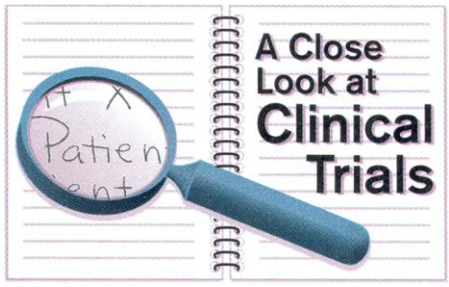
RESEARCH METHODS

Problems in Clinical Trials Go Far Beyond Misconduct



Last June, AIDS researcher Margaret Fischl of the University of Miami got a rude reception when she presented the results of a large trial of anti-HIV drugs at the ninth international AIDS conference in Berlin: AIDS activists greeted her talk with jeers and catcalls. The statisticians in the conference hall were more polite, but some of them were probably equally upset by what they heard.

Overall, the Fischl trial found that a combination of two anti-HIV drugs—AZT and ddC—offered no greater benefit than AZT alone. But Fischl and her colleagues chose to emphasize that the combination of drugs appeared to benefit a small number of patients, whom the researchers had identified by comparing many different subgroups of patients (such as patients at different stages of the disease). To activists and statisticians alike, this smacked of trickery: statistical arm-wrestling to force the data to yield a positive result.

The Fischl team may be the only one to be publicly taunted for this practice, but it's far from alone in departing from what's considered good clinical trial practice. Although results of randomized controlled clinical trials (RCTs) have transformed medical practices—and remain the gold standard for testing new treatments, drugs, and devices—many clinical trials are substandard, critics say, largely because few researchers are trained in the basics of clinical trial design and execution (see box).

"Most people are expected to pick it up on the job," says Christopher Williams, a medical oncologist at the University of Southampton in England and chair of the cancer therapy committee of the Medical Research Council, "but the people they work with aren't well trained either, so they pick up bad habits." Biostatistician Thomas Fleming of the University of Washington, Seattle, who sits on numerous Food and Drug Administration (FDA) advisory committees and data monitoring committees for clinical trials sponsored by industry and the National Institutes of Health, says the majority of RCTs "have flaws in their design that [if not caught] can affect the integrity and reliability of the trial." As a result, even in the absence of anything that would usually be called misconduct, many clinical trials turn up ambiguous, contradictory, or misleading results (see table).

A *Science* survey of experts on clinical trials found that the key problems are failure to guarantee randomization (which ensures that patients are assigned to a treatment by chance); enrolling too few participants to detect treatment differences; inappropriate analysis of trial subgroups, as in Fischl's case; post-hoc removal of data from the final analysis; and misleading substitution of "surrogate" biological markers for clinical endpoints such as improved health or survival.

Clinical Trials 101

Some of the most common flaws are also the most basic—such as failing to ensure randomization. About one third of RCTs pub-

Ignorance Is Not Bliss

What is the most prevalent form of misconduct in clinical trials? Accepting ineligible patients, as Roger Poisson of St. Luc Hospital in Montreal did in a now-notorious breast cancer trial? Altering elements of the patient's data? Fudging results? Richard Peto, Oxford University's renowned clinical trialist, suggests that the answer is much simpler. When it comes to clinical trials, says Peto, "ignorance is the biggest form of misconduct," because it can lead to problems that invalidate a trial's conclusions.

That should come as no surprise if one considers how much formal training in conducting clinical trials physicians receive: in many cases, none. The number of medical schools in the United States that offer anything related to training in clinical trials is "abysmally small," says Domenic Sica, deputy chair of the American Society of Clinical Pharmacology and Therapeutics' education committee and head of clinical pharmacology at the Medical College of Virginia, Richmond. Most "physicians are not trained in basic scientific principles, let alone clinical trials," Stephen George of Duke University, chair of the statistics committee for the National Cancer Institute's Cancer Clinical Cooperative Groups, told attendees at the 15th annual Society for Clinical Trials meeting, held 8 to 11 May in Houston.

