if *P* is less than $\alpha = .05$, the *combined* probability of Type I Error would be greater than .05, making it too easy for the trial to succeed. Therefore, in order for the combined risk to be .05, the rejection levels for the individual analyses have to be adjusted. (Statisticians will be familiar with methods such as the one of Lan and DeMets for designing these studies.)

Next, note that, for example, outcomes *b, c,* and *e* all have a total of 2 heads. They also all have the same probability: 1/8. The only difference is the timing of the 2 heads. For the same total number of heads, it is merely a matter of random chance whether those heads occur early or late. Similarly, in a clinical trial the same total number of successes might with equal probability tend to occur among the patients enrolled early or those enrolled late. One of the reasons for choosing a sequential design in preference to a fixed sample size is the hope that the successes may occur early and the interim analysis may thus allow the trial to end early. On the other hand, if the interim analysis does not show many early successes this is not necessarily a cause for discouragement.

Finally, the example shows that changing rules in the middle of a trial changes the odds, and since many people know this, weakens the study's credibility with the FDA and in the scientific community.

In short, a *sequential* statistical design can be valid provided appropriate steps are taken to adjust the *P*–value. There are two types of advantage in using one. First, there is the probabilistic effect described above. As in the World Series, where there is no point in continuing after one team has won 4 games, also in a clinical trial agreeing to stop as soon as one treatment appears decisively superior results in a shorter trial *on the average*, although the *maximum* sample size is slightly larger. The second reason is that one may be uncertain beforehand about the size of the drug effect to be expected. Maybe a modest effect size (meaning that the drug is worth while and commercially important but that the sample size will have to be large) is perfectly possible, but there is also evidence that the effect size could be larger than this (so that a smaller sample would be enough). Often the second of these advantages is the more persuasive.

GOLDEN RULE 8: *If you have pilot data that indicate a good likelihood that the actual effect size may be much higher than the minimum size calculated under* GOLDEN RULE 6*, then consider a sequential design.*

Unfortunately, this option is not always available. For example, if the primary measure of efficacy is the extent of recovery at the 1–year follow–up exam and the planned recruitment period is 2 years, then by the time complete data are available for the first year cohort, the second cohort is already enrolled.

Sequential designs require a different approach to efficacy analysis. As other authors have argued, in a sequential trial there is an asymmetry between showing that the new treatment is better than placebo and showing that it is worse. Once it becomes clear that it is unlikely that any large benefit of the new treatment will be proven, it is unethical and infeasible to continue the trial merely to establish whether the new treatment is actually worse or only not much better. Therefore, there are two one–sided tests:

efficacy — Can we show that the new treatment appears to be effective: i.e., can we disprove the hypothesis that the new treatment is no more effective than placebo?

sufficiency — Can we show that the new treatment does not appear to be sufficiently better than placebo to merit continuing the trial?

A positive result on either of these tests stops the trial: in the first case the drug is shown more effective; in the second the trial is pointless.

PLAYING THE WINNER: MAXIMIZING THE ETHICAL BENEFIT TO THE PATIENTS YOU STUDY.

Another possibility that can arise, in parallel with the opportunity for a sequential design, is that preliminary evidence may suggest that the new compound is much better than placebo. Whether it is indeed so, will only be established *after*, and as a result of, the trial, but this is cold comfort to the subjects who are actually in the trial and to whom you have an ethical obligation. Shouldn't there be a way of randomizing so that most of the patients in the trial will receive the better treatment?

In answer to this, the statistician Marvin Zelen proposed the classic *play the winner* randomization scheme. Flip a coin to see whether to give the first patient placebo or the new compound. Observe the outcome. If the outcome is good, then assign the same treatment to the second patient; if not, then switch. Keep playing the winner as new patients arrive. There is an adaptive effect: if the new compound is indeed much better, then most of the sample will be allocated to it; if not then the allocation will be about 50:50.

There are studies where practical difficulties prevent this. If the observations require long follow–up, then a patient's outcome may not be known by the time the next is ready to enter. Also, if treatment is administered by a small group of physicians to whom the treatment or the outcome is

obvious, they may be able to guess the assignment of the next patient and manipulate entry in ways that can bias the conclusions.

A response to these difficulties is the modified play the winner scheme, of which there are many varieties. You might start out by writing "placebo" on 10 slips of paper and "new compound" on 10 more, putting the 20 slips in a hat and stirring them. When any new patient enters he is randomized by drawing a slip from the current contents of the hat. The slip remains out until the result for that patient is known, after which accordingly either the same slip or one of the opposite denomination is replaced in the hat.

GOLDEN RULE 9: *If you have preliminary evidence that the new compound is much better than the standard therapy, think about using a play the winner randomization.*

WHY THERE CAN BE ONLY A SINGLE, PROSPECTIVELY CHOSEN, PRIMARY EFFICACY ANALYSIS.

As our discussion of sequential trials indicated, doing multiple tests can unfairly increase probabilities beyond their nominal values. Indeed, it's intuitively obvious: if I continue to take .05 chances (1 in 20), I shall eventually get a hit. If I can find enough ways to analyze my data — if I can perform enough tests — I can eventually "prove" the therapeutic efficacy of rainwater.

