Review

Synthèse

Dr. Sackett is Director of the Trout Research & Education Centre at Irish Lake, Markdale, Ont.

This article has been peer reviewed.

CMAJ 2001;165(9):1226-37

Return to October 30, 2001 Table of Contents

Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!)

David L. Sackett

> 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[1]

From the underside

It was the seventh time I had taken a course in basic biostatistics, and I vowed that this time I was actually going to understand it. The year was 1974, and we were on sabbatical in London. The Beatles had started to beat on each other, Ali beat Frazier, the good guys beat Nixon, Lord Lucan beat his nanny, and I could still beat the odds of getting run over while bicycling between NW3 and St. Thomas's Hospital. I had the time, I had the "coal-face" experience (by then I'd been a PI on several randomized controlled trials [RCTs]), and I already knew the English, Greek and Latin bits (I'd aced my 6 previous courses). Moreover, I'd had the good fortune to have worked with biostatisticians who were not only brilliant methodologists but also outstanding teachers.

Alas, history repeated itself. At the end of my studies I was as incapable of applying this course to my current phase III RCTs as I had been incapable of applying my medical school biostatistics course to my patients 18 years earlier. Chastened, I returned home to find that the new crop of would-be trialists who also had successfully completed graduate courses in biostatistics were as confused as I when they tried to integrate how sick their prospective study patients might be, how well their outcomes might be ascertained, and how powerful their interventions might be with how many patients they might need to enrol and how certain they could be about any of their conclusions. Even today, 26 years later, the young clinical-practice researchers who come to our Trout Workshops up here in the woods still find it difficult or impossible to see the practical forest among the statistical trees.

Causes

I've concluded that the fault here lies neither with the teachers nor the would-be trialists, but with an irreconcilable mismatch between what's judged necessary to be learnt about biostatistics and who's to learn it. The myriad statistical formulas that appear in textbooks and articles about how to do phase III RCTs possess 5 drawbacks for the clinician–would-be-trialist:

- *They are frightening to behold*. I reckon numerophobia is as prevalent among clinician-trialists as are refractive errors.
- *They are tough to remember*. Forty-four years after my (mostly irrelevant) anatomy course I can still name the cranial nerves ("Ole Olson Ought To Take A Fling At *****"), but I've never encountered a mnemonic for recalling the formula for the 95% confidence interval around a difference in proportions.
- *They require an understanding of mathematics and statistics far beyond most would-be trialists' background knowledge and expertise*. Graduate statistical programs have recognized that an extensive prior knowledge of mathematics is a prerequisite for later competency as a practising biostatistician. However, most health pro-

fessional schools have recognized the irrelevancy of the traditional, "hard" basic sciences to the provision of effective and compassionate clinical care. Very few successful applicants to medical school would be accepted into graduate training in biostatistics.

- *Time taken to master their nuances is at the expense of maintaining clinical competence, a social life, a positive self-image and a sense of humour*. While writing a recent essay on the fall of clinical research,[2] I tried to think of any clinical colleague who, after taking sufficient time away from clinical practice to master statistics (or, to be fair, molecular biology), I'd trust to take care of a really sick patient with multisystem disease. I could think of only one.
- *They exist in isolation, without relation to each other*. Wisdom in designing RCTs requires both the thoughtful integration of several statistical principles and sound clinical judgement, neither of which is to be found in individual statistical formulas. When this fact dawns on learners who have struggled to master them, it shouldn't surprise us if they heap criticism on their statistics courses and act in ways described by Bokonon: "Beware of the man who works hard to learn something, learns it, and finds himself no wiser than before. He is full of murderous resentment of people who are ignorant without having come by their ignorance the hard way."[3]

Given the foregoing, I find it far more remarkable that some clinicians succeed in integrating their first course in biostatistics than that most of them fail to do so.