Things are no better on the other side of the Atlantic. In Europe, as in the United States, the most that's usually offered for the budding physician-cum-researcher is a course in biostatistics, covering everything you ever wanted to know about statistics, epidemiology, and the science of clinical trials in a single packed semester, says Christopher Williams, chair of the cancer therapy committee of the United Kingdom's Medical Research Council. "It gives them a broad brush," says Williams, "but it's not adequate training for someone who is going to run a clinical trial."

One bright spot in this gloomy picture is provided by elective courses offered to qualified doctors by a small number of universities and teaching hospitals, including Bowman Gray Medical School in Winston-Salem, North Carolina, McMaster University in Hamilton, Canada, and the London School of Hygiene and Tropical Medicine. Soon to be added to that list are courses at the National Institutes of Health's (NIH's) clinical center (in 1994) and Oxford University's Medical School in England (in 1995). These courses range from part-time lecture programs to full-blown master's degree courses and attempt to cover every aspect of clinical trial execution.

As good as those courses are, however, according to the educational reformists, voluntary training for a few highly motivated doctors isn't enough. "I fantasize that in the future [training] will become a requirement for running clinical trials," says John Gallin, director of NIH's clinical center, who steers the committee that will be developing the center's clinical-trial curriculum. Others argue that every physician needs such training, since during her working life a physician in any specialty is likely to be called upon to enroll patients in a clinical trial, collect data, and determine treatments on the basis of clinical trial results.

Says Thomas Chalmers, currently of Tufts University School of Medicine and former dean and president of Mount Sinai School of Medicine in New York, "We're ready for a big change in medical training. It should be based on clinical trials."

–R.N.

lished in elite medical journals appear not to ensure that patients are assigned to different treatments by chance, according to a survey reported in *The Lancet* in 1990 by Douglas Altman of the Imperial Cancer Research Fund in London and Caroline Doré of London's Royal Postgraduate Medical School. That defect vitiates the reason RCTs were developed in the first place—to prevent bias in how doctors treat patients from upsetting the results. Without randomization, for example, a doctor might put healthier patients on the experimental treatment due to an unconscious desire to see the treatment vindicated.

Unfortunately, many common methods attempting randomization are open to abuse. Take the traditional practice of asking doctors to assign their patients to a treatment according to the order in which they enroll—the first patient gets the placebo; the second gets the test drug; the third, the placebo, and so on. If a physician wants a particular patient in the treatment group, she may simply enroll the patient when the treatment is next in line.

According to clinical trialists attending a special session on bias in trials at the 15th annual meeting of the Society for Clinical Trials, held in Houston, 8 to 11 May, attempts by physicians to circumvent randomization are not isolated events; they're part of an endemic problem stemming from ignorance (or, less often, from a doctor's desire to provide a patient with what he believes is the best available treatment). "It seems that a lot of people don't understand the basic principles of clinical trials," says Kenneth Schulz of the Centers for Disease Control and Prevention in Atlanta; "they don't realize how these little things bias the comparison" between treatment groups.

At the Houston meeting, Schulz reported data showing that the experts' fears about ignorance are justified. Schulz examined 250 reports of clinical trials in perinatal medicine published between 1955 and 1992 and came up with an intriguing finding. Papers that did not describe adequate safeguards to ensure that treatment allocation was kept secret until randomization was carried out generally reported larger differences between treatment and control groups—indicating that bias had crept in and undermined the validity of the results.

In an attempt to rectify this problem, the *British Medical Journal* has established a new policy of refusing to publish results of any clinical trial that fails to guarantee proper randomization. That doesn't mean the trial must be "blinded" (keeping the doctor from knowing which treatment the patient is receiving), but simply that the doctor must not know the treatment until the patient has been assigned to it. The journal now only accepts papers from trials that include such safeguards as sealing randomly ordered treatment allocations in numbered, opaque envelopes.

Even that method isn't foolproof, warns Schulz, who says he personally knows three doctors who admit they have used radiology "hot lights" to read treatment allocations in their opaque envelopes. A better way of randomizing, says Schulz, is "central telephone randomization," in which the doctor must phone a central number and enroll the patient before treatment is assigned.

