7 The event of interest

In modern laboratory experiments, the efforts to reproduce predicted outcome events reliably are largely successful. The doubts that I mentioned in the last chapter need to be put into perspective. Accuracy in laboratory measurements has improved to the point that experimental errors in many procedures are regarded as merely nuisances that can be controlled by careful attention to detail. And, as I have suggested, such technical capability often leads to a dilemma when we must choose an outcome criterion in a pragmatic clinical trial. Should we record what we can measure with minimum error or measure what we think is directly relevant with the highest accuracy possible? In medicine, we often find ourselves in the position of the drunk who dropped his key in a dark hallway and was observed looking for it under the street lamppost: 'The light is better here,' he explained.

SURROGATE OUTCOME

Proxy outcome-events are often chosen in medical trials because of technical limitations, the impracticality of prolonged observation, and also because of moral restraints.

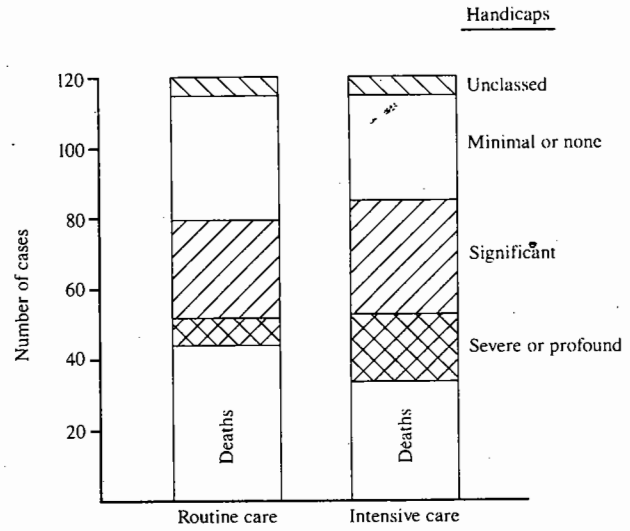
Survival as an outcome indicator

Concerns about mixed end results of modern care of premature infants have grown in recent years as life-support methods have become more intensive. A prominent question is, Have recently introduced, highly developed techniques of diagnosis and treatment improved the outlook for these small babies? The ongoing uncertainty has not been resolved, since much of the difficulty centers around the matter of deciding what outcome events should be chosen as measures of 'improved outlook.'

For many years the efforts to better the prospects for premature babies were measured by a decrease in neonatal mortality rate (deaths during the first 28 days of life among liveborn infants weighing less than 2.5 kilograms). Surveys in the 1960s, however, indicated that the relationship between rapidly falling mortality rate among the smallest neonates and the risk of survival with major handicap was complex.

Death and survival-with-handicap in small premature infants

An alternate-assignments trial of intensive care



Frequency of death and handicap at 8 years of age among 238 prematurely born children (birthweight 1-1.5 kilograms) who had been assigned, by Kitchen and co-workers, to routine or intensive neonatal care on an alternating basis on arrival in the hospital. Infants with gross abnormalities at birth were not enrolled in the trial. Severe or profound handicaps included major sensory, intellectual, and motor disabilities; significant handicaps included motor incoordination, epilepsy, and serious visual problems. (Redrawn from the figure of Kitchen and co-workers.)

An alternate-assignment clinical trial to compare the results of routine versus intensive care of premature infants was conducted by W.H. Kitchen and co-workers of the University of Melbourne between 1966 and 1970. The Australian researchers observed that an increased survival attributable to vigorous techniques of treatment may have been achieved at the expense of an increased number of severely handicapped children. Surveys conducted in the 1970s suggested that both mortality and the frequency of major

handicap have decreased, but the controversial issue about the full impact of modern life-support techniques in the management of very small babies has not been subjected to further testing with concurrent controls. Ethical conflicts make it virtually impossible to conduct rigorous tests of the questions about opposing end points—in this case, death versus disability—of different policies of treatment.

A conflict of perspectives When there is a conflict in outcomes, which event of interest is appropriate in a planned trial? It is impossible to avoid a value-oriented point of view. It happens, not infrequently, that the perspective of the medical profession differs from that of the community at large or from specific groups in a plural society.

