Clinical Trials, Overview

Trial is from the Anglo–French *trier*, meaning *to try*. Broadly, it refers to the action or process of putting something to a test or proof. *Clinical* is from *clinic*, from the French *cliniqué* and from the Greek *klinike*, and refers to the practice of caring for the sick at the bedside. Hence, narrowly, a *clinical trial* is the action or process of putting something to a test or proof at the bedside of the sick. However, broadly it refers to any testing done on human beings for the sake of determining the value of a treatment for the sick or for preventing disease or sickness.

The broad definition of clinical trial includes definitions allowing for use of the term in references to studies involving a single treatment (e.g. as in most
CAP002- **Phase I trials** and some **Phase II drug trials**) and
CAP003- for studies involving use of an external control (e.g. studies involving historical controls) [66]. However, use herein will be in the stricter sense of usage; that is, to refer to trials involving two or more treatment groups comprised of persons enrolled, treated, and followed over the exact same time frame.

The *treatment* can be anything considered to hold promise in caring for the sick, in the prevention of disease, or in the maintenance of health. The term, in the context of a trial, refers to the experimental variable – the variable manipulated by the trialist. The variable may have just two states (e.g. as in a trial involving a single test treatment and single control treatment) or three or more states (e.g. as in a trial involving several different test treatments and one or more control treatments). The variable, in the case of drug trials, may serve to designate different drugs, different doses of the same drug, or different forms or routes of administration of the same drug. In other contexts, it may variously refer to different kinds or forms of surgery, different kinds or forms of care or management regimens, different kinds or forms of diagnostic tests, different kinds or forms of medical devices, different kinds or forms of counseling regimens to achieve some desired end, or combinations of the above.

The clinical trial, in its simplest form, involves the application of the experimental variable – treatment to a person or group of persons – and observation during or following application of the treat-
CAO001- ment to measure its effect. That measure **(outcome measure)** may be death, occurrence or recurrence of some morbid condition, or a difference indicative of change (e.g. difference in blood pressure measured for each person just prior to the start of treatment and again at some point during or after treatment).

There is no way to "test" a treatment or to "prove" its effectiveness in the absence of some absolute or relative measure of success. Trials are said to be *controlled* if the effect of a treatment is measured against a comparison treatment administered over the same time period and under similar conditions. That comparison treatment may be another test treatment or, depending on circumstances, a *control treatment* consisting of an accepted standard form of therapy, a placebo (*see* **Blinding or Masking**) or sham treatment, or observation only (no treatment).

A trial is said to be *uncontrolled* if it does not have a comparison treatment or if the enrollment to and administration of the test and comparison treatments is not concurrent (e.g. as with use of *historical controls* for evaluation of a treatment). The Book of Daniel (Chapter 1, verses 12–15) provides an account of what amounts to an uncontrolled trial involving a diet of pulse – edible seeds of certain pod-bearing plants, such as peas and beans (*see* **History, Overview**).

> Prove thy servants, I beseech thee, ten days; and let them give us pulse to eat, and water to drink. Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the King's meat: and as thou seest, deal with thy servants. So he consented to them in this matter, and proved them ten days. And at the end of ten days their countenances appeared fairer and fatter in flesh than all the children which did eat the portion of the King's meat [1].

Fortuitous events can produce conditions reminiscent of the features of a trial. One such account is that given by Ambroise Paré (surgeon, 1510–1590) during the battle in 1537 for the castle of Villaine. The treatment for gunshot wounds in Paré's time was boiling oil poured over the wound. Because of the intensity of the battle, Paré ran out of oil and resorted to using an ointment made of egg yolks, oil of roses, and turpentine. The result of his "trial" is summarized by his observation the morning after the battle:

> I raised myself very early to visit them, when beyond my hope I found those to whom I had applied the digestive medicament, feeling but little pain, their

> wounds neither swollen nor inflamed, and having slept through the night. The others to whom I had applied the boiling oil were feverish with much pain and swelling about their wounds. Then I determined never again to burn thus so cruelly the poor wounded by arquebuses [72].

Many of the essential elements of the modern day *controlled trial* are contained in Lind's account of a trial performed aboard the *Salisbury* at sea in 1747:

> On the 20th of May 1747, I took twelve patients in the scurvy, on board the *Salisbury* at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz., watergruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times puddings, boiled biscuit with sugar, etc; and for supper, barley and raisins, rice and current, sago and wine, or the like. Two of these were ordered each a quart of cyder a day. Two others took twenty-five gutts of elixir vitriol three times a day, upon an empty stomach; using a gargle strongly acidulated with it for their mouths. Two others took two spoonfuls of vinegar three times a day, upon an empty stomach; having their gruels and their other food well acidulated with it, as also the gargle for their mouth. Two of the worst patients, with the tendons in the ham rigid, (a symptom none of the rest had), were put under a course of seawater. Of this they drank half a pint every day, and sometimes more or less as it operated, by way of gentle physic. Two others had each two oranges and one lemon given them every day. These they eat with greediness, at different times, upon an empty stomach. They continued but six days under this course, having consumed the quantity that could be spared. The two remaining patients, took the bigness of a nutmeg three times a-day, of an electuary recommended by an hospital surgeon, made of garlic, mustard-seed, rad raphan, balsam of Peru, and gum myrrh; using for common drink, barley-water well acidulated with tamarinds; by a decoction of which, with the addition of cremor tartar, they were gently purged three or four times during the course. ... the most sudden and visible good effects were perceived from the use of oranges and lemons, one of those who had taken them being at the end of six days fit for duty [62].

The Treatment Protocol

The treatment protocol (the general term, *study protocol* or *trial protocol* (*see* **Protocols**) has broader meaning and refers to the constellation of activities involved in conducting a trial) of the trial specifies the treatments being studied, the manner and method of usage and administration, and conditions under which other treatments are called for when needed for the well-being of those enrolled. The treatment may be administrated in one application or multiple applications. The period of treatment may be short (e.g. as in trials involving a single application of treatment such as surgery) or extended (e.g. as in trials involving the treatment of a chronic condition with drugs) over a period of weeks, months, or years. The treatment, in the case of drug trials, may involve a fixed dose administered according to some schedule or dose titration in which each person ultimately receives the amount needed to achieve a desired effect (e.g. the amount of a hypoglycemic agent needed to bring blood glucose levels to within the normal range).

Protocols for all research involving human beings are subject to review and approval by institutional review boards (IRBs) or ethics review boards (ERBs) before implementation and at periodic intervals thereafter until the research is finished (*see* **Ethics**). Therefore, investigators undertaking trials have the obligation and responsibility to obtain IRB or ERB review and approval prior to initiation of a trial, and to seek its review and approval prior to implementing amendments to the protocol of the trial. They also have a responsibility to inform IRBs and ERBs of record of any untoward events in the conduct of the trial and to report to such boards any conditions or events believed to change the risk–benefit ratio for persons enrolled into the trial or still to be enrolled.