Table 7 shows the result of performing multiple, independent tests. If the tests are not independent (if they overlap so that they are partially retesting the same thing) then the effective P–value

TABLE 7: THE COMBINED *P*–VALUE FOR MULTIPLE INDEPENDENT TESTS, EACH AT A NOMINAL *P*–VALUE OF .05.

Number of tests	1	2	4	7	10	20
Combined *P*–value	.050	.098	.185	.302	.401	.642

doesn't climb as fast. We often see in medical journals, even the most prestigious, tables where entire columns of results are compared to a baseline column, altogether perhaps 20 to 100 tests. Such tables are always decorated with asterisks and *P*–values that statisticians find as meaningless as splattered paint.

GOLDEN RULE 10: *Each statistical test (each P–value) represents a risk of error. As I perform more tests, my combined risk of error rises above the nominal value and eventually becomes almost certain. Therefore, fishing expeditions in which everything in sight is tested are <u>not</u> "good science;" they are ridiculous and possibly fraudulent. Good science is a small number of prospectively, and thoughtfully, chosen tests.*

The FDA is well aware of the possibilities of fishing in your data until you find a good looking result, and they will take steps to ensure that the combined risk in accepting a confirmatory study is no higher than what they consider acceptable.

GOLDEN RULE 11: *A clinical trial can have only a single primary efficacy analysis, and it must be prospectively chosen and specified in detail. That analysis can have multiple components — population subgroups, treatment regimens, or endpoints —, but the probabilities must be combined and adjusted explicitly.*

The next few sections of this chapter are about how to approach these three types of problems: multiple subgroups, treatments, and endpoints.

WHAT TO DO WITH MULTIPLE SUBGROUPS IN YOUR PATIENT POPULATION.

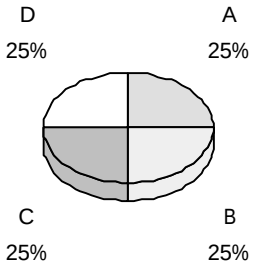
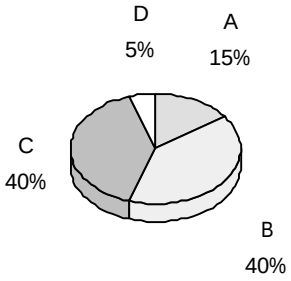
Table 5 showed us that the required sample size depends strongly on the effect size. For example, if the drug can boost the recovery rate from 25% to 45% then the table shows a combined sample size of 304 to get 95% protection against Type II Error. However, if the recovery rate on placebo is only 15% and the drug only raises it to 25% then we need a combined sample with 860 subjects. If we planned for the first of these, and it was really the second, we would have a problem: our 304 subjects would only give us 53.6% protection. The question we would like to look at now is what if our population were inhomogeneous and there were two distinct subgroups (say, young versus old, or incomplete versus complete injury) corresponding to these two realities. What would that mean for us?

TABLE 8: A HYPOTHETICAL POPULATION WITH 4 SUBGROUPS.

Group	Recovery rate, placebo	Recovery rate, drug
Type A	0%	0%
Type B	15%	25%
Type C	25%	45%
Type D	90%	90%

To make the illustration more effective, let's imagine that there are four subgroups, as in Table 8. To be concrete, we might imagine that the Type A patients had an anatomical transection, so that nothing could help them; that Type B patients were quadriplegic; that Type C patients were incomplete paraplegics; and that Type D patients had good function in their arms and in one leg so that, with good physical therapy, they would be likely to recover in any case.

TABLE 9: STATISTICAL CHARACTERISTICS FOR TWO HYPOTHETICAL MIXTURES OF THE SUBGROUPS FROM TABLE 8.

SCENARIO I	SCENARIO II
	
Pooled Recovery Rate Placebo: 32.5 % Drug: 40.0 % Protection against Type II: 44.8% (with total sample of 600)	Pooled Recovery Rate Placebo: 20.5 % Drug: 32.5 % Protection against Type II: 90.4% (with total sample of 600)

The effect of having these subgroups depends on their relative prevalence within the target population from which we sample. Table 9 illustrates two possible scenarios. In Scenario I, the native prevalence of the four subgroups is equal. In the pooled mixture, the overall recovery rates are 32.5% for placebo and 40% for drug. These are close. If we use the total sample size of 600 that was suggested above, we shall have only 44.8% protection against Type II Error. In Scenario II there is less representation from Types A and D. This leads to pooled recovery rates of 20% for placebo and 32.5% for drug. These are more separated, and our protection against Type II is a healthier 90.4%.

If we thought about it we would recognize the plausibility of the idea that, when we pool, the larger subgroups will drive the outcome as a whole. Let's look carefully at the consequences. These scenarios illustrate two basic possible outcomes. In either we're likely to make a mistake unless we're careful and self–disciplined in interpreting of our results. In Scenario I there is a large dilution by Types A and D. This leads to having a less than even chance of detecting oa treatment effect, although there is one in Type B and a rather large one in Type C. If we run the trial and we aren't lucky enough to detect, what then? If we follow GOLDEN RULE 5(B) we shall realize that we have not proven that there's no treatment effect, we won't give up, and we'll try again. By contrast, in Scenario II the sample is weighted towards Types B and C. There is an excellent chance of detecting, but it goes along with a danger of over–interpreting. We are likely to think, and to advertise, that the drug "is effective in raising the recovery rate among spinal injury patients," cruelly holding out false hope to patients in Type A.