Preventive strategy: a 3-part solution and introduction to "physiological statistics"

Whatever success I've had as a clinician-trialist has been the result of good luck, great statistical colleagues and the development of 3 strategies to overcome these 5 drawbacks. I offer them here in the possibility that they may be of use to others:

1. *Forget the formal formulas*. I know fewer of them now than when I designed my first RCT in 1963.
2. *Never work alone*, but always in collaboration with statisticians (and whatever other experts could contribute to the success of the enterprise). Complex statistical analyses, far beyond the competency of clinician-trialists, are occasionally vital in understanding the result of an RCT. A statistical collaborator will know when and how to use them. The great majority of clinician-trialists I've encountered know enough statistics to get into trouble, but not enough to get out again.
3. *Employ "physiological statistics."* As noted above, the importance of statistical formulas lies not in their individuality 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 physiological terms, analogous to combining the determinants of systemic arterial blood pressure. Just as 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), the confidence we have in an RCT's results (that is, the narrowness of the confidence interval around the effect of the experimental treatment or, in the old-fashioned terms that most trialists have abandoned, the trial's "statistical significance") is the net result of the interaction of patients, treatments and study factors that, as you'll see, also can behave synergistically and antagonistically. Invoking "physiological statistics" to combine the formulas gives us licence to borrow from the time-honoured tradition of employing physiological "stories," which, although they have great explanatory power, are not quite true. In similar fashion, the price for clarity in the rest of this essay is the occasional stretching of statistical truth (I apologize in advance to statistical purists and will no doubt do penance for my statistical hubris in the letters-to-the-editor department).

The "only formula" of physiological statistics

The formula is ridiculously simple, and looks like this

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

(Equation 1):

Expressed in words, the *confidence* in the conclusion of an RCT is the ratio of the magnitude of the *signal* to the magnitude of the *noise* times the square root of the *sample size*.

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 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 been involved, the most useful signal in understanding their design, execution, analysis and interpretation has been the (absolute) arithmetic difference obtained when you subtract the rate (or average severity) of events among experimental patients from the rate of events among control patients. When, as in most RCTs, these outcomes are discrete clinical events such as strokes, bleeds or death, I'll call this arithmetic difference (the control event rate minus the experimental event rate) the *Absolute Risk Reduction* (ARR). Why don't I prefer the more frequently reported Relative Risk Reduction (which is the absolute risk reduction divided by the control event rate)? 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%). Despite the criticism that, in some circumstances, the absolute risk reduction can be influenced more by patients' underlying risks (their control event rates) than by the proportion by which that risk is reduced with treatment (the relative risk reduction), I prefer the former. 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 0–100 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.

Finally, *sample size* is the number of patients in the trial. Note that its influence on confidence intervals is as its square root. As you'll see later, this means that, 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 detect an absolute risk reduction of 0.10 needs to quadruple its sample size in order to detect an absolute risk reduction of 0.05 (half as great).

For a quick appreciation of the "physiology" described by this formula, I suggest that readers pause at this point and perform a simple experiment. Place an audiotape player next to a radio. Ask a friend to insert one of your favourite melodies (the signal) into the former but not tell you which one it is. 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 and note the "confidence" with which you can identify the melody as you vary the volume of the audiotape (signal), the radio static (noise) and the amount of time (analogous to sample size) it takes you to discern the former amidst the latter.

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[4] and systematic reviews.[5] Their effect of 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) and considerable noise (as long as the latter does not result from bias). However, most trials, even when carried out in multiple centres, are of small to moderate size, and they must confront and solve the challenges of small (but useful) signals, large amounts of noise and scarce patients.

Table 1 summarizes the effects of changes in each of these 3 elements on the confidence interval around a trial's absolute or relative risk reduction when the other 2 elements are held constant (if any of its contents are confusing, I suggest that you repeat the audiotape/radio experiment until they make sense).

You can now identify and understand the factors that raise or lower confidence in an RCT result by acting on each of these elements. Since your objective is to maximize everyone's confidence in that result, the remainder of this essay will describe these 3 elements in terms of how they help or hinder the achievement of this goal. Because this pursuit of confidence may involve restricting the entry of certain sorts of patients into your RCT, it may start to shift away from a "pragmatic" orientation ("Does offering the treatment to all patients do more good than harm under usual circumstances?") toward an "explanatory" one ("Can rigorously applying the treatment to just some subgroup of patients do more good than harm under ideal circumstances?"), and I will 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, these factors may also affect noise). They are the "baseline" or control group's risk of an outcome event, the responsiveness of experimental patients to that treatment, the potency of the experimental treatment, and the completeness with which outcome events are ascertained and included in the analysis. Understanding how these determinants operate begins and ends with the 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 are present in every group of individuals being initially considered for, or later invited to join, a phase III RCT. Sometimes they are already optimum (in terms of maximizing the signal) within all potential study patients, and no restrictive eligibility criteria need to be applied on their account. More often, however, they are optimum only in certain subgroups of these pa-

