 0889–8588/00 $15.00 + .00

A BRIEF HISTORY OF THE RANDOMIZED CONTROLLED TRIAL

From Oranges and Lemons to the Gold Standard

Marcia L. Meldrum, PhD

The science of medicine in the last 2 centuries has expanded and enhanced the knowledge of the body and of the pathologic conditions which threaten its well-being. The practice of medicine, however, remains an art, because the patient does not always respond to treatment in the way the physician expects. This unpredictability is a greater problem when the treatment itself carries certain risks or deleterious side effects. Even more challenging to simple logic is the converse: sometimes the patient will respond to a *placebo*, a treatment which produces no known physiologic effects at all. To attribute these paradoxes of medical practice to individual variation, whether in genetics or in psychologic conditioning, may be theoretically sound but leaves the problem of constructing "a rational therapeutics" unsolved.[29]

In the twentieth century, to discover the hidden causes of unpredictable and unknown responses to treatment, medical researchers, with the aid of statisticians, have developed a mathematical model to describe and calibrate the complex responses of the human body to therapeutic interventions. The basic principles of this model are (1) comparison, under controlled conditions, of two or more therapeutic regimens (one of which may be a traditional treatment, a placebo, or the exclusion of

From the Pain and Neurosensory Mechanisms Branch, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, Maryland

active treatment), and (2) statistical analysis of the possibility of error. The recognized methodology is the randomized controlled trial (RCT), with its associated features of (1) control groups, (2) randomization, and (3) blinding.

The RCT is by no means a straightforward solution, as even its advocates agree. On the one hand, the principle of comparison often means that one set of subjects will receive a less effective treatment, or possibly none at all, a situation which may sometimes be ethically questionable. On the other, the logistics of designing and carrying out a trial within the real-world constraints of cost, time, and personnel require that the investigators select certain subjects for treatments, specify outcome measures and criteria, and set limits to the duration of treatment and follow-up. These necessary choices and exclusions may affect the statistical result of the RCT or cast doubts on its external validity. As John McKinlay has written,

> Recognizing the legitimacy of certain objections, researchers often attempt to accommodate them in the design of an RCT. . . . In making these accommodations and implementing a study in the real (sometimes hostile) world, certain methodological allowances must be made. . .the researcher here has been forced by circumstances to depart from the ideal textbook design. . . . But without these methodological accommodations, the RCT would never have been permitted in the first place. These allowances, which are forced on researchers by practical considerations, are seized upon by critics to discredit the entire RCT. . . . It is analogous to someone saying they will not attend a party unless they can decide who is to be invited, and then complaining after the party that the company left much to be desired![30]

Nevertheless, the RCT remains the "gold standard." Its power as a model for good practice rests on its imposition of experimental order on the clinical setting and its production of numerical results that may not be absolutely accurate but that are unquestionably precise. As Theodore Porter has argued, the value of the precise quantitative result is that it is readily translated outside its original experimental setting, for replication, comparison, and adaptation elsewhere.[38]

The inferential authority of the RCT has been such that it is accepted as a standard for "rational therapeutics" by physicians and regulatory authorities and also by patients and populations at risk. In the late 1980s, for example, groups such as the Institute for Research on Women's Health documented the exclusion or artificial restriction of women from clinical trials, even when the disease in question affected both sexes, and the scarcity of trial evidence on problems specific to women, such as menopause. In 1991, the National Breast Cancer Coalition challenged the cancer research establishment to carry out trials on new and innovative treatments. In effect, women demanded inclusion in clinical trials and the production of trial evidence specific to their needs.[13, 21] The following decade saw the creation of the National Institutes of Health (NIH) Office for Research on Women's Health and the institution of a number of gender-specific and gender-comparative trials. Today, guidelines for

Public Health Service (PHS) grant applications specifically require the inclusion of women unless there is a valid reason for exclusion.

AIDS activists have also demanded RCT expansion, with the Treatment Action Group of ACT UP lobbying for more trials, for the inclusion of more subjects and more minorities, and for the use of active drugs rather than placebo controls. ACT UP proposed the redesign of trial methodologies to make them more sensitive to patient needs, in particular the replacement of life-or-death outcome criteria with surrogate markers of therapeutic efficacy, such as CD4 cell counts. ACT UP's emphasis on "participatory knowledge making" has been adapted by advocacy groups for Lyme disease, breast cancer, and chronic fatigue syndrome.[7, 8] Although all these groups may criticize trial procedures, they do not reject trial evidence; rather, they seek to participate in its production and to find ways to combine statistical rigor with sensitivity to patient needs. As one activist told sociologist Steve Epstein, "It's about having good science that develops good therapies so that we may have a cure or therapy someday."[8]

As the clinical trial has evolved in the last 100 years, physicians and scientists—and subjects as well—have faced the same challenge: how to develop "good therapies" based on "good science," science that imposes order on, but neither distorts nor devalues, individual human experience. The RCT is a dynamic methodology, and its present and future are informed by its history.

