ANALYSIS AND INTERPRETATION

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

Analysis and Interpretation Check List:

Before you begin your trial, and based on your study question:	
1 ❍	Draft "Table 1" summarizing the entry characteristics of experimental and control patients.
2 ❍	Specify your primary and secondary data analyses
3 ❍	Specify your subgroup analyses
4 ❍	Select your analytic methods for deciding whether your treatment effect is "real".
5 ❍	Decide how to handle missing data in the analysis.
6 ❍	Decide how you will interpret your results to determine their "importance"
7 ❍	Establish interim analysis plans and statistical warning rules for efficacy, safety, and futility
After your trial is over	
8 ❍	Don't exaggerate your conclusions, especially about subgroups
9 ❍	Report your results regardless of their interpretation
10 ❍	Update the systematic review that justified your trial
11 ❍	Formulate the logical question for your next trial

Preface: Is this section really necessary?

In the opening paragraphs of this chapter, I stressed the importance of recruiting a statistician as co-principal investigator right at the start of formulating the question for your RCT. Why, then, intrude on their turf with a section on the analysis and interpretation of the trial? My reasons are three. First, as you can see from the checklist, several of the issues are not strictly statistical (specifying what to analyze, interpreting your results, and the like). Second, co-PIs are precisely that, and everyone with that title should collaborate in the discussion and debate at every step in the trial. Accordingly, this section's first function is to provide non-statisticians with a sufficient introduction to RCT analysis to help you contribute to those discussions and debates. Third, when some non-statistician trialists get their feet wet in statistics, they discover (to their surprise and mine) that they enjoy learning more about it. So, this section's final function is to whet some appetites.

We used parametric (t-test) and nonparametric (chi-squared) tests to analyze base-line differences among study groups, associated hematologic investigations, and compliance. We used the log-rank life-table method suggested by a team of experts led by Richard Peto[1] for our primary analysis. This primary analysis assessed the overall benefit of aspirin and sulfinpyrazone in all patients. However, we also judged it important to examine the relative efficacy of these drugs among clinically sensible subgroups. We advised readers to interpret these secondary analyses with caution since true significance levels are affected by repeated challenges of the data.

We monitored withdrawals to detect possible drug toxicity.

Our first examination of the data for efficacy occurred in April 1976, when we had entered 569 patients into the study. At that time there was a trend favoring aspirin that was not statistically significant. We decided to continue admitting patients until June 30, 1976, by which time we expected to reach the target of 600 patients. We would then follow all patients for a further 12 months, analyze, and interpret our results.

In our primary analysis, aspirin achieved a statistically significant (P<0.05) reduction in the composite hierarchy of TIA, stroke and death. In a secondary analysis that excluded TIAs, aspirin still achieved a statistically significant (P<0.05) reduction in stroke and death. Sulfinpyrazone was not effective. Aspirin produced a relative risk reduction of 31% and an absolute risk reduction of 7.2% for stroke and death. Thus, the Number of patients one Needed to Treat (NNT) with aspirin for 2 years to prevent another stroke or death was 14. The Number of Patients one Needed to treat to Harm one of them (NNH) with a major gastrointestinal bleeding over that same period was 48.

I later decided that one of our planned analyses was a bad idea. Can you guess which one that was? Read on.

Let's proceed with the checklist.

1. <u>Draft "Table 1" summarizing the entry characteristics of experimental and control patients.</u>

The first table in your RCT report describes and compares the entry characteristics of your experimental and control patients. We suggest that you show "empty" drafts of this table to potential clinical collaborators, including especially those who you hope to influence with its results. Typically, such tables include characteristics likely to influence risk or responsiveness to treatment, plus sociodemographic items. We already showed you Table 1 for men in the RRPCE trial as Table 3-6-2, and it was accompanied by similar tables for women and for the occurrence and timing of their qualifying TIAs. As we noted back in Table 3-6-1, a baseline imbalance between treatment groups for important prognostic characteristics can damage a trial's credibility in ways that multivariate statistical adjustments can never rehabilitate. Accordingly, when you create your first draft of your Table 1 before you start the trial, you should decide whether to take steps (such as stratification-before-randomization or minimization) to be sure that your trial is credible as well as valid.

In answer to the question posed at the end of the scenario, I think we erred in applying significance tests to our Table 1, and recommend against it. In a small trial, minimallyimportant differences might be statistically non-significant. But they should be prevented, not documented after the fact. I've described their preventives (minimization or stratification prior to randomization) in section 3-06 on allocation. Conversely, in a large trial, trivial differences will routinely be statistically significant, and will suggest important imbalance when it is absent.

Many trialists seek statistical reassurance that baseline imbalances didn't affect their trial results. They do this by performing multivariate outcome analyses in which they adjust for one or more baseline factors. Some statisticians disagree with this approach[2]. When Stuart Pocock's team reviewed 50 RCT reports in general medical journals, 72% of them included such analyses (but gave reasons for doing so only about half the time)[3]. When performed, the "covariate adjusted" analyses received more emphasis than the unadjusted analysis about a third of the time. However, only one report changed its conclusion (incorrectly, in the Pocock team's opinion) on the basis of an adjusted analysis.

2. <u>Specify your primary and secondary data analyses.</u>

Which events, at what point or over what period of time, will answer your trial's primary question? If you did a good "PICOT" job of specifying your question back at the beginning, this should be an automatic decision. Even if your question addresses equivalence or non-inferiority (but greater safety or lesser cost), the issues are the same. Later portions of this section will focus on this primary analysis.

My co-authors and I carry out 5 sorts of secondary analyses in our RCTs. They deal with determining safety, subdividing a composite primary outcome, assessing secondary outcomes, confirming homogeneity across clinical subgroups and centers, and generating hypotheses for our next RCT. Most of these analyses are straightforward and trustworthy, but others can mislead. And subgroup analysis is tricky enough to deserve its own section below.

i. Determining Safety: These document the magnitude, timing, severity, and outcomes of adverse responses to your experimental therapy. Such "safety" secondary analyses are routine in planning and conducting RCTs, and often must adhere to rigorous external regulations in reporting.

ii. Sub-dividing a composite primary outcome: As you learned on page xxx, in order to generate enough events to achieve a statistically significant result, many RCTs create composite primary outcomes. Some of these combine primary events (such as death or heart attack), and some may add predicaments (such as the need for hospitalization for unstable angina or heart failure, or the need to perform angioplasty). Others are "hierarchies" that combine frequent mild events with their less common but more severe sequalae. As already reported, in the RRPCE study, our primary outcome was a composite of continuing TIA, stroke, or death. Although TIAs are clinically important, their inclusion in the primary outcome was, in part, a sample-size "hedge." Because they occurred much more frequently than stroke or death, they contributed lots more events to the primary analysis, and increased the trial's power (i.e., its ability to find, and label as statistically significant, a beneficial effect of aspirin). Once we had our positive primary outcome, we removed patients who only had continuing TIAs and did a secondary analysis on just strokes and deaths. If you're wondering why we didn't analyze TIAs all by themselves, you'd better (re)read about diagnostic hierarchies in Section 3-08 on Events.

iii. Assessing secondary outcomes: Lots of RCT outcomes are secondary by design. For several of our cerebrovascular trials, Brian Haynes developed measures that captured how well our patients retained their independence in communicating, dressing, eating, toileting, shopping, and the like[4]. These measures provided important functional confirmation of the consequences of the clinical events in these RCTs. Because every patient contributes a functional "event" in the analysis, we were confident that we would have ample sample size to demonstrate any minimally important differences in function, and we were right.