GLOSSARY

Randomization—A process that prevents bias by secretly and arbitrarily assigning patients to treatment and control groups

Intention-to-treat protocol—Analyzes data from every patient assigned to a treatment whether or not the patient complies with the treatment

Surrogate markers—Measurements of a drug's biological activity that substitute for clinical endpoints such as death or pain relief

Meta-analysis—A statistical process for pooling data from many clinical trials to glean a clear answer

Large simple trials (megatrials)—Massive randomized clinical trials that test the advantages of marginally effective experimental drugs by enrolling 10,000 patients or more

Dredging through subgroups

Of course, even when investigators have designed their study properly and ensured randomization, they may still find no difference between treatment groups. In that situation, some researchers resort to zealously analyzing data from smaller and smaller subgroups in the hope of showing that at least some types of patients benefited from the experimental treatment.

That practice—the one for which Fischl received catcalls in Berlin—courts disaster, according to David Sackett of McMaster University in Hamilton, Canada. Although it's acceptable practice to analyze a few subgroups that were defined before the study began, if investigators go beyond that, Sackett says, to "look at subgroup after subgroup after subgroup, by the laws of probability one in 20 of these comparisons is going to come up bingo even when nothing is going on at all." (In defense of the Fischl analysis, Harvard's Kenneth Stanley, senior statistician for the trial, says the subgroup analysis was planned ahead of time and was intended—although not perceived—as an exploratory exercise, not as a means of proving AZT and ddC's efficacy.)

Many experts think inappropriate subgroup analysis is a recurring problem. "Fischl was a classic case, but it appears in all clinical areas that are new to clinical trials," says Fleming. "Everybody thinks their disease is different—the exception. We are seemingly unwilling to learn from history, from other disease settings." That frustrates Fleming and his statistician colleagues, who think all researchers who run clinical trials should have learned these lessons from fields such as cardiology and oncology.

As an example of how confusing subgroup analysis can be, Fleming points to two NCI-sponsored trials he participated in as a member of the statistics team. In a 1989 trial, subgroup analysis showed that a combination of the drugs levamisole and 5-FU is a particularly effective therapy for young patients and for females with Duke's stage C colon cancer. In a 1990 trial, subgroup analysis indicated that the combination is most effective in older patients and males. As a result of the contradiction, the subgroup analyses were ignored, and on the basis of the global results, which found a 30% reduction in death rate, levamisole and 5-FU became standard therapy for all patients with Duke's stage C colon cancer.

Although there's a consensus that overzealous subgroup analysis is wrong, not all problems involving data analysis in clinical trials are simple to resolve, since in some cases even experts don't agree on where to draw the line between analyzing data and massaging it. One disputed area is the issue of "intention to treat." The majority of biostatisticians believe that, to prevent bias from creeping into the analysis, RCTs should be analyzed according to the treatment the patient is assigned to—whether the patient complies with the treatment or not.

To explain why such an intention-to-treat analysis is the best way to go, its advocates cite the case of clofibrate. In 1975, investigators for the Coronary Drug Project (CDP), a multicenter trial sponsored by the National Heart, Lung and Blood Institute (NHLBI), announced that lipid-lowering clofibrate, one of their test drugs, had no impact on 5-year survival following heart attack. Many of the patients in the study, however, had taken fewer than four fifths of their pills—but these patients were not excluded from the analysis. Eventually, under pressure from their statistically less sophisticated peers, a team of CDP statisticians led by Paul Canner, then of the University of Maryland at Baltimore, did a reanalysis on the basis of the actual—rather than intended—treatment. As predicted by those in favor of the reanalysis, the death rate among patients who faithfully took their pills was almost 40% lower than that of patients who took less of the drug, strongly suggesting that

clofibrate helped heart attack victims.

But the second part of the Canner analysis dampened any resurgence of enthusiasm for the drug. Patients who routinely took their placebo pills, it turned out, were also far more likely to survive than patients in the placebo group who didn't comply with the regimen. To this day, no one knows why, when it comes to heart attacks, faithful compliance to the trial protocol is a marker for improved prognosis. Nonetheless, says biostatistician Paul Meier of Columbia College of Physicians and Surgeons, the Canner analysis illustrates "the fallacy of an as-treated analysis."

