8 Avoiding entrapment

Decision theory strategists look upon problems in the biological sciences as 'games against nature', and they plan tactics to outwit this capricious adversary. Prudent clinical investigators adopt a similar outlook; they assume that the 'opponent' has planted innocent looking 'traps' to foil planned efforts to separate effects in a comparative study. The suspicious attitude is justified; as I will emphasize by a few more examples, circumstances that mislead in medicine lie in wait to entrap the unwary.

An example of confoundment

In an observation of an effect of detergent on the health of a dog

A woman asked her veterinarian whether it would be safe to wash her dog with a particular detergent. He advised against it, saying it would make the dog sick.

That evening the woman called to say that the dog was ill and asked the vet to come. The woman admitted that, against advice, she had washed the animal with the detergent. The vet shook his head and said, 'I knew it would harm the dog.' 'Oh, I don't think it was the detergent,' said the woman, 'it was putting him in the spin dryer afterward.'
(Related by Edmond A. Murphy of Johns Hopkins University)

TIME-RELATED TRAPS

Intervals before and after start of treatment

Pitfalls related to the element of time are encountered often in studies of the effect of treatments on the duration of life. Here, two segments of time need to be taken into account: the first interval begins with the onset of illness and runs to the beginning of treatment, the second marks the time from initiation of treatment to death. We may be badly misled if we direct our attention solely to a change in the latter period, and fail to consider the marked influence of the timing and efficiency of diagnostic procedures on the elapse of time before treatment. It is not difficult to envisage a change in proportions of the two intervals with no change in the total

duration of time from illness onset to death. Age of illness is sometimes more important than chronological age of patients when comparing like with like in clinical trials.

Consider, for example, the results of treatment with an ineffective anticancer agent. When a diagnosis of cancer is made early and the useless treatment begins soon after the onset of illness, the drug will appear to prolong life simply because the interval from the start of treatment to death has been made to *seem* relatively long.

Age on enrollment and mortality

Temporal effects are also seen in studies of treatments in newborn infants. Relatively small differences in age on enrollment are associated with exaggerated changes in duration of life. Most deaths occur soon after birth; thus, the probability of survival rises sharply with each passing hour after delivery. When an infant is transferred from the hospital of birth to a special care facility, age on admission may provide a reliable clue to a forecast of the length of life.

Crude versus 'adjusted' mortality rate A spurious improvement once observed in survival rates of very small babies (under 1.0 kilogram) reared in a large referral center is a case in point. In 1950, only 18 per cent (9/50) of these infants survived in the hospital; two years later the rate rose to 38 per cent (20/53). The improvement was encouraging, and there was much speculation about the cause. However, when corrections were made for the distorting effects of differences in age on admission, the explanation was revealed. The apparent improvement was accounted for by the admission of older, relatively robust infants in the second period of comparison. In fact, the 'adjusted' rates indicated that the survival experience of *comparable* babies was 6 per cent lower in the latter year. (The statistical methods used in the calculation of so-called standardized event-rates—adjusted for age, sex, and so forth—are described in epidemiology textbooks.)

In the same hospital experience, the survival of relatively large infants (1.0–1.5 kilograms) remained unchanged over the period of time under scrutiny (59/81 versus 71/97, 73 per cent survival rate for both years). A comparison of the 'adjusted' rates, however, indicated that there had been a 30 per cent improvement among comparable babies, a finding that was hidden by the discrepancies of age on admission.

Duration of observation

One of the most common 'traps' in medical studies can be found in the details of duration of observation. Most pathological processes take some time to evolve, and patients must remain under observation for a specified period if the manifestation of an abnormal state is to be recognized. When

patients are discharged from observation too soon or fail to return at appointed times, the irregularities play havoc with efforts to interpret the frequency of medical events. The distorting effect can be quite subtle.