For example, in NASCET we let our collaborating surgeons decide the dose of aspirin given to trial patients at the time of their surgery. In dredging our data, we found, to our surprise, that patients taking 650-1300 mg of aspirin daily at the time of surgery were much less likely to suffer perioperative stroke or death (1.8%) than patients taking 0-325 mg of aspirin (6.9%). As with any other clinical observation, we could come up with a biologic explanation that would tidily explain this finding[5]. But these were Level 2b cohort data discovered while looking for the pony. Wayne Taylor decided they weren't a sound enough basis for clinical practice, and led a subsequent RCT that asked: "Among patients undergoing carotid endarterectomy, would giving them 81, or 325, or 650, or 1300 mg of aspirin, starting before their surgery and continuing thereafter, reduce their risk of stroke, myocardial infarction or death at 30 and 90 days?" Some surgeons were convinced that high-dose aspirin was efficacious and refused to join this second trial. However, enough of them and their patients did join to give us the startling and important answer that low-dose aspirin, not high-dose aspirin, was best at preventing stroke and death following surgery[6].

iv. Generating hypotheses for your next RCT: You shouldn't hesitate to perform "exploratory data analyses" or "data-dredging" to look for subgroups of patients who display major differences in their responses to therapy. However, the purpose for this search must never be to draw conclusions about subgroup efficacy. Rather, it is to generate questions to ask in your next trial. This quest shouldn't be undertaken until you understand how to wield the two-edged sword of subgroup analysis, so I'll take that up here.

3. <u>Specify your subgroup analyses</u>:

You will want to determine whether your primary result is consistent across clinically sensible subgroups. Our confidence (and our readers' confidence!) in our positive NASCET result was raised when we found consistent efficacy in clinically sensible subgroups based on sex, site and type of qualifying TIA, and comorbidity. In multicenter trials, you will often look for similar results across centers and countries. For example, in the RRPCE trial we found that 14 of our 24 centers (contributing 75% of our patients) agreed with the overall result; 5 centers (contributing 10% of our patients) showed no trend, and 5 (contributing 15% of our patients) showed a reverse trend. A test for heterogeneity across centers was not statistically significant, but if it had been, we'd have performed a "sensitivity analysis" to see whether excluding the centers with the most extreme results affected our conclusion about efficacy.

You shouldn't be surprised to find minor differences in the degree of therapeutic responsiveness of different subgroups of patients in your trial. I'll call these "quantitative" interactions, to denote that they represent differences in the degree of efficacy, such that one clinical subgroup is slightly more or less responsive to experimental therapy than another. For example, in the NASCET trial the relative risk reduction (RRR) for ipsilateral stroke rose with increasing symptomatic carotid stenosis (from 12% in patients with 70-79% stenoses, to 18% in patients with 80-89% stenoses, and 26% in patients with 90-99% stenoses).

But alarm bells should sound when your secondary analysis suggests a "*qualitative*" difference in efficacy between subgroups. By "*qualitative*" difference, I mean finding that experimental treatment is clearly efficacious in one subgroup and clearly (and statistically significantly) harmful or "confidently ineffective" in another. By "confidently ineffective," I mean that the 95% confidence interval for efficacy in that subgroup excludes anyhumanly useful benefit.

Secondary analyses among clinical subgroups can mislead you, for if you carry out enough of them, you are guaranteed to find one by chance alone. Even when supported by statistical tests for an interaction between efficacy and the presence or absence of a subgroup's identifying characteristic, these sorts of secondary analyses can mislead. Furthermore, the risks of over-interpreting subgroup analyses go beyond mere mischief. They include withholding efficacious treatment from subgroups who need it, forcing useless treatments on subgroups who don't, and wasting millions of dollars on research to clean up the messes. For example, in the RRPCE trial we concluded that aspirin worked in men but not women (wrong!), and that it didn't work among diabetics (wrong again!) or in patients with a past history of myocardial infarction (wrong yet again!). The same Christmas story about "looking for the pony" that helped us explain the dangers of performing multiple diagnostic tests on patients in the 2nd edition of this book[7] is useful here:

"Looking for the pony" comes from a Christmas tale of two brothers, one of whom was an incurable pessimist and the other, an incurable optimist. On Christmas day, the pessimist was given a roomful of shiny toys and the optimist, a roomful of horseshit. The pessimist opened the door to his roomful of toys, sighed, and lamented, "A lot of these are motor driven and their batteries will run down; and I suppose I'll have to show them to my cousins, who'll break some and steal others; and their paint will chip; and they'll wear out. All in all, I wish you hadn't given me this roomful of toys." The optimist opened the door to his roomful of horseshit and, with a whoop of glee, threw himself into the muck, and began burrowing through it. When his horrified parents extracted his from the excrement and asked him why on earth he was thrashing about in it, he joyfully cried: "With all this horse shit, there's got to be a pony in here somewhere!"

There are 2 ways to safeguard against spurious "qualitative interactions." First, you can limit your secondary analyses of subgroups to just 1 or 2 of them, carefully pre-specified in the protocol. Second, if you think that you will find an important qualitative interaction between subgroups, you can design separate and simultaneous trials for each of them. Each of these trials should have a sufficient sample size to answer the question. For example, in NASCET we suspected that there should be a qualitative interaction between the efficacy of surgery and the degree of carotid stenosis. We thought that it probably would produce a big net benefit among patients with high-

grade stenoses but that it might be useless or even harmful among patients with only moderate stenoses). We therefore designed and carried out two simultaneous trials, one for each of them, but with the same study staff and follow-up apparatus.

In summary, it's fine to perform "exploratory data analyses" or "data-dredging" to look for the pony you might like to ride in your next trial, as long as you don't draw conclusions about subgroup efficacy.