PRECURSORS

The ideas of intentional comparison under controlled conditions and of elementary statistical analysis appear in medical history within two generations of the scientific revolution of the 1600s, that is, soon after "experimental philosophy" became current among the upper and middle classes in Europe. In the 1720s, James Jurin and others compared the proportionate mortality of cases of naturally occurring smallpox with that of cases occurring as a result of inoculation. Their comparison demonstrated the efficacy of the new practice.[23] In 1753, the naval surgeon James Lind published his famous account of the comparative treatment of 12 scurvy patients, "their cases as similar as I could have them," noting that "the most sudden and visible good effects were perceived from the use of the oranges and lemons."[24]

The expansion of scientific education and research in the nineteenth century, in particular the physiologic studies of physicians such as Xavier Bichat, Claude Bernard, and Rudolf Virchow, encouraged the application of experimental methods to clinical studies, and the large patient populations in the great city and military hospitals facilitated the use of controlled conditions and statistical comparison. Many early investigators recognized that they could draw valid inferences from clinical studies only if they chose their comparison groups to preclude the effects of chance and hidden bias. Pierre Louis' description of his "méthode numér-

ique," for example, emphasized the selection of cases "of as similar a description as you could find," which should then be "taken indiscriminately" for assignment to one of the treatment groups.[23, 25]

Other nineteenth century studies demonstrating a practical understanding of the importance of controlled comparison include John Snow's epidemiologic analysis of cholera in London neighborhoods served by different water companies[43]; Ignaz Semmelweiss' inquiry into the differences in morbidity between the medical and midwifery divisions at the Vienna Lying-In Hospital[41]; Louis Pasteur's 1881 demonstration at Pouilly-le-Fort of the positive effects of anthrax inoculation on two herds of animals[16]; and Walter Reed's experiment with Army volunteers in Cuba, showing that yellow fever was transmitted by mosquito bites, not through the breath or wastes of infected persons.[39]

The reader will note that each of these important studies validated a preventive method, not a therapeutic intervention. Although the one exception, Louis' demonstration of the inefficacy of bloodletting in pneumonia, was well known, individual physicians continued to take blood routinely until well after the American Civil War. Snow's identification of cholera as a water-borne infection saved many lives in the nineteenth century, but it offered no guidance to the doctor faced with actual cases of the disease, and cholera treatment remained inconsistent and generally ineffective. Some comparative data were available: in Paris in the 1840s, the water-based treatments of the homeopaths reduced mortality much more successfully than the purges and emetics administered by their orthodox colleagues. Such evidence was too contrary to theory, however, to be accepted by "scientific" practitioners.

Understanding the importance of clinical comparison was one thing; applying the information to the treatment of real patients was another. The large hospital wards were populated with the poor and malnourished; the individual practitioner usually saw patients of a very different class and economic status, and he saw them one at a time. In the literature, he found numerous reports of therapies that had worked well in individual cases, and he learned of others from his teachers and colleagues. Applying these therapies and making his own observations, each physician developed a personal set of standby drugs and rule-of-thumb guidelines. Individual clinical judgment might vary among communities and regions, but disease patterns differed as well, as did heredity, climate, diet, socioeconomic conditions, and a host of other contributing factors.

Where controlled comparisons were possible, in university hospitals, for example, or during epidemics, there were often strongly urged ethical and professional injunctions against setting aside one group of patients for a treatment which might be less effective than the accepted old or the promising new regimen. A lone exception at the very end of this period, perhaps the first example of a clinical trial with attempted randomization, was Johannes Fibiger's 1898 study of serum treatment on 484 diphtheria patients. Patients were allocated to treatment by day of admission: patients admitted to the Copenhagen hospital on alternate

days received serum injections, while their counterparts received traditional treatment. Fibiger's account argues for the need "to eliminate completely the play of chance and the influence of subjective judgment," showing a clear understanding of the hazards of uncontrolled comparisons, but his innovation seems to have had little effect even in Denmark. It seems likely that the methodologic purity employed in this instance was possible because Fibiger had the backing of a powerful superior, Professor Sorenson, who was dubious about the controversial new treatment.[20]

The problem was not a lack of drugs for testing. Ethical drug and patent medicine manufacturers proliferated in the second half of the century, and their products were readily available over-the-counter or through the mail to patients open to self-experimentation. Physicians might consider many of these preparations to be of dubious value, but their controlled study within a significant group of patients was not within the purview of individual practice. What was lacking was a strong professional rationale justifying clinical comparative studies and an organizational infrastructure to support these studies on a large scale.

JUDGMENT AND CHANCE

By the early twentieth century, the sheer abundance of drugs and patent medicines on the market, coupled with the extravagant claims made for them by manufacturers, advertisers, and salespeople, had created a climate in which a muckraker like Samuel Hopkins Adams could label the pharmaceutical industry "The Great American Fraud." One result was the Pure Food and Drug Act of 1906, which created a small bureau within the Agriculture Department to review the labeling and advertising of drugs. In the previous year, the American Medical Association had already established its own Council on Pharmacy and Chemistry to provide expert assessment of the plethora of drugs available to physicians and their patients.