Incidence of strawberry marks A trivial but instructive illustration of this kind of entrapment took place in connection with observations made on strawberry marks of the skin (capillary hemangiomas) during the RLF epidemic and immediately thereafter (when occurrence of the eye disorder fell sharply in 1954–5). These minor skin conditions are seen frequently in young infants. They usually appear some days and weeks after birth, grow for a time, and usually disappear spontaneously by the age of two or three years. The 'mark' consists of a tangle of newly-formed blood vessels; under the microscope the appearance is very similar to the exuberant growth of blood vessels in the early stages of RLF. Strawberry hemangiomas seemed to occur more frequently in premature than in full-term babies, and in the late 1940s it was suggested that RLF was nothing more or less than a postnatally-appearing hemangioma of the retina.

In an effort to examine this suggestion systematically, a large special care center began to keep a record of the time of appearance of the superficial marks and to correlate the skin findings with those noted in the eyes of premature infants. The associations were not particularly striking, nonetheless the joint observations were continued over a period of several years. When the occurrence of RLF fell dramatically in late 1954 and in 1955, it was found that the frequency of strawberry marks also declined sharply. These intriguing observations were collated and a draft was made of a report to announce the curious association in a medical journal. As the manuscript was readied for publication, an article was published revealing a 'trap' that had been overlooked.

The new report provided the results of a series of observations made in Sweden on the frequency and time of appearance of skin hemangiomas among 640 prematurely born infants and a control group of 186 mature infants. It was found that weekly examinations for several months after birth in *both* groups revealed no difference in the frequency of occurrence among the premature infants and full size babies. The previously held belief that skin marks were more common in the small infants seemed to be related to length of observation in hospitals. Small babies remained in the hospital for relatively long periods of time after birth and the marks came to medical attention regularly. The large infants left the hospital before the hemangiomas appeared, and frequency among the latter babies was less obvious because they were not assembled in one location.

On the basis of the new insight, the hospital discharge ages of premature infants in the long series about to be reported was examined: the median age at discharge dropped abruptly when the occurrence of RLF fell since

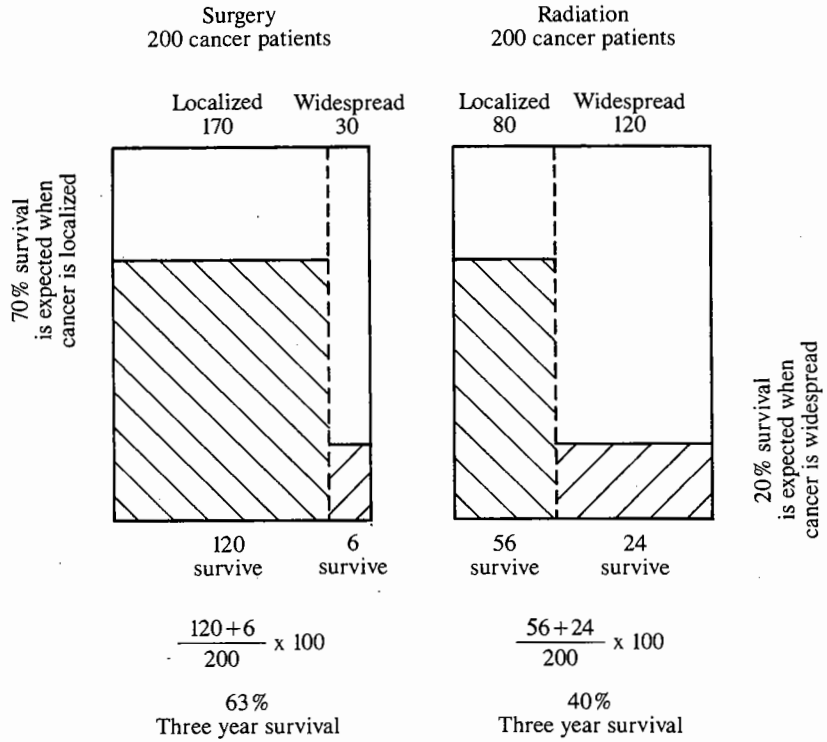
infants were no longer kept in the hospital for additional days and weeks to monitor the progress of abnormal eye changes. The report of the 'interesting' change in frequency of skin hemangiomas with restriction of oxygen administration was quietly shelved.