4. Select your analytic methods for deciding whether your treatment effect is "real."

I will provide you with only the "bare bones" of an approach to statistical analysis in this chapter. No readers in their right minds should undertake RCTs without biostatisticians as co-Principal Investigators, and I reckon that most of you already will have taken at least an introductory course in biostatistics. If any of the following ideas and suggestions are unclear, I suggest that you read Gordon Guyatt's chapter on statistics. If that doesn't help, consult your co-PI and/or your favorite statistical text (mine is Doug Altman's *Practical Statistics for Medical Research*[8])

Short-term parallel trials

As with other steps in executing your trial, the selection of the "right" analytic method depends on the question you posed. If yours is a short term trial (say a few days or weeks) in which your question calls for a simple comparison of homogenous groups of patients at the end of this time period, then simple statistical analyses will do. For short-term parallel trials with events (say, the occurrence of immediate side effects after taking an established drug and its newer, presumably better-tolerated nephew), a straightforward chi-squared test will serve just fine. For short-term parallel trials with continuous measurements (say, which of two bronchodilators produces a better improvement in the ease of breathing [FEV-1] 30 minutes later), your statistician co-PI will probably use an analysis of covariance. When you analyze a result as both an event (such as achieving goal blood pressure) and as a continuous measure (such as average blood pressure reduction), you need to specify up front which analysis will take precedence in answering your trial's question. Remember, however, that you will require far more patients to show a real difference in event rates (using the chi-squared family of statistics) than in averages (the t-test family).

Short-term crossover trials

Short term cross-over trials would use the analogous paired tests: the McNemar chi-squared or the paired t-test, and we show an example of the latter in Table 3-09-1, which displays treatment effects (Δ) in patients who have been allocated to receive treatment A or B in the first period and the other treatment in the second period:

Table 3-09-1: Treatment effects (Δ) in a crossover trial:

		Effects of Treatment (Δ)	
		Period 1	Period 2
Patient Allocation	A first	Δ A1	Δ B2
	B first		

		Δ B1	Δ A2

However, before you carry out the paired t-test on the data from a cross-over trial (Δ all A vs. Δ all B), you need to be sure that it is an unbiased analysis, unaffected by "carry-over" or "calendar":

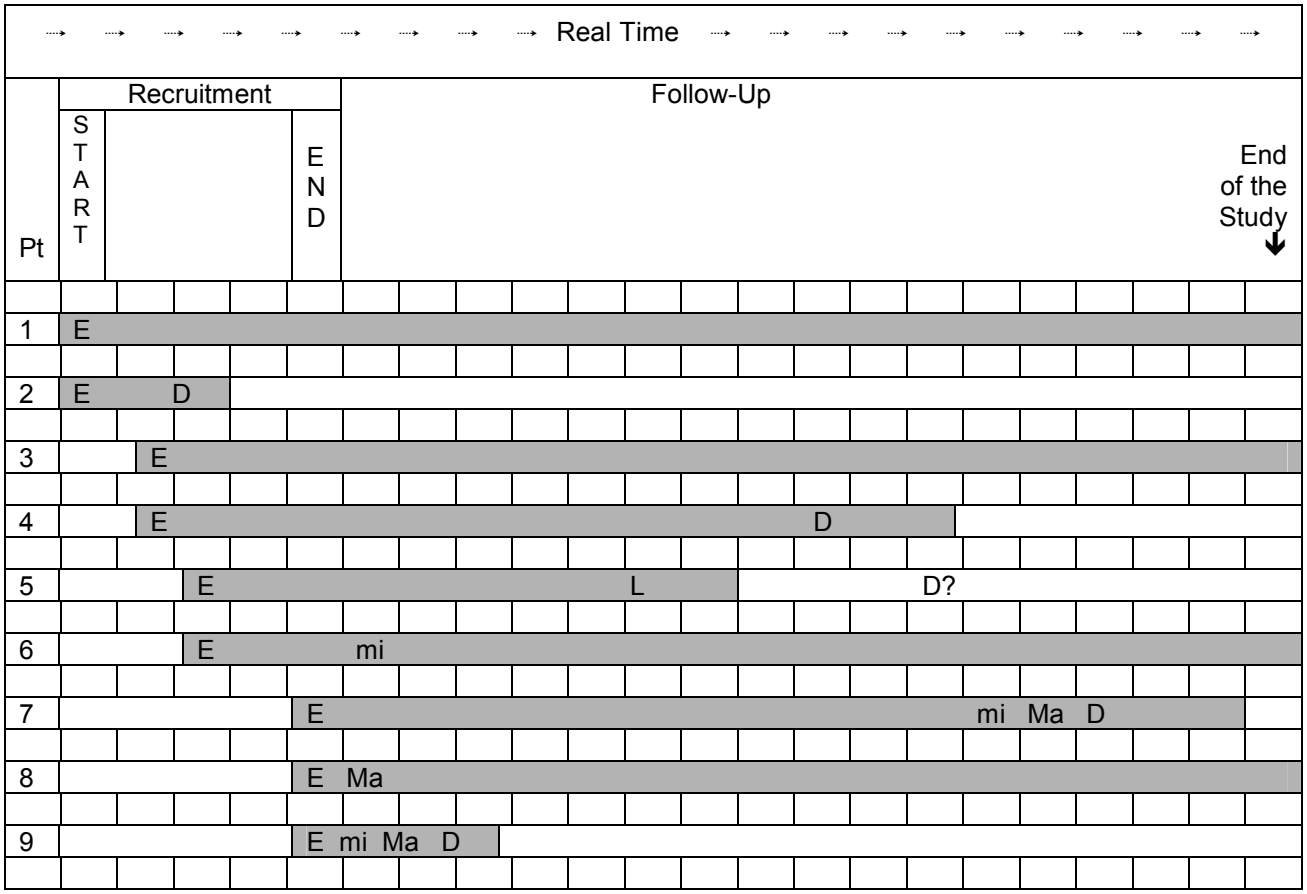
1. You'll need to find out whether there has been any "carry-over" effect of the treatment given in the first period into the second period. This is found by comparing the results within each treatment when it is given first and second: (Δ A1 vs Δ A2 and Δ B1 vs. Δ B2); you'll need to show that they are not statistically significantly different in the two periods before you can combine them.

2. You'll need to find out whether there has been any a "calendar" or "temporal" effect in which the patients' underlying illness is getting better (i.e., recovering) or worse with the simple passage of time. You can do this by comparing the differences between treatments in the first and second periods: ([Δ A1 minus Δ B1] and ([Δ A2 minus Δ B2]). Again, you'd need these to be roughly equal before carrying out the paired t-test on the overall result.

Long-term trials:

Most of the trials I've carried out have been long-term ones (lasting from two to several years) in which we were hoping to either prevent, or at least postpone, bad outcomes among patients at varying risk for these outcomes. Two special features of these trials have to be taken into account in their analyses, and these are illustrated in Figure 3-09-1 for experimental patients who enter a trial at point (E), and thereafter may (or may not) go on to minor events (mi), major events (Ma), die (D) or become lost-to-follow-up (L).

