PHYSIOLOGICAL STATISTICS[1]

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

> "Because statistics has too often been presented as a bag of specialized computational tools with morbid emphasis on calculation it is no wonder that survivors of such courses regard their statistical tools as instruments of torture [rather] than as diagnostic aids in the art and science of data analysis."
> -George W. Cobb[2]

The myriad statistical formulae that appear in textbooks and articles about how to do Phase III RCTs are frightening to behold. They are tough to remember, and exist in isolation without relation to each other. In addition, they require an understanding of mathematics and statistics far beyond most would-be trialists' background knowledge and expertise. Finally, they take so much time to master that clinicians who do so risk losing their clinical competence, social life, positive self-image, and sense of humour

All the foregoing is true until you realize that the importance of these statistical formulae lies not in their individual application but in their thoughtful combination. Although it's possible (and in statistical circles, mandatory) to describe this combination in mathematical terms, clinicians might understand them far better by thinking of them in physiologic terms, analogous to combining the determinants of systemic arterial blood pressure.

A patient's blood pressure represents the net effects of multiple cardiac, central nervous system, endocrine, renal, and vascular factors (that can interact both synergistically and antagonistically). By wonderful analogy, the confidence we have in an RCT's results (that is, the narrowness of the confidence interval around the effect of the experimental treatment[1]) is the net result of the interaction of patients, treatments, and study factors that, as you'll see, also can behave synergistically and antagonistically.

The only formula you'll ever need

The "only formula" of physiological statistics is ridiculously simple, and looks like this:

$$\text{Confidence} = \frac{\text{Signal}}{\text{Noise}} \times \sqrt{\text{sample size}}$$

Expressed in words, the confidence you have in the conclusion of an RCT is the ratio of the size of the signal generated by your treatment to the size of the background noise, times the square root of sample size. Let's define these terms so that they are clear.

Confidence describes how narrow the confidence interval is (the narrower the better) around the effect of treatment, whether expressed as an absolute or relative risk reduction or as some other

[1] or, in the old-fashioned terms that most trialists have abandoned, the trial's "statistical significance"

measure of efficacy. For readers still imprisoned by P-values, this sort of “Confidence” becomes greater as the p-value becomes smaller.

The **Signal** describes the differences between the effects of the experimental and control treatments. In the RCTs in which I’ve taken part, the most useful Signal in understanding their design, execution, analysis, and interpretation has been the (absolute) arithmetic difference between the rate (or average severity) of events in experimental and control patients. When, as in most RCTs, these outcomes are “discrete” clinical events such as strokes, bleeds, or death, we’ll call this arithmetic difference (the control event rate minus the experimental event rate) the Absolute Risk Reduction (ARR).

Why don’t we prefer the more frequently reported Relative Risk Reduction (which is the absolute risk reduction divided by the control event rate)? This is because the relative risk reduction doesn’t distinguish important treatment effects from trivial ones (slashing deaths from 80% down to 40% generates the same relative risk reduction [0.5] as teasing them from 0.008% down to 0.004%).

Finally, in some RCTs the outcomes are “continuous” measures such as blood pressure, elapsed time on a treadmill before chest pain occurs, or location on a 0 to 100 scale of disease activity or functional status. In these latter cases, the signal is best represented for me by the Absolute Difference (AD) in this continuous measure.

The **Noise** (or uncertainty) in an RCT is the sum of all the factors (“sources of variation”) that can affect the absolute risk reduction or absolute difference. Why might patients’ responses to treatment, or our measurements of them, vary? Some of these sources are obvious but others aren’t, so I’ll use plenty of examples along the way.

Finally, **sample-size** is the number of patients in the trial. Note that its influence on confidence occurs as its square root. Accordingly, if you want to cut the confidence interval around a study’s absolute risk reduction in half by adding more patients to it, you need to quadruple their number. Alternatively, an RCT designed to confidently detect an absolute risk reduction of 0.10 needs to quadruple its size in order to confidently detect an absolute risk reduction of 0.05 (half as great).

A simple way to demonstrate the formula at work or play

For a quick appreciation of the “physiology” described by this formula, I suggest that you perform a simple experiment. Place a CD player alongside a radio. Ask a friend to Insert one of your favorite melodies (the signal) into the former but not tell you which one it is. Next, tune the radio to a spot between stations where you hear only static (the noise) and turn up the volume. Then start the audiotape at low volume. Note the “confidence” with which you can identify the melody within, say, 2 seconds. Then vary the volume of the CD (signal), the radio static (noise) and the amount of time (analogous to sample size) it takes you to discern the former amidst the latter.

Large, simple trials

In order to generate extremely small and highly convincing confidence intervals around moderate but important benefit signals, a very strong case can and has been made for really large, really simple RCTs[3] and systematic reviews[4]. Their success in revolutionizing the treatment and improving the outcomes of patients with heart disease, cancer and stroke attests to their success. When study patients number in the tens of thousands they can overcome, by the brute force of numbers, the negative influences of small but highly important absolute risk reductions (e.g., the polio vaccine trials that required hundreds of thousands of study individuals) in the presence of considerable noise (as long as the latter does not result from bias). They are described and discussed on page xx. However, most trials, even when carried out in multiple centers, are of small to moderate size, and they must confront and solve the challenges of small (but useful) signals in the face of lots of noise and a shortage of eligible patients.

Effects of signal, noise, and sample size on the confidence or our conclusions

Table 3-02-I-1 summarizes the independent effects of changes in each of these three elements on the confidence interval around a trial's absolute or relative risk reduction when the other two elements are held constant. If any of its entries are confusing, I suggest that you repeat the CD/radio experiment until they all make sense.