INEQUALITIES IN COMPARED GROUPS

Unequal prognoses

I have noted the strategy of prognostic stratification (p 50) used to sort patients according to susceptibility of suffering an outcome event. Here I wish to emphasize the potential for confusion when unequal groups are compared by relating a hypothetical example given by Feinstein.

Co-mingling patients with unequal prognoses



The outcomes of two treatments (surgery versus radiation for cancer, say), when neither has any effect on the natural course of disease, are biased, Feinstein points out, in favor of treatment used in a group of patients in which there is a disproportionate number with relatively good outlook.

Surgery versus radiation for cancer Suppose, Feinstein argues, the expected three-year survival rate for patients with a particular cancer is quite high in those with a growth localized to one site in the body and that the rate is sharply reduced when the malignant process has spread widely to many sites. In a comparison of surgical versus radiation treatment, we would immediately recognize a problem in interpreting the results if surgical treatment was used exclusively for those with localized cancer and radiation was administered to those with widespread involvement. However, the garbled effect is not immediately obvious when both kinds of patients are represented in the two treatment groups. The bias will be felt unless the 'mixture' is proportionately equal in the subgroups with unequal prognostic expectations. A mistaken conclusion about the effectiveness of surgery may come about solely as a result of our failure to recognize prognostic differences and disproportionate co-mingling of patients.

Unequal compliance with treatment

The same kind of distortion occurs, Feinstein points out, when there is unequal compliance with treatments of equal effectiveness in a planned trial. Suppose, for instance, that success under treatment occurs in many patients when either A or B maneuver is faithfully maintained and some improve when treatment is abandoned. If B has such an unappealing taste, appearance, schedule of administration, and the like, so that it is maintained by only half of the patients assigned to this regimen, the poor showing of this group of patients may be attributed, mistakenly, to an inferiority of the pharmacologic action of the drug.

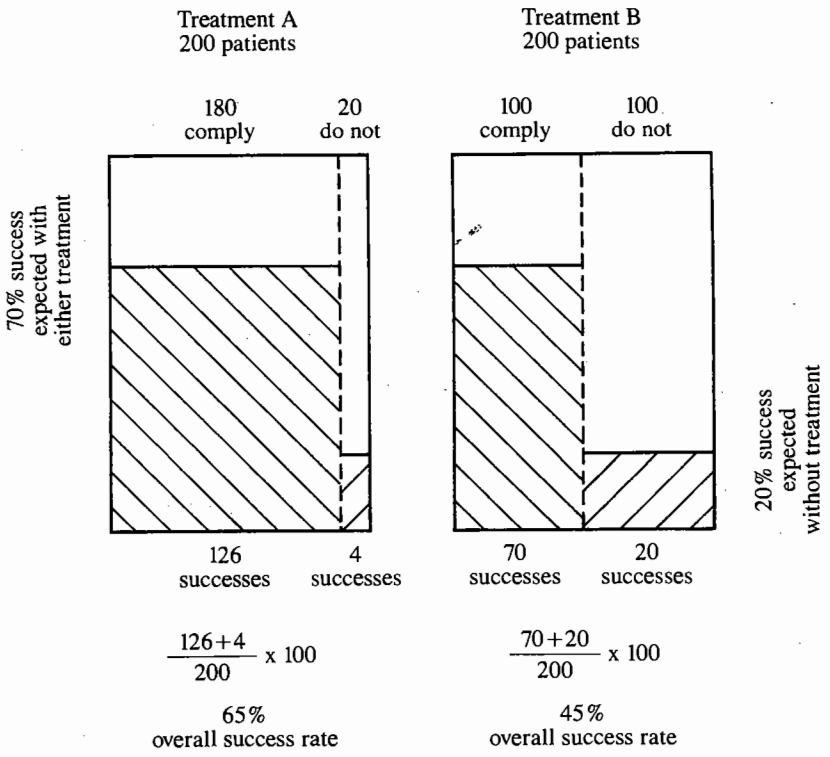
The adage, there's many a slip 'twixt the cup and the lip (which dates to the Greek legend of the Argonauts in search of the Golden Fleece) is applicable here. The issue of compliance is a formidable problem in medical trials. If it cannot be documented that patients did, in fact, take the drugs prescribed for them, it would be wise to use some variation of this cautious wording (and emphasis) in a heading that describes a comparison trial: 'The outcomes after *prescription* of drugs A and B.'