The Council, composed of academic researchers and led by Torald Sollmann, professor of pharmacology at Western Reserve University, saw medical practice as based on the integrity of professionals dedicated to an ideal of truth and sought to make the truth about drugs readily available through the *Journal of the American Medical Association* and through its own publications such as *Useful Drugs*.[29] As Sollmann wrote, "clinical experimentation should follow the canons of other scientific experimentation."[44]

Council members felt secure in providing knowledge based on laboratory evidence about the composition of a drug, its known pharmacologic actions, and its observed effects in animals, but they acknowledged that clinical evidence of a drug's effectiveness for particular indications was often inadequate. The Council sought to rely on the experience of reputable academic clinicians but often encountered "honest differences of opinion . . . among responsible observers."[29] However

carefully and impartially an investigator might have assessed a drug, whatever controls he might have used, the patients, settings, conditions, and methods were impossible to replicate exactly; therefore, most findings were not truly generalizable.

The result, as noted by the British physician Richard Doll when he surveyed the recommended treatments for peptic ulcer in 1948, was often a long "list of treatments beginning with each letter of the alphabet."[5]

A notable attempt to address the problems of therapeutic evaluation, the Cooperative Clinical Group's 7-year study of syphilis treatments (1928–1935), sheds light on the complexities encountered. A distinguished group of six experts on the disease, each with access to a significant patient base, attempted to carry out a systematic comparison of the multiplicity of regimens and treatments available. Participants included John Stokes of the University of Pennsylvania, Joseph Earle Moore of Johns Hopkins University, and Thomas Parran, Commissioner of Public Health for New York. The Public Health Service provided clerical and statistical assistance. The project suffered from a lack of resources and made heavy demands on the senior investigators' time, but the ultimate stumbling blocks were the participants' inability to agree on common protocols and methods and their failure even to consider the syphilis treatments in common use among general practitioners, which they had prejudged to be ineffective. The lengthy study produced much data but little uncontested knowledge: "behind each fact lay a series of decisions, often controversial and sometimes inconsistent."[28]

An alternative to the large collaborative trial was a small trial design using two carefully matched groups under identical conditions and tight controls. In 1930, John Wyckoff and his colleagues reported the use of an alternate-control design similar to Fibiger's to evaluate the efficacy of digitalis for treatment of pneumonia.[23, 50]

When the Medical Research Council (MRC) of Great Britain formed its Therapeutic Trials Committee to evaluate new drugs for manufacturers in 1931, it employed alternate controls on several occasions (e.g., in a study of serum treatment in lobar pneumonia). The MRC's statistician, Austin Bradford Hill, argued in his classic 1937 series of articles in the *Lancet* that the alternate-control method ensured the comparability of the experimental and control groups and eliminated investigator bias.[18] Physicians who wished to challenge the findings of a trial, however, could suggest that the experimenter's awareness of the design had affected his assessment of the outcome and might even have enabled him to modify patient assignment.[5]

A slightly more sophisticated variant of alternation was the Michigan study which effectively discredited the use of sanocrysin for tuberculosis. J. Burns Amberson and his colleagues carefully allocated their 24 patients to two individually matched groups and tossed a coin to select one group for treatment with the gold compound. This trial has often been described as the first use of a formal method of randomization,

although it was still open to experimental manipulation and offered only a limited degree of generalizability.[1, 5, 23]

In 1930, Torald Sollmann suggested a different approach to the problem of investigator bias: the use of a blinded observer and a placebo control.[45] At Cornell University, Harry Gold and his colleagues refined the double-blind method and use of placebos over a period of 5 years (1932–1937) in their studies of ether and xanthines for the treatment of angina. According to Gold's colleague, Nathaniel Kwit, the idea of a *blindfold test* was adapted from the use of a blindfold in advertising comparisons of Old Gold cigarettes.[42] Although Gold's techniques facilitated the objective comparison of the different results among groups, they were inadequate to ensure that those differences were not attributable to chance variance, for example, that the placebo group was not simply more resistant to the treatment.

THE IMPRIMATUR OF STATISTICS

By the beginning of World War II, therefore, the problem of clinical experimentation had been clearly defined. The preferences, judgments, and biases of individual clinicians for or against a particular treatment, often soundly based on personal experience, compounded by the high degree of patient variability in many disorders, made each trial result unique, impossible to replicate perfectly or to translate into a generalized guideline.

One of the most erratic disorders and one highly liable to spontaneous remission, tuberculosis was also one of the most deadly. It is not coincidental that many innovative trials were studies of "the white plague." In reporting their use of Amberson's coin-toss method of randomization in an early study of streptomycin on tubercular patients at the Mayo Clinic, William Feldman and Corwin Hinshaw proposed a set of principles for good clinical trial design: careful selection of cases, blinding of observations, and "some procedure of chance" in allocating patients to treatment groups.[19, 51] As they suggested, it seemed possible to develop a convincing trial design by combining several of the proposed techniques; but it was not known whether clinical investigators would comply readily with such a structured and standardized model.