Table 1: Effects of changes in a single element on our confidence in the randomized controlled trial (RCT) result

	Effect on our confidence in the RCT result*	
Element	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

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

tients, and the trialist needs to decide whether to selectively enrol just these optimum subgroups. As we shall see, manipulations of eligibility criteria to accomplish this selective enrolment can result in large, indeed definitive, increases in the signal produced by the trial. On the other hand, the opportunity costs of examining, lab testing and imaging all patients in order to find just the optimum subgroup of them may be prohibitive. Moreover, as noted in the previous section, eligibility criteria might shift an RCT away from its intended "pragmatic" orientation ("Does offering this treatment to all patients do more good than harm under usual circumstances?") toward an "explanatory" one that is more difficult to apply ("Can rigorously applying the treatment to just some subgroup of patients do more good than harm under ideal circumstances?"). With those caveats in mind, we can now consider each of the determinants and how they convert into strategies for maximizing the signal.

Selectively enrol "high-risk" patients

Restricting eligibility to patients who are at higher than average "baseline" risk of outcome events leads to higher "Control Event Rates" (CER) among those receiving placebo or standard therapy. Because the absolute risk reduction signal is equivalent to the product of this control event rate and the relative risk reduction from therapy (ARR = CER × RRR)[6] it follows that, if the relative risk reduction achieved by the experimental treatment is both true and 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 Fig. 1. 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

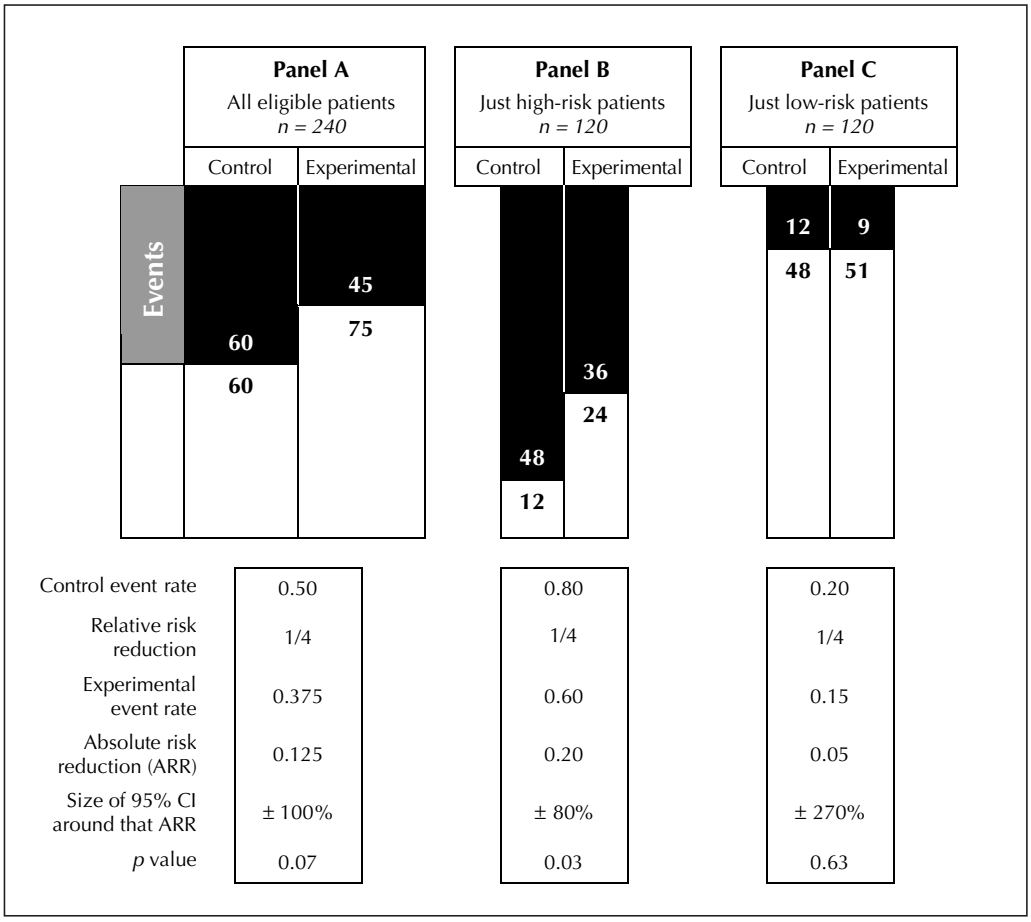

Fig. 1: Effect of enrolling only patients with higher control event rates ("high-risk" patients). In panel A we have randomly assigned 240 patients into equal-sized control and experimental groups (and have lost none to follow-up). Although their overall risk of an event if given conventional therapy is 50% (control event rate 0.50), they are a heterogeneous lot: 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 (CIs) shown here are calculated as the CI for a difference in absolute risk reductions.[6]