The combined effects of differences in duration of observation and in compliance with prescribed treatments have also been described by Feinstein. He warns of compound distortions that may result when precautions are not taken to avoid potential sources of confoundment.

PRECAUTIONS IN CLINICAL TRIALS

Early in the experience with clinical trials Mainland gave up trying to catalogue the numerous difficulties that may come up. Instead, he proposed a list of precautions that might reduce procedural errors (p 68) that occur so commonly. The guidelines included these categories: choice of investi-

Unequal compliance with prescribed treatments



A hypothetical example given by Feinstein of distortion of results caused by differences in patient compliance with prescribed treatments in a trial of two treatments of equal effectiveness.

gators, time for planning, realism in planning, carrying out the plan, and policing the trial.

Although more than twenty years have gone by since he proposed these safeguards, his advice remains timely. The need to improve external relevance requires the involvement of many doctors and more than one hospital in clinical studies—the very situation in which Mainland found that problems multiplied.

Choice of investigators

The personal characteristics of physician investigators, such as motivation, commitment, and acceptance of regimentation are not easily bared, but they are pertinent in large scale trials. The projects cannot be carried out

successfully unless the doctors involved are willing to 'freeze' their activities into prescribed molds for a fairly substantial period of time. During this period they are almost certain to be buffeted by stormy forces. Patients, relatives and friends of patients, other physicians, and well-wishers may exert pressure to change treatment when results of other studies become available or when preliminary findings and hopeful leads are widely publicized.

The difficulties are considerable and physicians who are to become involved should have an opportunity to discuss these matters before they agree to participate. The enrolled patients are entitled to an unwavering commitment on the part of their caretakers to carry out a clinical trial faithfully in order to produce interpretable results. Doctors should bow out if there is any question about fulfilling these obligations wholeheartedly.

Additionally, those in charge of a large-scale trial must recruit the support of referring physicians to reduce sampling distortions that may occur over time. As I will discuss later, the active involvement of the medical community is essential if the results are to have an impact on everyday practice at the conclusion of a trial.

Unfortunately, I can do little more than sound these ominous warnings about physician related problems that are a potential source of entrapment in bedside studies. Although there is no ready solution, the difficulties cannot be ignored; it is usually necessary to review the issues periodically during the course of study.

The planning period

A considerable amount of time is needed to plan a medical trial. In one multicenter cancer treatment trial, for instance, one year was allowed for drawing up the protocol (the document containing the agreed upon plan) and the corresponding record-gathering documents. During the year, the planners (physicians and statisticians) met together for one eight-hour day each month. For every hour spent in formal meeting, the planners spent several more in preparation for the working sessions. This pace may seem leisurely and the attention to detail pedantic in the face of the everyday problems of sick and dying patients who cannot wait for an idealized plan, but the worldly pressures simply must be resisted during the hatching period.

Realism in planning

On the other hand, the real world should not be shut out when projecting the proposed conduct of a bedside exercise. The opinions and the behavior of nurses, laboratory technicians, and other health care workers need to be taken into account since their cooperation is indispensable. It is unrealistic to expect that a trial can be carried out without a hitch unless there is fairly

complete understanding of the rationale and approval of the goals of an experiment involving a team of caretakers who have a strong sense of duty to protect the welfare of their patients.

Resistance of caretakers A poignant illustration of the need for understanding and cooperation took place in 1952 when the first effort was made to examine the effect of oxygen restriction on RLF occurrence by means of a formal controlled trial. The study was conducted in a large municipal hospital in Washington, D.C. The hospital nursery was staffed by experienced nurses who had been using oxygen liberally in the supportive care of small babies for many years, and they were convinced of the importance of this form of preventive treatment.