Only patients judged eligible (as determined by specified **eligibility criteria**) may be enrolled, and among those, only those who consent to participate in the trial. Persons are under no obligation to enroll or to continue once enrolled, and must be so informed prior to being enrolled. A person must be informed, as well, of what is entailed by enrollment, of the risks and benefits that may accrue by enrollment, and of such matters and details that might cause a reasonable person to decline enrollment when so informed (e.g. that treatments are randomly assigned and that they will be administered in masked fashion).

CAE001-

All trials involve data collection at various time points over the course of enrollment and follow-up of persons. The amount collected per person depends on the nature of the disease or condition being treated

and on the nature of the treatment process implied by the study treatments being used. The requirement for repeated observation of a person, as a rule (except for trials done in hospital or other settings involving resident populations or in which enrollment, follow-up, and treatment is directed or managed by telephone or mail), obligates a person to a series of visits to the study site. Usually, the purpose of the first visit or series of visits will be to determine eligibility, collect necessary baseline data, obtain consent, and initiate treatment. Visits thereafter will be to fine-tune or continue treatment and to collect necessary follow-up data. The schedule of follow-up visits
CAR001- will be timed from the point of **randomization** or initiation of treatment and, as a rule, will be on a defined time schedule (e.g. once every week) with provisions for interim (unscheduled) visits when necessary for the care of those enrolled.

Comparison of the different treatments tested for effect is done in different ways depending on the outcome measures used to assess effect. The comparison, in the case of an event, such as death or occurrence of a morbid event, will be based on the event rate (or the raw percentage of persons experiencing the event) as seen for the different treatment groups. In the case of a continuous variable, such as weight or blood pressure, the change from entry to some defined point after enrollment will be determined for each person studied and then summarized in some fashion (e.g. by calculating a mean or median). The treatment effect will be estimated by the difference obtained by subtracting the summary measure for the comparison treatment from the indicated test treatment.

Judging the safety or efficacy of a treatment is problematic in trials not involving a designed comparison group – often the case in Phase I, II, and I/II trials (see below for definitions). The problem is compounded by the typically short duration and small size of these trials. The problem is most acute in the testing of drugs in people having a life-threatening disease when the drugs themselves carry their own morbidity and increased risk of death. Are the morbid events observed the result of the disease or the drug? Even deaths become difficult to interpret in the presence of a high background death rate from the disease. Was a death the natural outcome of the disease, or was it induced by the treatment? The issue is rarely clear until sufficient information has accumulated to cause one to discount natural causes as the likely explanation, or to allow one to recognize an unusual clustering of deaths and morbid events, as with the case of a trial of fialuridine (FIAU) [63].

Classes of Trials

Most clinical trials involve *parallel treatment designs*, i.e. designs where an assignment unit (usually a person) is assigned to receive only one of the treatments under study. The word *parallel* indicates that two or more groups of assignment units are proceeding through the trial side by side, with the only ostensible difference (other than baseline differences in the composition of the groups) being the treatment administered. The goal in trials with parallel treatment designs is for each person enrolled to receive the assigned treatment and to have no exposure to any of the other treatments under study in the trial (except where the requirements for proper care are overriding and make such exposure necessary).

The assignment unit (randomization unit in randomized trials), in the case of parallel treatment designs, is usually a person but can be an aggregate of persons (e.g. members of the same household) (*see* **Group-randomization Designs**) or a subpart of a person (e.g. an eye, as in the Glaucoma Laser Trial [34].

The treatment design in **crossover** trials is differ- CAC014-
ent. In this class of designs a person or treatment unit receives two or more study treatments in a specified order. Crossover trials are classified by the number of treatments to be administered to a person or treatment unit and by whether a given person or treatment unit receives all (complete or full crossover) or just some (partial or incomplete crossover) of the study treatments. For example, a two-period crossover design is one in which each person or treatment unit receives two study treatments in some order, usually random. An n-way crossover design is one in which a person or treatment unit receives n of the treatments represented in the design.

The utility of crossover designs is limited to settings in which it is feasible to administer different treatments to the same person or treatment unit, each for a short period of time, and in which it is possible to measure the effect at the end of each treatment period. They are not useful in settings in which the outcome of interest is a clinical event that can occur at any time after enrollment.

In a trial with a parallel treatment design, assignment determines the treatment to be administered

(except to the extent that other treatments are needed for proper care) whereas, in a crossover trial, assignment determines the order of treatments to be used. Typically, each treatment is administered for a designated period of time (e.g. 4 weeks). Often the last administration of one treatment and the first administration of the next treatment are separated in time (e.g. 1 week) to allow the effect of the preceding treatment to "wear off" ("washout period") before administering the next treatment.

Imagine a trial involving three study treatments (A, B, and C) with the same (uniform) assignment probabilities and 54 people. In a trial with a parallel treatment design, 18 people would be assigned to receive treatment A, 18 would be assigned to receive treatment B, and 18 would be assigned to receive treatment C. In a trial involving a complete (full) crossover of treatments, each of the 54 people would receive treatments A,B, and C. Assuming treatments are arranged in all possible orderings, there would be six different orderings of the treatments (ABC, ACB, BAC, BCA, CAB, and CBA), and nine patients would be randomly assigned to receive a given ordering.

While the goal of the two designs is the same, to find the most effective treatment, the methodology differs. With the parallel treatment design, the treatment is evaluated in comparable groups of treatment units (usually persons), and with the crossover treatment design, the treatment effect is evaluated within the same treatment unit (usually a person).

Trials involving parallel treatment designs are of two general types with regard to sample size design – fixed or sequential. The majority are of the fixed type. That is, the sample size is specified at the outset, as determined by pragmatic considerations (e.g. by the amount of money available for the trial) or by a formal **sample size** calculation. Trials are considered to have a fixed sample size even if they do not proceed to the desired sample size, e.g. are stopped early because of a treatment difference. The sample size is fixed in the sense that the intent is to enroll and follow the specified number of assignment units unless indicated otherwise by events transpiring during the course of the trial.

CAS002-

In sequential trials (also of two types – open and closed), enrollment and observation continue until a stopping boundary, constructed for the outcome of primary interest (usually a binary "success" or "failure" type event), is crossed. Open sequential designs involve two boundaries, one indicative of superiority and the other indicative of inferiority of a test treatment relative to a comparison treatment. Enrollment continues until the observation function for the outcome measure of interest crosses one of the two boundaries. The design has the advantage of providing a test of the null treatment hypothesis for given type I and II error levels that, on average, requires a smaller sample size than that for a fixed sample size design.

However, the actual sample size required for a boundary crossing can be larger (in theory, sometimes much larger) than that for a fixed sample size design. The possibility of the final sample size being much larger is ruled out with the closed sequential design. That design, in addition to the two boundaries mentioned above, involves a third boundary serving to place an upper bound on enrollment. If that boundary is crossed, because neither of the other two boundaries is crossed (signifying a difference in favor of one of the treatments), then the treatments being compared are considered to be of equivalent value as measured by the outcome observation function.

Sequential designs have limited utility in the context of clinical trials, partly because they require rigid adherence to a stopping rule. Use is limited to instances where the "success" or "failure" of a treatment can be determined shortly after administration. They are not useful in settings involving long-term treatment and with outcome measures requiring weeks, months, or years of observation. In general, more flexible methods of monitoring trials are more appropriate (*see* **Data and Safety Monitoring**).