Table 3-02-I-1: Effects of changes in a single element on Confidence

Element that is changed	Effect on our Confidence in the RCT Result	
	When this element increases	When this element decreases
Signal (ARR)	Confidence rises	Confidence falls
Noise	Confidence falls	Confidence rises
Sample size	Confidence rises	Confidence falls

Note: Confidence increases as the confidence interval around the absolute risk reduction (ARR) signal narrows.

You are now ready to understand how each of these elements can raise or lower the confidence in your RCT result. But first a cautionary note. Because this pursuit of confidence may involve restricting the entry of certain sorts of patients into your RCT, it may shift it away from a "pragmatic" orientation ("Does offering the treatment to all patients do more good than harm under usual circumstances?") towards an "explanatory" one ("Can rigorously applying the treatment to just some subgroup of patients do more good than harm under ideal circumstances?"). I'll discuss the implications of this shift as they arise.

Determinants of the signal, and how they can be manipulated to maximize it

Four determinants affect the magnitude of the signal generated in an RCT (as you will see later, they can also affect noise). They are: the "baseline" or control group's risk of an outcome event, the potency of your experimental treatment, the responsiveness of experimental patients to it, and the completeness with which you detect outcome events. Your understanding of how these determinants operate begins and ends with your realization that the important number in an RCT is not the number of patients in it, but the number of outcome events among those patients.

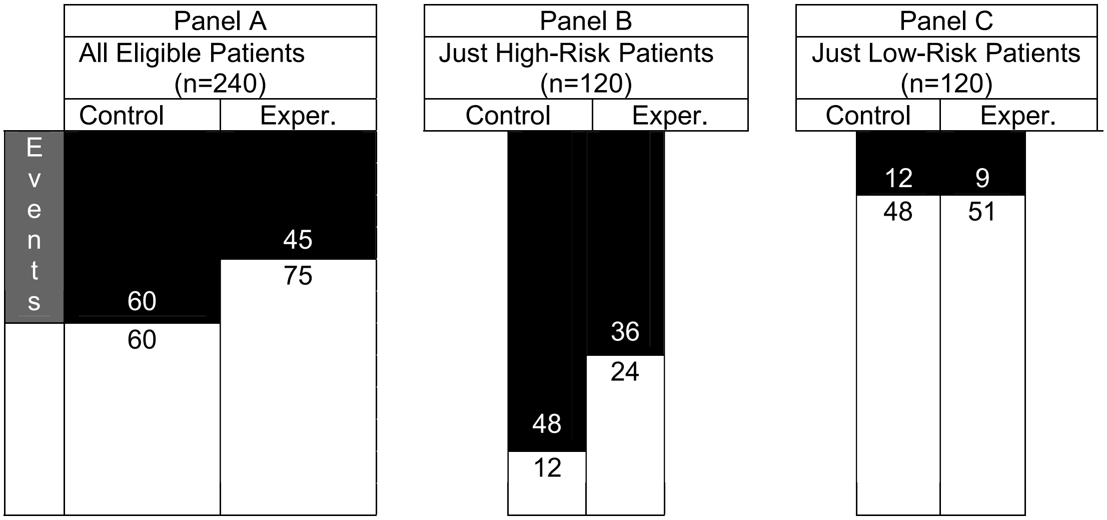
All 4 determinants operate in every group of individuals you consider for, or later invite to join, a Phase III RCT. Sometimes they are already at optimum levels: your patients are at high risk, your experimental treatment is powerful, all your patients can respond to it, and you can capture every outcome event. In that case, you won't need to apply any restrictive eligibility criteria on their account.

More often, however, these determinants are optimum only in certain sub-groups of potential study patients. Accordingly, you'll need to decide whether to selectively enroll just these optimum subgroups. As I'll show you in a moment, changing the eligibility criteria to achieve this selective enrolment can result in large, indeed definitive, increases in the signal you produce in your trial. On the other hand, the opportunity costs of examining, lab testing, and imaging all patients in order to find that optimum subgroup can be prohibitive. Moreover, as noted earlier, restricting patient eligibility criteria might shift your RCT away from its intended "pragmatic" orientation towards an "explanatory" one. With this caveat in mind, I'll now demonstrate each of these four determinants and how they can be manipulated to maximize your treatment signal.

Maximizing the signal by selectively enrolling “high risk” patients.

Restricting eligibility to patients who are at higher than average “baseline” risks of outcome events leads to higher “Control Event Rates” on control (and experimental) therapy. The absolute risk reduction signal (ARR) is the product of this control event rate (CER) and the relative risk reduction from therapy (RRR). In terms of simple maths, ARR = CER x RRR[5]. If the relative risk reduction constant over different control event rates, the experimental treatment will generate a larger absolute risk reduction signal when the control event rate is high than when it is low. This is illustrated in Table 3-02-I-2. If the relative risk reduction is 1/4 for all patients in the RCT (regardless of their control event rates), notice the different impacts on the absolute risk reduction signal and the corresponding confidence in the trial result when we enrol all patients and when we restrict enrolment to just the subgroups at high and low baseline risk. Recruiting and randomizing just the subgroup of 120 high-risk patients in Panel B generated both a higher absolute risk reduction (up from 0.125 to 0.20) and a 20% narrower confidence interval around it (from +/- 100% to +/- 80%) than randomizing all 240 patients in Panel A. An examination of the low-risk patients in Panel C shows how they inflate the confidence interval around the absolute risk reduction signal. In fact, every low-risk patient admitted to this trial makes the need for additional patients go up, not down!