Figure 3-09-1: Patients in a long-term trial, in real time:



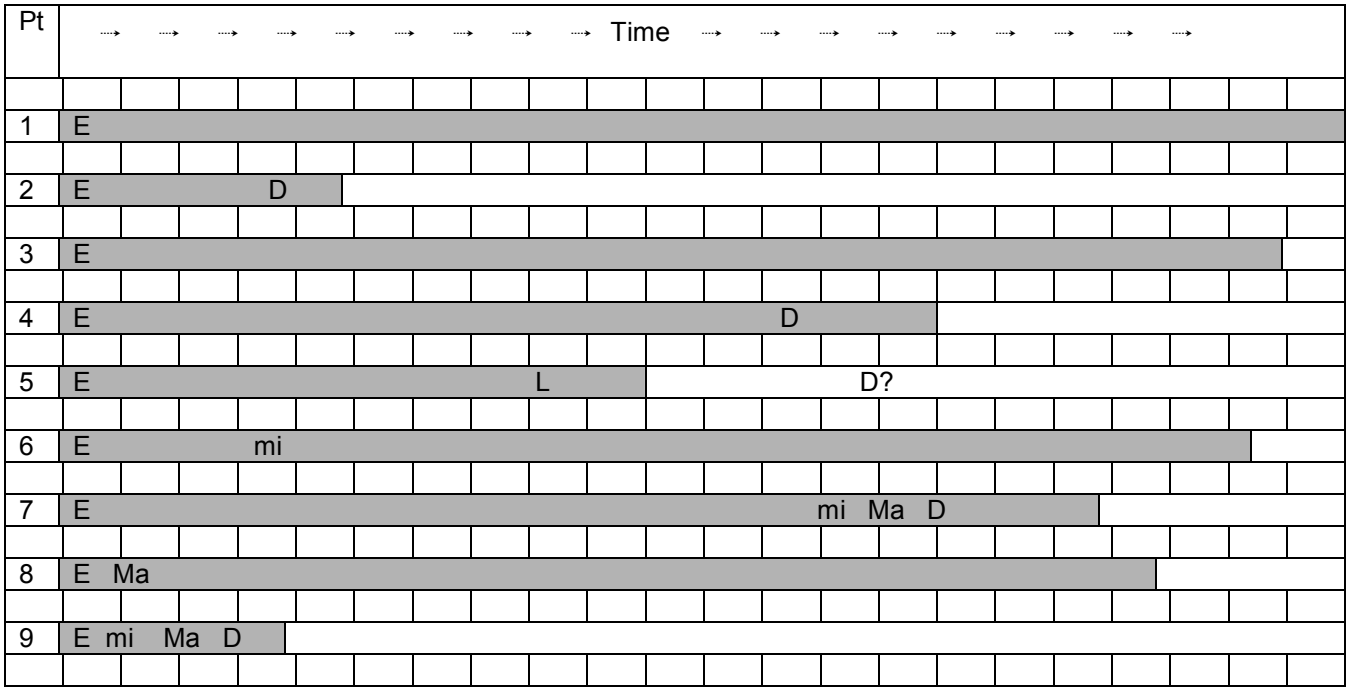
Legend: E=enters the study mi=minor event Ma=major event D=death L=Lost-to-Follow-Up

The first of these special features in long-term trials is that patients enter them throughout a recruitment period that can last for years, and finish these trials either at the variable time of their terminating event (such as death) or at a common stopping time at the close of the trial. As a result, individual patients are in the trial for widely different periods of time or "durations of follow-up." In Figure 3-09-1, Patient 1 survives the entire trial, but Patient 2 is dead even before the end of the recruitment period; Patients 1, 3, 6 and 8 are all present at the end of the study, but their follow-up times differ; Patients 1 and 7 sail through the trial event-free, but Patient 1 is followed for much longer.

The second special feature of long-term trials is that their treatment objectives include postponement of disabling or fatal outcomes (such as severe stroke or death) and not just their prevention (after all, if the trial went on for decades, everyone in it or working on it would eventually die). For example, both Patients 2 and 4 die, but Patient 4 lives event-free much longer. Both Patients 7 and 9 suffer minor, major, and fatal events, but Patient 9 has them in a cluster. Finally, Patient 5 is lost to follow-up, but still has more time in the study than Patients 2 or 9; moreover, a surveillance of the national death registry detects Patient 5's death some time later. Our analyses of such trials have to take these special features of variable length of follow-up and outcome-postponement-as-well-as-prevention into account.

The tactics of doing so begin with ignoring calendar time and thinking of patients as if they all entered the study at the same, common starting point, as shown in Table 3-09-2.

Figure 3-09-2: Patients in a long-term trial, taken back to a common starting point:



Legend: E=enters the study mi=minor event Ma=major event D=death L=Lost-to-Follow-Up

[Note to Editor: the symbols keep changing every time I "save" tables 3-09-1&2, so I'll need to be sure to check them when they are drafted. DLS]

The result, which is one variant of a "life-table," now more clearly reveals the differences in "survival" and follow-up, as well as the relative timing of events between patients in the trial. From such a table, we can calculate the proportion of experimental patients who are event-free at any given time after entry, and the resulting graph is called a "survival curve." One variant of a survival curve breaks follow-up time into small intervals, say days, and considers the probability that patients alive at the start of each day are likely to survive event-free to the start of the next day. As it happens, the probability of surviving event-free to day #120 is the "conditional" probability of surviving on day #120, given that you've already survived day #119. Survival curves generated in this fashion don't require assumptions about the nature of any theoretical "underlying distribution" of "true" probabilities we are guessing at in the trial, and are called "nonparametric" or "distribution-free." The one we've just generated is called a "Kaplan-Meier" survival curve after Edward Kaplan and Paul Meier, the statisticians who described this useful way of thinking about survival[9]. By this method we can calculate the probability of "surviving" event-free at any point in the trial or throughout it for experimental and control patients, and can generate a "noise" factor (say, the standard error) for each probability. The Kaplan-Meier curves for any major stroke or death in the surgical and medical arms of the high-grade stenosis NASCET trial are shown in Figure 3-09-3.

Figure 3-09-3. Kaplan-Meier Curves for major stroke or death in the surgical and medical arms of the high-grade stenosis NASCET trial.