when we restrict enrolment to just the subgroups at high and low baseline risk. Recruiting and randomly assigning 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 randomly assigning 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!

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 ASA and the secondary prevention of cardiovascular disease,[7] and of both ACE inhibitors[8] and β-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]

> The important number in an RCT is not the number of patients in it, but the number of outcome events among those patients.

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 patients with initially severe disease than among those with less severe disease (if this isn't clear, consider how much "room for improvement" there is in a patient who already is doing pretty well v. one who is doing poorly).

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 patients (or those with severe disease) 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.

Selectively enrol 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 2 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-randomization "faintness-of-heart" task, rejecting those who are unwilling or unable to comply with it. This is because, once patients are randomly assigned, 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. Fig. 2 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 is identical to panel A of Fig. 1. If, as in panel B, just the highly compliant subgroup is recruited, the resulting confidence interval around the absolute risk reduction is

narrower than that 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 in subgroups with progressively higher relative risk reductions. Once again, although there isn't much documentation of control event rates in 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 highly responsive patients has to be both accurate and inexpensive if it is to decrease the total effort necessary for achieving a definitive trial result.

The foregoing elements of risk and responsiveness can usefully be combined as shown in Table 2, where I have summarized the "attractiveness" (in terms of maximizing the absolute risk reduction signal and the confidence in a positive trial result) of different sorts of patients whom you might consider enrolling into your RCT. This will come home to haunt you if, toward the end of your recruitment

Table 2: The attractiveness of different sorts of potential RCT patients

Risk (control event rate)	Responsiveness to (compliance with) the experimental treatment (relative risk reduction)	
	High	Low
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!