When two young investigators carried out the controlled study they discovered, belatedly, that the plan had been contaminated. The nurses believed that the doctors were going to jeopardize the lives of babies assigned to the 'curtailed oxygen' group. At night, some of the older nurses turned the oxygen valves on for babies who were not receiving the gas, then stopped the flow when they went off duty in the morning. Well-meaning resistance of this kind (the word 'sabotage' is unfair) is not at all rare, and it may go undetected if a trial is not closely monitored.

Practical limit At an early stage in the step by step development of information about two treatments, reality often imposes a practical limit on the specificity of the question addressed in a comparative trial. For example, Peto and co-workers noted that a realistic question may be, Is it better to adopt a *policy* of Treatment A if possible, with deviations if necessary, or a *policy* of Treatment B if possible, with deviations if necessary, for patients who seem to have a particular disorder?

This wobbly question concerning management may be more relevant for the moment than a higher order question, Is A better than B for the disorder? The refined question must simply wait until physicians and patients are ready to follow narrowly prescribed policies in future trials.

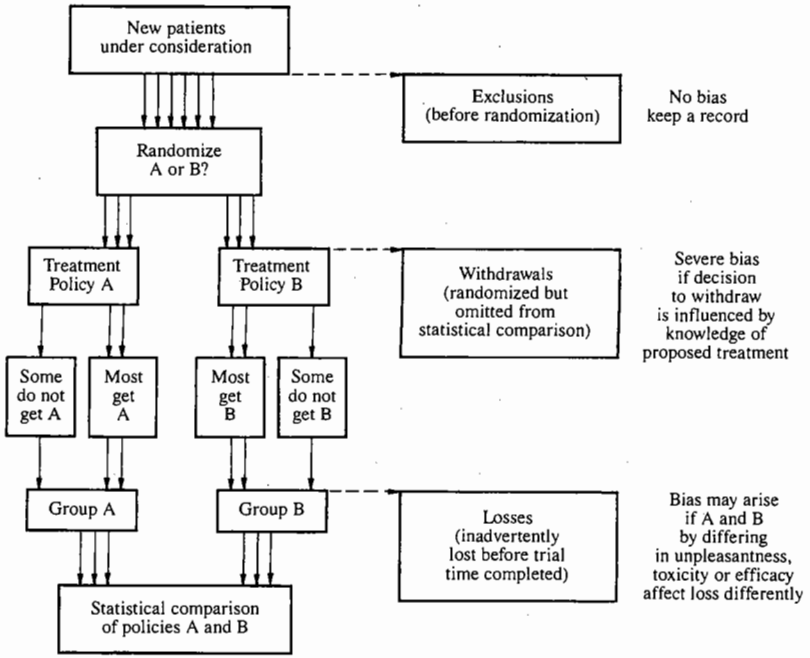
Irregularities in the trial plan

In a pragmatic trial, an accounting should be made of all patients who are considered for enrollment, including those who do not satisfy the requirements for prescribed treatments. I mentioned the problem of 'irregulars' earlier, now I wish to distinguish between exclusions, withdrawals, and losses.

Exclusions Patients excluded before randomization do not bias the treatment comparison. However, it is important to report the number of patients

Exclusions, withdrawals, and losses

In a pragmatic (management policy) trial



(Redrawn from Peto and co-workers' report to Britain's Medical Research Council)

left out and the reasons given for exclusion in order to make comparisons with other series.

Withdrawals Patients deliberately removed from a trial after randomization cannot be dismissed from consideration easily. Results may be distorted by the absence of such post-allotment withdrawals. It is necessary, for instance, to adopt an explicit rule to deal with conditions which are difficult to diagnose. If a review indicates that a patient was enrolled on the basis of an error in diagnosis, the participant may be withdrawn and classified as an 'exclusion'. However, we need to be assured that this decision was not influenced by knowledge of the treatment that the patient would have received.

The working rules about all withdrawals (especially partially treated patients who defect) need to be carefully thought out and specified in some detail to guard against systematic bias.