Altman explains that "you can't assume that people who don't comply are a random sample." For example, patients who have a bad reaction to a drug might stop taking it, but that bad reaction could indicate that the patient had a different chance of survival than others in the trial. "The strongest aspect of an RCT is randomization, and when you start leaving people out, you destroy that," says Altman.

But not everyone is impressed with that logic. The intention-to-treat concept "started out as a simple idea, but it's got out of hand," says John Lewis of the Institute of Mathematics and Statistics at the University of Kent in England. Rather than being rigid about intention-to-treat, with the attendant risk of missing evidence of a valuable new therapy, says Lewis, "you should start to form judgments, to look at the reason [people stop complying] and establish whether it's random or due to the treatment."

That is the approach some clinical trialists have taken. In the 1985 National Surgical Adjuvant Breast and Bowel Project RCT investigating surgical treatment for breast cancer, women who refused their assigned treatment were excluded from the published analysis. By and large, when a researcher makes an informed decision to remove some patients from the analysis and clearly states so in published reports of the trial, it's deemed acceptable by her peers—even though the impact on the conclusions is still open to debate.

That situation, however, is quite different from the one in which inexperienced clinical trialists resort to post-hoc data removal in order to squeeze out the result they desire—and then fail to describe their tactics in their written results. Researchers "in many fields don't even know what an intention-to-treat analysis is," complains CDC's Schulz. "They don't tell you who they include in the final analysis, so you end up with a mistaken impression of the results of the trial."

Surrogates under siege

The issue of "intention to treat" is contentious, with experts lined up on both sides, but the vast majority of clinical trialists contacted by *Science* agree on one thing: The most potentially damaging flaw in clinical trials today is the inappropriate substitution of "surrogate" markers for well-defined clinical endpoints such as survival or pain relief.

Surrogate markers are measures of biological activity that seem to correlate with clinical outcome. For example, in AIDS, the number of CD4 cells, a key immune system cell, declines as the disease progresses; in heart disease, high cholesterol is a risk factor for death. As a result, many clinical trials look for effects on CD4 counts and serum cholesterol levels rather than monitoring hard clinical endpoints. Most experts in clinical trials agree that it's fine to use surrogate markers to identify promising new agents in the early stages of drug develop-

Teething Problems for Two Innovations

Back in the 1940s, when streptomycin took on tuberculosis, all that was needed to show the new wonder drug saved lives was a 100-patient randomized controlled clinical trial, or RCT. Since that first RCT, however, it's become much tougher for an experimental drug to prove its mettle. The problem is that few new drugs produce the astounding results chalked up by the first antibiotics. Today, clinical trials generally test therapies that offer small improvements over existing treatments, and under these conditions conventional RCTs often yield confusing, ambiguous results.

To start getting clear-cut answers again, clinical trialists have come up with two innovations: meta-analysis and megatrials (or large simple trials). Both methods improve the statistical power of RCTs by increasing their size. Meta-analysis is a statistical procedure that pools raw data from small RCTs. The megatrial is a giant RCT, enrolling 10,000 patients or more, that relies on sheer weight of numbers and strict randomization to increase sensitivity. Both meta-analysis and megatrials can boast successes, but recently it became clear that they were suffering teething problems when the two approaches provided contradictory answers to the same question: Does magnesium therapy save the lives of heart attack victims?

Beginning in the 1980s, several meta-analyses showed that infusing heart attack patients with magnesium salts saves lives. The huge ISIS-4 megatrial, however, led by Oxford University's Richard Peto, has recently concluded that magnesium salt treatment is useless. When rumors of the negative ISIS-4 results (due to be published this fall) began filtering out, clinical trialists assumed the meta-analyses were at fault. They speculated that they were too small (under 3600 patients compared with 58,000 for ISIS-4) or that they had fallen prey to the "file-drawer problem," in which a meta-analysis becomes skewed because negative or neutral RCT results have been relegated to a file drawer, never to be published—or included in a meta-analysis.

