3 Representative patients

Explanatory and pragmatic goals in clinical trials are so intertwined that attempts to make a clear separation between them seem contrived. And yet, as I have noted, a public-value-guided choice to emphasize one or the other of the two aims must be made. They cannot be pursued with equal vigor in a single trial. French biostatisticians Daniel Schwartz and his co-workers have argued that a choice between the two purposes must be taken into account at every stage in the design of a comparative trial and in analysis of results. For example, a purely explanatory trial is conducted with idealized patients under restricted conditions in an effort to demonstrate specifically predicted biologic effects of a new treatment. The pragmatic trial seeks to assess the practical value of a new treatment when given to a wide range of patients in usual medical settings. In this book I have focused on trial formats which emphasize the latter approach.

Pragmatic trials are carried out in the hope that results will provide a reliable guide for future treatment policy. But such generalization requires a leap that must not be undertaken lightly. I have indicated that forecasting in medicine requires an appeal to experience. Now we need to examine an important requirement of experience used to predict results that doctors can expect when they prescribe for their patients. The experience must be representative.

RANDOM SAMPLING MODEL

The idealized model for establishing the ground rules that enable us to make inferences about unexamined cases comes from direct experiments in random sampling. The underlying idea is that we can obtain information about the whole by examining only a part. The statistical reasoning is the same as that used for forecasting outcomes in games of chance. In the classic example, a jar is filled with black and white marbles in some undisclosed ratio and we remove a sample set. From the proportions observed in the sample, we make statements about the actual distribution in the jar

(that is, the ranges within which the proportions are expected to lie); and we assign a graded quantity of belief to these assertions.

Assumptions

The entire operation hinges on the matter of representative sampling. Thus, we need some assurances before a test set of marbles is removed. We must assume that the marbles do not have inherent characteristics, such as 'stickiness', differential weights, and internal 'motors'; and that they cannot be maneuvered by external forces, such as 'magnetism' which would favor systematic aggregation. We must also be assured that there is an unequivocal distinction between black and white marbles (no grays). And the jar must be well stirred to assure random scattering before a sample is removed.

The word 'random' The connotations of the word 'random' deserve some attention here. The term frequently implies dynamic movement or occurrence that takes place without aim, purpose, or fixed principle. Another connotation—at liberty and free from restraint or control—is also relevant in the present context. The scientific meaning of 'random distribution' was made clear by Venn who pointed to the arrangement of drops of rain in a shower. 'No one can guess', he said, 'where a drop will fall at any one instant, but we know that if we put out a sheet of paper, it will gradually become uniformly spotted all over. And equal areas on the paper will gradually tend to be struck equally often.'

From sample observation to population estimate

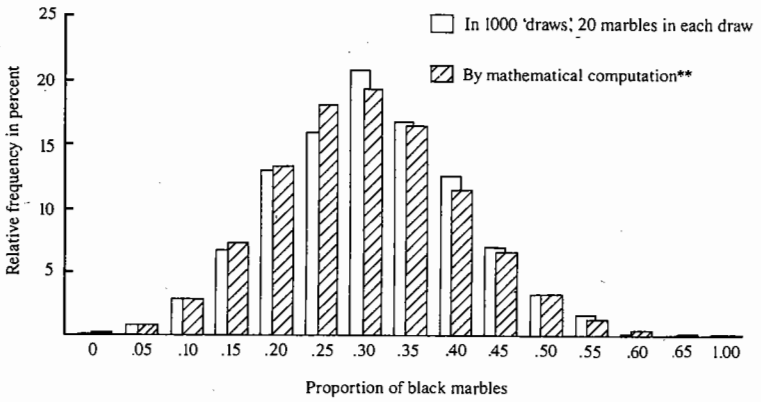
If the pre-sampling assumptions are reasonable, batch testing may proceed with confidence. The proportion observed in a random sample will provide a calculable estimate of the true ratio of black and white marbles in the jar which is termed the 'target population', 'parent population', or 'population of interest'. Notice that this population is related only vaguely to the 'universe' of marbles which exists outside of the jar.

GROUPS OF PATIENTS

The ideas that lie behind the simple marble-sampling model should be kept in mind when considering the recruitment of a group of patients who will serve as representatives in a planned clinical study, but it is important from the outset to recognize the limitations of the analogy as we enter the complicated real world. The inherent characteristics of patients which influence their systematic choices of physicians and hospitals, their susceptibility to 'magnetic' external forces represented by the medical system, the blurred distinctions between categories of patients, and their resistance to thorough 'stirring' virtually guarantee sampling difficulties. The qualities that distin-

ESTIMATION OF PROPORTIONS IN A PARENT POPULATION SAMPLING

Empirical Results* versus Computation** of the Binomial Probability Distribution



* Frequency of occurrence (in per cent) of 20-marble samples with specified proportions of black marbles.