Losses Patients who do not return for specified follow-up examination present an awkward problem in medical studies. Although it is the common practice, for instance, to accept the length of life prior to loss in a comparison (using a life-table correction for duration of observation), this is not completely satisfactory because the lost patients may turn out to be a special class. No effort should be spared to document the fate of patients who disappear. The considerable time and expense needed to pursue missing patients should be regarded as essential investments.

Conventions concerning deviations

In discussing procedural biases, I noted that it is unrealistic to expect error-free treatments in a planned trial and I presented the argument that in a pragmatic trial the results of the treated-by-error patients should be included in the results for assigned treatment, despite the mistake. Disagreement about this point is common, but the logic is fairly straightforward. Peto pointed out that, when all the deviants are retained, the conclusions will be changed only if deviations are more numerous in one treatment group than in another and if the outcome in the deviants differs markedly from that in protocol-adherents. But in both of these circumstances, disclaiming the deviants is certainly not justifiable. Thus, removal of protocol-deviants becomes either irrelevant or invalid, and the safe general rule is to retain the results of all randomized patients in the final numerical analysis.

It is often of interest to describe the results among those who received treatments exactly as prescribed. But this is a data-dredging operation; there is no valid way in which the laws of probability may be invoked to help with the interpretation of any differences found in such a selected comparison.

Standard criteria for deviations I should point out that efforts must be made to regularize departures from described treatments whenever possible. For instance, an 'index of failure' of treatment may be devised and adopted as a practical end point. The convention avoids the confusion introduced by auxilliary treatments given when death is thought to be imminent. When the criteria for treatment 'failure' have been satisfied, the patient has completed his or her trial time; treatments and outcomes beyond this point are described, but they are not reckoned in formal results.

Unfortunately, this is only a partial solution to the problem of irregularity. If we are unable to guard against unequal application of the 'failure' criteria, the results may become hopelessly tangled.

Policing the trial

Mainland emphasized the need for a highly visible overseer to reduce the risk of unexpected entrapment in the conduct of treatment trials. The on-site coordinator may be given various titles, but his duties should be those of a policeman: to prevent trouble, to catch it quickly when it starts, and to be always ready to give immediate help when it is needed. Moreover, he must snoop into every detail of the on-going trial as he 'walks his beat'.

Randomized clinical trials are exceptionally difficult to carry out because they require restrained behaviour on the part of a profession that is licensed by society to act in an unrestrained manner to relieve human suffering. The duties of 'law enforcement' are sometimes assigned to a statistician, but I suggest that this is quite unfair. The 'night stick' belongs in the hand of a physician who is thoroughly familiar with all of the details of the current study and the deliberations that led to the wording of the protocol. And, most importantly, he must have an intimate knowledge about the setting in which the studies are carried out, including good insight into the character of each of the doctors, nurses, and others involved in the project, if he is to be effective in the prevention of systematic errors.

Coordination center In a multicenter trial, an executive unit must be designated to coordinate the exercise. This coordination center, which is usually physically separate from the participating institutions, receives, processes, edits, and analyzes data generated in the trial; and it provides epidemiological, statistical, and computational guidance. A central laboratory may be required to insure uniform observations of defined criteria (these may include chemical analyses, coding of x-rays and electrocardiograms, and so forth). A large-scale trial also requires a committee of experts, a detailed organizational structure (such as a policy board, steering committee, data and safety monitoring committee, and quality control group), and a clearly defined chain of command to control the multifold activities of the complex operation. Patients who agree to participate in a planned experiment deserve the assurance that 'law and order' are policed on their behalf.

Clinical trial committee
Specialist members in a complex multicenter trial

Expert clinician for field surveillance
Clinical pharmacologist
Epidemiologist
Biostatistician
Pathologist
End point specialist (e.g. electrocardiography . . .)
Laboratory specialist
Computer specialist
Expert in classification of disease
Medicolegal expert
Ethicist
Patient representative
(From Christian R. Klimt, University of Maryland)