{This should be taken from Figure 1-F on page 449 of: North American Symptomatic Carotid Endarterectomy Trial (NASCET) Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med 1991;325:445-53.}

The next step is to compare the survival curves generated for experimental and control patients. The method that my co-PI statistical colleagues routinely use compares the number of events we Observe in each treatment group during any interval (say, any given day) with the number of events we'd Expect to observe if there was no difference in efficacy between the experimental and control treatments. We then cumulate, over each interval of time, the (O – E) / E and the result is our old friend chi-squared, with (the number of intervals – 1) degrees of freedom. We call this form of analysis the "log-rank life-table" method, and credit for its elegant simplicity goes to Richard Peto. It tells us whether any difference in the events we observe between experimental and control patients is real; that is, whether that difference in events is "statistically significant."

A final point before we move on. The foregoing discussion is about events. However, it can be carried to a higher statistical plane by considering not just whether an event occurred, but the time elapsed between entry and that event. This "time-to-event" analysis is more powerful, but beyond the scope of this chapter. Similarly, many long-term trials include continuous outcome measures such as functional capacity and quality of life. They also present analytic difficulties, especially when patients die or otherwise stop contributing to these measures early in the trial. Analyses of continuous measures are taken up by Gordon Guyatt in Chapter 6.

I've already beaten to death the analysis of 1-sided superiority and noninferiority trials back on pages xx-xx.

5. Decide how to handle missing data in the analysis.

You need to decide, before the trial begins, how to handle patients like #5 in Table 3-09-1, who is lost to follow-up part way through the trial. It is tempting to treat them as if they were lost on the final day of the trial, and "censor" anything bad that might have happened afterward. This policy is risky for two reasons. First, if they left the trial because their target condition was deteriorating, ignoring their possible bad outcome would bias your conclusion. Second, ignoring them could lower the credibility of your conclusion.

Beyond the obvious solutions of not losing any study patients in the first place, and scouring mortality registers for them in the second, what can you do? I suggest that the most convincing way to handle them is a "worst-case scenario" in which you arbitrarily assign them the outcome that will make it hardest for you to answer your study question with a "yes." Suppose you are asking, "In patients of a particular sort, does experimental treatment E, when compared to control treatment C, reduce the risk of death?" In the "worst-case scenario," experimental patients who are lost get assigned the outcome of death, but control patients who are lost are assumed to have survived to the end of the trial. You then present the analysis in two parts. Part 1 "censors" lost patients from the moment they are lost, but in Part 2 they are returned in a "worst-case scenario." Second only to the trial that doesn't lose any patients at all, the trial with the highest credibility is the one in which the Part 2 worst-case scenario analysis reaches the same conclusion as the Part 1 censored analysis.

Many trialists believe that the "worst-case scenario" approach is too harsh. Various statistical "modeling" procedures have been proposed for assigning "appropriate" (but fictitious) outcomes to lost patients, and I leave it to you to discuss them with your statistical co-PI.

6. Decide how you will interpret your results to determine their "importance"

Chi-squared, t-tests, and the log-rank test are great at telling you whether your treatment effect is real (that is, unlikely to be due to chance). However, they can't tell you whether your treatment effect is great enough to be useful for patients. Recalling that trivial treatment effects become statistically significant when trials enroll huge numbers of patients, the next step is to determine

whether the statistically significant difference your trial generated also exceeds some "minimally important difference" (MID) that is deemed important by the trial participants (who, as the study "subjects," could comprise patients, providers, teachers, administrators, etc). This is a two-stage process; the first stage is mathematics, and the second stage is a judgment call.

The mathematics of minimally important differences (MIDs)

The mathematics are straightforward for a short-term parallel trial with events as outcomes, as you don't need to adjust for the relative times at which patients entered the trial or had events. In such trials, you simply determine the frequency of events in the control group (the Control Event Rate or CER) and in the experimental group (the Experimental Event Rate or EER). For example, in our ACE trial of high-dose (experimental) vs. low-dose (control) aspirin to prevent stroke, myocardial infarction or death in the month following carotid endarterectomy, the Experimental Event Rate (EER) among high-dose patients was 8.2% and the Control Event Rate (CER) among low-dose patients was 3.7% (P=0.002)[10].

In a long-term parallel trial with events as outcomes, the same principle holds but we derive the Control Event Rate (CER) and Experimental Event Rate (EER) as "failure" probabilities from the Kaplan-Meier curves I've already shown you. Kaplan-Meier Control Event Rates (CERs) and Experimental Event Rates (EERs) take into account both the fact that events are occurring throughout the trial and that their denominators are constantly changing as patients enter and leave the trial. As a result, they are larger than the Control Event Rates (CERs) and Experimental Event Rates (EERs) you'd calculate (incorrectly) at the end of the trial if you simply divided the numbers of events by the numbers of patients enrolled. In these long-term parallel trials, we generate the Kaplan-Meier Control Event Rate (CER) and Experimental Event Rate (EER) for some clinically sensible time (we selected 2 years after entry for the NASCET trial), and also generate their accompanying "noise" in the form of, say, a standard error. Thus, in NASCET, the Kaplan-Meier estimates of the Control Event Rate (CER) for major stroke or death at 2 years was 18.1% but the Experimental Event Rate (EER) was only 8%.

In trials with continuous outcomes, you have two choices. You can stick with the absolute differences between control and experimental groups that you used for determining statistical significance. Alternatively, you can convert these absolute differences into "events" by determining, for example, the rates at which control and experimental patients achieved some pre-set change in the continuous measure, or the rate at which they achieved a 50% reduction in a continuous measure of symptoms or disability. I refer you to Gordon Guyatt's section on continuous outcome measures for a complete discussion of the appropriate approaches.

The judgements that determine which differences are minimally important

In the second, judgment step we decide whether these differences are important. This chapter will provide one example each from the perspectives of clinicians and patients, respectively. Patient's perspectives on their minimally important differences will receive its major attention in the chapter on Outcomes (starting on page xx).

A clinician's MID: The example here will be the number of patients the clinician will need to treat in order to prevent one more bad outcome (NNT) or cause one more harmful adverse event (NNH). The results for all patients in an RCT generate a Control Event Rate (CER) and Experimental Event Rate (EER) for the average patient in the trial. The resulting Absolute Risk Reduction (ARR) and Number Needed to be Treated to prevent one more event (NNT) also apply to the average patient in the trial.

In the RRPCE trial, the Absolute Risk Reduction (ARR) for major stroke or death (CER – EER) among all patients treated with aspirin for 2 years was 7.2%, with its 95% Confidence Interval (CI)

running from 0.8% to 13.6%. Inverting these Absolute Risk Reductions (ARRs) we find an NNT of 14 with a confidence interval from 8 to 125.

These same methods apply to judging the importance of harm. The Absolute Risk Increase (ARI) and its reciprocal, the Number needed to be Treated to Harm one more of them (NNH) can be generated from the side-effect rates in the experimental and control patients. In the RRPCE trial, 2.1% of patients who took aspirin for 2 years had severed gastrointestinal bleeds (with a 95% confidence interval from 0.97% to 4.4%); none of these bleeds were fatal. Thus, the NNH for a severe bleeding episode was 48 with a confidence interval from 23 to 103.