Drug Trials

Compounds, no matter how promising or impassioned the pleas for use, have to go through a series of tests in animals before they can be tested in humans. Those considered to lack promise after animal testing do not come to testing in humans.

Typically, the testing in humans is done in a time-ordered sequence, as suggested by the phase label affixed to trials as defined below. However, in truth, adjoining phases overlap in purpose. Hence, the label, at best, serves only as a rough indicator of the stage of testing, especially when, as is often the case, drug sponsors, at any given point in time, may have several trials under way carrying different phase labels. The

definitions of the different phase labels follow:

Phase I: Usually the first stage of testing performed in anticipation of an Investigational New Drug Application (INDA); done to generate preliminary information on the chemical action and safety of the indicated drug and to find a safe dose; usually not randomized.

Phase II: Usually the second stage of testing; generally carried out on persons having the disease or condition of interest; done to provide preliminary information on efficacy of the drug and additional information on safety; may be designed to include a control treatment and random assignment of patients to treatment.

Phase I/II: A trial having some of the features of Phase I and II trials; designed to provide preliminary information on safety and efficacy.

Phase III: Usually the third and final stage in testing, prior to submission of an INDA; concerned with assessment of dosage effects, efficacy, and safety; usually designed to include a control treatment and random assignment to treatment. When the test is completed (or nearly completed), the drug manufacturer or sponsor may request permission to market the drug for the indication covered in the testing by submission of an INDA.

Phase II/III: A trial having some of the features of Phase II and III trials; designed to provide information on safety and efficacy.

Phase IV: A fourth stage of testing, sometimes carried out. Usually controlled and performed after approval of the INDA. Typically done under circumstances approximating real-world conditions; usually has a clinical event as a basis for sample-size calculation and provides for extended treatment (where appropriate) and long-term follow-up, with efficacy and safety of the drug being measured against a control treatment.

Drugs, after marketing approval, remain under surveillance for serious adverse effects. The surveillance – broadly referred to as **postmarketing surveillance** – involves the collection of reports of adverse events via systematic reporting schemes and via sample surveys and observational studies.

CAP005-

Sample size tends to increase with the phase of the trial. Phase I and II trials are likely to have sample sizes in the 10s or low 100s compared to 100s or 1000s for Phase III and IV trials.

The focus shifts with phase. The aim in the early phases of testing is to determine whether the drug is safe enough to justify further testing in human beings. The emphasis is on determining the toxicity profile of the drug and on finding a proper, therapeutically effective dose for use in subsequent testing. The first trials, as a rule, are uncontrolled (i.e. do not involve a concurrently observed, randomized, control-treated group), of short duration (i.e. the period of treatment and follow-up is short), and conducted to find a suitable dose (usually via some traditional or **Bayesian** dose escalation design) for use in subsequent phases of testing. Trials in the later phases of testing, for the most part, involve traditional parallel treatment designs, randomization of patients to study treatments, a period of treatment typical for the condition being treated, and a period of follow-up extending over the period of treatment and beyond.

CAB001-

Most drug trials are done under an INDA held by the sponsor of the drug. The "sponsor" in the vernacular of the Food and Drug Administration (FDA) is typically a drug company, but can be a person or agency without "sponsorship" interests in the drug. Regulations require investigators to report adverse events to the FDA. The general guidelines regarding consent are similar, but not identical, to those promulgated by the Office for Protection from Research Risks (OPRR) for IRBs.

The Randomized Trial

A *randomized trial* is a trial having a parallel treatment design in which treatment assignment for persons (treatment units) enrolled is determined by a randomization process similar to coin flips or tossings of a die (*see* **Randomized Treatment Assignment**). The trialist's purpose in randomization is to avoid selection bias in the formation of the treatment groups. The bias is avoided because the treatment

to which a person is assigned is determined by a process not subject to control or influence of the person being enrolled or those responsible for recruiting and enrolling the person. The comparison of one group to another for treatment effect will be biased if, for whatever the reason, one group is "healthier" or "sicker" on entry than the other. Schemes in which one knows or can predict treatment assignments in advance of issue are open to such bias. Clearly, that is the case with assignment schemes posted in a clinic and open for all to see prior to issue. The bias is likely as well with systematic schemes, such as those in which every other person is assigned to the test treatment or in which persons seen on odd-numbered days receive the test treatment and those seen on even-numbered days receive the control treatment.

The goal is to create groups that provide a valid basis for comparison. To achieve that end one has to ensure that the groups are similar (within the range of chance) and to avoid bias in the assignment process. The usual method for achieving both ends is randomization.

Randomization does not guarantee comparability of the treatment groups with regard to the various entry characteristics of interest. Indeed, one can, by chance, have differences among the treatment groups. A large difference (one yielding a small P value) can arise by chance and, hence, cannot be taken as prima facie evidence of a "breakdown" (e.g. "peeking" or other purposeful acts aimed at determining assignment before issue) of the randomization process, unless supported by other evidence of a "breakdown".

The hallmarks of a sound system of randomization are: reproducible order of assignment; documentation of methods for generation and administration of assignments; release of assignments only after essential conditions satisfied (e.g. only after a person has been judged eligible and has consented to enrollment); masking of assignments to all concerned until needed; inability to predict future assignments from past assignments; clear audit trail for assignments; and the ability to detect departures from established procedures [67] (*see* **Audit and Quality Control**).

The randomization may be simple (complete) or restricted. The purpose of restriction is to force the assignments to satisfy the specified assignment ratio at intervals during enrollment. Those restrictions are typically referred to as blocking. For example, suppose a trial involves two treatments, A and B, and the desired assignment ratio is one-to-one. A simple (unrestricted) randomization scheme would involve the equivalent of repeated flips of an unbiased coin with a head leading to assignment to treatment A and a tail to treatment B. The design would, on average, yield the desired assignment ratio, but allows for wide departures from the desired mix, depending on the "luck" of the flips.

If such departures are of concern, then the randomization scheme can be restricted by blocking so as to ensure the desired mix after a specified number of assignments. For example, imposition of a blocking requirement after every eighth assignment would have the effect of "forcing" the randomization to yield the desired mix of one-to-one after every eighth assignment. The purpose of the blocking is to ensure a near desired assignment ratio so as to protect treatment comparisons against secular trends in the mix of patients as the trial proceeds.

The randomization also may be stratified. The purpose of stratification is to provide treatment groups comprised of persons or treatment units having identical (within the limits of the stratification) distributions of the stratification variable. It is useful only in so far as the variable used for stratification serves to influence or moderate the outcome of interest. The stratification has the effect of "controlling" the influence of the stratification variable on outcome by ensuring the same distribution of the variable across the different treatment groups. For example, suppose one wishes to stratify on gender in the trial described above (because, perhaps, of a belief that the treatment effect will be different in women than in men). The stratification would be achieved by creating two randomizations schedules, each with a one-to-one assignment ratio and with blocking to satisfy the assignment ratio after enrollment of the 8th, 16th, 24th, etc. person in each stratum. The effect of the stratification would be to ensure the same gender mix (within the limits of the blocking) for the two treatment groups, regardless of the underlying gender mix of the population to be studied. For example, suppose 96 patients are to be enrolled from a population with a 1:2 mix of males to females. In that case, one would expect to enroll 32 males and 64 females, and to have 16 males and 32 females in each treatment group. If the underlying mix is one-to-one, then there would be 24 males and 24 females in each of the two treatment groups.