But by this year's 15th annual meeting of the Society for Clinical Trials, held 8 to 11 May in Houston, the tide had turned against the megatrial, partly because the protocol ran counter to magnesium's mechanism of action. Recent studies in animals suggest magnesium protects heart muscle from damage that occurs when blocked blood vessels reopen, either spontaneously or in response to clot-busting drugs. This finding suggests that magnesium must be given to heart attack victims before the blood vessels reopen. But the ISIS-4 protocol, planned in the mid-1980s before magnesium's mechanism was known, advised providing routine therapy (which can include clot-busting drugs) before enrolling the patient into the trial and providing magnesium.

By chance, the majority of RCTs in the meta-analysis started magnesium salt treatment the moment patients entered the hospital. That difference could explain the contradiction between the meta-analysis and ISIS-4, says Jean Pierre Boissel, head of clinical trials at the Neuro-cardiology Hospital in Lyon, France, who chaired the Houston session on megatrials and meta-analysis. Peto, who didn't attend the meeting, doesn't agree. Even with clot-busters, he says, it takes 1 to 2 hours for blocked blood vessels to open, by which time many ISIS-4 patients would already have received magnesium. He thinks the field's original verdict was correct: "The old meta-analyses are totally wrong," he says.

However this debate is resolved, for Kent Woods of the University of Leicester, England, there's a valuable lesson to be learned. "A megatrial should not be planned without a clearly formulated mechanism of action," he told the Houston meeting. If that's not done, he said, "there's a serious risk that you'll have a very precise answer to a question that is not the relevant question...while putting participants in the trial at unknown risk." Meanwhile, many clinical trialists—including Peto—believe megatrials and meta-analysis are both needed. Says Peto: "We need large-scale randomized evidence, and it doesn't matter how we get it. We can have bigger trials, we can string together smaller trials [with meta-analysis], or preferably we can do both."

–R.N.

A SAMPLING OF PROBLEMS IN MAJOR CLINICAL TRIALS

Trial (Sponsor)	Result	Problem	Recommendation
MRFIT—the Multiple Risk Factor Intervention Trial, 1982 (NHLBI)	Reducing smoking, blood cholesterol, and blood pressure in men at high risk of heart disease doesn't reduce the death rate.	Some researchers associated with MRFIT believe that a drug used to reduce the patients' blood pressure was toxic, canceling out the overall benefits of the interventions.	Despite the negative findings, NHLBI continued to recommend quitting smoking and reducing blood cholesterol and blood pressure.
North Central Cancer Treatment Group Study, and Cancer Intergroup 0035, 1989 and 1990 (NCI)	Levamisole and 5-FU combination therapy reduces by 30% death due to Duke's stage C colon cancer.	Treatment was especially effective for young females according to subgroup analysis of the 1989 trial, but in the subgroup analysis of the 1990 trial was especially effective in older males.	The contradictory subgroup analyses were deemed misleading. Levamisole and 5-FU became standard therapy for Duke's stage C colon cancer for all patients.
Protocol 019, 1990, and Concorde, 1993 (NIAID, and British-French collaboration	AZT delays the onset of AIDS symptoms, according to Protocol 019, but doesn't according to the Concorde trial.	Both trials may be too small and brief to accurately assess AZT, which can at best offer only a tiny, short-lived benefit.	In 1990, NIH recommended AZT to HIV-infected, symptom-free patients. Three months after Concorde, NIH recommended that doctors and patients decide when to start taking AZT.
Alpha-Tocopherol, Beta-carotene Cancer Prevention Trial, 1984 (NCI)	Contrary to expectations, the antioxidant beta-carotene actually increases incidence of lung cancer among heavy smokers.	Some say the surprise finding was a chance event, and beta-carotene may still reduce cancer risks. Others say the rationale for thinking so was weak to start with.	NCI recommended that people keep off the antioxidants and eat lots of vegetables.

ment. But they say surrogate markers are an untrustworthy way of proving a drug's clinical benefits.