At the time we reported these results, there were no other treatments that had been shown in RCTs to reduce the risk of stroke among patients with TIAs. It is therefore understandable why clinicians decided these results clearly exceeded their minimally important difference in stroke reduction, but did not exceed their minimally important difference for harm. Aspirin use soared.

But is aspirin for everyone? There are clinically relevant subgroups of patients in most trials, and they might have important differences in their control and experimental event rates (CERs and EERs). How might their risk and responsiveness be estimated by clinicians for extrapolation to similar groups of patients outside the trial?

Most treatments (e.g., most drugs) are designed to prevent or slow the progression of disease. The RRPCE trial is a typical example. And, there is growing empirical evidence that in trials of these "delaying" treatments, the Relative Risk Reduction (RRR) tends to be constant over a wide range of Control Event Rates (CERs)[11,12]. That being so, a case can be made for using the trial's overall Relative Risk Reduction (RRR} and applying it to groups of control patients at different baseline risks (CERs) to estimate their Absolute Risk Reductions (ARRs) and Numbers Needed to be Treated (NNTs) to prevent one more event. The authors of this book disagree with each other (a little, not a lot) about how much credence should be given to the subgroup Control Event Rates (CERs) in a trial. I am less willing to accept them than my coauthors, especially when they have large Confidence Intervals (Cis), and would wait for the meta-analysis (ideally, on individual patient data) of several similar trials before I'd trust them.

Anecdotal evidence suggests that Relative Risk Reductions (RRRs) might not be constant for treatments designed to reverse (as opposed to slow) the progress of disease. I provided one anecdote in an earlier paragraph on subgroup analysis, where I described the rising Relative Risk Reduction (RRR) from carotid endarterectomy as it was performed on patients with progressively more severe carotid stenosis. We don't know enough about the behavior of other Relative Risk Reductions (RRRs) for "reversing" treatments among different subgroups to offer any firm advice on their extrapolation, save to say that you may have to rely on subgroups within the trial.

A patient's MID: This same RRPCR trial can be used to provide patients with information that they can use to determine their own minimally important difference. Dr. Sharon Straus has pioneered a strategy for helping patients accomplish this[13]. Her method combines the benefits of therapy with its accompanying risks, and then adds the patient's own judgement about the relative severity of the bad outcome prevented by treatment (in this case, the stroke) and the bad outcome caused by treatment (in this case, the bleeding episode).

The key step here is the patient's decision about how much worse or better it would be to have a stroke than a bleeding episode. Suppose a patient was at "average" risk of both the stroke and the bleed. Suppose further that her health preferences and values were such that she considered having a stroke to be 4 times as bad as having a bleed. With an NNH to cause a bleed of 48, an NNT to prevent a stroke of 14, and her judgment about severity (S) that having a stroke would be 4 times as bad as having a bleed, we can calculate the likelihood that she would be helped vs. harmed on her own terms by taking aspirin. The formula for doing this is NNH x S / NNT. In her case, and incorporating hew own heath preferences and values, she is 48 x 4 / 14 or over 13 times as likely to be helped vs. harmed over the next 2 years if she starts taking aspirin.

For one final wrinkle in this example of creating a patient's MID, we needn't even assume that she is an "average" patient. Suppose that her risk of a stroke was only half that of the average patient in the RRPCE study, but that her risk of a bleed was twice that of the average trial patient. By applying Richard Cook's modification[14] in which these relative risks are expressed as decimal fraction and placed in the denominator of the corresponding NNT or NNH, her NNT to prevent a stroke rises from 14 to 14/0.5 or 28, and her NNH for suffering a bleed falls from 48 to 48/2 or 24. Even then, given that she considers a stroke to be 4 times as bad as a bleed, the likelihood that she'll be helped vs. harmed by taking aspirin for the next two years is 24 x 4 / 28 or over 3 to 1.

7. Establish interim analysis plans and statistical warning rules for efficacy, safety, and futility

Assume that you are conducting a trial with a 3-year follow-up, but are employing 95% confidence intervals (or $P < 0.05$) to its emerging "interim" results once a month. You are at great risk of inappropriately stopping the trial for no good reason. In the first place, if your pre-trial estimate of efficacy (say, an Absolute Risk Reduction of 5%) is accurate, then trials of it that stop early will be biased toward overestimating that efficacy, and you certainly don't want to do that. In the second place, trends in the early, unstable portions of trials can flirt with, or even cross, conventional boundaries of statistical significance for harm as well as benefit. Finally, in the eyes (or entrails) of many statisticians, the performance of multiple interim analyses increases the risk of an ultimately false-positive conclusion (that the experimental treatment works, when in fact it doesn't) at the end of the trial. How can you avoid these pitfalls and still stop the trial as soon as your results are both statistically and, more important, clinically convincing?

The first step is to withhold your first interim analysis until you have followed enough patients long enough for the real trends in safety and efficacy to become established. This is a judgment call, based on your patients' likely risk and the timing of their likely responsiveness (good and bad) to your experimental treatment. For example, you might want to perform your first interim analysis when 50% of your projected sample size should have been treated long enough to display the effects of experimental treatment.

The second step is to set the confidence intervals or P-values for your interim analyses at quite stringent levels, so that you minimize your risk of wrongly triggering your statistical warning rules. For example, in the HOPE trial we used the original Haybittle[15]-Peto[16]approach, and set the interim warning rule to trigger for benefit at two consecutive differences of 4 standard deviations (one-sided $P \leq 0.00003$) during the first half of the trial and at 3 standard deviations (one-sided $P \leq 0.002$) in the second half. As you can see, even with penalties for "multiple looks," we retained plenty of "P" for the final analysis. If these interim P values strike you as unattainable, reducing this "warning rule" exercise to mere window dressing, I'm happy to report that they were, in fact, triggered, the PI was unblinded, and the trial was stopped 8 months before its scheduled close.

We set a less rigorous warning trigger for safety at 3 standard deviations (one-sided $P \leq 0.002$) in the first half of the trial and 2 standard deviations ($P \leq 0.23$) in the second half, retaining our option of unblinding the investigators much sooner if we observed even a few severe unanticipated adverse events ("SUAEs").

In a similar fashion, you could set statistical limits for determining when a trial is simply not going to show any minimallyimportant benefit from experimental therapy. You can do this by specifying a very strict confidence interval around your treatment effect and watching to see whether it excludes, on the "ineffective" side, the minimally important benefit.

