SAMPLE SIZE

From the Therapy chapter for the 3rd edition of Clinical Epidemiology, by DL Sackett
17 April 2004 (day 108)

Sample Size Check List:

1 ❍	Based on your study question, calculate your sample size requirement
2 ❍	Based on your eligibility criteria, estimate (with appropriate skepticism) the availability of appropriate patients
3 ❍	If (when) necessary, apply strategies to increase (effective) sample size

Preface: Is this section really necessary?

Throughout this chapter, I've stressed the importance of recruiting a statistician as co-principal investigator right at the start of formulating the question for your RCT. Why, then, intrude on their turf with a section on sample size? As with the previous section on analysis, my reasons are three. First, as you can see from the checklist, two of its three entries are not strictly statistical. Second, especially early in thinking about your trial, you may want to do some sample size "doodling" to understand the effects of, for example, recruiting high- vs. low-risk patients. Accordingly, this section's first function is to provide non-statisticians with a sufficient introduction to sample size determinations to be able to roughly estimate them without bugging your statistical co-PI with every new idea. Third, as with data analysis, when some non-statistician trialists get their feet wet in statistics, they discover (to their surprise and mine) that they enjoy learning more about it. So, this section's final function is to whet some appetites.

In the RRPCE trial, we estimated that the annual incidence of stroke among our study patients would be 7%, that their annual death rate would be 4%, and that one or both drugs would halve these rates. We decided to limit our risk of concluding that either or both drugs were better than placebo when, in fact, they weren't (the Type I error) to 0.05 (α). We also decided to limit our risk of concluding that neither drug was better than placebo when, in fact, one or both were (the Type II error) to 0.20 (β).

Although a spot survey at our Clinical Principal Investigator's center predicted that his team would see 52 eligible patients each year, we urged our neurologic collaborators at the other other 23 to be very pessimistic in predicting patient availability. Together, they predicted that they could recruit 150 patients per year.

They recruited less than half this number (78) in year 1, and our Clinical Principal Investigator began to visit every center at frequent intervals. Recruitment rose to 148 patients in year 2 and 164 patients in year 3. Our Clinical Principal Investigator' team recruited almost a third of all our study patients.

The scenario nicely describes what happens when collaborators' high-flying, rosy predictions of patient availability come to earth and land in the swamps where recruitment really happens. In this section, I'll do my best to guide you through these swamps.

The first part of the sample size checklist applies to the planning stages of your trial.

1. <u>Based on your study question, calculate your sample size requirement</u>

Let’s begin by considering a superiority trial that uses discrete events as outcomes. You can usefully think about it as a diagnostic test for the truth about efficacy, as shown in Table 3-10-1.

Table 3-10-1: The RCT as a diagnostic test for the truth about efficacy

		The truth about efficacy	
		Experimental treatment really is superior	Experimental treatment really is not superior
Results of your RCT	Experimental treatment appears to be superior	1 – β Power!	Type I error Risk = α P-value!
	Experimental treatment appears not to be superior	Type II error Risk = β	

The rows of this table describe the conclusions you draw at the end of your trial, and the columns describe the truth (that you are trying to "diagnose" with your trial). If your conclusions match the truth, all is well. What you want to avoid is drawing incorrect conclusions. As it happens, you can specify the risks you are willing to run of drawing these wrong conclusions before you start your trial. In a superiority trial, you want to minimize the risk of drawing the false-positive conclusion that your experimental treatment is superior when, in fact, it is not. Statisticians call this a Type I error and trialists pre-specify this risk, typically at 0.05 , and call it α. After the trial is over, tests for the "statistical significance" between the experimental and control event rates describe the actual risk you ran of drawing the false-positive conclusion, and present it as a P-value. If it hasn't already occurred to you, this is why we want P-values to be very small.

Similarly, you want to minimize the risk of drawing the false-negative conclusion that your experimental treatment is useless when, in fact, it is superior. Statisticians call this a Type II error and trialists pre-specify it, typically at 0.2, and call it β. You are probably more familiar with its complement, $1 - \beta$, which we call power. Again, if it hasn't already occurred to you, this is why we want power to be very large. And you may find it useful to think of the power of an RCT the same way that you think about the sensitivity of a diagnostic test. It tells you the probability that you will find superiority, and label it statistically significant, if it really exists.

a. <u>If a difference in the occurrence of events will answer your question</u>:

There are two ways to calculate your sample size requirements here. These are the forward "<u>patients I need</u>" approach and the backward "<u>patients I can get</u>" approach. The "<u>patients I need</u>" approach tells you just that: how many patients you need per treatment group. But the "<u>patients I can get</u>" approach tells you how big a bang (in terms of power) you will get from the patients you can get. I'll describe them in sequence.