Table 3-02-I-2: Effect of enrolling only "high-risk" patients with higher control event rates (CERs).



Control Event Rate	0.50	0.80	0.20
Relative Risk Reduction	1/4	1/4	1/4
Experimental Event Rate	0.375	0.60	0.15
Absolute Risk Reduction	0.125	0.20	0.05
Size of the 95% Confidence Interval around that Absolute Risk Reduction	+/- 100%	+/- 80%	+/- 270%
P-value	0.07	0.03	0.63

Note: In Panel A we have randomized 240 patients into equal sized control and experimental groups (and have lost none of them to follow-up). Although their overall risk of an event if left on conventional therapy is 50% (control event rate = 0.50), they are a heterogeneous lot and half of them (Panel B) are at high risk if left untreated (control event rate = 0.80) and half (Panel C) are at low risk (control event rate = 0.20). The relative risk reduction (1/4) is the same in all groups. Confidence intervals shown here are calculated as CI for a difference in absolute risk reductions, as described by Douglas Altman[6].

Remember that this strategy works only when the relative risk reduction is either constant or increasing as control event rates increase. Although there isn't much documentation about this, and there are some exceptions, I've concluded that relative risk reduction is pretty constant over different control event rates when the treatment is designed to slow the progression of disease and prevent its complications. This has been observed, for example, in meta-analyses of aspirin and the secondary prevention of cardiovascular disease[7], and of both ACE-inhibitors[8] and beta blockers[9] in heart failure. Moreover, in an examination of 115 meta-analyses covering a wide range of medical treatments, the control event rate was twice as likely to be related to the absolute risk reduction as to a surrogate for the relative risk reduction (the odds ratio), and in only 13% of the analyses did the relative risk reduction significantly vary over different control event rates[10]. When the treatment is designed to reverse the underlying disease, I've concluded that relative risk reduction should increase as control event rates increase, exemplified by carotid endarterectomy for symptomatic carotid artery stenosis where the greatest relative risk reductions are seen in patients with the most severe stenosis (and greatest stroke risks)[11].

When outcomes are "continuous" you can look for evidence on whether the experimental treatment will cause the same relative change in a continuous outcome (say, treadmill time) for patients with severe starting values (awful exercise tolerance, analogous to high-risk patients for discrete events) and good starting values (good but not wonderful exercise tolerance, analogous to low-risk patients for discrete events). If this evidence suggests a consistent relative effect over the range of the continuous measure, I hope it's clear why the absolute difference signal generated by experimental treatment is greater (and its confidence interval narrower) among the initially severe patients than among the less severe ones (if this isn't clear, consider how much "room for improvement" there is in a patient who already is doing pretty well vs. one who is doing poorly).

The harsh truth

Harsh as it may sound, you need people in your RCT who are the most likely to have the events you hope to prevent with your experimental treatment (e.g., myocardial infarctions, relapses of a dreadful disease, or death). And, as long as the relative risk reduction from treatment is constant or rises with increasing control event rates, these high-risk patients also have the most to gain from being in the trial. Finally, to be practical this "high-risk" strategy requires not only solid prior evidence that high- and low-risk patients exist, but also that their identification is easy and cheap enough to make their inclusion and exclusion cost-effective in conducting the trial.

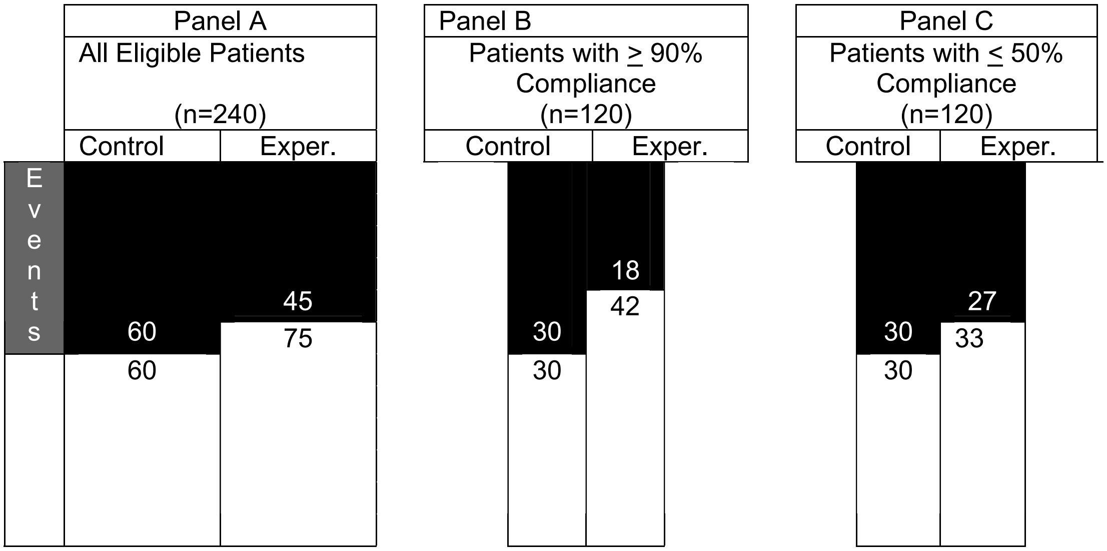
The foregoing should cause second thoughts among trialists who are considering arbitrary upper age limits for their trials; they may be excluding precisely the high-risk patients who will benefit the most, raise the absolute risk reduction and make the largest contribution to the confidence in a positive result. On the other hand, if high-risk (or severe) patients are too far-gone to be able to respond to the experimental therapy, or if competing events (e.g., all-cause mortality) swamp those of primary interest in the trial, the absolute risk reduction's confidence interval will expand and its signal might decrease. This discussion introduces a second element, responsiveness.