Although, we could avoid the first of these errors by spending extra for an unnecessarily large sample, there are better means that we should use. (It is, in any case, good to use restraint in interpreting results.)

Our first recourse is to use some thought in deciding which subgroups to admit into the sample. Suppose, for the minute that we are in Scenario I of Table 9. The implications of some possible design choices are detailed in Table 10. If we admit everybody, we shall need a total combined sample of 2176.

TABLE 10: TOTAL COMBINED SAMPLE SIZE REQUIRED FOR 95% POWER UNDER VARIOUS DESIGN CHOICES IN SCENARIO I OF TABLE 9.

Design	Pooled Recovery Rate	Sample Size Needed
Admit all Types	Placebo: 32.5 %, Drug: 40%	2176
Admit only B and C	Placebo: 20 %, Drug: 35%	474
Admit only B	Placebo: 15 %, Drug: 25%	860
Admit only C	Placebo: 25 %, Drug: 45%	304

Excluding Types A and D drops the required sample size way down to 474. This exclusion seems mandatory.

GOLDEN RULE 12: *If the patient population has subgroups in which <u>either</u> everybody will improve, regardless of treatment, <u>or</u> in which nobody can improve, regardless of treatment, then these groups should be excluded. Including them is expensive: it dilutes the treatment effect, pushing the sample size up. At the same time, the only reward you can hope for by going to this expense is the possibility of being able to claim falsely that the drug is effective in these groups. Big expense; negative reward.*

A further step could be to admit only Type C, dropping the sample size down to 304. This is a harder decision, there are advantages either way. Including both B and C gives the potential for listing both groups on the package insert, helping more people, and getting more sales. On the other hand, focusing on C may mean a significant cost saving, and we might like to wait and see if there are encouraging results from Group C before we try to tackle Group B. If we could show that the drug works in C, we might be even able to start generating revenue that could pay for further studies in B. In considering whether to omit or delay Group B, we should bear in mind that, since physicians are permitted to prescribe products outside of the approved labeling, the FDA will be hesitant to approve one indication while other obvious related indications and populations are left unstudied. The relative importance of these two arguments changes with the relative prevalence of the two subgroups (the assumptions made in Table 10), and also with the disparity between groups in the magnitude of the treatment effects (the assumptions made in Table 8). With different assumptions, one argument or the other could be much more compelling. Note that, since we are pooling groups, the length of the recruitment period is the same for either choice.

Once we have made our decisions about the design of the sample, our second recourse is to think carefully about the analysis plan. Thus far we have been speaking of pooling any subgroups into a single statistical test. Instead, we might well want to search for separate answers in the separate subgroups. There are a number of ways to do this.

One thing we cannot do is simply to perform a separate test, looking for *P* less than $\alpha = .05$ in each group. Since these would be independent tests, the numbers in Table 7 exactly reflect the way in which the effective *P*–value increases with the number of subgroups. We would need a correction. A simple approximate formula, *Bonferroni's Inequality*, shows that it would be safe to divide α by the number of groups. However, when the number of groups is large this can be needlessly conservative. For independent tests, it is bettter to use the exact formula: for *n* groups and for any given value of the overall α, we need to look for *P* less than the nominal value $\alpha^* = 1 - (1-\alpha)^{1/n}$ in each group.

Doing two separate tests gives a higher quality answer: we know in greater detail whether and to what degree the drug has an effect in each of the groups. However, we might wonder if there isn't too high of a price to be paid, in sample size, for this quality of service. Let's work an example that illustrates the issues involved in deciding whether to pool or to do two separate tests.

This time we will imagine that the spinal injury population has two subgroups E and F comprising, respectively, 75% and 25% of the whole. When placebo is given (best standard care), the recovery rate in Group E is 5% and in group F is 20%. I shall discuss 4 different design possibilities. The first, Design I, is the simple strategy that we have been looking at so far: recruit all eligible comers in their natural proportions and, when the data are in, pool them into a single analysis. This plan will be more or less successful depending on how big the drug effect is in each of the two groups, i.e. on how great the odds for recovery on GM–1 are compared to the odds for recovery on placebo.

TABLE 11: POWER CALCULATIONS
TWO PATIENT SUBGROUPS, E AND F: 75% IN GROUP E AND 25% IN GROUP F
DESIGN I: RECRUIT E AND F IN NATURAL PROPORTIONS; ANALYZE POOLED DATA, $P < .05$
N = 798
RECOVERY RATE, PLACEBO: 5% IN GROUP E, 20% IN GROUP F
POOLED RECOVERY RATE, PLACEBO: 8.8%