Some of the misunderstanding about medicine's mission becomes evident when we examine the word 'lifesaving' as used to describe a medical procedure. The idea of 'saving lives' is deeply ingrained in the medical thinking (it is uncomfortably close to the evangelist's conception of 'saving souls'). But death is inevitable; it can merely be postponed, even by the most successful treatments.

Quality of life as an outcome indicator

The outcome of interest to each individual and to the community at large is life prolongation, rather than death. The former 'event' is difficult to measure. It is not a discrete function, it is a continuous variable with a concrete temporal dimension and innumerable value-defined qualities. In addition to short-term considerations, there are long-term effects of medical interventions that must be taken into account.

In a review of assessments in studies of surgical treatments John P. Gilbert and associates of Harvard University found most concern with immediate outcomes. Information about the quality of life of patients was

> *Euphranor*: Tell me, Alciphron, can you discern the doors, window and battlements of that same castle?
> *Alciphron*: I cannot. At this distance it seems only a small round tower.
> *Euphranor*: But I, who have been in it, know that it is no small round tower, but a large square building with battlements and turrets, which it seems you do not see.
> *Alciphron*: What will you infer from thence?
> *Euphranor*: I would infer that the very object which you strictly and properly perceive by sight is not that thing which is several miles distant.
> *Alciphron*: Why so?
> *Euphranor*: Because a little round object is one thing, and a great square object is another. Is it not so? ... Is it not plain, therefore, that ... the castle, *which you see there* [is not that] real one which you suppose exists at a distance?
>
> Bishop George Berkeley

usually missing. For proper evaluation of alternative surgical treatments, the reviewers argued, there is a need to assess the patient's residual symptoms, state of restored health, feeling of well-being, limitations, new or restored capabilities, and responses to these advantages or disadvantages. In the case of infants and children, for example, the immediate consequences of treatments are often dwarfed by those that become manifest during the long lifetime ahead.

Limitations of short-term studies The important weaknesses of studies of short-term effects of treatments are self-evident, but it is difficult to devise stronger approaches. The passage of time introduces confounding influences and imposes major impediments. It must be supposed, for example, that long-term outcomes are related to an early intervention, not to some intervening influences. Enormous organizational efforts are required to sustain follow-up study of highly mobile modern populations, and even the limited life span of investigators conspires against the best laid plans to conduct prolonged observations. Finally, the discovery of an unexpected outcome after years of observation, marks the beginning, not the end, of investigation.

Hippocrates was well aware of the constrictions imposed by time in medical study. The opening passage in his first book of *Aphorisms* notes: 'Life is short and Art long; the occasion fleeting; experience fallacious, and judgment difficult.'

ACCOUNTING OF TIME-RELATED EVENTS

The 'flight of time'—duration of various states and timing of events—needs to be examined in considerable detail so that the limitations of a circumscribed clinical trial are clearly in view.

The 'trial time' of each patient

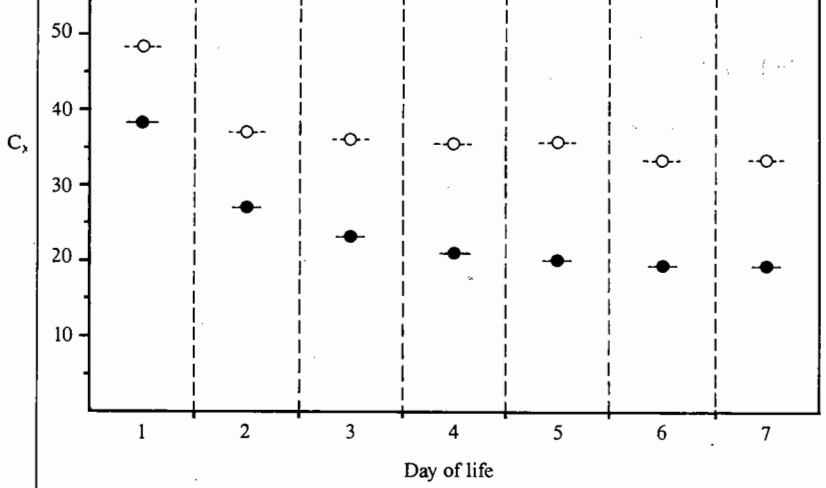
A number of complexities are introduced when patients are recruited over a designated period of time (they are rarely available all at once). In changing proportions over the course of a study, the enrolled population normally consists of individuals who have completed treatment, others who are undergoing treatment, and newly enrolled patients who are about to be treated. The duration of a study may be a fixed interval marked off by the calendar or a span of time that is determined by enrollment of a specified number of individuals.