** Expansion of the binomial formula $(p+q)^{20}$
where $p=0.3$ = the proportion of black marbles in the jar,
$q=0.7$ = the proportion of white marbles in the jar, and the exponent ^{20} is the number of marbles in each sample drawn from the jar.
In the sampling experiment conducted by Mainland, 300 black marbles and 700 white marbles were placed in a jar, the contents thoroughly mixed, 20 marbles were drawn, the result noted, and the marbles returned to the jar. The sampling procedure was continued until 1000 samples of 20 were obtained. The frequency distribution of proportions of black marbles obtained by the empirical approach is very close to the values obtained by mathematical computation. In this example, samples which contained fewer than 10 per cent black marbles (zero or one black marble in twenty), and more than 55 per cent (more than 11 black marbles in twenty) occurred rarely; less than 5 per cent of the samples fell into these extreme tails of a normal distribution.

An estimate of the sampling 'error' may be expressed by the formula:

$$\text{S.D.} = \sqrt{\frac{p \times q}{n}}$$

where S.D. = the standard deviation (in some contexts it is termed, the standard error) of a percentage of a dichotomous (either/or) property,
p = the percentage of black marbles observed in the sample
q = the percentage of white marbles in the sample
n = the number of marbles in the sample
The percentages in the parent population are estimated within given 'confidence limits'. In a hypothetical example, we draw a single sample of 20 marbles from the jar of 300 black and 700 white marbles, and count 5 black marbles (25 per cent) in the set, thus

$$\text{S.D.} = \sqrt{\frac{25\% \times 75\%}{20}} = 9.7\%$$

In repeated samples of 20 drawn from the jar, we expect that the percentage of black marbles will lie within $p+2$ S.D. and $p-2$ S.D. (25 per cent $\pm$ 19.4 per cent), and that values will be found outside these limits about once in twenty draws. Thus, we estimate that the percentage of black marbles in the jar (as judged by this hypothetical single draw) lies somewhere between 5.6 per cent and 44.4 per cent; and we expect to be wrong about 5 per cent of the time. If we increase the sample size tenfold to 200 marbles, the range of the 'confidence limits' is reduced by a factor of 3 (now 25 per cent $\pm$ 6.1 per cent estimates that the percentage in the jar lies between 18.9 per cent and 31.1 per cent; and in this hypothetical instance the percentage in the jar, 30 per cent, lies within the range estimated by the single sample draw of 200 marbles).

guish *all* groups of patients are far removed from the connotation 'at liberty and free from restraint or control.'

Haphazard sampling in medicine

There is widespread belief in a disturbing misconception that if the source of the patients is not known, the sample is a random one. The misunderstanding stems from failure to make a distinction between everyday use of the word 'random', which leans toward the idea of 'without aim, purpose, or fixed principle,' and the use by statisticians. The latter are concerned with the process of random sampling in a precise technical sense in order to fulfill the assumptions of probability theory. The term 'probability samples' is sometimes used by statisticians to avoid confusion.

Edmond A. Murphy of Johns Hopkins University has pointed out that it is not an uncommon practice when attempting to establish 'normal values' of blood constituents, to collect specimens from groups of people in blood banks, from technicians working in a laboratory, or from people walking down a hospital corridor. A sample of this kind is not, as investigators may imagine, a random sample of the population. It is, he emphasized, merely a haphazard one. If the conclusions about 'normal values' are to be extended to a wide population, say the entire country, we must go to some lengths to satisfy the probability sampling requirement that each person in the entire population is *equally likely* to be chosen for the sample. In a formal study, sampling difficulties should be recognized and shortcomings should be reported.

Target population

To deal with the problem of recruiting representative patients, we are first obliged to define the target population: To whom are the results of the study meant to apply? The answer, which spells out the *external relevance* of the investigation, is usually more parochial than hoped for. Limitations are imposed because of the way in which patients aggregate in various parts of the medical system.