The forward "<u>patients I need</u>" approach is the classical, theoretical one that appears in most beginning courses and textbooks. In the "<u>patients I need</u>" approach, you simply pick your α and β, specify the event rates you expect to observe among your experimental and control patients, and calculating your sample size requirement is simple maths. However, I never trust my or my students' hand-calculations of sample size (too often we get it wildly wrong). Instead, I go to a

statistical website that will do it right (the one I use is run by Rollin Brant at the University of Calgary: http://www.ucalgary.ca/~brant/stats/ssize/. [1], but you should surf until you find the one that's best for you). Regardless of how you do it, you should avoid the common mistake of thinking that the answer you get is for the total number of patients you require for your trial; in fact, it's usually the number of patients you need per treatment group.

Moreover, any way you do it, you shouldn't be satisfied with a single "patients I need" sample size calculation. Just suppose that you've overestimated the true control event rate (CER) and/or the true relative risk reduction (RRR). As a result, the absolute risk reduction (ARR) you'll observe in your trial will be smaller than you'd predicted, and you will have too few patients to generate a statistically significant result with a nice, tight confidence interval. To avoid this pitfall, you should plug in all the clinically sensible control event rates and relative rtisk reductions, as shown in Table 3-10-2.

Table 3-10-2: The number of **"patients I need"** into each treatment group if α =0.05 and β =0.2 (80% power) and the CERs and RRRs are as shown.

		If the Relative Risk Reduction (RRR) is:				
		20%	25%	30%	35%	40%
If the Control Event Rate (CER) is:	.60	270	173	120	88	67
	.55	324	205	148	105	79
	.50	388	247	170	124	93
	.45	466	298	203	146	111
	.40	564	356	244	176	133

Then you can ponder the table and decide what to do. If you could actually recruit the largest ("worst-case") sample size you're likely to require, that would be great insurance. Then, if the actual control event rate (CER) and/or relative risk reduction (RRR) you observe in your trial turn out to be higher than your "worst case," your statistical warning rule should be triggered early.

The crucial mistake you want to avoid is "finishing" your trial and shutting it down, only to discover that it's too small. Imagine your agony if the confidence interval around your moderate but still useful absolute risk reduction (ARR) crosses zero.

Special cases:

You can use this same approach to calculate the "patients I need" to answer 1-sided questions about superiority and non-inferiority. All that you do differently is to set your α at 0.10 rather than 0.05. In Table 3-10-2, this reduces the number of patients needed in each cell by about 20%.

You can use this same strategy for 2x2 factorial designs if you assume that the effects of the 2 interventions will be similar and additive (that is, the response to one of them is unaffected by receiving the other, and there is no "interaction"). This factorial design permits you to "do two trials for the price of one," because you use each study patient twice, once for each treatment. The sample size calculation will tell you the numbers of patients you need at the end of each row and column. You can then simply split them between the cells making up that row or column.

I'm warning you right now that your first look at the results of your "patients I need" calculation is likely to chill your blood and bring up your lunch. What you thought was a short and simple single-center RCT can morph before your eyes into a multi-center mega-trial.

[1] In using this site, pick the option that is labelled: "Comparing Proportions for Two Independent Samples"

For this reason, lots of trialists ignore the forward-looking "patients I need" approach altogether. They reckon it's much more realistic and honest to apply the backward-looking "patients I can get" approach. In this approach, sample-size is an input, not an output. You start by estimating the number of patients you are confident you can enroll in the study. Let's say there are 300 patients "you can get," or 150 per group. As before, you specify a range of reasonable control event rates (CERs) and relative risk reductions (RRRs), and pick your α. The rest is simple maths, and another visit to Rollin Brant's website, <http://www.acs.ucalgary.ca/~brant/ssjava.html>, will generate Table 3-10-3.

Table 3-10-3: The power I can generate when the **"patients I can get"** is 150 per group, α =0.05, and the CERs and RRRs are as shown.

