Anesthesiology
80:923–928, 1994

J. B. Lippincott Company, Philadelphia

Clinical Research

A Simple Recipe for Doing It Well

*Warren S. Browner, M.D., M.P.H.**

SUPPOSE a physician observes that five of the last six patients seen with postoperative myocardial infarction were taking β-blocking drugs and wants to determine whether this is a "cause–effect" association (*i.e.*, β-blocking drugs cause postoperative myocardial infarction). How can a study of this topic be designed and analyzed?

Answering this question requires an understanding of both the anatomy and the physiology of clinical research.[1] The anatomy of research includes the different types of study designs, sampling, and measurements. The physiology is how these ingredients work together to produce a coherent whole: the study result. In this essay, I provide a basic introduction to the anatomy and physiology of clinical research, emphasizing study designs like the randomized trial and statistical issues like confidence intervals that may be of special interest to the readers of, and aspiring writers for, ANESTHESIOLOGY.

Study Designs

The Randomized Trial

The most rigorous study design is the randomized blinded placebo-controlled trial (RCT).[2,3] In an RCT, subjects are randomly assigned to receive either a treatment (*e.g.*, β-blocking drugs) or a placebo control, and the outcome (myocardial infarction) is subsequently assessed by investigators blinded to the group to which a subject was assigned.

Why has the RCT become the standard for excellence in clinical research? We must first look at the goal of most clinical research, which is to show that an association between two variables is due to cause–effect: that a certain *predictor* (*e.g.*, use of β-blocking drugs) causes a given *outcome* (myocardial infarction). Showing that an association is cause–effect involves choosing a research design and using an analytic technique that make less likely the four other explanations for an association: effect–cause, effect–effect, bias, and chance. "Effect–cause" involves a reversal of the temporal sequence: patients with myocardial infarctions may complain of chest pain and be treated with β-blocking drugs. "Effect–effect," or confounding, occurs when the treatment and the outcome are causally related to a third condition, called the *confounder*. Use of β-blocking drugs and postoperative myocardial infarctions, for example, are both more common in patients with coronary artery disease. "Bias" can be generously thought of as an inadvertent mistake; only a perfectly designed, executed, and analyzed study is entirely free of bias. One common mistake—detection bias—occurs if the investigator looks harder for a disease in one group than another; for example, if patients treated with β-blocking drugs postoperatively are more likely to be in intensive care units, in which infarctions are more readily detected. Finally, an association may have occurred by "chance"; investigators can use statistical tests to estimate how likely it is that chance explains a study's findings.

RCTs, if designed and executed properly, can eliminate the possibilities of effect–cause, effect–effect, and bias. That is why they are so important in clinical research. In an RCT, patients are assigned to receive the treatment or the control before the outcome has occurred, thereby eliminating effect–cause. Randomization should also eliminate effect–effect: assigning sub-

* Associate Professor and Chief, General Internal Medicine.

Received from the General Internal Medicine Section, Department of Veterans Affairs Medical Center, and the Departments of Medicine, Epidemiology and Biostatistics, and Anesthesia, University of California, San Francisco, California. Accepted for publication December 1, 1993.

Address reprint requests to Dr. Browner: General Internal Medicine Section 111A1, Department of Veterans Affairs Medical Center, San Francisco, California 94121.

Key words: Case-control studies. Cohort studies. Randomized controlled trials. Research methods. Sampling.

jects randomly to the groups ensures that the only variable is the treatment. The distribution of all the other characteristics of the subjects, including possible confounders like underlying coronary artery disease, should be similar in the two groups. A comparison of the characteristics of patients assigned to the treatment and control groups, usually the first table in the results section, indicates whether or not randomization was successful. "Blinding" means that the investigator cannot tell to which group a subject has been assigned; it is especially important when determining whether a subject had an adverse outcome or not. In drug trials blinding involves using a "placebo control" that has an appearance identical to that of the active drug. Placebo controls also eliminate the possibility of a placebo effect, in which a patient responds nonspecifically to any intervention. If blinding and randomization are successful, they eliminate bias. Thus all that is left to explain an effect of the treatment, if there is one, are chance (which can be quantified with a *P* value) and cause–effect. If the *P* value is small, then cause–effect becomes the most likely explanation.