'Available' patients For example, studies are conducted most commonly in groups of patients chosen from those who are 'available'; they turn up in hospitals, out-patient departments, and doctors' offices. Referral patterns

Some biases in specifying and selecting the study sample

Admission rate (Berkson) bias
: If hospitalization rates differ for different exposure/disease groups, the relation between exposure and disease will become distorted in hospital-based studies.

Centripetal bias
: The reputations of certain physicians and medical institutions cause individuals with specific disorders or exposures to gravitate toward them.

Detection signal (unmasking) bias
: An innocent factor may become suspect if it causes a sign or symptom which sets off an intensified search for the disease.

Diagnostic access bias
: Individuals differ in their geographic, temporal, and economic access to the diagnostic procedures which label them as having a given disease.

Diagnostic suspicion bias
: A knowledge of the subject's prior exposure to a putative cause (ethnicity, taking a certain drug, having a second disorder, being exposed in an epidemic) may influence both the intensity and the outcome of the diagnostic process.

Diagnostic vogue bias
: The same illness may receive different diagnostic labels at different points in space or time.

Membership bias
: Membership in a group (the employed, joggers, etc.) may imply a degree of health which differs systematically from that of the general population.

Non-contemporaneous controls bias
: Secular changes in definitions, exposures, diagnoses, diseases, and treatments may render non-contemporaneous controls non-comparable.

Non-respondent bias
: Non-respondents (or 'late-comers') from a specified sample may exhibit exposures or outcomes which differ from those of respondents (or 'early-comers').

Popularity bias
: The admission of patients to some doctors' practices, medical institutions, or procedures (surgery, autopsy) is influenced by the interest stirred up by the presenting condition and its possible causes.

Prevalence-incidence (Neyman) bias
: A late look at those exposed (or affected) early will miss fatal or other short episodes, plus mild or 'silent' cases and cases in which evidence of exposure disappears with disease onset.

Previous opinion bias
: The tactics and results of a previous diagnostic process on a patient, if known, may effect the tactics and results of a subsequent diagnostic process on the same patient.

Procedure selection bias
: Certain clinical procedures may be preferentially offered to those who are poor risks. (Selection of patients for 'medical' vs 'surgical' therapy.)

Referral filter bias
: As a group of ill are referred from primary to secondary to tertiary care facilities, the concentration of rare causes, multiple diagnoses, and 'hopeless cases' may increase.

Unacceptable disease bias
: When disorders are socially unacceptable (V.D., suicide, insanity) they tend to be under-reported.

Volunteer bias
: Volunteers or 'early comers' from a specified sample may exhibit exposures or outcomes (they tend to be healthier) which differ from those of non-volunteers or 'late-comers'.

(Described by Sackett)

of physicians, special interests in particular hospitals, admission practices (when hospitals are crowded, only the most severe cases of a particular disease are admitted and when many beds are free, milder cases may be admitted also), and the hierarchical organization of hospitals (into primary, secondary, and tertiary care centers of increasing specialization) all tend to bring about an irreproducible sorting of patients. As a result, the characteristics of hospital patients may vary widely from institution to institution. Generalizations based on observations in hospitalized groups are notoriously untrustworthy. In one study, hospitalized patients with respiratory diseases had arthritis complaints twice as often as those not affected with respiratory difficulties. This 'increased risk of arthritis' was not observed in comparable patients who were not in hopspitals. The spurious association is an example of what is called 'Berkson's Bias', one of a long list of selection biases collected by Sackett.

Dissimilar babies The 1949 survey of the occurrence of RLF among infants reared in hospitals throughout the United States (p 16) is another example of the difficulties in obtaining groups of 'typical' patients for formal study. As I have said, the occurrence of blindness and treatment practices varied considerably from hospital to hospital and in the same hospital over time. This confusion was further compounded by dissimilarities in the at-risk populations. In addition to the relatively small numbers of observations reported by some of the hospitals, the birthweight distributions were mismatched. In one institution, the proportion of infants weighing less than 1.5 kilograms at birth was 13 per cent of the total, while in another it was 'reported' to be 100 per cent. Since the risk of developing RLF is inversely