The dizzying array of time designations is best expressed in terms of a basic unit, the trial time, determined for each patient. This interval of time begins at the moment of assignment-by-lot to a treatment category and

The life-table method of expressing outcome
Abbreviated version

A life-table is an efficient accounting form for summarizing the survival experience of an at-risk population over a specified period of time (newborn infants in the first week of life, for instance). For this approach, the number of individuals who were alive at the beginning of a specific age interval (l_x) and the number who died during that interval (d_x) are set out as follows:

l_x	226	85	62	51	47	46	44
d_x	141	23	11	4	1	2	1
p_x	0.38	0.73	0.82	0.92	0.98	0.96	0.98



The table indicates the survivorship of 226 male infants (birthweight ≤1.5 kilograms born during the years of oxygen curtailment, 1955–7) in the first 7 days of life. On the first day of life there were 141 deaths; thus the survival rate for this age interval is:

$$p_1 = 1 - \left(\frac{d_1}{l_1}\right) = 1 - \left(\frac{141}{226}\right) = 0.38, \text{ the } \textit{observed survival rate}$$

and on day 2 there were 23 deaths (d_2) among $226 - 141 = 85$ survivors (l_2) thus,

$$p_2 = 1 - \left(\frac{23}{85}\right) = 0.73$$

$p_3 \ldots p_7$ are calculated similarly

These observed rates provide a rough estimate of the probability of survival for each of the days after birth. We may then argue that the way to live 7 days after birth is to be alive for 6 days and then live one more day. Thus, the probablity of living 7 days is the probability of living 6 days multiplied by the chance of surviving day 7:

$$C_1 \times C_2 \times C_3 \ldots \times C_7$$

where,

C_1 is the chance of surviving at least one day after birth
C_2 is the chance of surviving a second day *after* surviving the first day of life
$C_3 \ldots C_7$

The true values for these cumulative probabilities (*C*s) are unknown, but we may estimate any one (C_7, for example) by calculating the product of the *observed* survival rates:

$$C_7 = p_1 \times p_2 \ldots p_7 = 0.38 \times 0.73 \ldots \times 0.98 = 0.19$$

The life-table estimates of C_1 through C_7 for males are plotted in the graph –●– (The values for females are plotted for contrast –○–; the base numbers are not shown.)

The life-table approach is a useful way to express survival after treatment (trial time, not age, is used for this application). Importantly, it draws attention to the issue of duration of exposure to risk and durations of observations.

ends with the occurrence of an event of interest (such as time of death, disappearance or appearance of an arbitrary manifestation).

In most medical trials, all patients do not experience the event of interest; the trial time runs to the pre-designated termination of the period of observations. Inevitably, there are incomplete observations because enrollment terminates on an agreed upon date before the latest participants have completed the prescribed period of observation, because patients are withdrawn or otherwise default, and because patients fail to return for follow-up in a long-term study. These unequal time durations need to be accounted for in some orderly fashion.

Life-table accounting of events

The life-table method of 'bookkeeping' (used to make actuarial calculations) provides a very useful solution to the maddening time-related problems of accountancy in medical trials.

The basic idea behind the life-table approach to the expression of event rates is found in the following statement: to survive a whole week, a patient must survive each of the 7 days comprising it. Although seemingly trivial, this simple tautology is the key to an efficient scheme for expressing outcome. For example, among 439 very small babies born over a three-year period (1955–7), only 116 survived the first week of life: 44 of 226 males (19 per cent) and almost twice that proportion among females, 72 of 213 (33 per cent). This overall summary provides no estimation of the gradations of risks in the two sexes that occurred from birth through the seventh day. But they are clearly expressed by the use of life-table bookkeeping to record the experience.

We begin with 226 boys at birth, for instance, and record the losses incurred by this cohort on each day of life. During the age interval 0–24 hours (day 1) there were 141 deaths, leaving 85 who were available to undertake the risks of the second day. Of these, 23 succumbed on day 2; and the decimations continued until there were 44 boys left to face the risks of the seventh day of life. Daily outcome rates are now calculated and these are used to calculate estimates of cumulative probabilities in actuarial fashion.