The actual process of randomization is simple. First, each enrolled patient is given a study number; this ensures that everyone who was enrolled in the study is forever after considered a subject in the study. Next, each subject is randomly assigned to either the intervention or control groups. To do so, the investigator can generate a random number with a computer program; if that number is odd, the subject is assigned to the control group. If a preprinted list of random numbers or assignments is used, two persons should be involved—one who is enrolling subjects, and the other who keeps the list. This assures that the subject's group assignment is not known at the time of enrollment. Alternatively, one can have a pile of numbered, sealed (and opaque!) envelopes containing group assignments.

Unless blinding is perfect, pseudorandomization techniques, such as assigning a subject to treatment or control by odd or even Social Security numbers, should be avoided. The problem here—as well as with similar schemes such as randomizing by day of the week—is that the investigator knows in advance the group to which a particular subject will be assigned.[4] This provides the opportunity to reject a sick patient who has, for example, an even Social Security number and who would have been in the intervention group. There can also be a problem if subjects are randomized in pairs (one each to intervention and control) and it is not possible to maintain blinding—the investigator then knows that if the first patient in the pair was in the intervention group, the next patient will be in the control group.

Not all randomizations need to be 50:50, with approximately equal numbers of subjects in each group. For some studies, there may be more than one intervention that is of interest, say, a large dose of a medication *versus* a small dose *versus* a placebo. Occasionally, there may be reasons to randomize more patients to the intervention group, such as if one can enroll a limited number of patients, and wants to be sure to have enough patients in the intervention group to be able to rule out rare side effects. However, the investigator must recognize that these alternatives to 50:50 randomization may make it more difficult to find a difference between the intervention and control groups. For any given total number of subjects, power (see below) is greatest when comparing two groups of equal size.

Sometimes a patient is assigned to receive a therapy but does not, for example, because a contraindication develops after the randomization process. RCTs should follow the rule "once randomized, always analyzed": a subject is always considered a member of the original randomization group. This rule means that RCTs are comparing assignment to a particular therapy, rather than to the therapy itself. The alternative—to analyze just those subjects who actually receive the treatment or the control—introduces a potential bias, because subjects who are lost to follow-up or who refuse treatment are likely to be different in important ways from the other subjects. Because of this rule, before a subject is randomized, the investigator should be absolutely certain that the patient is eligible, has given informed consent, and can be followed for the length of the study.

RCTs are not perfect. Sometimes it is impossible to blind, as in a trial of surgical *versus* medical therapy for angina. Even placebo-controlled trials can be difficult to blind; for example, it is easy to distinguish subjects who take β-blocking drugs from those given placebo. In these situations, the investigator must decide whether failure to maintain blinding necessarily invalidates the results of the study. That usually depends upon the outcome: if one is comparing the effect of two treatments on a endpoint like death, presumably the inability to blind the investigators will not bias the ascertainment of results. But other endpoints—like the specific cause of a death, or the diagnosis of bronchospasm—may be influenced by knowledge of a patient's treatment assignment.

RCTs are also subject to "co-intervention effects," in which an unintended effect of the intervention is responsible for the outcome. For example, β-blocking drugs may appear to cause an increase in myocardial infarctions that on further examination is found to be due to a co-intervention: patients who received β-blocking drugs were more likely to develop bradycardia, and therapy for the bradycardia (atropine and sympathomimetic agents) may have been responsible for the increase in infarction risk.

There is another problem with RCTs, especially if the total number of subjects is small. By chance alone, there may be maldistribution of an important characteristic: for example, all ten patients with severe underlying coronary artery disease may have been assigned to the control group. If the investigator knows that there is a particularly important confounder, it may be wise to use a stratified randomization, having separate randomization schemes for each of the two strata (those with and those without severe coronary disease). This ensures that there will be roughly equal numbers of those patients in both the intervention group and the control group.

RCTs bring up an important ethical dilemma. At the start of the study, the investigator presumably does not know which therapy is better: that is the purpose of the study.[5] But what if it turns out—based on early data from one's own study or from another study—that one of the therapies appears beneficial or harmful? Elaborate procedures, known as "stopping rules," have been developed to establish guidelines for stopping a study before the anticipated end of enrollment. These guidelines are intended to facilitate decision-making by a group of independent investigators who have access to the unblinded data.[2]

Other Designs

RCTs are not the only way to do clinical research. RCTs are inappropriate for many questions, and impractical for others. For example, one cannot randomize patients to smoke or not smoke before surgery. When the outcome of interest is rare, such as intraoperative myocardial infarction, an RCT to determine whether the risk is greater with one anesthetic agent than another might require enrolling every patient who undergoes surgery at ten medical centers for the next 5 yr. These research questions require other sorts of designs.