Clearly, the number of variables that can be controlled by stratification is limited. The more variables, the more subgroups for randomization and the less useful the process is as a reliable means of variance control [35]. In addition, there are logistic difficulties associated with use of variables whose values have to be determined by performing laboratory tests or other diagnostic procedures during the enrollment process. Even if one stratifies on a few selected variables, other variables may well be considered to be important determinants of outcome. Hence, the experienced trialist strives to "remove" the effect of such differences via analysis procedures, e.g. by assessing the treatment effect within defined subgroups (subgroup analysis, *see* **Treatment–Covariate Interaction**); or by providing estimates of treatment effect that are adjusted for differences in the distribution of important demographic or baseline variables via regression procedures [90].

Masking

Masking is the purposeful concealment of some fact or condition and is done to keep knowledge of that fact or condition from influencing the behavior, observation, or reporting of persons so masked. Masking, in the context of trials, is imposed to reduce the likelihood of a treatment-related bias due to knowledge of treatment assignment (*see* **Blinding or Masking**).

That bias, after a person is enrolled, occurs whenever knowledge of that person's treatment assignment serves to color the way he or she is treated, followed, or observed. One way of reducing it is by masked treatment administration. In one form of such administration (single-masked), only one member of the subject–treater pair is masked to treatment, usually the subject. Another form of masking is one in which both members of the pair are masked – double-masked treatment administration. As a rule, double-masked treatment administration means that all persons in a clinic are masked and, therefore, that those responsible for data collection and generation are masked to treatment as well.

Generally, it is not possible or prudent to mask treatment administration in trials involving treatments requiring different routes or modes of administration (e.g. as in a trial involving a medical vs. a surgical form of treatment), where knowledge of treatment assignment is part of the effect being tested (e.g. as in trials aimed at modification of one's eating habits via different modes of dietary consulting), or where the masking carries risks for those enrolled. Therefore, the opportunities for double-masked treatment administration are limited largely to trials of drugs considered safe and that are reasonably free of side-effects and that can be administered at fixed dose levels. It is usually not wise or practical to administer treatments in a double-masked fashion when treatment doses are to be titrated to achieve desired effects.

Masked treatment administration has been used as a mark of "quality" for trials. There is, therefore, a tendency to view results from masked trials as more reliable than those from unmasked trials. In truth, however, masked treatment administration is rarely 100% effective. All forms of treatment, and especially those involving drugs, produce side-effects and telltale signs that may serve to unmask treatment. Hence, the protection provided by masking can be illusory. As a result, it is better to make assessments of "quality" in terms of the risk of treatment-related bias and the likely effect of such bias, if present, on the results reported. The risk of treatment-related bias is low for "hard" outcome measures and with explicitly defined treatment protocols, even in the absence of masked treatment administration.

The second line of defense, in the absence of double-masked treatment administration, is to mask as many groups of persons involved in the trial as is possible within the limits of practicality and safety. Hence, even if it is not possible to mask patients or those who treat them, it may be possible to mask those responsible for data collection or data generation (e.g. as with an arrangement as in the Glaucoma Laser Trial [34], where intraocular pressure was measured by masked readers, or as with laboratory personnel or readers of X-rays, ECGs, or fundus photographs masked to treatment assignment).

With or without treatment masking, trialists strive for objectively defined treatment and data collection procedures and for outcome measures as free from observer or respondent bias as is humanly possible. In addition, they are inclined toward continuing effort over the course of a trial aimed at maintaining the training and certification of study personnel in regard to required study procedures, and toward establishing and maintaining standards of performance via ongoing monitoring and quality control surveillance (*see* **Audit and Quality Control**).

Analysis

The protection provided against treatment-related bias by the assignment process is futile if the analysis is biased. Treatment comparisons, to be valid, must be based on analyses that are consistent with the design used to generate them. In the case of the randomized trial, this means that the primary analyses of the outcomes of interest must be by assigned treatment (also
CAI001- known as *analysis by* **intention-to-treat**). It means, for example, that observations relating to a morbid event are counted to a patient's assigned treatment regardless of whether or not the patient was still on the assigned treatment when the event occurred.

Analyses involving arrangements of data related to treatment administered may be performed, but only as supplements to the primary analyses. They should not and cannot serve as replacements for those analyses.

Analyses by treatment assignment, as a rule, serve to underestimate the treatment effect. Usually, analyses in which the requirement is relaxed will yield a larger estimate of the treatment difference than seen when evaluated under the intention-to-treat mode of
CAU001- analysis (e.g. as in the case of the **University Group Diabetes Program (UGDP)** trial) [90].

Designs allowing for termination of data collection when a person can no longer receive or be maintained on the assigned treatment are open to treatment-related bias. The goals of the primary analyses cannot be met when data collection for a person ceases when that person experiences a nonfatal "endpoint" or when the person's treatment is stopped or changed. The analysis requirement implies continued follow-up of all persons enrolled into a trial to the scheduled close of follow-up regardless of their treatment or outcome status.

Monitoring Treatment Effects

The randomized trial depends on a state of equipoise – a state of legitimate doubt regarding the test treatment relative to the control treatment(s) [4, 30, 61]. It cannot be undertaken without a proper ethical climate characterized by such a state of doubt (*see* **Ethics**). It does not matter whether that state has been dispelled by observation and data, by declaration, or in other ways. For example, it would not be possible to assess the value of coronary care units for persons appearing to be having a myocardial infarction (MI), even if their value has not been demonstrated by controlled trials. They are considered to be required for good care and, hence, the window of opportunity for testing via designed randomized trials has closed. Once closed, it may remain closed, or may open again years later if people start questioning the merits of the treatment. When the oral hypoglycemic agents appeared on the scene in the early 1950s, they were widely regarded as safe and effective and, hence, became a part of the armamentarium for care of the adult-onset diabetic. However, doubts raised in the late 1950s as to their value led to a climate of doubt suitable for initiation of the UGDP trial [89].

Trials are done because of the prospect of benefit associated with a new treatment, or to test the efficacy of an existing treatment. They are not undertaken to prove a treatment to be useless or harmful. Indeed, a trialist is obligated to stop a trial prior to its scheduled completion if the accumulated data indicate that the treatment of interest is inferior to the control or comparison treatment. In fact, some argue that there is an obligation to stop if it becomes clear that the test treatment is no better than the comparison treatment, even if one is uncertain whether it is harmful. Hence, for example, investigators in the UGDP opted to stop use of tolbutamide in that trial once they were certain it was no better than the control treatment – the usual antidiabetic dietary recommendations and placebo medication.