Statistical warning rules, not stopping rules

Note that I've called them statistical warning rules, not stopping rules. That's because trialists like me think that decisions to stop a trial should never be based on statistics alone. Accordingly, the third step is to decide what additional information you will use to interpret a statistical warning rule when it is triggered. Typically, this interpretation involves the clinical and biologic sense that might support or refute a decision to stop the trial. For example, in NASCET, we began monthly interim analyses 2 years into the trial, and a recommendation to stop the trial early required a demonstration of efficacy (an RRR of at least 10% for stroke or death) at 3 standard deviations in each of 6 clinically sensible subgroups every month for 6 months. Despite these stringent rules, we stopped the trial among patients with high-degree carotid stenosis early. More about how monitors and monitoring committees ought to work begins on page xxx.

Once you've designed your statistical warning rules, their review and acceptance by your collaborators and monitor(s) should be completed before you start the trial. You can start reading more about warning/stopping rules in Curtis Meinert's heavily referenced RCT text[17].

The final four items on the checklist come into play after you've completed your trial and are polishing off your analysis.

8. Don't exaggerate your conclusions.

If you've followed our advice so far, you can carry out a valid RCT. Congratulations! Don't blow it at the end by exaggerating your conclusions in ways that mislead your audience, leave you open to legitimate criticism, and damage your credibility. The 3 most common exaggerations we encounter are reporting only the "sexiest" efficacy measure, looking for the pony, and calling an indeterminate trial "negative."

Don't report only your sexiest measure of efficacy

Ironically, the first exaggeration of reporting only the "sexiest" efficacy measure (typically a Relative Risk Reduction (RRR) rather than an Absolute Risk Reduction (ARR) or Number Needed to be Treated to prevent one more event (NNT)) is increasingly common in cardiovascular trials, a field that is not only justifiably proud of its past accomplishments, but also a victim of its past successes. A steady progression of positive cardiovascular trials has validated an ever-expanding combination of effective treatments. These combinations thus became "established effective therapy." As a result, the question in today's cardiovascular trial is some form of: "Among patients with unstable angina, does the addition of drug X to established effective therapy achieve a further reduction in the risk of myocardial infarction or death?" In operational terms, this becomes a "placebo add-on" trial in which both groups receive established effective therapy. To do it, we add the promising new drug to the regimen of experimental patients, and add its placebo to the regimen of control patients. As you learned back in Section 3-2-I on physiological statistics, the established effective therapies pull the Control Event Rate (CER) toward zero, and even new drugs that generate large Relative Risk Reductions (RRRs) will generate only small Absolute Risk Reductions (ARRs) and large Numbers Needed to be Treated to prevent one additional event (NNTs). Thus, in reporting the effect of ramipril on the risk of stroke in the HOPE trial, the investigators confined themselves to the impressive Relative Risk Reduction (RRR) of 32% for all stroke and 61% for fatal stroke[18]. It was after several letters of protest[19] that they provided a table of Numbers Needed to be Treated (NNTs) for the entire HOPE trial, including an NNT of 111 to prevent a stroke and a very impressive NNT of only 8 to prevent one of the composite cardiovascular events[20].

This is not academic nit picking. There is increasing evidence that Relative Risk Redeuctions (RRRs) create higher opinions about efficacy among both physicians[21] and health policy makers[22] than their corresponding Absolute Risk Reductions (ARRs) or Numbers Needed to be Treated to

prevent one additional event (NNTs). Moreover, when Stuart Pocock reviewed 45 trials reported in the BMJ, the Lancet, and the New England Journal of Medicine, he concluded: "Overall, the reporting of clinical trials appears to be biased toward an exaggeration of treatment differences.[23]" I strongly support the CONSORT recommendation that trialists report "For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision.[24]" I take this to mean that, regardless of whether trialists focus on Relative or Absolute Risk Reductions (RRRs or ARRs), they must provide readers with the Control and Experimental Event Rates (CER and EER) that would permit readers to calculate the efficacy measure they find most informative. This is the editorial policy of our "evidence-based" journals. It is also appropriate, when reliable data are at hand, to report at least the Numbers Needed to be Treated to prevent one additional event (NNTs) for clinically identifiable low, medium, and high-risk subgroups.

Don't go looking for the pony

As you've already read, subgroup analysis is a two-edged sword. In the design phase of the RRPCE trial, I went looking for the pony and pushed for subgroup analyses of efficacy based on the nature and location of the qualifying TIAs, by age and sex, and by several comorbid conditions. This led to more than a dozen subgroup analyses, with some of them further divided by sex. I therefore must take the blame for our statement, based on one of these subgroup analyses, that "Aspirin was of no benefit in reducing stroke or death among women." We didn't base this erroneous statement merely on an effect that was statistically significant effect in men but not in women. Women were 42% more likely to suffer stroke or death on aspirin, and we showed a statistically significant ($P<0.003$) difference in the relative risk reductions between the sexes. Despite this extremely positive subgroup analysis the first time anyone tested aspirin for TIAs, later trials proved our conclusion was wrong.

Looking for the pony has caused countless trialists to emerge from subgroup analyses appropriately besmirched, even when (as in the RRPCE trial) they demonstrate a statistically significant qualitative interaction. My advice to fellow trialists: never draw a conclusion (especially in print) about efficacy from any subgroup analysis that produces an unanticipated qualitative interaction in which a treatment that is effective in one subgroup of patients is either harmful or powerfully useless in another. I suggest that the only appropriate response to such a finding is replication (in an independent study), not publication. If you're not convinced by this tough stance, revisit our aspirin vs. perioperative stroke and death experience described in Check list item 2-vi on forming hypotheses for your next trial.

When that admonition is ignored, Andrew Oxman and Gordon Guyatt have warned the readers of RCT reports not to accept conclusions based on subgroup analyses unless they are big, highly statistically significant, specified prior to analysis, replicated in other independent trials, and supported by other evidence[25].

Never, ever label an indeterminate trial "negative" or as showing "no difference"

The third exaggeration is reporting an "indeterminate" trial as "negative;" that is, reporting that an intervention has "no effect" just because the 95% Confidence Intervals (Cis) for its Relative Risk Reduction (RRR) and Absolute Risk Reduction (ARR) cross 1 and 0, respectively. I've already described this problem in Figure 3-09-3, but it deserves repeating. This problem is as old as RCTs, and 25 years ago Jenny Freiman, Tom Chalmers, Harry Smith, and RR Kuebler examined 71 "negative" trials and found that 94% of them had a greater than 10% risk (power less than 0.9) of missing an RRR of 25% (the sort of effect observed among many efficacious treatments)[26]. Alas, 16 years later David Moher and his colleagues documented that this problem had not gone away[27]. As described above, whether in a planned "debunking" trial of a treatment thought to be

useless, or as an unexpected result of a superiority trial, the issue is not the (nonsignificant) difference that you found, but the difference, of significance to patients, that you can rule-out.