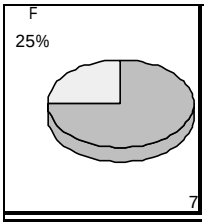
	GM–1 odds equal to placebo odds in Group E	GM–1 odds 1.5 times placebo odds in Group E	GM–1 odds 2 times placebo odds in Group E	GM–1 odds 4 times placebo odds in Group E
GM–1 odds equal to placebo odds in Group F			Pooled recovery rate, drug: 12.1% Power: .291	Pooled recovery rate, drug: 18.0% Power: .962
GM–1 odds 1.5 times placebo odds in Group F			Pooled recovery rate, drug: 14.0% Power: .597	Pooled recovery rate, drug: 19.7% Power: .991
GM–1 odds 2 times placebo odds in Group F	Pooled recovery rate, drug: 12.1% Power: .291	Pooled recovery rate, drug: 13.8% Power: .565	Pooled recovery rate, drug: 15.5% Power: .800	Pooled recovery rate, drug: 21.4% Power: .998
GM–1 odds 4 times placebo odds in Group F	Pooled recovery rate, drug: 16.3% Power: .874	Pooled recovery rate, drug: 18.0% Power: .962	Pooled recovery rate, drug: 19.6% Power: .990	Pooled recovery rate, drug: 25.5% Power: >.999

Table 11 shows the power calculations for different effect sizes. As we go to the right the effect in Group E increases; as we go downwards the effect in Group F increases. The sample size has been adjusted to 798 to make the power equal to .80 provided that the drug increases the odds of recovery by 2 times in both of the groups. If the drug effect is larger in both groups, then we have lots of power; if it is smaller in both, then we have a problem. If one of the groups has a large drug effect while the other doesn't, then the worse group drags down the overall power distressingly.

TABLE 12: POWER CALCULATIONS
TWO PATIENT SUBGROUPS, E AND F:
75% IN GROUP E AND 25% IN GROUP F
DESIGN II:
RECRUIT F ONLY, AND ANALYZE AT $P < .05$
N = 798
RECOVERY RATE, PLACEBO: 20% IN GROUP F

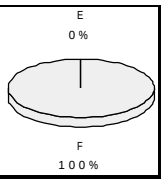
	
GM–1 odds equal to placebo odds in Group F	Recovery rate, drug: 20.0%
GM–1 odds 1.5 times placebo odds in Group F	Recovery rate, drug: 27.3% Power: .652
GM–1 odds 2 times placebo odds in Group F	Recovery rate, drug: 33.3% Power: .987
GM–1 odds 4 times placebo odds in Group F	Recovery rate, drug: 50.0% Power: >.999

Table 12 shows the result of applying a different strategy, Design II, to the same scenario. This time, we accept only patients from Group F, the group with the higher recovery rate on placebo. (Note that the recruitment period will likely be longer than if we accept both groups.) Power is high, and we have no worries about Group E. If we knew beforehand that the drug was effective in Group F, and if we were willing, at least temporarily, to concentrate our attention on it, Design II would be a better choice than Design I.

Next, Table 13 shows Design III, which uses only Group E. This is not as strong as Design II, because Group E has less potential for detecting an effect than Group F does. Due to the difference in the recovery rate on placebo, Group E needs the odds on drug to be a larger multiple of the odds on placebo to get the same power as Group F.

Finally, Table 14 shows Design IV, in which we recruit both groups in natural proportions but analyze in separate tests, agreeing to look for $P < 1 - \sqrt{1-.05} \cong .025$. We see that, while our protection is a bit less when the effect size is the same in the two groups, it's a bit better when one group is worse. Overall, our protection is about the same, but we're buying insurance in case one group doesn't work out.

To some degree, I have, for illustrative purposes, rigged this example to accentuate the possible benefits of Design IV, which are greater when (1) there is a large disparity in the recovery rates on placebo between the 2 groups, and when (2) the group with the larger potential for detecting an effect is one that is naturally less numerous. If you are in the position to meet condition (1), and find Design IV otherwise attractive, then improve your position by actually running it as two separate trials, linked only

by the common adjustment in the target *P*–value. One advantage is that you can force the recruitment to meet condition (2). The second advantage is that if one group finishes early, you can go to the FDA early.

TABLE 13: POWER CALCULATIONS
TWO PATIENT SUBGROUPS, E AND F: 75% IN GROUP E AND 25% IN GROUP F
DESIGN III: RECRUIT E ONLY, AND ANALYZE AT $P < .05$
N = 798
RECOVERY RATE, PLACEBO: 5% IN GROUP E

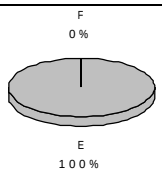
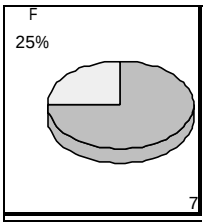
	GM–1 odds equal to placebo odds in Group E	GM–1 odds 1.5 times placebo odds in Group E	GM–1 odds 2 times placebo odds in Group E	GM–1 odds 4 times placebo odds in Group E
	Recovery rate, drug: 5%	Recovery rate, drug: 7.3% Power: .226	Recovery rate, drug: 9.5% Power: .641	Recovery rate, drug: 17.4% Power: >.999

TABLE 14: POWER CALCULATIONS
TWO PATIENT SUBGROUPS, E AND F: 75% IN GROUP E AND 25% IN GROUP F
DESIGN IV: RECRUIT E AND F IN NATURAL PROPORTIONS; ANALYZE SEPARATELY, $P<.025$ EACH
N = 798
RECOVERY RATE, PLACEBO: 5% IN GROUP E, 20% IN GROUP F