The need for ongoing monitoring exists for any trial in which the treatments carry risk of harm, and in which it is possible to reduce that risk by timely monitoring (*see* **Data and Safety Monitoring**). That need makes it necessary for the trialist to aim for an orderly and timely flow of data from the site of generation or collection to the processing and analysis site. Clearly, the best systems in this regard are those having real-time or near-real-time flows (e.g. as with systems requiring transmission of data related to a patient visit on completion of the visit or on occurrence of an outcome of interest).

Typically, treatment effects monitoring is entrusted to a group of people that together have the necessary skills and expertise to monitor effectively (*see* **Data and Safety Monitoring Boards**) [38, 67, 70, 94]. The group is usually comprised of 5–12 people with expertise in the disease under treatment, in the design, conduct, and analysis of clinical trials, or in other specialty areas. When the group comprises

a mix of people from within the trial (e.g. the officers of the trial, such as the chair and vice chair, the director of the coordinating center, etc.) and outside the trial, the votes concerning recommendations for change, generally, are vested in those outside the study. The restriction is imposed, typically, because of concerns that persons associated with the trial may have conflicts of interest that could serve to influence their votes [17].

Monitoring proceeds under different constructs, depending on the philosophy of those doing the monitoring. Some constructs require stopping rules and restrictions on the type of data that may be monitored and the number of interim "looks" that can be made in relation to the monitoring. Other groups consider such restrictions unnecessary and rely instead on the collective judgment of the monitoring group [28].

The OPRR (an office within the National Institutes of Health (NIH) responsible for the promulgation and administration of regulations regarding institutional review boards) and the set of rules relating to research on human beings obligates IRBs to be satisfied that risk to subjects is minimized. As part of this assurance in regard to clinical trials, investigators must have "adequate provision for monitoring the data collected to ensure the safety of subjects" (Section 46.111) [71] and must provide participants with information on ... "significant new findings developed during the course of the research which may relate to the subject's willingness to continue participation will be provided to the subject" (Section 46.116) [71]. This requirement makes it necessary to inform patients of results during the trial that bear on their willingness to continue. This requirement pertains to information from inside or outside the study, if the information is likely to cause patients to reconsider their decision to be enrolled in the trial. Formal reconsent procedures may be required if the treatment effects monitoring committee recommends changes to the treatment protocol (e.g. as discussed in [67]).

Representativeness, Validity, and Generalizability

Representativeness, in the context of a trial, refers to the degree or extent to which those enrolled can be considered representative of the general population of persons to whom the treatment may be applied, if shown to be useful. *Validity*, in the context of a treatment difference, refers to the extent to which that difference can be reasonably attributed to treatment assignment. *Generalizability* refers to the degree to which the findings of the trial can be extended to the general population of eligible persons.

The concepts of validity and generalizability are different. Validity derives from the design of the trial and from the way it is carried out, whereas generalizability is largely a matter of judgment. A treatment comparison is valid if it is based on comparable groups of persons treated and observed in such a way so as to make treatment assignment the most likely explanation of the result observed. "Representativeness" is deduced by comparison of the demographic and other host characteristics of the study population to that of the general population of eligible persons (or by comparison with all persons screened for enrollment).

The desire for representativeness arises from the belief that conclusions from a trial will be strengthened by having a broadly "representative" study population. The drive for demographic representativeness has been propelled in recent years by the belief that women and persons of ethnic minorities have been "underrepresented" or "understudied" relative to men and the prevailing ethnic majority in trials and other areas of clinical research. Those concerns have been sufficient to cause the US Congress, in the NIH Revitalization Act of 1993, to impose requirements on trials aimed at ensuring adequate numbers of women and ethnic minorities to determine whether the treatments being studied in a trial work differently in men than in women or in an ethnic minority than in the ethnic majority [88].

There is no way to ensure "representativeness" in the absence of a sampling frame for the eligible study population and a related sampling scheme aimed at providing a representative sample of that population. However, even if one were able to develop a sampling frame (usually impossible because to do so one would have to screen the general population to identify persons eligible for study), the population ultimately enrolled, even if selected by sampling, would, at best, be representative only of those able and willing to be enrolled, because of the requirements of consent.

Hence, trials, by nature of their design, involve select, nonrepresentative populations. Even if a treatment is found effective in a trial, one has no direct way of knowing if it would be effective for those

patients not agreeing to be studied. If the issues of consent and lack of a sampling frame were overcome, then one would still be left with the fact that most clinics, for practical and ethical reasons, have to rely on those who come to them. They do not have the ability or moral authority to go and seek out suitable patients for study, especially if doing so means that those who routinely come to them would be turned away. Such a "selective" approach would be viewed as violating the principle of justice as set forth in the Belmont Report [69].

That one needs to generalize is obvious. The need arises in regard to the route of treatment, amount of treatment, type of treatment, and type of patients. For example, if a trial involved a single fixed dose of a drug and failed to find a difference (e.g. as in the UGDP trial, regarding tolbutamide) [90] does one conclude that use of the same drug, under a different, more flexible dosing scheme would produce a more favorable result? Similarly, if one compound produces a benefit, does one conclude that other sister compounds will show the same effect? Or conversely, if one member of a drug family has a bad effect (e.g. FIAU) or fails to show a benefit (e.g. tolbutamide), does one shy away from other related compounds? Also, if a trial involves mildly diseased people and shows a beneficial effect for the test treatment, does one conclude that the test treatment will have a similar effect in sicker people?

Last, if the drug tested works for the disease or condition being treated, is it not likely that it would be useful as well for a related condition or disease? So-called "off label" use (from the fact that drugs are approved for designated indications) accounts for a large number of treatment prescriptions [12, 36, 91, 93].

Whenever one generalizes, whatever the nature or direction, one is in effect answering one or more of the above questions. If as a treater, one chooses to use a sister compound of a drug shown to be ineffective in a trial, then one is in effect saying that the result from the trial, for whatever reason, is not generalizable. Generalizations depend on judgments regarding the trials and on prior beliefs regarding the treatment in question.

A trial can provide a valid basis for comparing one treatment to another if the differences in outcomes for the treatment groups being compared can be attributed reasonably to treatment. The general "laws of science" and "principles of parsimony" require that one defaults to the simplest explanation – usually the one requiring the fewest assumptions. Hence, in the case of the trial in which treatments are selected by the patient or physician, one is as a rule more inclined to attribute the difference to selection factors than to the test treatment. By the same principle, one should be more inclined to attribute a treatment difference to bias on the part of the observer rather than to the treatment when the opportunity for such bias exists. The degree of "reasonableness" of such an explanation will depend on the nature of the outcomes and whether one can reasonably ascribe it to biased observation. It becomes progressively more difficult to do so, even if the observer is not masked, the "harder" the outcome measure. For example, it is not reasonable to expect that one's opinion regarding the merits of a treatment will influence one's ability to report reliably whether a person is alive or dead, but such opinion may influence how one sees or reports on a person's quality of life. There is a responsibility on the part of trialists to rule out other lesser explanations of results before ascribing them to treatment.