The introduction and evaluation of penicillin during the World War II offered no new model for trial design. Under the conditions of scarcity and emergency, and with the drug's dramatic action against bacteria, concurrent controls were impractical and inhumane; investigators compared outcomes with historical cases. The Committee on Medical Research (CMR) controlled the civilian supply, but physicians who were allocated the precious drug often devised and followed their own protocols.[29] Therapeutic reformers nevertheless hailed the CMR's work as a model for large-scale cooperative studies and standardized procedures that should be applied to the next "wonder drug" on the horizon, streptomycin.

In the late 1940s, three factors merged to enable the effective use of a new statistically based model for comparative clinical trials. First, Ronald Fisher's *The Design of Experiments,* published in 1935, provided a cogent argument for the use of strictly randomized allocation. Second, the wartime climate of scarcity and government financing assisted national agencies in implementing a standardized model. Third, three large streptomycin trials showed that the statistically based design could validate clinical experimentation, verify conformity of protocol, and strengthen the generalizability of findings.

High expectations surrounded the first of these trials, begun in June 1946, because it was sponsored by the US Veterans Administration (VA) and was conducted in the extensive network of VA hospitals. The large number of tubercular patients, the bureaucratic organization, and the established follow-up procedures at the VA encouraged investigators to believe that they could carry out a well-controlled and persuasive experiment. To ensure sufficient enrollment, however, the researchers had first to abandon the idea of an untreated control group, using historical case comparisons instead. As drug supplies and subject numbers increased during the year, it became more and more difficult to restrict VA physicians to the approved protocol or to prevent clinical judgment from influencing case selection. Despite the committed effort, the VA trials ultimately failed to provide generalizable evidence of the efficacy of streptomycin in treating tuberculosis.[28, 29]

The failure of the VA trial contrasted with contemporary studies, identical in objective but more rigorous in approach, run by the PHS and by Great Britain's MRC. The MRC trial of 1947–1948, generally considered the first published instance of a randomized and blinded clinical trial, allocated its 107 patients to experimental and control groups using a system of random number assignments devised by Bradford Hill. As noted in an editorial in the *British Medical Journal,* the method "removed personal responsibility from the clinician" for selecting which patients would benefit. Further, to ensure the objective assessment of patient status, the radiologists who interpreted the radiographs were also blinded.[5, 51]

In 1937, Hill had been doubtful about the ethics of random (rather than alternating) allocation, but he justified its use in 1947 because the United States had made only a limited supply of streptomycin available to its ally, and the trials were the sole means by which most British patients could receive the drug. He became a strong and persuasive advocate for the use of randomization in clinical trials, arguing that it ensured the comparability of the test groups and precluded the biases introduced by "our personal idiosyncracies, consciously or unconsciously applied, [or] our lack of judgment."[5, 29, 51] Only in later years did R. A. Fisher's primary argument, that randomization allowed the experimenter to make a precise statement of the likelihood of error, to "know how often his chance arrangement will coincide with the devil's," become the accepted doctrine among RCT specialists.[29]

The success of the MRC trial was replicated by a third streptomycin

trial initiated by the US Public Health Service in 1947. Carroll Palmer, Corwin Hinshaw, and their associates on the Tuberculosis Study Section Steering Committee used the financial clout of the PHS to ensure that investigators receiving funds to participate in the trial would adhere to a "rigidly controlled project" that entailed "the collection of uniform observations," the use of a preselected control group assigned "by proper random device," and the evaluation of outcomes with blinded assessment of radiographs. To ensure that investigators were shielded from criticism, patients also remained unaware of their assignment to a treatment control group.[28, 29]

The PHS trials demonstrated the efficacy of streptomycin in treating tuberculosis and also the superior credibility of the tangible answers achieved with a consistent design. The use of the randomized, blinded model made possible precise statements of confidence and error, but as historian Harry Marks has explained, that was only its most obvious contribution. By removing decisions about patient selection and allocation from the physician and forcing the use of standardized, nonqualitative criteria to assess outcome, the RCT model eliminated opportunities for deviation based on physician judgment or bias while providing a powerful basis for conformity.[28, 29] Although a physician might still argue, as Thomas Lewis had in 1934, that such a method made no allowance for fine distinctions among individual patients,[5] the statistician could respond that the findings were nevertheless tangible; that is, their meaning could be understood, translated into other settings, and widely applied.

The question remained whether the method would gain acceptance outside the small circle of physicians who acted as government advisors. The writings of Bradford Hill and the reports of several successful trials conducted in the 1950s by the MRC and the NIH were certainly influential within the medical community. Above all, the massive polio vaccine field trial of 1954 most clearly demonstrated the superior credibility of the RCT.