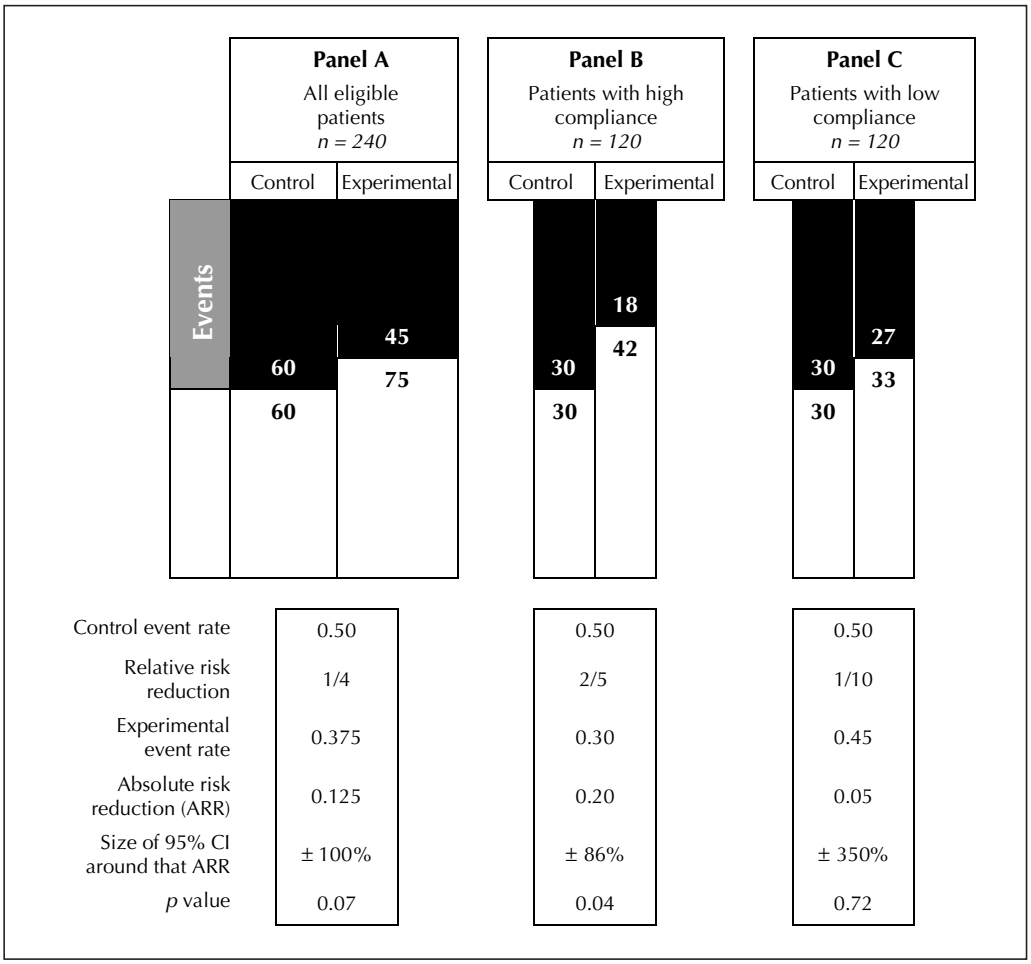

Fig. 2: Effect of enrolling only patients with higher relative risk reductions (highly responsive patients) in an RCT. In panel A we have randomly assigned 240 patients into equal-sized control and experimental groups (and have lost none to follow-up). Although their overall compliance rate is great enough to achieve a relative risk reduction of 1/4, they are a heterogeneous lot: half of them (panel B) are highly compliant and achieve a relative risk reduction of 2/5, and half (panel C) display low compliance and achieve a relative risk reduction of only 1/10. The control event rate (0.50) is the same in all groups.

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 Figs. 1 and 2, 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 a sufficiently narrow confidence interval around it.

> 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.

Use a potent experimental treatment and give it a chance to exert its effect

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 II 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 "standard therapy" (rather than a placebo) be provided to 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 in panel C of Fig. 1), and even if relative risk reductions are maintained at their former levels, the resulting absolute risk reductions will fall and their confidence intervals 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, β-blockers, ASA 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 generalizability 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 generalizability 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. I'll address this straw man in a later essay in this series.

Identify and record (ascertain) every event suffered by every patient in the trial

This is the fourth way that you can maximize an absolute risk reduction signal and the confidence in a positive trial result. 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

> Relative risk reduction should increase as control event rates increase when the treatment is designed to reverse the consequences of the underlying disease.

displayed in Table 1 and Fig. 1 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. 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 Fig. 3.

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 Fig. 3 would occur, with a loss in the strength of the absolute risk reduction signal even though the relative risk reduction is preserved. But what if the accuracy of ascertainment differs between control and experimental patients, such as might occur in nonblinded 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 leads to missing only 5% of events in the experimental group while continuing to miss 25% of events in the control group? This situation is shown in Fig. 4. Missing more events among control patients than among 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