Contrary to lay perceptions, trials and comparisons of treatments within the trial are made robust to selection bias and the consequences of "nonrepresentative" study populations by randomization. The assessment of treatment effect is achieved by having comparable groups of patients in the different treatment groups and by having procedures for observing and following patients that are independent of treatment assignment. The comparison is valid regardless of the study population and provides information on the relative value of one treatment to another. Hence, from the perspective of the trialist, it is far more important to have comparable treatment groups than to have "representative" treatment groups.

The drive for "representativeness", while perhaps of some social value, does little to make generalizations less risky or to increase the validity of trials. There are sound practical reasons to design trials with as few exclusions to enrollment as possible. The fewer the restrictions the easier and faster it is to recruit. Any effort to make them more "representative" by selective recruitment and enrollment will make them more costly and will increase the time required to enroll them. The imposition of recruitment quotas to achieve a desired sample size for gender, age, and ethnic origin groups poses a far more complicated and costly recruitment effort than one

involving the enrollment of all comers regardless of gender, age, or ethnic origin.

The goal of the trialist should be to strive for demographic neutrality in enrollment. That is to say, the trialist should not exclude potential participants on the basis of gender, ethnic origin, or age unless justified on scientific grounds. Scientific grounds include the knowledge or expectation of a qualitative treatment by demographic interaction (i.e. where treatment is believed to be beneficial for one demographic group and harmful for another) (*see* **Treatment–Covariate Interaction**).

Another reason for exclusion is contraindication of treatment in a particular demographic group. If any one treatment is contraindicated in a trial involving multiple treatments, then the restriction has to apply to all treatments. For example, this requirement was one of the reasons why the Coronary Drug Project (CDP) involved only men. Two of the five test treatments in the trial could not have been administered to premenopausal women; thus this demographic group could not be included without making the trial much more complex [22].

As a rule, an anticipated low number in a specified demographic group is not a reason to exclude. Disease and extent of disease are much more likely to affect the response to treatment than are "demographic" characteristics. Analyses of treatment effects across the various demographic subgroups represented in a trial can help determine whether there are treatment by demographic interactions. In general, interactions, when noted in the context of treatment trials, are more likely to relate to disease characteristics than to demographics [32, 42, 45, 46, 65].

Table 1 References on methods and procedures of clinical trials

Topic	References
Specialty journals	
Applied Clinical Trials	[2]
Controlled Clinical Trials	[37]
Statistical Methods in Medical Research	[85]
Statistics in Medicine	[19], [20]
Textbooks	
Clinical trials	[15], [33], [44], [47], [67], [75], [82]
Data analysis	[40], [41]
Ethics	[51], [60]
History	[87]
Dictionaries/Encyclopedias	
Clinical trials	[66]
Epidemiology	[58]
Statistics	[53]
Journal articles	
Analysis	[18], [24], [25], [50], [74], [83]
Bayesian methods	[6], [21], [31], [84]
Cost and efficiency	[95]
Design	[18]
Equipoise	[4], [30], [61], [76]
Ethics	[3], [4], [69]
Forms design and data management	[39], [81], [96], [97], [98]
History	[13], [62]
Meta-analysis and overviews	[5], [9], [14], [40], [48], [59], [86], [99]
Philosophy	[78], [79]
Randomization and stratification	[11], [35], [49], [56], [64], [73], [80], [92]
Sample size	[7], [8], [10], [27], [54], [55], [57], [77]
Subgroup analyses	[6], [23], [100]
Treatment effects monitoring	[3], [16], [26], [28], [29], [38], [43], [52], [59], [76]

The above list is due the efforts of Susan Tonascia, ScM, Johns Hopkins School of Public Health, Department of Epidemiology.

The mind-set regarding selection is different in
CAP007- **prevention trials**, where the goal is to determine whether a proposed prevention strategy works. One has to find a population suitable for testing the proposed strategy. Hence, unlike the treatment trial, risk factors predisposing to a disease and risk of an event are important. In this setting, one has to pay attention to both factors in trying to design a cost-effective trial. Considerations of this sort led, for example, the designers of the Multiple Risk Factor Intervention Trial (MRFIT) [68] to exclude females from enrollment. The risk factors targeted (high blood pressure, high cholesterol, and smoking) occur less frequently in women than in men. Consequently, the effort required to find women for study would have been much greater than that required to find men. Further, for the age range studied, women have a markedly lower myocardial infarction rate (the outcome of primary interest) than do men. This lower event rate would have meant that the planned sample size with women included would have to have been considerably larger to detect the same relative difference at the power level specified for the trial. As it was, the trial required a sample size of 12 866 men.

Readings

The literature on the design, conduct, and analysis of clinical trials is ever-expanding. Students of trials need to monitor the literature of reported trials as they appear in medical journals and to read specialty journals, such as Biometrics, Statistics in Medicine,
CAC008- ***Controlled Clinical Trials***, *Applied Clinical Trials*, and Statistical Methods in Medical Research. The list of citations given in Table 1 is but a snapshot of selected references dealing with the methods and procedures of trials.

References

[1] American Bible Society (1816). *The Holy Bible: Old and New Testaments, King James Version (1611).* American Bible Society, New York.

[2] *Applied Clinical Trials* (1992–97), Astor, Eugene.

[3] Ashby, D., ed. (1993). Conference on methodological and ethical issues in clinical trials, 27–28 June, 1991, *Statistics in Medicine* **12**, 1373–1534.

[4] Beecher, H.K. (1966). Ethics and clinical research, *New England Journal of Medicine* **274**, 1354–1360.

[5] Berlin, J.A., Laird, N.M., Sacks, H.S. & Chalmers, T.T. (1989). A comparison of statistical methods for combining event rates from clinical trials, *Statistics in Medicine* **8**, 141–151.

[6] Berry, D.A. (1985). Interim analysis in clinical trials: classical vs Bayesian approaches, *Statistics in Medicine* **4**, 521–526.

[7] Blackwelder, W.C. (1982). "Proving the null hypothesis" in clinical trials, *Controlled Clinical Trials* **3**, 345–353.

[8] Blackwelder, W.C. & Chang, M.A. (1984). Sample size graphs for "proving the null hypothesis", *Controlled Clinical Trials* **5**, 97–105.

[9] Boissel, J.P., Blanchard, J., Panak, E., Peyrieux, J.C. & Sacks, H. (1989). Considerations for the meta-analysis of randomized clinical trials: summary of a panel discussion, *Controlled Clinical Trials* **10**, 254–281.

[10] Bristol, D. (1989). Sample sizes for constructing CI's and testing hypotheses, *Statistics in Medicine* **8**, 803–821.

[11] Brown, B.W., Jr (1980). Designing for cancer clinical trials: Selection of prognostic factors, *Cancer Treatment Reports*, **64**, 499–502.

[12] Brosgart, C.L., Mitchell, T., Charlebois, E., Coleman, R., Mehalko, S., Young, J. & Abrams, D.I. (1996). Off-label drug use in human immunodeficiency virus disease, *Journal of Acquired Immune Deficiency Syndrome and Human Retrovirology* **12**, 56–62.

[13] Bull, J.P. (1959). The historical development of clinical therapeutic trials, *Journal of Chronic Diseases* **10**, 218–248.