The National Foundation for Infantile Paralysis found itself compelled to use a double design to test the efficacy of Jonas Salk's polio vaccine in grade-school aged children. In 33 states, it used an observed-control trial, comparing polio incidence in vaccinated children with that among their unvaccinated peers. This large-scale demonstration was necessary to attract news coverage and to maintain the Foundation's support among its many local volunteers. Eleven states, however, consented to the use of a strictly randomized and blinded placebo-control design, despite the ethical and logistic difficulties involved. Thomas Francis, the public health expert from the University of Michigan who supervised the field trial evaluation, and his scientific allies considered the rigorous methodology essential to show the efficacy of the vaccine, which was sponsored, not by a government or by an academic medical group, but by a lay volunteer organization, and which faced considerable opposition from expert virologists.[31]

Perhaps more surprising than Francis' position was the informed

support given the blinded design by many participating parents and children. The children of Public School (PS) 61 on New York's Lower East Side, for example, quizzed reporters about the ideas behind the experiment. "Did we understand, they asked, about vaccines? About controls? About immunity? Their parents had said . . . . Their teacher had said . . . . Their principal had said . . ."[37]

The polio vaccine trials showed that statistical theory and method had the power to demonstrate the validity of a therapeutic innovation, even one less than fully accepted by the scientific community. The NIH, the PHS, the VA, and major voluntary groups, such as the American Cancer Society and Planned Parenthood, sponsored randomized controlled trials during the 1950s. By the end of the decade, at least within the literature, the RCT had become the standard for therapeutic evaluation.

THE POLITICAL SOLUTION

The 1950s also saw the beginnings of the "chemotherapeutic revolution," with more than 400 new drugs introduced each year. Pharmaceutical companies seeking to establish the credibility of their products sponsored clinical trials of new drugs and packed their advertising literature with citations. Many of these reports, however, failed in one respect or another to follow the textbook RCT design. A 1970 review of the literature on dextropropoxyphene (Darvon), one of the best-selling drugs of the previous decade, for example, found many of the published studies "inconclusive" or "of questionable validity"; only about half were "worthy of critical review."[33] Also, independent investigators might adapt particular elements of the RCT without understanding how these elements fit into the methodology as a whole. Trial specialists Walter Modell and Raymond Houde felt compelled to warn in the *Journal of the American Medical Association* that, despite "the magical quality it appears to have . . . [use of] the double-blind technique . . . will not validate otherwise poorly designed experiments."[34]

The translation of statistical design from the literature into routine clinical investigation took different routes. In the United Kingdom, the MRC adapted a gradual policy of involving groups of physicians in multicenter trials under its guidance. In 1955, for example, this process was used in cooperation with the American Heart Association to evaluate corticotropin and cortisone for the treatment of rheumatic fever.[5] According to Richard Doll of Oxford's Clinical Trial Service Unit, "It was many years before randomization was accepted as such a normal procedure." The success of this long effort, in his view, was seen in the collaborative International Study of Infarct Survival (ISIS) studies (published in 1988), which used an RCT model to evaluate the long-term survival of more than 17,000 patients with myocardial infarctions.[5] In Germany, the government and the medical profession continued to depend more on medical consensus than on formal evaluation; when

the German Drug Law of 1978 mandated premarket testing, it made no specific requirements for randomization or blinding.[4]

In the United States, the RCT achieved a more official status. Since 1938, the Food and Drug Administration (FDA) had had the authority to review new drugs for safety, which usually meant scrutinizing animal studies and small human volunteer trials for any signs of serious hazard. In the summer of 1962, the shocking reports of thalidomide-damaged babies in England and Germany coincided with Congress' consideration of a new bill amending the 1938 Food, Drug, and Cosmetic Act. Enacted in response to public outrage, the Kefauver-Harris Amendments gave the FDA the power to approve or disallow the introduction of new drugs and the continued marketing of established compounds, based on its consideration of "substantial evidence" of their therapeutic efficacy, as well as safety. "Substantial evidence" was described in the legislation as "adequate and well-controlled investigations." Congress allowed the FDA considerable discretion to define the parameters.[46]

The agency did not immediately propose the RCT as the basis for acceptable evidence, knowing that most of the drugs then on the market lacked such validation. Its first strategy was to seek advice from the biomedical community, through advisory committees and through the Drug Efficacy Study (DES), a major review of previously approved drugs, undertaken from 1966 to 1969 by 180 panelists selected by the National Research Council. Where trial evidence was not available or was inadequate for a full evaluation of a drug, which was usually the case, the DES relied on "the informed judgment of the panel[s]," based "on the clinical experience of members [most of whom were academic physicians] . . . [and] a consensus of the experience of their peers."[36]

Only when the FDA's orders removing certain drugs from the market, based on DES recommendations, were challenged in the courts by their manufacturers, did the agency publish its "interpretive regulations announcing the essentials of an adequate and well-controlled investigation,"[22] which were accepted by the US Sixth Circuit Court of Appeals "as a proper application of the statutory definition of substantial evidence."[48] These regulations, as amended in May 1970, specified the use of criteria for patient selection, exclusion of bias, "comparability of variables," identification of a control group, and statistical analysis of the data. The FDA and the courts thus made controlled clinical trials, using a statistical model, a matter of regulatory law and legal precedent.[9, 22, 46, 48] The United States government thus used the RCT to establish a scientific rather than political basis for its regulatory authority over the drug industry.