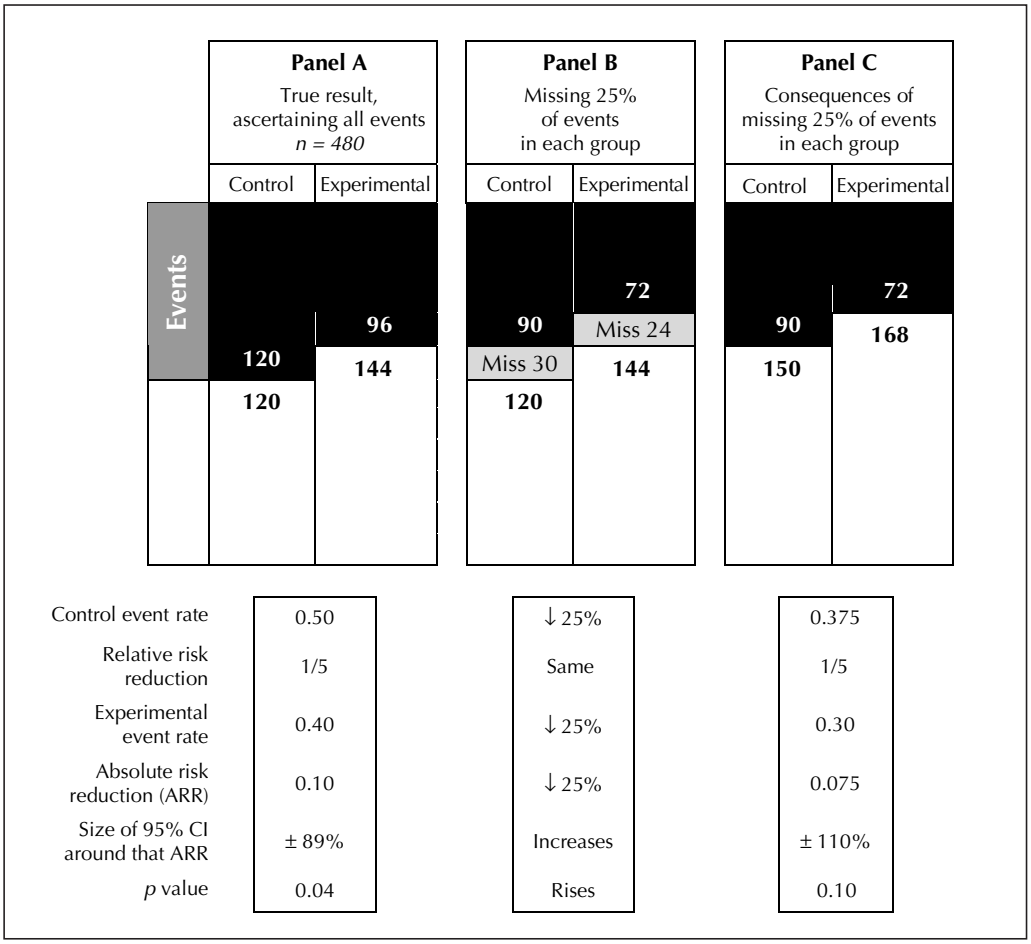

Fig. 3: Effect of equally incomplete ascertainment of events in both control and experimental patients. Panel A displays the true effect of the experimental treatment: a relative risk reduction of 1/5, generating an absolute risk reduction signal of 0.10 whose confidence interval excludes zero. If the experimental and control patients are equally affected by this incomplete ascertainment (missing, say, 25% of events in each group) the misclassification of events depicted in panel B would occur. As a consequence (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.

argument for blind RCTs (since they maintain equal scrutiny of experimental and control patients and equal ascertainment of their outcome events).

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 (e.g., blood pressure, functional capacity or quality of life) rather than discrete events (e.g., 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.

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 in the same patient at the same visit will generate different results (depending on whether, for example, it's the first or the fourth measurement at that visit, whether they are inhaling or exhaling, whether they are