	GM–1 odds equal to placebo odds in Group E	GM–1 odds 1.5 times placebo odds in Group E	GM–1 odds 2 times placebo odds in Group E	GM–1 odds 4 times placebo odds in Group E
GM–1 odds equal to placebo odds in Group F			Recovery rate, drug Group E: 9.5% Group F: 20.0% Power: .409	Recovery rate, drug Group E: 17.4% Group F: 20.0% Power: .994
GM–1 odds 1.5 times placebo odds in Group F			Recovery rate, drug Group E: 9.5% Group F: 27.3% Power: .465	Recovery rate, drug Group E: 17.4% Group F: 27.3% Power: .995
GM–1 odds 2 times placebo odds in Group F	Recovery rate, drug Group E: 5.0% Group F: 33.3% Power: .415	Recovery rate, drug Group E: 7.3% Group F: 33.3% Power: .465	Recovery rate, drug Group E: 9.5% Group F: 33.3% Power: .636	Recovery rate, drug Group E: 17.4% Group F: 33.3% Power: .996
GM–1 odds 4 times placebo odds in Group F	Recovery rate, drug Group E: 5.0% Group F: 50.0% Power: .989	Recovery rate, drug Group E: 7.3% Group F: 50.0% Power: .990	Recovery rate, drug Group E: 9.5% Group F: 50.0% Power: .993	Recovery rate, drug Group E: 17.4% Group F: 50.0% Power: >.999

Altogether, the moral of this story is a little complex. GOLDEN RULE 13: *If you have 2 subgroups in your population, the choice you make among designs depends on how much you know going in, and on what you want to accomplish.* *Recruiting both groups and pooling the data:* *Works best if you have good reason to suspect that the compound works equally in both groups, you want both groups indicated in your package insert, and you don't need separate answers.* *Recruiting one group only:* *Works best if you have reason to suspect there is better potential for detecting the treatment effect in one group than the other. (For Fisher's Exact Test, this means that the difference between placebo and drug is greater in that group, or , approximately, that it is the same but the recovery rate on placebo is closer to 50%..) For the same number of patients recruitment is slower, but power is greater so sample size is lower. You get more power for the same dollar investment. The downside is that the wording on the package insert is restricted. Depending on your financial picture, it may make sense to get confirmation in one group and*

then study the other; or it may not. It's a business decision, up to you. *Recruiting both groups, but doing separate tests:* *Works well if you really have no idea which group is more likely to have a larger effect size, since it buys insurance in case one group is much worse. Also it gives you a higher quality, more detailed, answer. If you decide to do this, then conduct two separate trials, adjusting the common target* P–*value and, if possible, adjusting recruitment so the group with the best potential for detection has smaller sample size.*

An alternative procedure, instead of analyzing as two separate tests, might be to think of the data as a 2x2x2 table and use the corresponding generalization of Fisher's exact test. (With continuous data, one would analogously use analysis of variance, ANOVA.) If this test establishes significance of the table as a whole, one then would use some procedure to compare treatments within groups. (With continuous data, one would use a procedure, such as Scheffé's, for simultaneous contrasts, rather than doing *T*-tests at .025, or worse at .05.) This alternative isn't incorrect, it's just roundabout, and slightly pointless. One isn't usually interested in whether there is significance in the table as a whole; one wants to compare treatments within groups, and that's what the two separate tests do, elegantly and economically.

When you do plan to pool data from multiple subgroups, it is important to avoid any imbalance in the randomization among the groups. For example, if we do a simple randomization, the *expected* proportion of patients from Group E among the GM–1 patients should be about the same as it is in the combined Group E and F population, namely 75%. Most of the time the actual proportion will be the same as the expected, but it doesn't have to be. Table 15 shows what can happen, even when 50% of the total patients are assigned to each of the two treatments. The probability is .5841 that they will be equally distributed. There is a .2225 chance that Group E will have more patients on drug than on placebo, and Group F will have fewer. This raises the pooled recovery rate on placebo, since the placebo group has a disproportionate number of Group F patients with better prognosis. It also lowers the pooled recovery rate on drug, making it much harder to detect the drug effect. In the left column, there is a small, .0088, chance that the imbalance will be so great that the probability *among sample patients* for recovery on drug is lower than on placebo even though *among the original population* the odds are twice as great. Although the chances of this are a bit less than 1 in 100, with so many clinical trials being conducted it will happen to somebody.

TABLE 15:THE EFFECTS OF RANDOMIZATION IMBALANCE
ASSUMING 200 TOTAL PATIENTS, 75% IN GROUP E
RANDOMIZED 50% TO EACH TREATMENT
RECOVERY RATE, PLACEBO: 5% IN GROUP E, 20% IN GROUP F
ODDS FOR RECOVERY ON DRUG ARE 2 TIMES ODDS FOR RECOVERY ON PLACEBO IN BOTH GROUPS

Probability	.0088	.2225	.5841	.1790	.0055
Number of patients: Group E - Placebo	55	65	75	85	95
Number of patients: Group E - Drug	95	85	75	65	55
Number of patients: Group F - Placebo	45	35	25	15	5
Number of patients: Group F - Drug	5	15	25	35	45
Pooled recovery rate: Placebo	.118	.103	.088	.073	.058
Pooled recovery rate: Drug	.107	.131	.155	.178	.202

GOLDEN RULE 14: *A randomization imbalance can change the apparent magnitude of the drug effect considerably, or even eliminate it. Stratify your sample on as many prognostic factors as you feasibly can. Since this will tend to even out these effects, this means that you do not have to model these covariates explicitly in the efficacy analysis.*

WHAT TO DO WITH MULTIPLE TREATMENT REGIMENS.