The 1970 regulations created a clinical trial industry in the United States Pharmaceutical manufacturers had protested vigorously against the FDA's new policies and rules of evidence. Once that fight was clearly over, however, they quickly used their significant resources to streamline and standardize drug evaluations to meet the changed requirements. Biostatisticians, already filling advisory roles in many medical schools, found their workloads doubled and their opinions given more weight;

few serious studies could now be performed without a statistician; and the discipline flourished. Consumers also quickly understood the rules. By 1969, feminists were challenging the safety of oral contraceptive drugs on the grounds that clinical trials had not been sufficiently extensive or rigorous.[49] A new age of trial-based therapy seemed to have arrived.

DEPARTURES FROM THE DESIGN

The first constituency of an RCT are the physicians who supervise clinical investigations and who rely on the results. Since at least the 1960s, the medical profession has recognized the value of statistical evidence in evaluating therapeutic innovations, in discriminating among available regimens, and in resolving thorny controversies. The 1990s brought increasing emphasis on the practice of evidence-based medicine.[10] The individual practitioner can find published meta-analyses or pooled results of multiple studies, and online databases, such as the Cochrane Collaboration,[3] can help sort through the mass of available information. Although these new resources are open to the inclusion of "convincing nonexperimental evidence," the RCT is "the fundamental source" of reliable data.[11]

Trial data, however, are rarely translated directly into clinical protocols. Physicians continue to assert the validity of clinical judgment and their duty and right to rethink RCT design, to re-interpret evidence within the context of patient care, and to consider "how competing clinical questions were prioritized for each case and how the evidence obtained was particularized to reflect the needs and choices of the individual patient."[17]

In examining how clinicians have made allowances for the patient's "needs and choices" in the era of the randomized controlled trial, medical historians have identified several types of departures from RCT protocol. These variations include challenging trial protocols which failed to consider all aspects of the clinical problem, altering trial design to extend access to experimental treatment, and reinterpreting trial evidence in the context of practice needs.

In 1969, the principal investigators of the NIH-funded University Group Diabetes Program (UGDP), a 10-year multicenter evaluation of insulin, the oral hypoglycemic tolbutamide, and placebo, decided to end the trial early because of a statistical finding that the group treated with tolbutamide exhibited a significantly higher mortality rate from vascular disease. Their published report, suggesting that the hypoglycemic agent was "less effective than diet alone or than diet and insulin" in prolonging life, met harsh criticism from practicing physicians.[47] These doctors had found that tolbutamide therapy offered their patients greater comfort and a better quality of life and reduced the risk of insulin dependence and of coma or shock. They argued that the UGDP investigators had been more interested in measuring blood sugar levels and other laboratory values than in qualitative issues and that the patient-

selection process had inadequately randomized for severity of disease. Despite repeated statistical review and confirmation of the findings, physicians continued to prescribe oral hypoglycemic drugs.[29]

In the 1970s, practicing oncologists involved in cancer chemotherapy trials under the supervision of the National Cancer Institute (NCI) deviated routinely from the textbook RCT design. The drugs under investigation were highly toxic and were justified for use only in the case of fatal and agonizing disease. Physicians used experimental compounds that had not gone through the prescribed regimen of tests on laboratory animals, failed to carry out complete trials with compounds that showed little benefit in initial tests, and, most significantly, made the most promising drugs available to the entire patient cohort, without using a control group. "Ideal experimental design," a joint FDA/NCI Task Force concluded in a review of these chemotherapy studies in 1982, "must be compromised to achieve the best possible patient care."[40]

Analgesics are among the most difficult drugs to evaluate, because the only available measure is the patient's verbal report, and patients seem to vary greatly in their subjective response to pain and to pain relievers.[2] In 1972, a Mayo Clinic team designed a comparative study of eight analgesics and placebo in treating cancer patients, using a crossover methodology that tested each drug sequentially against the others in the same patient. The drugs were coded and taken by the patients in randomized, blinded sequences. The investigators reported that high dosages of aspirin had proven significantly more effective in relieving pain than many newly introduced drugs.[35] Many physicians, however, preferred to prescribe the mild narcotic dextropropoxyphene, the "prescription-pad friend." Although dextropropoxyphene ranked only sixth among the nine compounds on the Mayo list, the doctors knew that their patients expected them to prescribe something safe but more "scientifically advanced" than aspirin.[12] As William Beaver, the dextropropoxyphene reviewer for the DES, told the FDA, trials like the Mayo study reported the *averaged* responses of patient groups; given the subjectivity of pain experience, they could not exclude the possibility that a particular drug, although not the most effective overall, would not be the best choice for "a particular type of pain."[32]

In contrast with these equivocal episodes stand those clinical trials which have cut through inconsistent and anomalous evidence to produce a clear consensus. In the 1980s and 1990s, for example, several major trials completely altered the standard treatment of breast cancer. For most of the century, women diagnosed with this disease had undergone total mastectomy, suffering physical and emotional disfigurement as a result. This radical procedure, developed by the Johns Hopkins surgeon William Steward Halsted in the 1890s, had saved many lives. In the 1960s and 1970s, although self-breast examination facilitated early detection and radiotherapy and chemotherapy extended survival rates, most oncologists still insisted that total mastectomy provided the best guarantee against long-term recurrence. The lumpectomy, advocated by some surgeons and by many women as equally effective and less detrimental to the patient's overall health, failed to gain wide acceptance until the