		If the Relative Risk Reduction (RRR) is:				
		20%	25%	30%	35%	40%
If the Control Event Rate (CER) is:	.60	55%	74%	88%	96%	99%
	.55	48%	67%	82%	92%	97%
	.50	41%	59%	75%	87%	95%
	.45	35%	51%	67%	81%	91%
	.40	30%	44%	59%	73%	85%

This time, because you have already specified the number of patients per group, the only thing left for the website to calculate is the power ("sensitivity") generated from each pair of control event rates (CER) and relative risk reductions (RRR). The shaded cells in this table spell trouble. They mark those unfortunate combinations of CERs and RRRs in which your 150 patients per group will fail to generate the 80% power (or "sensitivity") that most trialists (and granting agencies) require. What you need to do is get out of the shade and into the clear cells, and a revisit to the section on physiological statistics may help you.

Tables 3-10-2 and 3-10-3 are saying the same thing, but in different ways. You can confirm this by noting that the shaded, "underpowered" cells in Table 3-10-3 correspond to the cells in Table 3-10-2 that require more than 150 patients per group.

b. **If a difference in average values for a physiological, behavioral, or quality of life measure will answer your question:**

You decide about α and β as before. Then, you need to specify the smallest difference between experimental and control patients which, if observed at the end of the trial, your patients and (we trust) you would consider humanly important (often called the "minimum important difference" or MID). Usually, this will take the form of a minimum important difference between the changes in the measure from baseline to the end of the trial in the intervention and control groups. Finally, you need to plug in a description of how these continuous measures vary between patients and repeated measurements, typically in the form of a standard deviation of change. You can then proceed to one of the websites and crank out your sample size needs. As before, you should plug in all the reasonable values for the differences you'd like to detect and the standard deviations you're likely to observe.

As Gordon Guyatt and his colleagues have demonstrated, the "minimum important difference" (MID) can have tricky properties when applied to quality of life measures[1]. Several of the questionnaires they use employ 7-point scales. Patients rate their symptoms, function, or quality of life on these scales by matching their current state with the scale's verbal description. For example, in reporting how short of breath they've been in the past 2 weeks while climbing stairs they can choose from "extremely short of breath" at one end to "not at all short of breath" at the

other. Gordon Guyatt's team ~~team~~ have documented that the minimum difference that patients consider important (MID) is a change in score of >0.5. However, if some patients benefit greatly from the intervention (change >1.0) and others not at all (change = 0), an average MID <0.5 could still be important for the former. An alternative approach here is to assign a favorable "event" to every patients whose scores change by 0.5 or more. By doing this, you convert the analysis (and sample size determination) into the event strategy described earlier. Once again, I refer you to Gordon's chapter on developing and validating such measures. As noted in section 9, analysis of covariance may be a more appropriate pathway to follow in determining sample size in trials with "continuous" outcomes. This gets pretty complex pretty fast, and your statistician/co-principal investigator should determine which strategy is more appropriate.

c. **If you are planning a cluster randomized trial:**

A group of colleagues once asked me to help them with their protocol for testing a systemic treatment for preventing recurrent deep vein thrombosis (DVT). They proposed that every leg contribute to both the numerator and denominator of the recurrence event rate. Since each patient in the proposed trial would have approximately 2 legs, they reckoned that they could cut their sample size requirement in half (what, I wondered to myself, would they do if they were treating fingers and toes?). When I explored the mechanisms of recurrence with them, they readily agreed that whether a patient's left leg DVT recurred would be influenced a great deal by whether their right leg DVT had recurred the previous day. In other words, the responses of 2 legs belonging to the same patient would be much more similar than the responses of 2 legs belonging to 2 different patients. Accepting this pathophysiological dependence, the investigators readily changed their unit of analysis (and sample size calculation) from individual legs to pairs of legs clustered underneath whole patients.

This principle applies any time that study patients are allocated to treatments in clusters of 2 or more, such as families, practices, hospital wards, communities, provinces, and the like. The responses of study individuals within these clusters can be expected to be more similar (or "concordant") than the responses of individuals belonging to different clusters. Since individuals within clusters are not "independent," the traditional methods for determining sample size will underestimate the real sample size requirement (and the traditional methods of analysis may overestimate treatment effects).

Sample size determinations for cluster-randomized trials begin with estimating the degree of concordance within clusters. For continuous outcome measures like blood pressure, this concordance might be expressed as an intraclass correlation coefficient. For events like quitting smoking, concordance might be expressed as kappa[2]. These concordance factors are then used to determine the appropriate (increased) sample size requirement. The current authority on cluster-randomized trials is Alan Donner at the University of Western Ontario[3], who has developed methods that take concordance nicely into account.

d. If you are planning a crossover or "time to failure" RCT:

I won't discuss these less common and more complex RCTs here. If you're doing one of these without a statistician co-principal investigator, you deserve all the trouble you'll get. To talk intelligently with your co-P.I., you might want to consult one of the dedicated clinical trials books, such as the 3rd edition of the book by Lawrence Friedman, Curt Furberg, and David DeMets[4].