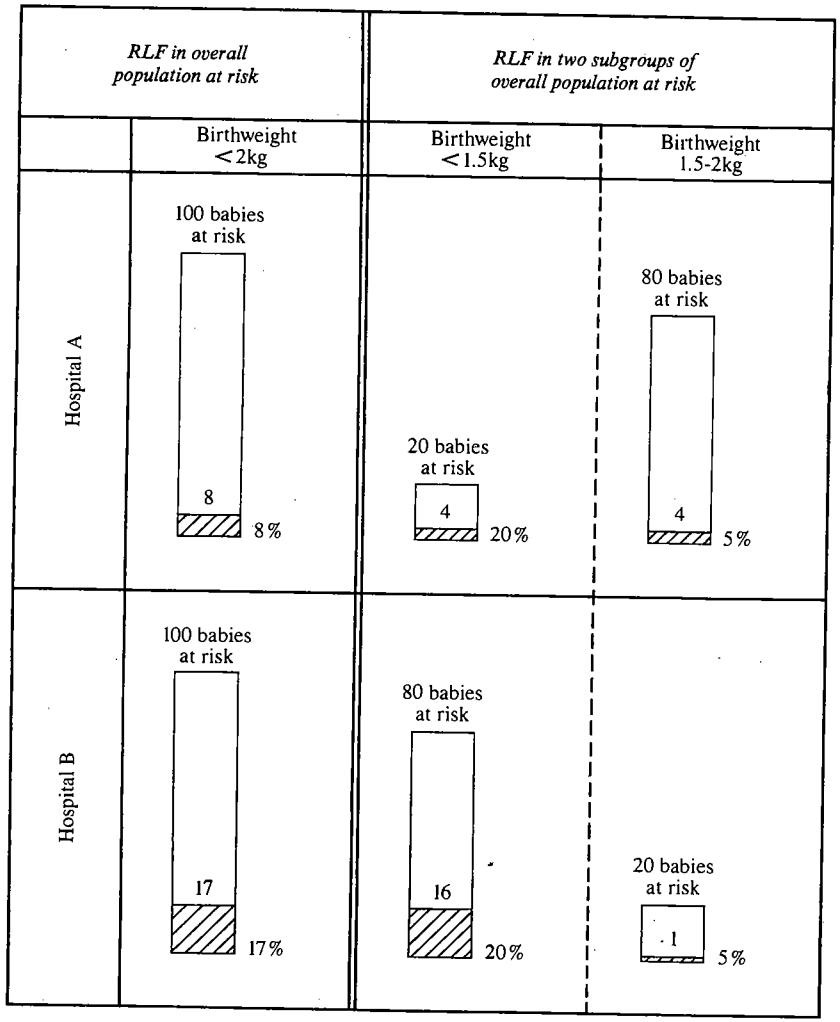
Proportion of very small babies in RLF surveys of various hospitals

By cities and time periods (1922–48)

City	*Years*	*No. of infants*	*RLF (per cent)*	*Birthweight under 1.5 kilograms (per cent)*
Boston	1938–48	150	7	29
,,	1943–7	200	22	19
Providence	1941–7	225	7	13
Baltimore	1935–44	86	0	27
,,	1945–7	72	7	46
Hartford	1948	35	23	20
New York	1939–46	207	3	31
Cincinnati	1943–7	96	7	23
Birmingham	1945	104	0	20
Denver	1948	14	14	71
Chicago	1922–47	211	2	100

(Taken from the data assembled by Kinsey and Zacharias)

RLF in two hospitals (hypothetical)



In Hospital A, RLF occurs in 8 per cent of 100 infants (birthweight under 2 kilograms); and the eye problem occurs in 17 per cent of 100 infants who are reared in Hospital B. The difference in distributions of birthweights within the under 2 kilograms class in these two hospitals accounts for the twofold *overall* difference in per cent occurrence of RLF; the rates within subgroup are identical in the two hospitals (under 1.5 kilograms = 20 per cent, 1.5–2 kilograms = 5 per cent).

related to the degree of prematurity as crudely judged by birthweight, relatively small differences in the distribution of birthweight would be expected to exert a magnified effect on the overall occurrence of RLF. When the experience in each hospital was examined in subgroups (by birthweight), some of the variation was reduced; but there was still considerable uncertainty about distortions related to other inequalities of babies in the various institutions.

Later surveys reporting associations between oxygen use and RLF were not consistent. Again, there was a strong suspicion that much of the inconsistency was related to differences in the compositions of the at-risk populations.

Circular definition of target population Under the bewildering selective distributions of patients that exist in clinical medicine, the most conservative way to define the target population is to describe the *actual* collection of patients enrolled in a study. In effect we say, by a circular definition, that the results of the study are meant to apply to patients like those encountered in the study. And we proceed to describe the characteristics of 'patients enrolled' in some detail. A full description allows others to see the limits of the experience and the need for more information (for instance, how patients became available and how they were enrolled) before overstepping the narrow bounds of the reported experience.