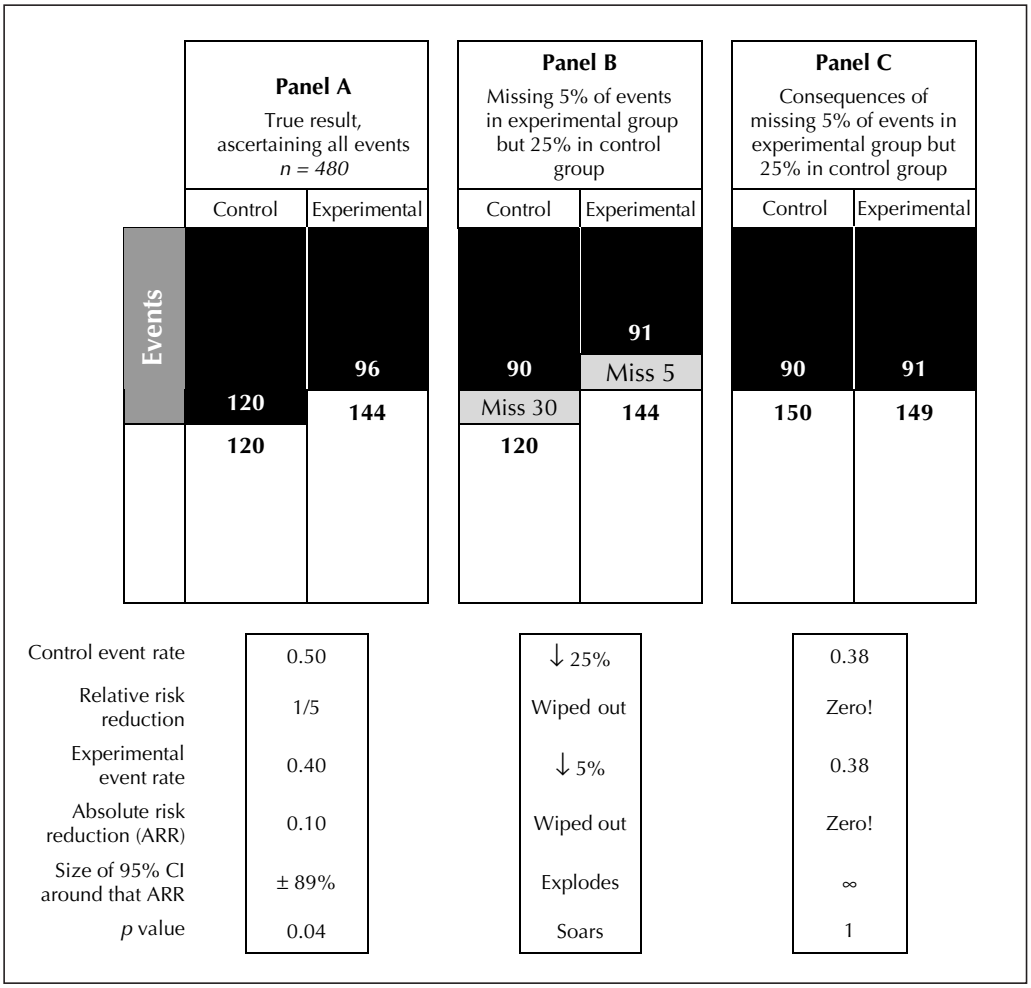

Fig. 4: Effect of better ascertainment of events in the experimental group than in the control group. Panel A displays the true effect of the experimental treatment: as in Fig. 3, there is a relative risk reduction of 1/5, generating an absolute risk reduction signal of 0.10 whose confidence interval excludes zero. If the 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 would occur. As a consequence (panel C), both the relative and absolute risk reductions are falsely reduced, and the trial draws a false-negative conclusion.

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 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.

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 in the results of the trial? In this case, the link between statistics and physiology is just about perfect. You reduce the noise element in your trial by eliminating or minimizing sources of uncertainty. I'll illustrate this with the blood pressure example.

- You can remove the uncertainty that arises from studying 2 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 treatment. This "within-patient" or "crossover" design, if feasible, removes 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 crossover trial). Although theoretically attractive, crossover 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, and the trial's noise continues unabated.
- You can reduce variations in the outcomes of study patients by making the patients 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 risks and responsiveness or by stratifying study patients for these features and then randomly assigning patients 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 stated, 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.
- You can reduce noise by making experimental and control patients as similar as possible. 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 whichever treatment group will minimize any differences between the groups).[15]
- In similar fashion, you can reduce noise by achieving similar (and high) compliance among all study patients.
- 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 2 or more observers who are blind to which treatment a study patient has received. In trials whose outcomes are measured in absolute differences (e.g., in hemoglobin levels), noise is reduced by analyzing the averages of duplicate or triplicate determinations of the outcome.
- 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.

Increasing sample size

Reducing confidence intervals by increasing the size of an RCT should be your last resort. There are 2 major reasons for this admonition. First, as I stated at the start of this essay, 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 Fig. 1, to halve the confidence interval for an absolute risk reduction from ±100% to ±50% demands a quadrupling of the 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 or exclusion criteria often leads to the recruitment of low-risk, low-response patients. Figs. 1 and 2 and Table 2 reveal that

adding patients of these sorts can paradoxically lower absolute risk reductions and increase the confidence intervals 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 to make the most of whatever sample size you do recruit. They come in 3 sets.

General strategies for increasing your sample size

- You can make it easier for clinical collaborators to approach and enter patients into the trial by shortening the entry forms to include just those items that are of immediate relevance. For example, some of the large, simple trials have used entry forms that take up less than a page.
- In similar fashion, you can reduce the complexity and time expended in deciding whether every patient is eligible for a trial by reducing its eligibility criteria to a bare minimum and employing the "uncertainty principle"[17] as the main determinant of an individual patient's eligibility.
- You can reduce the effort required of busy clinical collaborators by providing research assistants to help them with forms, baseline measurements, allocation and follow-up appointments. I vastly prefer this strategy to that of providing "bounties" to clinicians for every patient they enter.
- You can encourage "out-of-hours" recruitment by maintaining a randomization "hotline" on a 24/7 basis.
- 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.
- 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.