A final note

Many trialists add, say, 20% to their final sample size estimate to account for patients who don't comply with treatment, drop out, or are lost to follow-up. This is a double-edged sword. On the

one hand, it is comforting to have a sample-size cushion. On the other hand, losing anywhere close to 20% of your trial's patients will also lose you credibility when you report its results. Moreover, if the risk or responsiveness of the patients you lose differs from that of patients you retain, you will lose validity as well. It is far better to devote the resources required for recruiting another 20% of patients to keeping track of all those whom you've already recruited. I'll come back to this at the end of this section.

2. **Based on your eligibility criteria, estimate (with appropriate skepticism) the availability of appropriate patients.**

Four decades ago, we simply asked clinicians at the potential study site(s) to tell us the numbers of patients they thought they could recruit for the trial. It didn't take us long to realize that this approach led to hopelessly optimistic estimates of available patients. This realization had two effects. First, we adopted the aphorism: "The best way to eliminate a disease is to start an RCT on it." Second, we started applying "rules of thumb" which would divide these rosy estimates by 2, 4, or 8.

Nowadays we ask potential collaborators to make a list of every potential study patient they encounter for the next several weeks or months (while we are finishing the protocol). We ask them to ruthlessly distinguishing the minority of patients who meet all the eligibility criteria from the majority who, for one reason or other, don't. If you do this, you'll probably discover that you need a longer recruitment period or more clinical collaborators. This discovery is extremely annoying before a trial begins, but becomes catastrophic if it is made only after you have started the trial. This leads us nicely into the final d suggest that you read the next bit of this section while you're still in the early planning stage of your RCT.

3. **If (when) necessary, apply strategies to increase (effective) sample size**

As with any other potentially fatal disorder, the successful treatment of inadequate sample size begins with an accurate diagnosis. The reason you can't recruit enough patients may not be because they are rare. It might be the result of clinician- and patient-based barriers to participation. Sue Ross and her UK colleagues systematically reviewed 78 reports of these barriers and I have summarized their findings in Table 3-10-1[5].

Table 3-10-1 Barriers to participation in a randomised controlled trial

Clinician based	Patient based
Time constraints	Additional procedures and appointments for patient
Lack of staff and training	Additional travel problems and cost for patient
Worry about the impact on doctor-patient relationship	Patient preferences for a particular treatment (or no treatment)
Concern for patients	Worry about uncertainty of treatment or trials
Loss of professional autonomy	Patient concerns about information and consent
Difficulty with the consent procedure	Protocol causing problem with recruitment
Lack of rewards and recognition	Clinician concerns about information provision to patients

Taken from: Ross S, Grant A, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to participation in randomised controlled trials: a systematic review. J Clin Epidemiol 1999;52:1143-56.

Strategies for increasing patient numbers.

Based on your diagnosis, you can employ one or more of 12 strategies to either increase your sample size or make the most of whatever sample size you do recruit. These interventions come from several sources, including a Cochrane Methodology Review[6].

The first 7 are general strategies for increasing patient numbers:

1. You can make it easier for clinical collaborators to approach and enter patients into the trial by reducing the entry forms to just those items that are of high and immediate relevance. For example, entry forms for some of the large, simple trials I'll describe later occupy less than one side of one page.

2. In similar fashion, you can reduce the complexity and time expended in deciding whether every patient is eligible for a trial by both reducing its eligibility criteria to a bare minimum and by employing the "uncertainty principle[7]" as the major determinant of an individual patient's eligibility. The uncertainty principle is discussed on page xx.

3. You can reduce the follow-up effort required from busy clinical collaborators by providing Research Assistants to help them with forms, baseline measurements, allocation, and follow-up appointments. Trialists like me vastly prefer this strategy to that of providing "bounties" to clinicians for every patient they enter.

4. You can capture eligible patients who appear at night or on weekends by setting up a 24/7 randomization "hotline," perhaps via your hospital swithboard.

5. When a brand new drug or other treatment is not yet available to the public and has never been evaluated in a Phase III trial, many sponsors (especially health care providers who must pay for the innovation) will make the experimental treatment available only within an RCT.

6. You can explore collaboration with relevant organizations of patients and families who have come together to provide information, support and advocacy to the victims of the disorder you are studying. Growing numbers of such organizations have become strong and effective advocates for relevant RCTs.

7. You could write directly to your own patients, describing the trial and inviting them to learn more about it. This strategy has often been successful in recruiting patients in primary care.