In a "cohort study," the investigator does not intervene upon the subjects, but simply observes their natural history through time, as in studies of cardiac complications in patients undergoing noncardiac surgery.[6–8] Subjects are enrolled, predictor variables are measured, and then the occurrence of outcomes is ascertained. Cohort studies have several advantages: for one, they can answer a series of research questions, because the investigator need not be concerned that a new intervention has altered the natural history of the condition being studied. For another, the limitations on patient eligibility that plague many RCTs can usually be avoided, because the investigator is observing, not intervening. Finally, they can be done retrospectively, by doing a thorough medical records review. A main disadvantage of cohort studies is that, unlike RCTs, one cannot assume that the groups being compared (*e.g.*, those who use β-blocking agents, and those who do not) are alike in other ways. To the extent that potential confounders, such as severity of coronary disease, have been measured, they can be adjusted for in the analysis of the data. However, there are always unmeasured, and often inadequately measured, confounders; these are immune to statistical techniques for adjustment.[9]

In a "case–control study," the investigator begins with patients who had the outcome of interest. For example, one would identify the last 100 patients who had suffered a perioperative myocardial infarction. These are the "cases." Next, a "control" group of 100 patients who did not have a myocardial infarction is assembled. The investigator then determines whether more cases than controls were using β-blocking drugs. Case–control studies may be the best option when the outcome of interest is rare (such as intraoperative death); otherwise, one might have to enroll thousands of subjects in order to have enough outcomes. However, case–control studies are subject to many biases. Representativeness of the two groups is usually the key issue—for example, were the cases and controls in the study a fair sample of all the patients, and were they comparable in other respects aside from use of β-blocking drugs?

Samples

In addition to establishing that cause–effect is more likely than alternative explanations, the investigator has another task: to convince the reader that the study results will also apply to other patients. This brings us to the concept of a sample. The best sample is one that resembles the "target population," the group of patients to which the investigator wishes the results to apply. In our example, the target population is patients at risk of having perioperative myocardial infarction. This target population, of course, includes everyone who undergoes surgery, many of whom are at such a low risk of perioperative infarction that most investigators would wisely consider them not worth studying. Thus the investigator often restricts the target population to a higher risk group, such as patients with one or more cardiovascular risk factors undergoing surgery. The actual sample is chosen from a narrower but usually more convenient group of subjects, established according to precisely specified inclusion and exclusion criteria, such as patients at the investigator's medical center who are undergoing elective vascular surgery and who volunteer to participate.

Each difference between the actual sample and the target population reduces the ability to generalize the results. Much of what we know about perioperative ischemia, for example, has come from studying volunteer patients at a few academic medical centers[6–8]; the assumption is that those results will apply universally. All samples also contain an implicit chronologic assumption, that results generalize from the past and present into the future.

The investigator should make explicit the target population to which the results can be generalized; this is usually a judgment call. For example, if patients who had surgery on Tuesdays were excluded for logistic reasons, the investigator is probably justified in suggesting that the results would apply equally to that day. On the other hand, results in men do not automatically generalize to women, and results in academic medical centers may not pertain in the real world.

Measurements

No matter what the study design, it is essential that variables be specifically defined in an operations manual or study protocol before the study begins. This applies to all the important variables that are to be measured, whether the variable is a predictor (use of β-blocking drugs), an outcome (myocardial infarction), or a confounder (underlying coronary disease). How would a patient who received a single dose of a short-acting β-blocking drug the day before surgery be classified? Did a patient with a small increase in creatine kinase isoenzymes but no electrocardiographic changes have a myocardial infarction? Does an asymptomatic patient with a Q wave in an inferior lead have coronary artery disease? To the extent that definitions are indistinct, or are changed after the study is underway or completed, patients can be innocently moved from one category to the next, a form of bias that can render meaningless the results of the research project. Just as important, the measurement of the predictors should not depend upon the outcome variables, and vice-versa. This can be done by prospectively measuring the predictors before the outcomes have occurred, and by blinding the ascertainment of outcome.

The outcome should have "face validity": everyone should recognize it, and know what it signifies. The best outcomes, such as myocardial infarction or death, are those that are universally acknowledged as being important. Face validity, of course, varies from study to study. A physician examining whether a new anesthetic technique affects intraoperative renal blood flow may need to measure a few variables in a small number of subjects. However, it will not be possible to determine whether the new technique affects the incidence of postoperative renal failure.