It sometimes happens that we want to compare two different versions or doses of an experimental compound against a placebo. For example, in the spinal cord injury study it was desired to compare two dose levels, g and G, of GM–1 against placebo, p. In another trial, one might be unsure which of two methods of administration, or of two formulations, of a compound works best, and therefore want to compare both against placebo. As with multiple subgroups, there is a simultaneous inference problem, but here it's more complicated. The two tests — g versus p, and G versus p — are correlated because they both involve the same placebo group. This invalidates the assumption of independence that made the two separate test idea work. One way to approach this problem is to use the overall testing methods mentioned in the last paragraph of the preceding section.

TABLE 16: SOME HYPOTHETICAL DATA FROM A FUTURE TRIAL OF TWO DOSE LEVELS, g AND G, OF GM–1 AGAINST PLACEBO, p.

	improve	not	total
p	6	57	63
g	11	48	59
G	18	40	58
total	35	145	180

There is a method for using two tests and correcting for the overlap between them. This method is not widely available in statistical packages, and so the computations need to be done by a trained statistician. The ideas, however, are clear, and instructive. Table 16 shows some hypothetical data we can practice on.

As in Fisher's Exact Test, we would regard the totals (63,59,58) in the right–hand margins, as fixed constants. Further, we would condition on the observed total number, 35, of successes: this number, while important, does not directly help us to distinguish among the groups, p, g, and G. Thus we would regard the right and bottom marginal totals all as fixed for the rest of the discussion.

Given this convention, it is enough to examine only the observed numbers (11,18) of successes for g and G. All other numbers in the table can be found by subtraction: the 6 for group p equals 35–(11+18), and the second column equals the third minus the first. The pair (11,18) can be used as an appropriate test statistic, and we shall focus our attention on it.

What are the other *possible* test statistics that could have arisen with the same right–hand marginal totals in each group and with the same total number, 35, of successes? They can best be visualized as arranged schematically in Table 17. They form a triangle in the lower–left half of the

TABLE 17: POSSIBLE VALUES OF THE TEST STATISTIC FOR THE DATA OF TABLE 16

↑	(0,35)									
|	…	…								
|	(0,29)	…	…							
|	…	…	…	(11,24)						
G	…	…	…	…	…					
|	(0,18)	…	…	(11,18)	…	(17,18)				
|	…	…	…	…	…	…	…			
|	…	…	…	…	…	…	…	…		
|	…	…	…	…	…	…	…	…	…	
↓	(0,0)	…	…	(11,0)	…	(17,0)	…	(29,0)	…	(35,0)
	←	– – –	– – –	– – –	g	– – –	– – –	– – –	– – –	→

diagram.

As the shading indicates, (11,18) is in the column corresponding to 11 successes for group g and the row corresponding to 18 successes for group G. It is also in a diagonal stripe that corresponds to 11+18 = 29 successes for g and G combined, and therefore to 35–29 = 6 successes for group p.

The standard Neyman–Pearson theory of statistical hypothesis testing requires us to mark out a region, called *the critical region*. If the test statistic turns out to fall within this region then we agree that we will reject the *null hypothesis* that there is no difference between treatments. If the statistic is not in the critical region we will not reject.

We wish to make 2 comparisons: G versus p, and g versus p. We would like a set of critical regions that would give specific information on each of the comparisons while jointly containing the risk of Type II Error within the preassigned value.

The value (11,18) occurs in the row corresponding to 18 successes for G. *Within this row*, there is no information about G, but values to the right give evidence that g is superior to p and values to the left give evidence that p is superior to g. For this comparison we would like to form a pair of critical regions starting from the left and right of each row. This process of conditioning on staying within the row can be visualized as a 2×2 table, on the left in Table 18; the conditional probabilities are the same as in Fisher's Exact Test. Similarly, for the G versus p comparison we would stay within the column