The initial cohort of patients and the number alive at the beginning of

each age interval have been compared rudely to the contenders in each round of a steeplechase. Only the entrants who are present at the start of a round provide useful information about the risk of the up-coming circuit. And an estimate of the probability of completing the race is provided only by riders who successfully complete each round.

Life-tables may be elaborated to account for latecomers and withdrawals whose contribution to the estimates of risk is adjusted for curtailed periods and durations of exposure and observation. Feinstein has pointed out that the 'pat' solutions of the life-table must not be accepted uncritically. For example, the assumption that the age of infants transferred from various hospitals is the most important risk characteristic to be considered, should not go unchallenged. And the reasons for withdrawal may be more relevant than the age of withdrawal. Nevertheless, the life-table format makes it relatively easy for an 'auditor' to spot time-related problems of accountability which remain hidden in other forms of documenting outcomes in comparative trials.

CONFIRMING CRITERIA FOR OUTCOMES

Four requirements have been proposed for the characteristics of observations used to decide that an event of interest has, in fact, taken place. Individual diagnostic criteria or tests that are used in medical trials should be reproducible, discriminatory, accurate, and as simple as possible.

Reproducibility of confirming observations

The reproducibility (yet another term for precision, p 84) of confirming observations depends on the inherent consistency of the measurement or test in *everyday* circumstances, the constancy of the object or quality that is measured, and the ability of the observer to interpret and record what he or she has measured. Some of the general sources of variation in observations were discussed in Chapter 6. Here I wish to emphasize the variation in phenomena as they take place at the bedside.

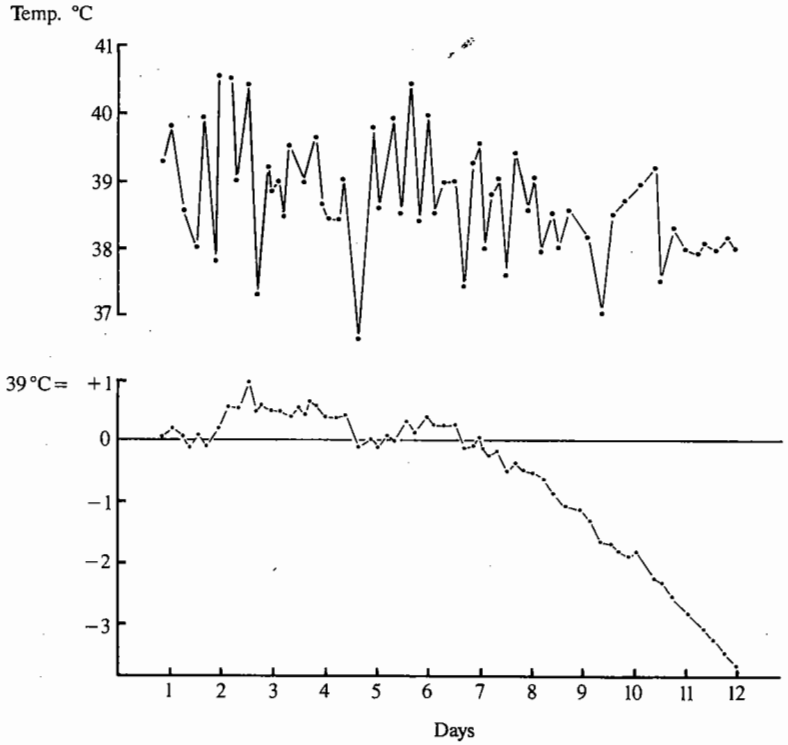
With few exceptions, the characteristics of patients change with time, many manifestations fluctuate (regularly and irregularly), and the variation in states is made to appear even greater when the conditions are measured indirectly by signs determined on physical examination. Thus, it is the pattern of dispersion of the diagnostic observations that is the replication sought in sets of clinical observations.

Observer error versus observer variation It is useful to distinguish between two kinds of misclassification made by observers. As we have seen (p 81), *observer error* refers to the mistakes that can be demonstrated either by

Cusum plot method of expressing trends

Serial measurements of characteristics which fluctuate widely (like body temperature in febrile states) are difficult to summarize, and trends are often obscured because of the scatter of values. The plotting of cumulative sums ('cusums') is a useful method of transforming unruly numbers.