Allan Detsky and I have suggested that there are two appropriate ways to evaluate apparently "negative" trials[28]. Both of them reject a priori sample size requirements and focus on results ("how many patients you needed depends on what you found"). First, we suggest that you simply generate a Confidence Interval around the effect that you did observe and see whether it excludes any minimally important effect (as in Column 1 of Figure 3-09-3). Second, we suggest an alternative in which you test your observed difference against the effect you hypothesized before the trial. Even if you rule out any minimallyimportant effect, I still advise against labeling your result "negative" because you may not have ruled out the effect somebody else considers worthwhile. For this reason, I have expanded Iain Chalmers' earlier proposal to ban the term "negative trial"[29] and have taken the position that the word "negative" should be banned in describing any result from any study[30]. I suggest that far more accurate and useful words are "inconclusive" or, if that is too bitter a pill to swallow in print, "indeterminate."

9. <u>Report your results, regardless of their interpretation</u>

Not to report trial results, regardless of what they show (or, especially, fail to show), is both bad science and bad ethics. Every trial result should be included in systematic reviews of that intervention, and avoiding publication bias (especially of trials with indeterminate results) is vital to their validity. Moreover, it is unethical to expose study patients to an RCT environment, with the expectation that they are contributing to medical science, and then suppress their outcomes.

In 1997, in recognition of the resistance to submitting and publishing "negative" trials, over 100 editors of medical journals declared an "amnesty for unpublished trials" and provided a free registration service for unpublished trials[31]. If you or the journals decide that your trial isn't fit to print, be sure to register it with the Medical Editors. I think this registration should extend even to fraudulent or "busted" trials in which adherence to the protocol was so awful that they had to be abandoned. Another means of reporting such trials is through "open access" internet-based publishers such as Biomed Central (http://www.biomedcentral.com).

10. <u>Update the systematic review that justified your trial:</u>

Your work won't be done until you incorporate your trial results (plus any other contemporaneous ones) in an updated systematic review of your intervention. Has your trial provided the necessary confirmation of efficacy or a useful narrowing of the confidence interval around the estimate of effectiveness? Besides contributing to the science of health care, there are three more personal benefits of updating the systematic review. First, if your trial results were indeterminate, they will nonetheless be incorporated into a systematic review, meeting your scientific and ethical obligations. Second, the updated systematic review contributes another publication to your CV. Finally, it tells you where to go next. Alas, as I wrote this chapter, Michael Clarke, Philip Alderson and Iain Chalmers reported that only 3 of 33 RCT reports published in May 2001 in the 5 major general medical journals even referred to relevant systematic reviews, and none of them presented any systematic attempt to "set the new results in the context of previous trials.[32]"

11. <u>Formulate the logical question for your next trial</u>:

With the possible exception of "debunking" trials that expose the uselessness of established treatments, the interpretation of an RCT result ought to lead to the formulation of the next logical question you should ask in your next RCT. Might an effective treatment for your study patients also benefit patients with a different but related disorder? Is the unexpected response of some subgroup of patients that you discovered while data-dredging of such potential importance that

you should test it in your next trial? Might a simpler, cheaper, or more easily tolerated regimen be non-inferior to the one you've just validated as efficacious? In addition to formulating the next logical question, you should incorporate everything you learned in designing, conducting and analyzing this trial into the design, conduct, and analysis of your next trial.

REFERENCES

[1] Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Br J Cancer 1976;34:585-612 & 1977;35:1-39.

[2] Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991. P 461.

[3] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in Medicine 2002;21:2917-30.

[4] Haynes RB, Taylor DW, Sackett DL, Thorpe K, Ferguson GG, Barnett HJ. Prevention of functional impairment by endarterectomy for symptomatic high-grade carotid stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. JAMA 1994;271:1256-9

[5] O'Brien JR, Etherington MD. How much aspirin? Thromb Haemost 1990;64:486.

[6] Taylor DW, Barnett HJM, Haynes RB, Ferguson GG, Sackett DL, Thorpe KE, Simard D, Silver FL, Hachinski V, Clagett GP, Barnes R, Spence JD for the ASA and Carotid Endarterectomy (ACE) Trial Collaborators. Low-dose and high-dose acetylsalicylic acid for patients undergoing carotid endarterectomy: a randomised controlled trial. Lancet 1999;353:2179-84.

[7] Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology. 2nd Edition. Boston: Little, Brown, 1991, page 13.

[8] Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991. (ISBN 0-412-27630-5).

[9] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. JASA 1958;53:457-81.

[10] Taylor DW, Barnett HJ, Haynes RB, Ferguson GG, Sackett DL, Thorpe KE, Simard D, Silver FL, Hachinski V, Clagett GP, Barnes R, Spence JD. Low-dose and high-dose acetylsalicylic acid for patients undergoing carotid endarterectomy: a randomised controlled trial. Lancet. 1999;353:2179-84.

[11] 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.

[12] Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol 2002;31:72-6

[13] 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.

[14] Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ. 1995;310:452-4.

[15] Haybittle, JL. Repeated assessment of results in clinical trials of cancer treatment. British Journal of Radiology 1971; 44: 793-7.

[16] Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observations of each patient. I. Introduction and design. British Journal of Cancer 1976; 34: 585-612.

[17] Meinert CL. Clinical Trials, Design, Conduct, and Analysis. New York: Oxford University Press, 1986. Pp 215-6.

[18] Bosch J, Yusuf S, Pogue J, Sleight P, Lonn E, Rangoonwala B, Davies R, Ostergren J, Probstfield on behalf of the HOPE Investigators. Use of ramipril in preventing stroke: double blind randomised trial. BMJ 2002;324:1-5.

[19] Letters. BMJ 2002;325:439

[20] Yusuf S, Bosch J, Sleight P. Responding to issues raised. BMJ Electronic Letter 30 October 2002.

[21] Bucher HC, Weinbacher M, Gyr K. Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. *BMJ* 1994;309:761-4.

[22] Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding results of clinical trials and systematic reviews. *BMJ* 1995;311:1056-9

[23] Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 1987;317:426-32.

[24] http://www.consort-statement.org/

[25] Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78-84.

[26] Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med 1978;299:690-4.

[27] Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-4.

[28] 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.

[29] Chalmers I. Proposal to outlaw the term 'negative trial'. Br Med J 1985;290:1002.

[30] http://bmj.com/cgi/eletters/325/7373/0/g#27016

[31] Medical Editors Trial Amnesty. Fax: 0171-383-6418. BMJ, BMA House, Tavistock Square, London WC1H 9JR., UK or send by e-mail to:Meta@ucl.ac.uk

[32] Clarke M, Alderson P, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals. JAMA 2002;287:2799-801.