Maximizing the signal by selectively enrolling "highly-responsive" patients

The second way that you can increase the absolute risk reduction signal and the confidence in a positive trial result is by selectively enrolling highly-responsive patients who are more likely (than average) to respond to the experimental therapy. Their greater-than average relative risk reductions translate to increased absolute risk reductions and higher confidence in positive trial results. This increased responsiveness can arise from two different sources. The first and most easily determined cause is patients' compliance with an efficacious experimental therapy. Those who take their medicine might respond to it, but those who don't take their medicine can't respond to it. No wonder, then, that so much attention is paid to promoting and maintaining high compliance during RCTs, and why some RCTs put patients through a pre-randomisation "faintness-of-heart" task, rejecting those who are unwilling or unable to comply with it. This is because, once patients are randomized all of them must be included in subsequent analyses, even if they don't comply with their assigned treatment. The second cause for increased responsiveness is the result of real biologic differences in the way that subgroups of patients respond to experimental treatment. This biologic difference may be much more difficult (and expensive) to determine among otherwise eligible patients. Table 3-02-I-3 illustrates how either cause works among another 240 patients, this time with subgroups at the same baseline risk but with differing degrees of compliance (or other aspect of responsiveness).

Panel A of Table 3-02-I-3 is identical to Panel A of Table 3-02-I-2. If, as in Panel B of Table 3-02-I-3, just the highly compliant subgroup is recruited, the resulting confidence intervals around the absolute risk reduction are narrower than those observed among all 240 patients. However, every patient with low compliance (Panel C) admitted to this trial made the need for additional patients go up, not down! Note that this high-response strategy works best when control event rates are either constant or increasing among subgroups with progressively higher relative risk reductions. Once again, although there isn't much documentation of control event rates among subgroups with different responsiveness, patients in our carotid endarterectomy trials with higher control event rates also enjoyed greater relative risk reductions with surgery[11]. As in the case of high-risk patients, the identification of high-response patients has to be both accurate and inexpensive if it is to decrease the total effort necessary for achieving a definitive trial result.

Table 3-02-I-3: Effect of enrolling only highly-responsive patients.



Control Event Rate	0.50	0.50	0.50
Relative Risk Reduction	1/4	2/5	1/10
Experimental Event Rate	0.375	0.30	0.45
Absolute Risk Reduction	0.125	0.20	0.05
Size of the 95% Confidence Interval around that Absolute Risk Reduction	+/- 100%	+/- 86%	+/- 350%
P-value	0.07	0.04	0.72

Note: In Panel A we have randomised 240 patients into equal sized control and experimental groups (and have lost none of them to follow-up). Although their overall compliance rate is great enough to achieve a relative risk reduction of ¼, they are a heterogeneous lot and half of them (Panel B) take 90% or more of their study medication and achieve a relative risk reduction of 2/5, whereas the other half (Panel C) take 50% or less of it and achieve a relative risk reduction of only 1/10. The control event rate (0.50) is the same in all groups.

Maximizing the signal by combining risk and responsiveness

The foregoing elements of risk and responsiveness can usefully be combined as shown in Table 3-02-I-4, where I have summarized the “attractiveness” (in terms of maximizing the absolute risk reduction signal and the confidence of a positive trial result) of different sorts of patients whom you might consider entering into your RCT. This will come home to haunt you if, toward the end of your recruitment phase, you are short of “ideal” patients and decide to relax your inclusion criteria and start admitting lower risk or less compliant individuals. As predicted in Tables 3-02-I-2 and 3-02-I-3, admitting such patients may increase, rather than decrease, the remaining sample size requirement (and administrative burdens) that must be satisfied to achieve a sufficiently large absolute risk reduction and sufficiently narrow confidence intervals around it.

Table 3-02-I-4: The attractiveness of different sorts of potential RCT patients.

		Responsiveness to (Compliance with) the Experimental Rx (Relative Risk Reduction)	
		High	Low
Risk (Control Event Rate)	High	Ideal!	Are they too sick to benefit? Admit with caution
	Low	Are they too well to need any treatment? Admit with caution	Keep out!

Maximizing the signal by giving enough treatment over enough time

The third way that you can tend to raise an absolute risk reduction signal and the confidence in a positive trial result is to employ a potent experimental treatment and give it a chance to exert its effect. You shouldn’t expect patients to experience better outcomes when their treatment regimens aren’t administered in a sufficient dose for a sufficient duration. Thus, an RCT to see whether drastic reductions in blood pressure reduce the risk of stroke must employ a drug that, in Phase 2 trials, really does reduce blood pressure to the desired level. This “be-sure-your-experimental-treatment-is-potent” strategy is dramatically demonstrated in surgical trials, where the principal investigators may restrict their clinical collaborators to just those surgeons with excellent skills and low perioperative complication rates. In similar fashion, you should be sure that the experimental treatment is applied long enough to be able to achieve its favourable effects, if they are to occur.