Strategies to ensure that all eligible patients are approached

The next 3 are strategies for overcoming the near-universal failure of participating centres (including your own!) to approach all eligible patients.

- You can increase recruitment from your current centre(s) by frequently exposing them 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 judgement. 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 and 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.
- You can increase recruitment by employing strategies that have been shown in other RCTs to change the behaviour of clinicians.[18–20] 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.
- 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.

Strategies for protecting against erosion of your sample size

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

- Minor gains can be made (or protected) by keeping the numbers of control and experimental patients equal. When hunches favouring one of the treatments are strong, it may be tempting to randomly assign a larger proportion of eligible patients to that arm of the trial. However, there is a price to pay. Randomly assigning twice as many patients to one of the treatments (2:1 randomization) requires 12% more patients overall; 3:1 randomization requires 33% more patients.[21]
- The most important admonition in this essay is to protect your sample size by not losing any study patients. Keeping track of all of them serves 2 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.

Gaining first-hand experience with 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 recreating the figures 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 audiotape 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 developed by Wayne Taylor and Eric Bosch[22] permits users to input whatever risks, responsiveness, compliance, loss to follow-up, ascertainment of outcomes, dropouts, crossovers, and so forth, they desire into the model and determine their joint effects on both the validity of their hypothetical trials and the confidence intervals around their signals.

I reckon that the more trialists use such audiotape-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.

Competing interests: Dave Sackett has been wined, dined, supported, transported and paid to speak by countless pharmaceutical firms for over 40 years, beginning with 2 research fellowships and interest-free loans that allowed him to stay to finish medical school. Dozens of his randomized trials have been supported in part (but never in whole) by pharmaceutical firms, who never received or analyzed primary data and never had veto power over any reports, presentations or publications of the results. He has twice worked as a paid consultant to advise pharmaceutical firms whether their products caused lethal side-effects; on both occasions he told them Yes. He has testified as an unpaid expert witness for a stroke victim who successfully sued one manufacturer of oral contraceptives, and has been paid by a second to develop "levels of evidence" for determining the causation of adverse drug reactions. His wife inherited and sold stock in a pharmaceutical company. While head of a division of medicine he enforced the banning of drug-detail personnel from clinical teaching units (despite the threat of withdrawal of drug industry funding for resident research projects). He received the Pharmaceutical Manufacturers' Association of Canada Medal of Honour (and cash) for "contributions to medical science in Canada" for the decade 1984–1994. His most recent award (the 2001 Senior Investigator Award of the Canadian Society of Internal Medicine) was sponsored by Merck Frosst Canada.

Acknowledgements: I sent drafts of this essay to 15 of the best trialists I know, and each of them sent back extremely helpful comments. To protect them from the wrath of their more traditional colleagues I am maintaining their anonymity.

References

1. Cobb GW. Introductory textbooks: a framework for evaluation. *J Am Stat Assoc* 1987;82:321-39.
2. Sackett DL. The fall of "clinical" research and the rise of "clinical-practice research." *Clin Invest Med* 2000;23:331-3.
3. Vonnegut K Jr. Books of Bokonon. In: *Cat's cradle*. New York: Dell; 1963. p. 187.
4. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409-22.
5. Baigent C: The need for large-scale randomized evidence. *Br J Clin Pharmacol* 1997;43:349-53.
6. Altman DG. Confidence intervals. In: Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine*. 2nd ed. Edinburgh: Churchill Livingstone; 2000. p. 111-3.
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;273:1450-6.
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 CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med* 1998;17:1923-42.
11. Barnett HJ, Taylor DW, Eliasziw M, Fox AJ, Ferguson GG, Haynes RB, et al. 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. *Control Clin Trials* 2000;21:82S
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. p. 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. 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. Available: www.cma.ca/cmaj/vol-162/issue-9/1311.htm
18. 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.
19. 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.
20. 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.
21. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall; 1991. p. 460.
22. Taylor DW, Bosch EG: CTS: a clinical trials simulator. *Stat Med* 1990;9:787-801.

Correspondence to: *Dr. David L. Sackett, Trout Research & Education Centre at Irish Lake, RR 1, Markdale ON N0C 1H0; sackett@bmts.com (No reprints are available.)*

Editor's note: The first article in this series of essays appeared in the May 2, 2000, issue of *CMAJ*, page 1311.