The next 3 strategies attack the universal failure of participating centres (including your own!) to approach all eligible patients.

8. You can increase recruitment from your current centre(s) by frequently exposing the people in these centers to your most charismatic and respected clinical collaborator. Our cerebrovascular trials succeeded in large part because our principal clinical investigator was willing to devote major time to national and international "circuit-riding" among the centres. His "outreach" visits began with grand rounds and bedside rounds, demonstrating and teaching clinical skills and evidence-based clinical judgment. Valuable in their own right, these sessions also dramatized the clinical relevance and

importance of the trial and gained the respect of the front-line clinicians (often in training) who were most likely to encounter eligible patients. Having established and reinforced the credibility of the study and its investigators, he then would turn to issues of recruitment and follow-up, encouraging, instructing, or admonishing as the situation dictated. His visits were almost always followed by dramatic increases in both recruitment and data quality. Equally dramatic are the numbers of trials without peripatetic clinical leaders that failed to recruit even a small portion of their projected numbers of patients.

9. You can increase recruitment by employing strategies that have been shown in other RCTs to change the behavior of clinicians[8, 9,10]. For example, keeping a "log" of all remotely relevant patients (both eligible and ineligible) at each centre provides the base for audit and feedback to the individual clinicians who had agreed to approach such patients for the trial.

10. You can increase recruitment by recognizing both the needs and contributions of individual participating centres. Providing continuing education (as well as study clarification) to local staff, recognizing their contributions in final reports, and providing them the opportunity to carry out and publish their own ancillary studies strengthens their commitment to the success of the parent study.

The final 2 strategies protect against erosion of your effective sample size by making the most of patients you already have enrolled:

11. You can make (or protect) minor gains by keeping the numbers of control and experimental patients approximately equal (but not exactly so if that would threaten allocation concealment. When hunches favoring one of the treatments are strong, it may be tempting to randomize a larger proportion of eligible patients to that arm of the trial. However, there is a price to pay. Randomizing twice as many patients to one of the treatments (2:1 randomization) requires 12% more patients overall; 3:1 randomization requires 33% more patients[11].

12. The most important admonition throughout this chapter is to protect your sample size by not losing any study patients. Keeping track of all of them serves two related purposes. First, it detects events that otherwise would be missed. Second, it increases your chances of being able to present a convincing "worst-case scenario" (in which all experimental patients lost to follow-up in a trial with a positive conclusion are assigned bad outcomes and all lost control patients a rosy one). When losses-to-follow-up are so few that absolute risk reductions and their confidence intervals remain convincing in worst-case scenarios, the credibility of a trial's positive conclusion is enhanced.

This section closes with an admonition. You should be very reluctant to relax your eligibility criteria in order to increase your sample size. This is especially dangerous when you are considering adding patients who are at lower risk or less responsive than your target study population. As you saw back in the section on "physiological statistics," every low-risk, low-response patient you admit to your trial can make your need for additional patients go up, not down.

REFERENCES

[1] Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. BMJ 1998;316:690-3.

[2] Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. 3rd Edition. New York: Springer, 1998. P 120 ff.

[3] Klar N, Donner A. Current and future challenges in the design and analysis of cluster randomization trials. Stat Med 2001;20:3729-40.

[4] Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. 3rd Edition. New York: Springer, 1998. P 112 ff.

[5] Ross S, Grant A, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to participation in randomised controlled trials: a systematic review. J Clin Epidemiol 1999;52:1143-56.

[6] Mapstone J, Elbourne D, Roberts I. Strategies to improve recruitment to research studies (Cochrane Methodology Review). In: *The Cochrane Library*, Issue 1, 2004. Chichester, UK: John Wiley & Sons, Ltd.

[7] Sackett DL: Why randomized controlled trials fail but needn't: 1. Failure to gain "coal-face" commitment and to use the uncertainty principle. CMAJ 2000;162:1311-4.

[8] Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Educational outreach visits: effects on professional practice and health care outcomes (Cochrane Review). In: The Cochrane Library, Issue 1, 2001. Oxford: Update Software.

[9] Thomson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback: effects on professional practice and health care outcomes (Cochrane Review). In: The Cochrane Library, Issue 1, 2001. Oxford: Update Software.

[10] Thomson O'Brien MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Local opinion leaders: effects on professional practice and health care outcomes (Cochrane Review). In: The Cochrane Library, Issue 1, 2001. Oxford: Update Software.

[11] Altman, Douglas. Personal communication to Dave Sackett, 2001.