Wide population base Collaborative studies, involving a variety of institutions in one linked effort, are used to reduce the problems of uneven distribution of patients by providing a broad-based source of recruits. This potential was exploited in 1946 (p 44) when the pioneering tuberculosis treatment trial was conducted in Britain; patients enrolled in the trial were chosen from a parent population in eight tuberculosis hospitals throughout the country. A multicenter strategy was also attempted in order to deal with the hospital-to-hospital inconsistencies concerning oxygen treatment and RLF. In the national controlled clinical trial of 1953–4, 18 American hospitals treated small premature infants according to a prescribed experimental protocol; liberal oxygen use was compared with restricted oxygen treatment. The results of this study, one of the earliest large-scale controlled trials carried out in the United States, supported the positive association between generous use of oxygen and an increased risk of RLF. Doubts quickly faded away and the use of oxygen in the management of *all* premature infants throughout the world was sharply curtailed.

Unrepresentative patients in a controlled trial I want to emphasize, since the point is often misunderstood, that random order of assignment in a treatment comparison trial does not solve the problem of *external* relevance.

The precaution improves the chances of internal representativeness (equal representation of participants in each of the treatment groups), but it does not increase the likelihood of drawing a representative sample of participants from the target population who are to be the beneficiaries of information acquired in the trial.

A striking example of the subtle ways in which systematic clustering of patients can occur was seen in the experience with a controlled study of the effectiveness of a treatment which was purported to reduce complications of pregnancy. Women enrolled in this trial were thought to represent a fair

Increased cancer risk among all participants in a controlled clinical trial

	Breast cancer (no.)	Occurrence[a] (per cent)	Risk ratio[b]
Exposed to DES	(32/693)	4.9	1.79
Unexposed to DES	(21/668)	3.1	
U.S. population[c]			1.00

[a] Participants in a double-blind study of the effectiveness of DES (diethylstilbestrol) in reducing the hazards of late complications of pregnancy (e.g. miscarriage), conducted in 1951–2. A total of 2162 enrolled in the study; 840 in the DES group and 806 in the non-medicated group completed the course of 'treatment'. During 1976–7, 693 mothers exposed and 668 not exposed to DES were interviewed by Bibbo and associates.
[b] Standardized incidence ratio adjusted for age distributions of women 'at risk' for breast cancer.
[c] The risk level of 1.00 was used as the base line for the general population as represented by the Connecticut Cancer Registry (incidence rates for the State of Connecticut—1963–5—are in good agreement with several United States and Canada data sources).

(Taken from the data of Bibbo and co-workers)

cross section of the at-risk pregnant population in the early 1950s. Years later it was found that the *daughters* of the women who received the treatment used in this trial (diethylstilbestrol, abbreviated DES) were developing cancer of the vagina at an alarming rate. As a result, the mothers who participated in the formal trial were contacted: Marluce Bibbo and coworkers at the University of Chicago found that those who had received DES treatment 25 years earlier experienced a relatively high rate of breast cancer. But this rate was only slightly higher than among concurrent controls who had been given no medication.

As it turned out, the entire group of women who were enrolled in this controlled trial had unrecognized (and inexplicable) similar characteristics which became manifest over the next quarter of a century: in the interim they developed breast cancer at a rate which was 1.79 times higher than the general population!

REPRESENTATION IN FOCUSED TRIALS

The uncertainties about representativeness can be narrowed if the scope of studies is scaled down to what John W. Tukey of Princeton University has called a 'focused clinical trial'. (This type of study is in contrast to a 'clinical inquiry' in which an intervention is postulated to help some class of patients, not specified in advance. From the collection of a great deal of information, the results are analyzed for each of many classes of patients—by age, sex, previous medical history, presenting symptoms, and prognosis, for example.) In the focused trial the class of patients to be considered is clearly specified *in advance*, Although the primary emphasis of the focused approach may remain pragmatic, a stated limit is placed on the practical goal.

Infants studied in the cooperative study of RLF

The 1953–4 national study of RLF, for example, was confined to the category of premature infants whose risk of blindness was most likely to be affected by a change in oxygen treatment policy: birthweight under 1.5 kilograms, age 2 days on enrollment (it was reasoned that early deaths would provide no information concerning the effect of treatment practices on RLF which develops many days and weeks after birth).