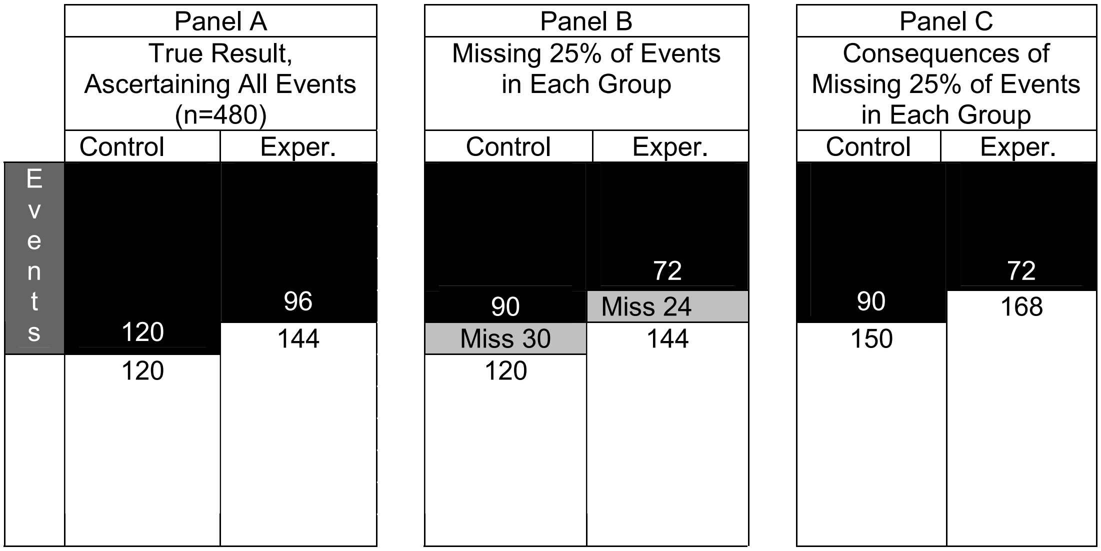
If you digested the foregoing, you’ll quickly grasp the incremental price of therapeutic progress that trialists must pay as they search for marginal improvements over treatments they already have shown, in previous RCTs, to do more good than harm. When today’s standard treatment is already known (through prior RCTs) to do more good than harm, clinicians and ethics committees should and will insist that this “Established Effective Therapy” (rather than a placebo) be provided to the control patients in any subsequent RCT of the next generation of potentially more effective treatments. As a result, the control event rates are progressively reduced in subsequent trials (they behave like the low risk patients described above), and even if an absolute risk reduction is maintained at its former level, its confidence interval will widen. No surprise, then, that RCTs in acute myocardial infarction have become huge and hugely expensive, not (only) because cardiologists are an entrepreneurial lot, but because they already are reducing control event rates with the thrombolytics, beta blockers, aspirin, and ACE-inhibitors they validated in previous positive trials.

As forecast in the introduction, the foregoing strategies for increasing the absolute risk reduction and narrowing its confidence interval by restricting trial participants to just the high-risk, high-response group, by maximizing compliance, by employing just the best surgeons, and so forth moves the resultant trial away from a “pragmatic” study question (“does offering the treatment do more good than harm under usual circumstances?”) toward an “explanatory” study question (“can rigorously applying the treatment do more good than harm under ideal circumstances?”)[12]. If the original question was highly pragmatic and intended to compare treatment policies rather than rigorous regimens, the strategies described above may be unwise and it becomes more appropriate to conduct a really large, simple trial. Similarly, these restrictive strategies may raise concerns (and not a few hackles) about the “generalisability” of the trial result. As I’ve argued elsewhere[13], it is my contention that front-line clinicians do not want to “generalize” an RCT’s results to all patients, but only to “particularize” its results to their individual patient, and already routinely adapt the trial result (expressed, say, as a “number-needed-to-treat” or NNT, which is the inverse of the absolute risk reduction) to fit the unique risk and responsiveness of their individual patient, the skill of their local surgeon, the patient’s preferences and expectations, and the like[14]. Moreover, cautionary pronouncements about generalisability have credibility only if the failure to achieve it leads to qualitative differences in the kind of responses patients display such that, for example, experimental therapy is, on average, unambiguously helpful for patients inside the trial but equally unambiguously harmful or powerfully useless, on average, to similar patients outside it. This issue is discussed in greater detail on page xx.

Maximizing the signal by ascertaining every event

The fourth way that you can maximize an absolute risk reduction signal and the confidence in a positive trial result is to make sure that you identify and record (that is, ascertain) every event suffered by every patient in the trial. Up to this point, I have assumed that all events have been ascertained in both control and experimental patients, and that the resulting absolute risk reduction signal, regardless of whether it is large or small, is true. In other words, although the absolute risk reductions displayed in Tables 1 & 2 are affected by the risk-responsiveness composition of the study patients, they nonetheless provide unbiased estimates of the effects of treatment. What happens in the real world of RCTs, where the ascertainment of events is virtually always incomplete? As you will see, this leads to systematic distortion of the absolute risk reduction signal away from the truth; that is, this estimate of the signal becomes biased. Suppose that the RCT’s follow-up procedures were loose, and many patients were lost. Or, suppose that the outcome criteria were so vague and subjective that lots of events were missed. If experimental and control patients are equally affected by this incomplete ascertainment, the situation depicted in Table 3-02-1-5 would occur, with a loss in the strength of the absolute risk reduction signal even though the relative risk reduction is preserved. Accordingly, the fourth way that you can increase the absolute risk reduction signal and the confidence in a positive trial result is by improving the ascertainment of events during the RCT. This is shown in Table 3-02-I-5.

Table 3-02-I-5: What happens with equally incomplete ascertainment of events in both control and experimental patients.