NCI sponsored a 12-year randomized evaluation of the two procedures (with and without radiotherapy) by Bernard Fisher and his colleagues at Allegheny University. Their results demonstrated that lumpectomy with axillary node dissection is as effective as total mastectomy in ensuring long-term survival in women with early diagnosed, operable cancers.[15]

Subsequent multicenter studies by the Early Breast Cancer Trialists in the United Kingdom and the National Surgical Adjuvant Breast Project (NSABP) in the United States showed that adjuvant chemotherapy could contribute significantly to disease remission and long-term survival in post-lumpectomy patients, and that preoperative drug administration could reduce tumor size and increase the chances of preserving the woman's breast.[6, 27] To make these findings possible, physicians' willing adherence to a rigorous protocol and sophisticated statistical analysis of the several patient groups involved, of the RCT's methodology, of its epistemologic basis, and of the infrastructure it had generated were all necessary.

The findings of NSABP have not been accepted without dissent. The 1998 NSABP report of the remarkable effects of the selective estrogen receptor tamoxifen in preventing breast cancer in women at high risk for the disease met with "a barrage of criticism." Specific failures cited were the early termination of the study and the alleged inadequate consideration of the risks of the drug, including pulmonary emboli, ovarian cancer, and depression, versus its benefits in premenopausal women.[14, 26] As principal investigator Bernard Fisher notes, the tamoxifen trial "obtained data that supported the concept being tested . . . the concept of breast cancer prevention has become a reality"—a significant finding somewhat lost in the controversy.[14]

The randomized controlled trial, has evolved to become both a standard and a tool of clinical evaluation. As a tool, it offers great precision in defining and assessing those concepts it is designed to test and considerable ease in extending their application to other settings. It may, however, either obscure or clarify for the astute observer those concepts which are less well adapted to statistical analysis: questions of quality of life, of special patient needs, or of subjective experience. Although a well-conducted trial may indicate a clear therapeutic preference, physicians (and their patients) can and do learn from thinking and arguing about the departures, the exclusions, and the standard deviations. The RCT helps impose order on medical knowledge by defining some areas as rationally based and worthy of consensus. From this starting point, the exploration of other, still chaotic or disputed areas can begin.

As have other implements in the medical toolbox—the stethoscope, the electrocardiograph, ultrasound scanning or MR imaging—the clinical trial has been adapted and refined to new uses, has acquired its own technicians and operating norms, and has gained its own place in medical culture. As with these other tools, however, real function of the RCT has been, and remains, to assist, not to replace, clinical judgment.

References

1. Amberson JB, McMahon BT, Pinner M: A clinical trial of sanocrysin in pulmonary tuberculosis. American Review of Tuberculosis 24:401–435, 1931
2. Beecher HK: The powerful placebo. JAMA 159:1602–1606, 1955
3. Chalmers I: The Cochrane Collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. Ann N Y Acad Sci 703:156–163, 1993
4. Daemmrich A: Drug cultures: Pharmaceutical regulation and social identity in the United States and Germany [dissertation]. Ithaca, NY, Cornell University, 2000
5. Doll R: Controlled trials: The 1948 watershed. BMJ 317:1217–1220, 1998
6. Early Breast Cancer Trialists' Collaborative Group: Systemic treatment of early breast cancer by hormonal, cytoxic, or immune therapy: 33 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Lancet 339:1–15, 71–85, 1992
7. Epstein S: Activism, drug regulation, and the politics of therapeutic evaluation in the AIDS era. A case study of ddC and the 'surrogate markers' debate. Social Studies of Science 27:691–726, 1997
8. Epstein S: The construction of lay expertise: AIDS activism and the forging of credibility in the reform of clinical trials. Science Technology and Human Values 20:408–437, 1995
9. Edwards CC: Hearing regulations and regulations describing scientific content of adequate and well-controlled scientific investigations. Federal Register 35:7250–7253, 1970
10. Evidence-based Medicine Working Group: Evidence-based medicine: A new approach to teaching the practice of medicine. JAMA 268:2420–2425, 1992
11. Feinstein AR, Horwitz RI: Problems in the "evidence" of "evidence-based medicine." Am J Med 103:529–535, 1997
12. Fergusson F: Letter to the editor. JAMA 214:763, 1970
13. Ferraro S: The anguished politics of breast cancer. New York Times Sunday Magazine: August 15, 1993, pp 25–27, 58–62
14. Fisher B: National Surgical Adjuvant Breast and Bowel Project breast cancer prevention trial: A reflective commentary. J Clin Oncol 17:1632–1639, 1999
15. Fisher B, Anderson S, Redmond CK, et al: Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer. N Engl J Med 333:1456–1461, 1995
16. Geison G: The Private Science of Louis Pasteur. Princeton, NJ, Princeton University Press, 1995
17. Greenhalgh T: Is my practice evidence-based? BMJ 313:957–958, 1996
18. Hill AB: Principles of medical statistics. I. The aim of the statistical method. Lancet 1:41–43, 1937
19. Hinshaw HC, Feldman WH: Evaluation of chemotherapeutic agents in clinical trials: A suggested procedure. American Review of Tuberculosis 50:202–213, 1944
20. Hróbjartsson B, Gótzsche PC, Gluud C: The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria. BMJ 317:1243–1245, 1998
21. Kornblum A: Are women being ignored? Newsday January 16, 1990, Section C, p 5
22. Ley HL Jr: Hearing procedure for refusal or withdrawal of approval of new drug applications and for issuance, amendment, or repeal of antibiotic drug regulations: Interpretative description of adequate and well-controlled clinical investigations. Federal Register 34:14596–14598, 1969
23. Lilienfeld AM: Ceteris Paribus: The evolution of the clinical trial. Bull Hist Med 56:1–18, 1982
24. Lind J: A Treatise of the Scurvy. Edinburgh, Sands, Murray, and Cochran, 1753
25. Louis PCA: Researches on the Effects of Bloodletting in Some Inflammatory Diseases and on the Influence of Tartarized Antimony and Vesication in Pneumonitis. Boston, Hilliard & Gray, 1836
26. Love RR: The National Surgical Adjuvant Breast Project (NSABP) breast cancer prevention trial revisited. Cancer Epidemiol Biomarkers Prev 2:403–407, 1993