TABLE 18: BREAKING THE DATA OF TABLE 16 UP INTO TWO SEPARATE TABLES CORRESPONDING TO THE ROW AND COLUMN IN TABLE 17

For g

	improve	not	total
p	6	57	63
g	11	48	59
total	17	105	122

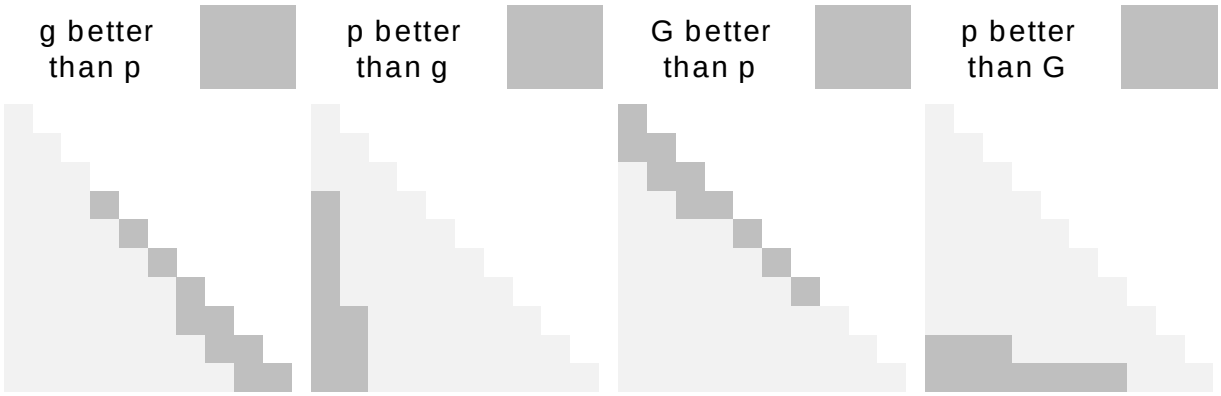
For G

	improve	not	total
p	6	57	63
G	18	40	58
total	24	97	121

corresponding to 11 successes for g, effectively forming a separate 2×2 table, on the right in Table 18. For this comparison, G versus p, we would like to form a pair of critical regions starting from the top and bottom of each column. How large should these two pairs of regions be, exactly?

The probabilities of each of the test statistics in Table 17 are defined mathematically by a distribution called the *Bivariate Hypergeometric Distribution*, a generalization of the Univariate Hypergeometric Distribution used in Fisher's Exact Test. To find the combined critical region, we would keep increasing the size of the 2 regions associated with each of the 2 comparisons — column and row —, keeping all 4 in step so that none was unnecessarily larger than the others, until their combined probability was as large as it could be without going over the allotted amount. This results in a system of 4 triangular regions — to the left, right, top and bottom of the big triangle of Table 17 — shown in Table

TABLE 19: SYSTEM OF CRITICAL REGIONS IN TABLE 17
FOR REJECTING THE NULL HYPOTHESIS OF NO TREATMENT DIFFERENCE



19. Observing whether the test statistic lies in any of the regions tells which comparisons between treatments are accepted, and in which direction.

Thus this procedure is, in its results, much like the two separate test procedure discussed in the previous section. It defines specific decisions depending on the region in which the test statistic lies. In the present case, the computations are made somewhat more intricate because of the correlation between the two component tests. Indeed, both of these procedures are applications of a single idea: the *Conditional Neyman–Pearson Algorithm* (CNPA).

There is also a rather different problem of this kind, in which, rather than testing two versions of the compound against one placebo, we want to test a single compound against two kinds of placebo. For example, two different compounds, A and B, might both be already known to be active against a disease, but with different mechanisms of action, leading one to hypothesize that a combination, AB, of the two might be more effective than either alone. In the previous example we can get approval if we can show that *either* g or G is better than p. The crucial difference here is that this time we need to show that the combination is better than *both* of its components.

A simple method analysis is by two separate tests, looking for $P < .05$ in each. This means that the combined level of the two tests is somewhere, unpredictably, between a minimum of .05 x .05 = .025 and a maximum of .05 x 1 = .05. Where it actually lies within this range depends on the degree of (necessarily positive) correlation between the two tests. Likely the correlation will be low to moderate, meaning that the effective value will be closer to .025 and therefore undesirably conservative. One might try to mitigate this problem by estimating, and correcting for, the correlation. Alternatively, there is a CNPA procedure, leading to a set of critical regions analogous to those in Table 19.

WHAT TO DO WITH MULTIPLE ENDPOINTS.

It very commonly occurs that we want to use multiple measures of recovery: either several different measures, or the same measure repeated at several timepoints after the start of therapy. Usually, it's best and simplest to avoid this, at least for the primary efficacy analysis, but not always.

The values of the multiple endpoints are likely to be correlated, and thus the analysis has a mathematical structure somewhat similar to the multiple treatment problem. Again, a fundamental decision is whether you need to win on *all* of the endpoints or on *at least one*.

One method of analysis is a CNPA procedure leading again to a system of critical regions. This time the possible test statistics form a rectangle, rather than a triangle as in Table 17. The probabilities in this rectangle would have to be estimated by using a *randomization test* considering all possible reassignments of the patients to the 2 groups.

L'ENVOI.

In this chapter, I have tried to show the general idea. Collaborate closely with your statistician starting very early in the design of the trial. He or she can save you expense by helping you with crossover or sequential designs. The statistician can also help you with the FDA by ensuring that you have a large enough sample to have a reasonable chance that the trial will demonstrate an effect and by ensuring that your thinking about issues like multiple subgroups, treatments or endpoints is well sorted out.

Don't think of statistics as an automatic procedure, but as a requirement and an opportunity for clearly modeling the effect of your compound. Collaborate closely with an energetic, imaginative, and empathetic statistician. Good Luck!

Or, as J.P. Getty said, "Rise early. Work hard. Strike oil."