The Research Hypothesis and Statistical Issues

The sample, the design, the predictor variable, and the outcome variable are the components of the "research hypothesis," which is a clear, simple, advance statement of what the investigator hopes to find: "We hypothesize that patients undergoing vascular surgery under general anesthesia at our hospital who are randomly assigned to receive prophylactic β-blocking drugs will have a lesser rate of postoperative myocardial infarction than those assigned to receive placebo."

The research hypothesis focuses the research as it is designed, executed, and analyzed. Put simply, the purpose of a research study is to determine whether the research hypothesis is true in the investigator's sample. Statistical tests help the reader make a judgment as to whether an effect that was found in a sample could have been due to chance (a "type I error"), or if no difference was found, how likely it is that an important effect could have been missed (a "type II error"). Before doing a study, the investigator determines how big an effect is anticipated, and what levels of error are acceptable. The smaller the effect the investigator wishes to detect, or the lower the levels of type I and type II errors, the larger the required sample size will be. By tradition, the maximum likelihood of a type I error (α) is usually set at 0.05. The maximum likelihood of a type II error (β) is set at 0.10 or 0.20, corresponding to a "power" ($1 - \beta$) of 90% or 80%.

How does one go about deciding how many subjects to study? With too many subjects, a study will be unnecessarily long and expensive. Without enough, the results will not be convincing, and they may not be statistically significant when there is a substantial difference between the groups being compared. Studies that are too large or too small waste the investigator's time and resources. They also impinge upon our ethical obligation to our subjects, who have agreed to participate in research in the belief that they are advancing science.

A general rule is that the sample size must be enormous if the outcome is rare (intraoperative infarction), moderate in size when the outcome is more common (intraoperative hypotension), and smallest when the outcome is a characteristic that can be measured on a continuous scale (intraoperative blood pressure). Investigators seeking more specific advice on sample size may wish to look at one of the available articles or texts on the subject.[1,2,10,11] Consultation with a statistically oriented colleague may be necessary; such advice will be most useful if the investigator has prepared a clear research hypothesis, and has some information about the incidence of the outcome of interest.

It is often tempting to do a small study, just to try out an idea. Small studies have a way of failing to be accepted for publication when they are negative. This becomes a problem when colleagues review the literature on a particular topic to do a metaanalysis: the literature is biased toward positive studies.[12]

Presenting the Results

Just as the study is designed around the research hypothesis, the manuscript should be written to provide the answer to that hypothesis, with whatever supporting information is needed. There should be a clear distinction between "data," which the investigator analyzes, and "information," the synthesized results of that analysis presented in a simple manner.[13,14] Statistical tests should be viewed as a means not an end, and should therefore illuminate, rather than obfuscate, the results. If the investigator does not understand how the statistical tests changed the data into meaningful information, there is little hope of explaining it to the reader.

Many investigators believe that purpose of statistical analysis is to find a few significant P values so that a manuscript will be publishable. But the P value simply indicates the likelihood of observing the study results (or ones more extreme) in the sample if there is really no difference in the population; it tells us nothing about the size of the difference. It is much better to emphasize the *effect* that was detected, using statistics in a supporting role.

The long-standing emphasis on P values is being replaced by reporting "confidence intervals."[15] Confidence intervals estimate the precision of the results, and are useful for both positive and negative studies. For example, two studies might find a difference of 10% in the risk of perioperative myocardial ischemia between those not receiving and those receiving β-blocking drugs. The first study (with a small sample size) has a 95% confidence interval from −20% to +40%. Its results are consistent with the possibility that use of β-blocking agents is associated with a substantial increase, or a substantial decrease, in the risk of perioperative ischemia. The second study, with a larger sample size, and a much narrower confidence interval of 8% to 12%, provides a much more precise estimate of the true effect of β-blocking drugs on ischemia.

Confidence intervals have a one-to-one correspondence with P values: 95% confidence intervals that include zero, as in the first example above, imply that the P value is greater than 0.05; how much greater depends upon how close zero is to an edge of the interval. Similarly, 95% confidence intervals that exclude zero are synonymous with a P value less than 0.05, as in the second example; how much less depends upon how far away zero is from the edges. But confidence intervals have a big advantage over P values. Because they indicate the range of values that are consistent with the results, they are especially useful in negative (nonsignificant) studies. If the confidence interval for a negative study includes a clinically important effect (in the first example above, a 40% decrease in ischemia risk), then the study has provided almost no information one way or the other.