Source of patients As I have noted, this trial was conducted in 18 hospitals to achieve the goal of wide representation, but as we look more closely at the details, there were some sampling problems that are now evident with the clear vision of hindsight. Although the hospitals were located in different parts of the country, there was little variety in the type of institution; 16 of the 18 were university hospitals. The population of pregnant women in these referral centers was relatively atypical (many women with pregnancy complications were referred for specialized care and delivery). Additionally, many prematurely born infants were transferred to the large centers for expert attention (fully half of the more than 700 infants in the trial were in this category). All of the eligible babies in the participating hospitals were enrolled, but there was no assurance that these individuals, who were selected systematically before and after delivery, were 'typical' representatives of their class. What was needed, we see in retrospect, was a description of the characteristics of small, two-day-old premature infants who were *not* available to allow a comparison with participants.

Distribution of 'marker' characteristics All descriptions of patients' characteristics are only more or less complete, because we can never be sure that we know all, or even the most relevant, prognostic factors to judge

'representativeness.' Nonetheless, there is an indirect approach to the problem that provides some useful insights. We count the frequencies of a few easily identified 'markers' (for instance, gender, ethnicity, and the like) and examine their distribution in the class of patients under consideration.

In the national study, for example, we may approach the question of representativeness by examining the records of all participants *plus* a sample from community hospitals not included in the trial. The proportion of

Distribution of a 'marker' characteristic in a focused trial: Example of a format

	Proportion of male, small two-day-old infants	
Within-hospital distribution (Participating hospitals)	Born in participating hospitals	Transferred to participating hospital from hospital of birth
Enrolled infants		

	Proportion of male. . . in one participating hospital	Proportion of male. . . in 17 other participating hospitals
Between hospital distribution (Participating hospitals)		
Enrolled infants		

	Proportion of male. . . in 18 participating hospitals	Proportion of male. . . in sample of small two-day-old infants drawn from community hospitals
Between-hospital distribution (Participating hospitals versus community hospitals)		
Enrolled infants versus Non-enrolled infants		

Overall proportion = proportion of *male* infants among infants enrolled plus non-enrolled sample from community hospitals

males, say, in the relatively wide-based population of at-risk babies provides an estimate of the proportion 'expected', and it may be compared with the proportions actually observed in each of the 18 participating hospitals. The distribution of the 'marker' provides a clue to the sorting of babies in the study; and we return for help with interpretation to the inanimate world of marbles and coins. The observed distribution of the proportion of male patients in each hospital may be compared with the frequency distribution pattern predicted by probability theory *if these groups were random samples* from the population of 'all' patients in the class. The differences between the proportions predicted and those observed may now be evaluated by reference to the laws of chance.

Vague warning of distortion

I should emphasize again that the arbitrarily chosen 'marker' characteristics may not be *the* important determinants with respect to outcome, but they serve as indicators of the amount of distortion that has occurred in the selection of patients. And unusual variations serve to warn that important (but unrecognized) variables may also be distributed in a way which distorts the results.

The sceptical approach is very much like that of the gambler who inspects a set of dice and finds that one is red and the other is green. Although the disparity in color may not influence the behaviour of the dice, a high-roller is warned to withhold his bets until he has some assurance that he is playing with unbiased 'bones'.

ELIGIBILITY AND ENROLLMENT

All of the eligible infants admitted to participating hospitals were enrolled in the 1953-4 RLF trial, but that situation was atypical. Usually only a fraction of eligible candidates, those who agree voluntarily to participate or who are volunteered by surrogates, are actually enrolled. Needless to say, this process of selection by personal decision (influenced by family, friends, physicians, or by wordless fear) further complicates the murky issue of 'typical' representatives in medical studies.

There is simply no satisfactory way to resolve all of the uncertainties introduced when studies are conducted in patients who are selected on the basis of subjective criteria. Each extrapolation of results (from 'patients enrolled', to 'patients eligible', to 'patients in this class') is hemmed in with difficulties. Nevertheless, it is important to recognize the problems and to make some attempt to describe their scope. A full accounting of eligible patients who were *not* enrolled should be provided in reports of clinical studies with a warning that such missing patients may differ in some systematic way from participants. When there are many exclusions prior to enrollment in a clinical trial, generalization of results is seriously weakened by uncertainty.

Following all of the difficulties and limitations I have described in the recruitment and enrollment of representative participants in human studies, yet another block must be overcome—random order of allotments in a treatment comparison trial. It is this move which plays the most important role in assuring that like will be compared with like. The ramification of this controversy-ridden step in a controlled clinical trial will be discussed in the next chapter.