GOLDEN RULE 1: *The valid use of probability theory to calculate the results of statistical tests depends on using data that are sampled randomly from the desired target population. If either (1) your data are not randomly chosen or (2) you cannot clearly identify a target population of which they are representatives, then you should not be using statistical tests.*

GOLDEN RULE 2: *Your object is to be able to inform a treating physician reading your results about the potential for the therapy to help the patient. Design your target population thoughtfully. In accordance with* GOLDEN RULE 1, *include in your primary efficacy analysis all and only the patients that fit the target population. Otherwise, the P-value cannot be legitimately computed by the laws of probability and the result cannot be scientific, despite the vigorous claims of advocates of "intent–to–treat" analyses that obligatorily include every randomized patient. Final decisions about inclusion and exclusion should be made by a blinded Adjudication Committee.*

GOLDEN RULE 3: *If the probability of an event is* n *out of* m, *then confidently expect both that it will happen about* n *out of* m *times and that it will not happen the other* m – n *times. Budget ahead for losses either way.*

GOLDEN RULE 4: *The only way to get a compound approved is to avoid Type II error (failing to detect a treatment effect when there is one). If you don't spend enough extra to get adequate statistical power, then the still rather large amount of money that you are investing in your trial is buying you nothing but the likelihood of no approval. If you're really dumb enough to gamble your compound away, at least don't pay for a clinical trial to do so: see a fortune teller instead; they're cheaper.*

GOLDEN RULE 5: *(a) There's nothing sacred about .05 except for the fact that it's traditional. It's a measure of risk. How much risk can you afford? (b) If you don't get P < .05, or whatever, it doesn't mean that the compound isn't effective, unless you have also controlled for Type II Error. It means that you have failed to find a treatment effect. How hard did you look?*

GOLDEN RULE 6: *To plan the sample size, use your pilot data, if available, to develop a guess about the whole range within which the effect size might possibly lie. Also, make a judgment as to what the minimum effect size is that would make further testing and development of the compound clinically and commercially worth while. Buy enough protection, enough sample, to cover yourself for the minimum effect size and also across a reasonable segment of the likely range.*

GOLDEN RULE 7: *Use a crossover design to reduce variability and sample size, if the chance presents itself, and if you're sure you know what you're doing.*

GOLDEN RULE 8: *If you have pilot data that indicate a good likelihood that the actual effect size may be much higher than the minimum size calculated under* GOLDEN RULE 6, *then consider a sequential design.*

GOLDEN RULE 9: *If you have preliminary evidence that the new compound is much better than the standard therapy, think about using a play the winner randomization.*

GOLDEN RULE 10: *Each statistical test (each P–value) represents a risk of error. As I perform more tests, my combined risk of error rises above the nominal value and eventually becomes almost certain. Therefore, fishing expeditions in which everything in sight is tested are not "good science;" they are ridiculous and possibly fraudulent. Good science is a small number of prospectively, and thoughtfully, chosen tests.*

GOLDEN RULE 11: *A clinical trial can have only a single primary efficacy analysis, and it must be prospectively chosen and specified in detail. That analysis can have multiple components — population subgroups, treatment regimens, or endpoints —, but the probabilities must be combined and adjusted explicitly.*

GOLDEN RULE 12: *If the patient population has subgroups in which either everybody will improve, regardless of treatment, or in which nobody can improve, regardless of treatment, then these groups should be excluded. Including them is expensive: it dilutes the treatment effect, pushing the sample size up. At the same time, the only reward you can hope for by going to this expense is the possibility of being able to claim falsely that the drug is effective in these groups. Big expense; negative reward.*

GOLDEN RULE 13: *If you have 2 subgroups in your population, the choice you make among designs depends on how much you know going in, and on what you want to accomplish. Recruiting both groups and pooling the data: Works best if you have good reason to suspect that the compound works equally in both groups, you want both groups indicated in your package insert, and you don't need separate answers. Recruiting one group only: Works best if you have reason to suspect there is better potential for detecting the treatment effect in one group than the other. (For Fisher's Exact Test, this means that the difference between placebo and drug is greater in that group, or , approximately, that it is the same but the recovery rate on placebo is closer to 50%..) For the same number of patients recruitment is slower, but power is greater so sample size is lower. You get more power for the same dollar investment. The downside is that the wording on the package insert is restricted. Depending on your financial picture, it may make sense to get confirmation in one group and then study the other; or it may not. It's a business decision, up to you. Recruiting both groups, but doing separate tests: Works well if you really have no idea which group is more likely to have a larger effect size, since it buys insurance in case one group is much worse. Also it gives you a higher quality, more detailed, answer. If you decide to do this, then conduct two separate trials, adjusting the common target* P*–value and, if possible, adjusting recruitment so the group with the best potential for detection has smaller sample size.*

GOLDEN RULE 14: *A randomization imbalance can change the apparent magnitude of the drug effect considerably, or even eliminate it. Stratify your sample on as many prognostic factors as you feasibly can. Since this will tend to even out these effects, this means that you do not have to model these covariates explicitly in the efficacy analysis.*

---

[1] Geisler, F.H., Dorsey, F.C., Coleman, W.P.: Recovery of motor function after spinal cord injury - a randomized, placebo-controlled trial with GM–1 ganglioside. The New England Journal of Medicine, 324(26):1829-1838, June 27 1991.