It may be tempting to dredge through the data, looking at many variables in the hope that if one combination of predictors and outcome does not work out, another will. Having succumbed, the investigator might seek evidence that β-blocking drugs reduce the risk of cardiac death, myocardial infarction, stroke, arrhythmias, ischemia, renal failure, or length of hospital stay. In clinical trials, a proliferation of outcome variables spells disaster. The more associations that are looked at, the greater the likelihood of finding something by chance alone. *Post hoc* analyses that find an effect in just one subgroup, for example, that β-blocking drugs affect myocardial infarction among those with baseline renal insufficiency, have the same problem. In both situations, a main advantage of the RCT—that the *P* value quantifies the likelihood of observing the study results by chance—is lost.

Analyses of multiple outcomes and subgroups must be evaluated in the light of the small prior probability of the tested hypotheses.[16] If at the start of the trial the investigator thought that a drug would work in just one subgroup or for just one endpoint, why bother studying anyone else or any other endpoints? The answer, of course, is that the investigator did not, but like the rest of us, is remarkably good at developing plausible explanations for almost any finding. A sensible policy is to require much stricter statistical criteria (smaller *P* values) for surprising results, and to recognize the absolute need for a confirmatory study.

Understanding the basic rules of research and the most common mistakes provides a framework for recognizing good research. But like a patient who does not present with classic signs and symptoms, research does not always follow the rules we have described. No study is without flaws. A minor bias may be inevitable, or a sophisticated statistical technique may be the only way to analyze the data. The investigator's overall task is to convince the reader that despite these problems, the study is still valid. In turn, the reader's most important job is to ask whether the study is well enough done to affect the way one understands and practices medicine.

References

1. Hulley SB, Cummings SR (editors): Designing Clinical Research: An Epidemiologic Approach. Baltimore, Williams & Wilkins, 1988

2. Friedman LM, Furberg CD, DeMets DL: Fundamentals of Clinical Trials. 2nd edition. Littleton, PSG Publishing, 1985

3. Kramer MS, Shapiro SH: Scientific challenges in the application of randomized trials. JAMA 252:2739–2745, 1984

4. Chalmers TC, Celan P, Sacks HS, Smith H Jr: Bias in treatment assignment in controlled clinical trials. N Engl J Med 309:1358–1361, 1983

5. Schafer A: The ethics of the randomized clinical trial. N Engl J Med 307:719–724, 1982

6. Slogoff S, Keats AS: Does perioperative myocardial ischemia lead to postoperative myocardial infarction? ANESTHESIOLOGY 62:107–114, 1985

7. Hollenberg MH, Mangano DT, Browner WS, London MJ, Tubau JF, Tateo IM for the SPI Research Group: Predictors of postoperative myocardial ischemia among men undergoing non-cardiac surgery. JAMA 268:205–209, 1992

8. Rhaby K, Barry J, Creager MA, Cook EF, Weisberg MC, Goldman L: Detection and significance of intraoperative and postoperative myocardial ischemia in peripheral vascular surgery. JAMA 268:222–227, 1992

9. Datta M: You cannot exclude the explanation that you have not considered. Lancet 342:345–347, 1993

10. Lachin JM: Introduction to sample size determination and power analyses for clinical trials. Controlled Clin Trials 2:93–113, 1981

11. Zar JH: Biostatistical Analysis. 2nd ed. Englewood Cliffs, Prentice-Hall, 1984

12. Thacker SB: Meta-analysis: A quantitative approach to research integration. JAMA 259:1685–1689, 1988

13. Wainer H: How to display data badly. Am Statistician 38:137–147, 1984

14. Jolley D: The glitter of the t table. Lancet 342:27–29, 1993

15. Braitman LE: Confidence intervals extract clinically useful information from data. Ann Intern Med 108:296–298, 1988

16. Browner WS, Newman TB: Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA 257:2459–2463, 1987

SAMPLE SIZE CALCULATION

$\alpha = 0.05$ $\beta = 0.20$ Two-tailed Test

Effect Size (%)	Placebo Rate (%)	Sample Size Per Group
50	30	133
	25	168
	23	185
	20	219
	18	247
	15	304
	12	389
	10	474
40	30	214
	25	270
	23	299
	20	353
	18	399
	15	492
	12	631
	10	770
30	30	389
	25	492
	23	546
	20	647
	18	733
	15	905
	12	1163
	10	1422
20	30	891
	25	1134
	23	1260
	20	1497
	18	1699
	15	2102
	12	2707
	10	3313