[14] Buyse, M. & Ryan, L.M. (1987). Issues of efficiency in combining proportions of deaths from several clinical trials, *Statistics in Medicine* **6**, 565–576.

[15] Buyse, M.E., Staquet, M.J. & Sylvester, R.J. (1984). *Cancer Clinical Trials: Methods and Practice*. Oxford University Press, New York.

[16] Canner, P.L. (1977). Monitoring clinical trial data for evidence of adverse or beneficial treatment effects, *INSERM – Essais Controles Multicentres: Principes et Problemes* **76**, 131–149.

[17] Chalmers, T.C. & Amacher, P. (1982). Conference on recent history of randomized clinical trials (with discussion), *Controlled Clinical Trials* **3**, 299–309.

[18] Colton, T, Freedman, L.S., Johnson, A.L. & Machin, D. (1989). Recent issues in clinical trials, *Statistics in Medicine* **8**, 401–516.

[19] Colton, T., Freedman, L.S., Johnson, A.L. & Machin, D. (1991). Cumulative decade index, *Statistics in Medicine* **1–10**, 1–179.

[20] Colton, T., Freedman, L.S., Johnson, A.L. & Machin, D. (1991). Tenth anniversary issue, *Statistics in Medicine* **10**, 1789–2027.

[21] Cornfield, J. (1969). The Bayesian outlook and its application (including discussion by S. Geisser, H.O. Hartley, O. Kempthorne & H. Rubin), *Biometrics* **25**, 617–657.

[22] Coronary Drug Project Research Group (1973). The Coronary Drug Project: Design, methods, and baseline results, *Circulation* **47**, Supplement 1, I-1–I-50.

[23] Coronary Drug Project Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project, *New England Journal of Medicine* **303**, 1038–1041.

[24] Cox, D.R. (1972). Regression models and life tables (including discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[25] D'Agostino, R.B., Lange, N., Ryan, L.M. & Selwyn, M.R., eds (1992). Symposium on longitudinal data analysis, Fort Lauderdale, FL, June 19-21, 1991, *Statistics in Medicine* **11**, 1797–2040.

[26] DeMets, D.L. (1987). Practical aspects in data monitoring: a brief review, *Statistics in Medicine* **6**, 753–760.

[27] Donner, A. (1984). Approaches to sample size estimation in the designing of clinical trials – a review, *Statistics in Medicine* **3**, 199–214.

[28] Ellenberg, S., Geller, N., Simon, R. & Yusuf, S. (1993). Practical issues in data monitoring of clinical trials (Proceedings), Bethesda, MD, January 27–28, 1992, *Statistics in Medicine* **12**, 415–616.

[29] Fleming, T.R. & DeMets, D.L. (1993). Monitoring of clinical trials: issues and recommendations, *Statistics in Medicine* **14**, 183–197.

[30] Freedman, B. (1987). Equipoise and the ethics of clinical research, *New England Journal of Medicine* **317**, 141–145.

[31] Freedman, L.S. & Spiegelhalter, D.S. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials, *Controlled Clinical Trials* **10**, 357–367.

[32] Freedman, L.S., Simon, R., Foulkes, M.A., Friedman, L., Geller, N.L., Gordon, D.J. & Mowery, R. (1995). Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993–The perspective of NIH clinical trialists (with discussion and response), *Controlled Clinical Trials* **16**, 277–312.

[33] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1985). *Fundamentals of Clinical Trials*, 2nd Ed. PSG Publishing, Boston.

[34] Glaucoma Laser Trial Research Group (1991). The Glaucoma Laser Trial (GLT): 3. Design and methods, *Controlled Clinical Trials* **12**, 504–524.

[35] Grizzle, J.E. (1982). A note on stratifying versus complete random assignment in clinical trials, *Controlled Clinical Trials* **3**, 365–368.

[36] Grossman, E., Messerli, F.H., Grodzicki, T. & Kowey, P. (1996). Should a moratorium be placed on sublingual nifedipine capsules given for hypertensive emergencies and pseudoemergencies? *Journal of the American Medical Association* **276**, 1328–1331.

[37] Hawkins, B.S. (1991). Controlled clinical trials in the 1980s: a bibliography, *Controlled Clinical Trials* **12**, 1–272.

[38] Hawkins, B.S. (1991). Data monitoring committees for multicenter clinical trials sponsored by the National Institutes of Health: I. Roles and memberships of data monitoring committees for trials sponsored by the National Eye Institute, *Controlled Clinical Trials* **12**, 424–437.

[39] Hawkins, B.S. (1995). Data Management for multicenter studies: methods and guidelines, *Controlled Clinical Trials* **16**, Supplement 2, 1S–179S.

[40] Hedges, L. & Olken, I. (1987). *Statistical Methods for Meta-Analyses*. Academic Press, Orlando.

[41] Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

[42] Holbrook, J.T., Meinert, C.L. & Gilpin, A.K. (1994). On the likelihood of finding subgroup differences in clinical trials, *Controlled Clinical Trials* **15**, 129S–130S.

[43] Hughes, M.D. & Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trials, *Statistics in Medicine* **7**, 1231–1242.

[44] Hulley, S.B., Cummings, S.R., Browner, W.S., Newman, T.B. & Hearst, N. (1988). *Designing Clinical Research: An Epidemiologic Approach*. Williams & Wilkins, Baltimore.

[45] Institute of Medicine, Committee on the Ethical and Legal Issues Relating to the Inclusion of Women in Clinical Studies (1994). *Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies*, Vol. 1. National Academy Press, Washington.

[46] Institute of Medicine, Committee on the Ethical and Legal Issues Relating to the Inclusion of Women in Clinical studies (1994). *Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies, Workshop and Commissioned Papers*, Vol. 2. National Academy Press, Washington.

[47] Johnson, F.N. & Johnson, S., eds (1977). *Clinical Trials*. Blackwell Scientific, Oxford.

[48] Jones, D.R. (1995). Meta-analysis: weighing the evidence, *Statistics in Medicine* **14**, 137–149.

[49] Kalish, L.A. & Begg, C.B. (1985). Treatment allocation methods: a review, *Statistics in Medicine* **4**, 129–144.

[50] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[51] Katz, J. (1972). *Experimentation with Human Beings*. Russell Sage Foundation, New York.

[52] Kiri, A. & Meinert, C.L. (1995). Treatment effects monitoring practices as viewed through the published literature (abstract), *Controlled Clinical Trials* **16**, (Supplement 3), 58S.

[53] Kotz, S., Johnson, N.L. & Read, B.C., eds (1982–1988). *Encyclopedia of Statistical Sciences*, Vols 1–9. Wiley, New York.

[54] Lachin, J.M. (1981). Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials* **2**, 93–114.

[55] Lachin, J.M. & Foulkes, M.A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance and stratification, *Biometrics* **42**, 507–519.

[56] Lachin, J.M., Matts, J.P. & Wei, L.J. (1988). Randomization in clinical trials: conclusions and recommendations, *Controlled Clinical Trials* **9**, 365–374.

[57] Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials, *Biometrics* **44**, 229–242.

[58] Last, J.M., ed. (1995). *A Dictionary of Epidemiology*, 3rd Ed. Oxford University Press, New York.