27. Mamounas EP: Overview of National Surgical Adjuvant Breast Project neoadjuvant chemotherapy studies. Semin Oncol 25 (suppl 2):31–35, 1998
28. Marks HM: Notes from the underground: The social organization of therapeutic research. *In* Maulitz RC, Long DE, (eds): Grand Rounds: One Hundred Years of Internal Medicine. Philadelphia, University of Pennsylvania Press, 1988, pp 297–336
29. Marks HM, The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1980. Cambridge, Cambridge University Press, 1997
30. McKinlay J: From 'promising report' to 'standard procedure': Seven stages in the career of a medical innovation. Milbank Memorial Fund Quarterly 59:374–411, 1981
31. Meldrum ML: 'A calculated risk': The Salk polio vaccine field trials of 1954. BMJ 317:1233–1236, 1998
32. Meldrum ML: Departures from the design: The randomized clinical trial in historical context, 1946–1970 [dissertation]. Stony Brook, NY, State University of New York Stony Brook, 1994
33. Miller R, Feingold A, Paxinos J: Propoxyphene hydrochloride: A critical review. JAMA 213:996–1006, 1970
34. Modell W, Houde RW: Factors influencing the clinical evaluation of drugs, with special reference to the double-blind technique. JAMA 167:2190–2199, 1958
35. Moertel CG, Ahmann DL, Taylor WF, et al: A comparative evaluation of marketed analgesic drugs. N Engl J Med 286:813–815, 1972
36. National Research Council, Division of Medical Sciences: Final Report of the Drug Efficacy Study to the Commissioner of Food and Drugs, Food and Drug Administration. Washington, DC, National Academy of Sciences, 1969
37. "O Pioneers!" New Yorker 30:24–25, 1954
38. Porter TM: Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton, NJ, Princeton University Press, 1995
39. Reed W: The propagation of yellow fever; observations based on recent researches. Medical Record 60:1901
40. Rothman DJ, Edgar H: Scientific rigor and medical realities: Placebo trials in cancer and AIDS research. *In* Fee E, Fox DM (eds): AIDS: The Making of a Chronic Disease. Berkeley, University of California Press, 1992, pp 194–206
41. Semmelweis I: The Concept of Childbed Fever. (Die aetiologie, der begriff, und die prophylaxis des kindbettfiebers) (Pest, Hatleben, 1861), KC Carter (ed and trans), Madison, WI, University of Wisconsin Press, 1983
42. Shapiro AK, Shapiro E: The Powerful Placebo: From Ancient Priest to Modern Physician. Baltimore, MD, Johns Hopkins University Press, 1997
43. Snow J: On the Mode of Communication of Cholera. London, J. Churchill, 1849
44. Sollmann T: Experimental therapeutics. JAMA 58:242–244, 1912
45. Sollmann T: The evaluation of therapeutic remedies in the hospital. JAMA 94:1279–1281, 1930
46. Temin P: Taking Your Medicine: Drug Regulation in the United States. Cambridge, MA, Harvard University Press, 1980
47. University Group Diabetes Program: A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II. Mortality results. Diabetes 19 (suppl 2):789–830, 1970
48. Upjohn Company v. Finch. 422 F2d 944 (1970):944–968
49. Watkins E: On the Pill. Baltimore, MD, Johns Hopkins University Press, 1998
50. Wyckoff J, DuBois EF, Woodruff IO: The therapeutic value of digitalis in pneumonia. JAMA 95:1243–1249, 1930
51. Yoshioka A: Use of randomization in the Medical Research Council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. BMJ 317:1220–1223, 1998

Address reprint requests to
Marcia L. Meldrum, PhD
18311 Lost Knife Circle
#202
Montgomery Village, MD 20886