Panel A
True Result, Ascertaining All Events (n=480)
Control
Exper.
Events
120
96
120
144
Panel B
Missing 25% of Events in Each Group
Control
Exper.
72
90
Miss 24
Miss 30
144
120
Panel C
Consequences of Missing 25% of Events in Each Group
Control
Exper.
72
90
168
150

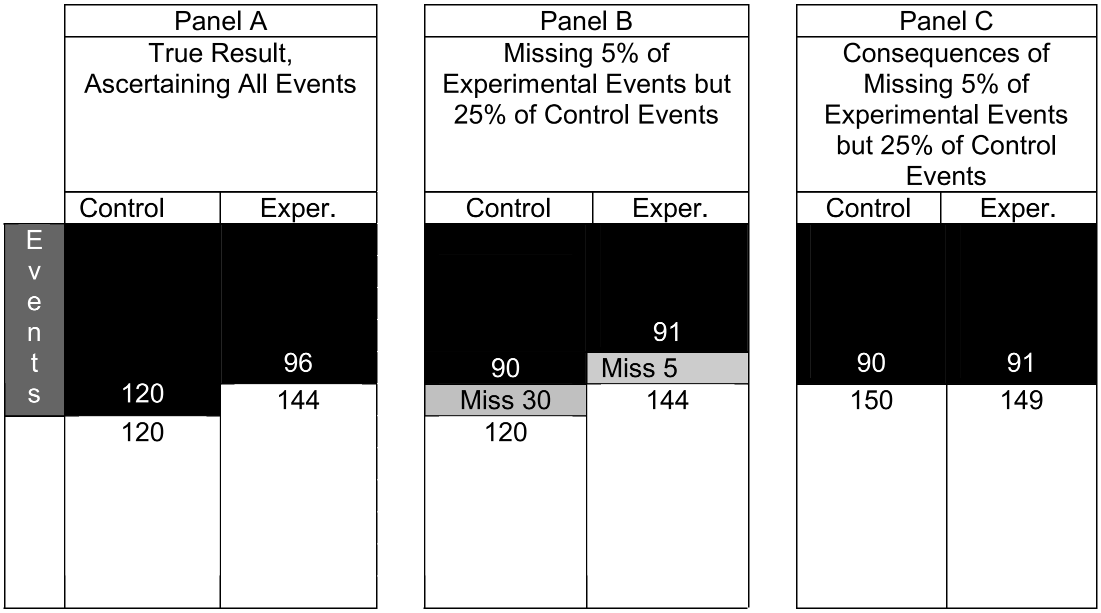
	Truth	Effect of this problem:	Resulting analysis:
Control Event Rate	0.50	⇩ 25%	0.375
Relative Risk Reduction	1/5	Same	1/5
Experimental Event Rate	0.40	⇩ 25%	0.30
Absolute Risk Reduction	0.10	⇩ 25%	0.075
Size of the 95% Confidence Interval around that Absolute Risk Reduction	+/- 89%	increases	+/- 110%
P-value	0.04	rises	0.10

Note: Panel A of Table 3-02-I-5 displays the true effect of the experimental treatment: a relative risk reduction of 1/5, generating an absolute risk reduction signal of 0.10 whose confidence intervals exclude zero. If experimental and control patients are equally affected by this incomplete ascertainment (missing, say, 25% of events in both groups) the misclassification of events depicted in Panel B would occur. As a consequence, shown in Panel C, although the relative risk reduction is preserved, the absolute risk reduction signal declines from 0.10 to 0.075, its confidence interval now crosses zero, and the trial result becomes indeterminate.

The price of unequal ascertainment among control and experimental patients

But what if the accuracy of ascertainment differed between control and experimental patients, such as might occur in non-blinded trials when experimental patients are more closely followed (e.g., for dose-management and the detection of toxicity) than control patients? What if that greater scrutiny of experimental patients led to missing only 5% of events in the experimental group while continuing to miss 25% of control events? This situation is shown in Table 3-02-I-6. Missing more events among control than experimental patients not only decreases the absolute risk reduction signal but also widens its confidence interval. In this case, the bias leads to a "conservative" type II error (concluding that the treatment may be useless when, in truth, it is efficacious), and presents a powerful additional argument for blind RCTs (since they maintain equal scrutiny of experimental and control patients and equal ascertainment of their outcome events).

Table 3-02-I-6: What happens with better ascertainment of events in experimental than control patients.


Panel A
True Result, Ascertaining All Events
Control
Exper.
Events
120
120
96
144
Panel B
Missing 5% of Experimental Events but 25% of Control Events
Control
Exper.
90
Miss 30
120
91
Miss 5
144
Panel C
Consequences of Missing 5% of Experimental Events but 25% of Control Events
Control
Exper.
90
150
91
149

	Truth	Effect of this problem:	Resulting analysis
Control Event Rate	0.50	⇩ 25%	0.38
Relative Risk Reduction	1/5	Wiped out	Zero!
Experimental Event Rate	0.40	⇩ 5%	0.38
Absolute Risk Reduction	0.10	Wiped out	Zero!
Size of the 95% Confidence Interval around that Absolute Risk Reduction	+/- 89%	Explodes	∞
P-value	0.04	Soars	1

Note: Panel A of Table 3-02-I-6 displays the true effect of the experimental treatment: as in Table 3-02-I-5, there is a relative risk reduction of 1/5, generating an absolute risk reduction signal of 0.10 whose confidence intervals exclude zero. If experimental and control patients are unequally affected by this incomplete ascertainment (missing 25% of events in the control group but only 5% of events in the experimental group) the misclassification of events depicted in Panel B of Table 3-02-I-6 would occur. As a consequence, shown in Panel C, both the relative and absolute risk reductions are falsely reduced and the trial draws a false-negative conclusion.