[59] Lee, Y.J., ed. (1996). Conference on meta-analysis in the design and monitoring of clinical trials, *Statistics in Medicine* **15**, 1233–1323.

[60] Levine, R.J. (1986). *Ethics and Regulation of Clinical Research*, 2nd Ed. Yale University Press, New Haven.

[61] Lilford, R.J. & Jackson, J. (1995). Equipoise and the ethics of randomization, *Journal of the Royal Society of Medicine* **88**, 552–559.

[62] Lind, J. (1753). *A Treatise of the Scurvy*. Sands, Murray, Cochran, Edinburgh. Reprinted in *Lind's Treatise on Scurvy*, C.P. Stewart & D. Guthrie, eds. Edinburgh University Press, Edinburgh, 1953.

[63] McKenzie, R., Fried, M.W., Sallie, R., Conjeevarm, H., Di Bisceglie, A.M., Park, Y., Savarese, B., Kleiner, D., Tsokos, M., Luciano, C., Pruett, T., Stotka, J.L., Straus, S.E. & Hoofnagle, J.H. (1995). Hepatic failure and lactic acidosis due to fialuridine (FIAU), an investigational nucleoside analogue for chronic hepatitis B, *New England Journal of Medicine* **333**, 1099–1105.

[64] Meier, P. (1981). Stratification in the design of a clinical trial, *Controlled Clinical Trials* **1**, 355–361.

[65] Meinert, C.L. (1995). The inclusion of women in clinical trials, *Science* **269**, 795–796.

[66] Meinert, C.L. (1996). *Clinical Trials Dictionary: Terminology and Usage Recommendations*. The Johns Hopkins Center for Clinical Trials, Baltimore.

[67] Meinert, C.L. & Tonascia, S. (1986). *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

[68] Multiple Risk Factor Intervention Trial Research Group (1977). Statistical design considerations in the NHLBI Multiple Risk Factor Intervention Trial (MRFIT), *Journal of Chronic Diseases* **30**, 261–275.

[69] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. US Government Printing Office, Washington.

[70] NIH Clinical Trials Committee (1979). Clinical activity, *NIH Guide for Grants and Contracts* **8**, 29 (R.S. Gorden, chair, June 5).

[71] Office for Protection from Research Risks (1991). *Code of Federal Regulations, Title 45: Public Welfare, Part 46: Protection of Human Subjects*, rev. June 18. Department of Health and Human Services, National Institutes of Health, Bethesda.

[72] Packard, F.R. (1921). *Life and Times of Ambroise Paré, 1510–1590*. Paul B. Hoeber, New York.

[73] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design, *British Journal of Cancer* **34**, 585–612.

[74] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. & Smith, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. Analysis and examples, *British Journal of Cancer* **35**, 1–39.

[75] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.

[76] Pocock, S.J. (1993). Statistical and ethical issues in monitoring clinical Trials, *Statistics in Medicine* **12**, 1459–1475.

[77] Rubinstein, L.V., Gail, M.H. & Santner, T.J. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation, *Journal of Chronic Diseases* **34**, 469–479.

[78] Sackett, D.L. & Gent, M. (1979). Controversy in counting and attributing events in clinical trials, *New England Journal of Medicine* **301**, 1410–1412.

[79] Schwartz, D. & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutic trials, *Journal of Chronic Diseases*, **20**, 637–648.

[80] Senn, S.J. (1989). Covariate imbalance and random allocation in clinical trials, *Statistics in Medicine* **8**, 467–476.

[81] Singer, S.W. & Meinert, C.L. (1995). Format-independent data collection forms, *Controlled Clinical Trials* **16**, 363–376.

[82] Smith, P.G. & Morrow, R.H., eds (1991). *Methods for Field Trials of Interventions against Tropical Diseases: A "Toolbox"*. Oxford University Press, Oxford.

[83] Souhami, R.L. & Whitehead, J., ed, (1994). Workshop on early stopping rules in cancer clinical trials (Robinson College, Cambridge, UK, April 13–15, 1993), *Statistics in Medicine* **13**, 1293–1499.

[84] Spiegelhalter, D.S. & Freedman, L.S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion, *Statistics in Medicine* **5**, 1–13.

[85] *Statistical Methods in Medical Research* (1992–1996). Edward Arnold. Hodder & Stoughton Ltd, Oxford, Vols 1–5.

[86] Stewart, L.A. & Clarke, M.J. (on behalf of the Cochrane Working Group in Meta-Analysis) (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data, *Statistics in Medicine* **14**, 2057–2079.

[87] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press, Harvard University Press, Cambridge, Mass.

[88] United States Congress (1993). *National Institutes of Health Revitalization Act of 1993*, §131, Pub. L No. 103–43, 107 Stat. 133 (codified at 42 USC §289a-2).

[89] University Group Diabetes Program Research Group (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: I. Design, methods and baseline characteristics, *Diabetes* **19**, Supplement 2, 747–783.

[90] University Group Diabetes Program Research Group (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: II. Mortality results, *Diabetes* **19**, Supplement 2, 785–830.

[91] US General Accounting Office (1996). *Prescription Drugs: Implications of Drug Labeling and Off-Label Use* (testimony before US House of Representatives, 12 September, by Sarah F. Jagger). General Accounting Office (GAO/T-HEHS-96-212), Washington.

[92] Wei, L.J., Smythe, R.T. & Smith, R.L. (1986). K treatment comparisons with restricted randomization rules in clinical trials, *Annals of Statistics* **14**, 265–274.

[93] Winker, M.A. (1996). The FDA's decision regarding new indications for approved drugs: where's the evidence?, *Journal of the American Medical Association* **276**, 1342–1343.

[94] Wittes, J. (1993). Behind closed doors: The data monitoring board in randomized clinical trials (with discussion), *Statistics in Medicine* **12**, 419–434.

[95] Wittes, J., Duggan, J., Held, P. & Yusuf, S., eds (1990). Cost and efficiency in clinical trials (workshop proceedings; sponsored by the National Heart, Lung, and Blood Institute, Bethesda, MD, January 18–19, 1989), *Statistics in Medicine* **9**, 1–199.

[96] Wright, P. & Haybittle, J. (1979). Design of forms for clinical trials (1), *British Medical Journal* **2**, 529–530.

[97] Wright, P. & Haybittle, J. (1979). Design of forms for clinical trials (2), *British Medical Journal* **2**, 590–592.

[98] Wright, P. & Haybittle, J. (1979). Design of forms for clinical trials (3), *British Medical Journal* **2**, 650–651.

[99] Yusuf, S., Simon, R. & Ellenburg, S.S., eds (1987). Proceedings of the workshop on methodologic issues in overviews of randomized clinical trials, May 1986. Sponsored by the National Heart, Lung, and Blood Institute and the National Cancer Institute, Bethesda, MD, *Statistics in Medicine* **6**, 217–409.

[100] Yusuf, S., Wittes, J., Probstfield, J. & Tyroler, H.A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials, *Journal of the American Medical Association* **266**, 93–98.

CURTIS L. MEINERT