Take heart failure. In patients who have recently survived a heart attack, irregular beating is a risk factor for death from a second attack. In the 1980s, cardiologists reasoned that two drugs capable of controlling irregular heart beat—encainide and flecainide—would also reduce the likelihood of a second life-threatening cardiac episode. Confidence in that logic was so strong that, in the U.S. alone, about 200,000 people a year received the drugs, and when an RCT to test the drugs' efficacy was suggested, many physicians balked, believing it was unethical to deprive patients in the control group of a supposedly beneficial therapy.

Despite the opposition, the Cardiac Arrthymia Suppression Trial (CAST) was started—only to be halted when preliminary results showed the drugs tripled the death rate. Yet many clinical trialists ignore the CAST lesson. If anything, says statistician David DeMets of the University of Wisconsin in Madison, "there's an increased tendency to use surrogate markers." DeMets believes misuse of surrogate markers, like many other problems in clinical trials, is the product of ignorance and the desire to get results quickly and cheaply by not waiting for clinical endpoints, which take longer to measure.

Nowhere has the pressure for results been more intense than in AIDS research. In a 1992 article in *Statistical Science*, Fleming warned against what he sees as the false and dangerous economy of using surrogate markers in AIDS clinical trials. "In AIDS," he wrote, "if surrogate markers are used to replace clinical endpoints, the public health consequences...[could be] staggering," as useful therapies go undetected while toxic or useless drugs pass muster.

But the statistician's assessment isn't always accepted in the AIDS field. Researcher Fred Valentine of the New York University Medical Center, for instance, hotly disagrees with Fleming. He maintains that the use of surrogate markers in AIDS clinical trials is the only answer to a key question: "How can you design a trial that can be done in a sensible number of years so that patients stick with it?" One grim reality of AIDS research, he says, is that participants stay in a trial for only about a year, then try other drugs. CD4 counts are far from perfect, he concedes, but says that they are "the best we currently have."

The FDA advisory committee that deals with new AIDS drugs appears to share Valentine's view. On the committee's recommendation, the FDA granted provisional approval to market ddI and ddC largely on the basis of the drugs' impacts on CD4 counts, says Fleming. In an upcoming issue of *Statistics in Medicine*, Fleming will again attempt to convince the FDA and researchers in the field of the folly of using surrogate markers to assess an AIDS therapy's effectiveness. He will report his analysis of 16 major AIDS RCTs, which "found that the effect [of the drug] on CD4 counts...tells you nothing about its effect on length of survival [of the patient] or the frequency of AIDS-related events."

AIDS clinical trialists are far from alone in using surrogate markers. "There's loads of them," says Edward Lakatos, director of statistics for G. D. Searle & Co. in Skokie, Illinois. For example, he says, three trials sponsored by NHLBI in the late 1980s—Trials of Hypertension Prevention, Dietary Intervention Study in Children, and Child and Adolescent Trial for Cardiovascular Health—relied on lowering of blood cholesterol or blood pressure as surrogate markers for cardiovascular mortality. Lakatos thinks that policy may be misguided: Although the relationship between high blood cholesterol and blood pressure and cardiovascular death holds up for some cholesterol-lowering interventions in some populations, it might not be universal. DeMets agrees. He refers to the reliance on surrogate markers as "the most disturbing, even threatening, issue today in clinical trials."

In fact, problems with surrogate markers, inadequate randomization procedures, and misleading analysis of data are such a major worry that the majority of experts contacted by *Science* believe they seriously undermine the credibility of many clinical trials. "Unless we give proper attention to these issues, [RCT] conclusions can be considered misleading or unreliable," says Fleming. For him, as for other experts, the best answer is further education of those who perform clinical trials. Williams even suggests one way of speeding up the learning process: "Rather than peer reviewing a trial when it's finished, journals should peer review the protocol, and if it meets certain minimum standards, guarantee its publication."

Even the sharpest critics of the way clinical trials are currently conducted, however, would not advocate that they not be done, for one simple reason: There's no better alternative. "Clinical trials are by no means perfect, but it's the best method we've got" for evaluating new drugs and therapies, says DeMets. Far from giving up on this valuable tool, he says, "we've got to stick with it. We've got to improve it."

–Rachel Nowak