A parallel lesson here is the need to achieve complete follow-up of patients in both explanatory and pragmatic trials. Remember, the important number in an RCT is not the number of patients in it, but the number of outcome events among those patients.

Having defined the determinants of the signal generated in an RCT and demonstrated how they can be manipulated to maximize that signal, it is time to consider how noise affects our confidence in the trial result, and how that noise can be reduced.

Determinants of the noise, and how they can be manipulated to minimize it

The effects of noise and its reduction are perhaps best understood by considering RCTs whose outcomes are continuous measures (blood pressure, functional capacity, quality of life, and the like) rather than discrete events (such as major stroke, brain metastasis, or death). The key to understanding noise is to think of all the sorts of factors ("sources of variation" or, better yet, "sources of uncertainty") that might affect the end-of-study result for this continuous measure, not just in the individual study patient but especially in the groups of patients that comprise the experimental and control groups in the RCT.

Consider blood pressure. You know from prior experience that you won't get the same blood pressure result for every patient in an RCT. Indeed, you know that repeat measurements of blood pressure in the same patient at the same visit will generate different results (depending on whether it's the first or the fourth measurement at that visit, on whether they are inhaling or exhaling, on whether they are talking, on whether you are supporting their arm and back, and so forth). At the group level you must add the variation in blood pressure that exists between study patients (based not only on differences in their individual endocrine, cardiovascular, and nervous systems and responses to therapy, but also depending on how well they know their examiner and the timing of their last cigarette, their last meal, their last conversation, their last void, and by which of several types of sphygmomanometers are being applied to them by which examiners with what hearing acuity and which preferences for the terminal digits 0, 2, 4, 6, and 8). These sources of variation in recorded blood pressure may, in combination, create so much noise that it becomes impossible to detect the signal (say, a small but important reduction in blood pressure) being generated by the experimental treatment.

Strategies for minimizing noise

How might you minimize this noise, recalling from the first section of this essay that decreases in noise are rewarded by decreases in confidence intervals around signals and, therefore, increases in our confidence about the results of the trial? In this case, the link between statistics and physiology is just about perfect. As summarized in Table 3-02-I-7, you reduce the noise element in your trial by eliminating or minimizing sources of uncertainty. I'll illustrate this with the blood pressure example.

1. You can make sure that every study patient actually has the target condition whose natural history you are attempting to change. Misdiagnoses at patient entry create subgroups of patients with the wrong conditions who may be incapable of responding to your experimental treatment, thus adding noise to the trial.

2. You can remove the uncertainty that arises from studying the two different treatments in separate, "parallel" groups of different patients (with their different baseline blood pressures and responses to treatment) by applying both treatments to every patient. This is accomplished by randomizing, for each patient, the order in which they receive the experimental and control regimens, separated by an intervening period of sufficient length to "wash-out" any effects of the previous regimen. This "within-patient" or

"crossover" design, if feasible, removes the effect of any variation between study patients and usually produces big reductions in noise that are reflected in big reductions in confidence intervals (ambitious readers can verify this by contrasting the results of paired and unpaired t-tests on a data set obtained from a cross-over trial). Although theoretically attractive, cross-over trials are not suited for disorders subject to irreversible events or total cures, and patients who withdraw or drop-out before completing both treatment periods are tough to analyze. Moreover, it is impossible to tell whether there is a "carry-over" of the effects of the first treatment into the second treatment period until the trial is over. When these carry-over effects are large, the data for the second period may have to be thrown away, the trial's noise continues unabated, and you are no better off than if the trial had been a more usual "parallel" design in the first place.

3. You can reduce variations in the outcomes of study patients by making them more homogeneous through the same strategies that you employed in the previous section: assembling study patients with similar risks (e.g., just those with the highest blood pressures) and similar responsiveness to the experimental treatment. This can be done either by "restricting" admission to the trial to just those patients with similar risk and responsiveness, or by stratifying study patients for these features and then randomizing from each stratum. The result is a narrower band of blood pressures and blood pressure changes with therapy (smaller standard deviations for these measures) and reduced noise. As previously mentioned, in explanatory surgical trials we routinely reduce uncertainty in responsiveness by drafting only those surgical collaborators who can document their high success and low complication rates.

4. You can reduce noise by making experimental and control patients as similar as possible in their risk and responsiveness. Although random allocation tends to create similar groups (and is our only hope for balance in unknown determinants of responsiveness), we can ensure similarity for known determinants by stratification prior to randomization or even by minimization (allocation of the next patient to which ever treatment group will minimize any differences between the groups)[15]. Minimization is described with an example on page xx.

5. In similar fashion, you can reduce noise by achieving similar (and high) compliance among all study patients.

6. You can minimize sloppiness and inconsistency in the ascertainment of outcomes. Not only should your outcome criteria be objective and unambiguous; they should be applied (or at least adjudicated) by two or more observers who are "blind" to which treatment a study patient has received. In trials whose outcomes are measured in absolute differences (say, in hemoglobin levels), noise is reduced by analyzing the averages of duplicate or triplicate determinations of the outcome.

Table 3-02-I-7: Strategies for reducing noise in an RCT

Strategy	Tactics
Validate eligibility	Make sure study patients have the target condition.
Crossover Trial	Give both treatments to every patient, in random order.
Homogenize	Restrict participants to a single risk-response subgroup.
Minimize	Render experimental and control patients as identical as possible in their risk and responsiveness.
Maintain high compliance	Monitor compliance with study regimens, and apply compliance-improving strategies when needed.
Ascertain all events	Achieve complete follow-up of all study patients and ascertain outcomes in every one of them.

Increasing Sample Size: the last resort

Reducing confidence intervals by increasing the size of an RCT should be a last resort. There are two major reasons for this admonition. First, as I stated at the start of this section, in order to halve the width of the confidence interval around the absolute risk reduction achieved by your experimental treatment, you need to quadruple the number of patients in your trial. For example, in Panel A of Table 3-02-I-1, to halve the confidence interval for an absolute risk reduction from +/- 100% to +/- 50% demands a quadrupling in sample size from 240 to 960 patients. Only after exhausting the foregoing strategies for increasing the signal and reducing the noise should you take on the daunting task of increasing your sample size. The second reason why it may be dangerous to attempt to rescue an RCT that is too small, is that scouring recruitment sites with relaxed inclusion/exclusion criteria leads to the recruitment of low-risk, low-response patients. Tables 2-4 reveal that adding patients of these sorts can paradoxically lower absolute risk reductions and increase the confidence interval around them. Of course, sample size requirements can be revisited during a trial (with care not to destroy blindness) and methods are available for determining the risk of drawing false negative conclusions after a trial is completed[16].

There are 11 strategies that you can employ either to increase your sample size or make the most of whatever sample size you do recruit. I'll present them in Section 3-10 of this chapter, when I discuss sample size.

Gaining first-hand experience with (and the "feel" of) physiological statistics

Just as the understanding of human physiology benefits from dynamic laboratory and bedside (real-life) observations of the effects of altering a single determinant (say, peripheral resistance) on a "final common pathway" (say, arterial blood pressure), aspiring trialists can increase their understanding of physiological statistics by creating the tables in this essay from their own protocols and data sets and examining the effects of altering these determinants, singly and in combination, on a final common pathway such as the confidence interval around an absolute risk reduction.

The simple experiment with the CD player and radio that opened this essay provided primitive insights. Better still, and analogous to what can be learned from interactive computer models of human physiology, aspiring trialists can study the combined effects of different signal strengths, different amounts of noise, and different sample sizes in computer models of randomized trials. For example, a Clinical Trials Simulator has been developed by an international consortium that is promoting and aiding the performance of pragmatic trials in low-income countries. Its current version can be accessed via the "PraCTIHC" website[17].

Users of these simulators can input whatever risks, responsiveness, compliance, loss to follow-up, ascertainment of outcomes, drop-outs, cross-overs, etc they desire. The simulator then carries out a few hundred simulations in a few seconds and displays the effects of these inputs on both the validity of the hypothetical trials and the confidence intervals around their signals.

I reckon the more that trialists use such CD and radio, pencil-and-paper or computer simulations to "massage" their assumptions before they start a trial, the less they'll have to "massage" their inconclusive data after it's over.

REFERENCES

[1] This section is condensed from a longer piece that appeared in the CMAJ 2001;165:1226-37.

[2] George W. Cobb: Introductory textbooks: a framework for evaluation. J Am Stat Assoc 1987;82:321-339.

[3] Yusuf S, Collins R, Peto R: Why do we need some large, simple randomized trials? Stat Med 1984;3:409-22

[4] Baigent C: The need for large-scale randomized evidence. Br J Clin Pharmacol 1997;43:349-53.

[5] Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB: Evidence-Based Medicine. 2nd edition. Edinburgh: Churchill Livingstone, 2000. Pages 111-3.

[6] Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB: Evidence-Based Medicine. 2nd edition. Edinburgh: Churchill Livingstone, 2000. Page 235.

[7] Antiplatelet Trialists' Collaboration: Collaborative overview of randomised trials of antiplatelet therapy--I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. BMJ. 1994;308:81-106.

[8] Garg R, Yusuf S. Overview of randomized trials of ACE-inhibitors on mortality and morbidity in patients with heart failure. JAMA. 1995; 273: 1450-1456.

[9] Heidenreich PA, Lee TT, Massie BM: Effect of beta-blockade on mortality in patients with heart failure: a meta-analysis of RCTs. J AM Coll Cardiol. 1997;30:27-34.

[10] Schmid C H, Lau J, McIntosh M W, Cappelleri J C: An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Statistics in Medicine 1998;17:1923-42.

[11] Barnett HJ, Taylor DW, Eliasziw M, Fox AJ, Ferguson GG, Haynes RB, Rankin RN, Clagett GP, Hachinski VC, Sackett DL, Thorpe KE, Meldrum HE: Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. N Engl J Med. 1998;339:1415-25.

[12] Sackett DL, Gent M: Controversy in counting and attributing events in clinical trials. N Engl J Med. 1979;301:1410-2.

[13] Sackett DL: Pronouncements about the need for “generalizability” of randomized control trial results are humbug. Cont Clin Trials 2000;21(2S):

[14] McAlister FA, Straus SE, Guyatt GH, Haynes RB: Users' guides to the medical literature: XX. Integrating research evidence with the care of the individual patient. Evidence-Based Medicine Working Group. JAMA 2000;283:2829-36.

[15] Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991. Pp 443-5.

[16] Detsky AS, Sackett DL. When was a "negative" clinical trial big enough? How many patients you needed depends on what you found. Arch Intern Med. 1985;145:709-12.

[17] The website for the “Pragmatic RAndomized Controlled Trials In Health Care” consortium is www.practihc.org