The first step is the selection of a reference value, such as the approximate mean of the original data points. In the example given by Herbert Wohl of the University of California (serial body temperatures in a patient with a serious blood disorder),



the plot in the upper panel indicates the observed fluctuations. Choosing 39°C as a reference point, it is subtracted from each data point in succession. Any remainder is added algebraically to the previous sum. If the temperature remained at exactly 39°, the plot would remain at the reference-zero line.

The cusum values are plotted in the lower panel and the transformed scale clearly demonstrates a downward sloping line beginning on day 8 (the trend is difficult to make out in the conventional fever chart in the upper panel). A change in slope represents a change in mean value; the distance from the reference value is disregarded—only a change in slope matters. The greater the change in slope, the greater the change in mean.

majority opinion of a number of observers, by the same evaluator at a subsequent re-examination, or by an independent criterion of assessment.

On the other hand, disagreement between observers or inconsistent evaluations by the same observer may not stem entirely from perceptual error. *Observer variation* may be due to the fact that a significant number of observations fall close to the boundary between categories. For example, in an effort to improve the reliability of observations, two or more ophthalmologists may conduct independent examinations of the eyes of premature infants. When there is disagreement, the disputed interior of the eye is re-examined and discussed until agreement is reached. This may simply result in deference to the most experienced or most domineering observer. Forced agreement may obliterate the very fact that the verdict on certain examinations *is* doubtful. When categorical judgments are made concerning a process that is continuous in nature, the fact that disagreements occur most often near a boundary should suggest the need for a finer scale of measurement.

Standardization of observations Observer error and variability may be reduced by efforts to 'tune' the skills of observers. This is accomplished by practice sessions in preparation for formal studies; it is particularly useful in preparing for a collaborative study that will involve observers in different hospitals.

A descriptive 'standard' is also a helpful device for improving the precision of observations in clinical trials. For example, a verbal description and set of drawings of the appearance of the retina at each of the stages of RLF was used in the national study of this disorder in an effort to achieve reproducibility of diagnoses from one institution to the next.

The Rumpelstiltskin effect

Richard Asher pointed out the power of words in medicine: however uninformative the name of his or her illness may be, a patient feels the foe is partially vanquished once the name is disclosed. A typical exchange sounds like this:

'I seem to have an inflamed tongue, doctor. Will you have a look at it?'

'Ah, yes. You've got glossitis.'

'Thank you, doctor. It's all right now that I know what it is.'

The phenomenon is called the Rumpelstiltskin Effect, after the nursery story of the miller's daughter who got into the clutches of a dwarf. Having pretended she could spin straw into gold, she was put to the test, and she was in despair until the little man came and did it for her. The dwarf then blackmailed the unfortunate girl and would only relinquish his claims if she could guess his name. She managed to get hold of his name, finally, by a trick and was freed.

But no wonder the dwarf was confident that he could not be undone: Rumpelstiltskin was the kind of name unlikely to occur to a nicely brought up maiden; it means 'crinkly foreskin'.

Glossary of terms It is also important to recognize problems that arise with terminology in formal studies. Everyday definitions of medical terms are often imprecise; additionally, the bestowal of a name on a concept, whether real or imaginary, may bring it into clinical existence. It is useful to prepare a glossary as an appendix to the protocol, providing a list of terms that will be used in the trial and definitions to serve as the common standard.

Discrimination and accuracy of diagnoses

The characteristics of discrimination and accuracy as applied to confirming observations in clinical trials refer to the correct sorting in a diagnostic classification (the nominal and ordinal operations of measurement).

Ideal observations and tests would place all the unaffected in one class and all affected in their correct positions in the remaining class or classes. In practice, lack of discrimination may result from poor correlation between the degree of abnormality as shown by test and the severity of the target condition in fact. For example, the eye changes (that occur in a relatively small proportion of infants exposed to supplemental oxygen) may turn out to be highly inaccurate criteria for separating oxygen-affected and unaffected babies if, say, late appearing neurological signs should indicate that the frequency of oxygen-induced damage to the brain is higher than has been appreciated.

Expressing the accuracy of diagnostic tests

Result of a test	Confirmed* status of a condition	
	Present	Absent
Positive	a	b
Negative	c	d

'True-positive' ratio $= a/a+c$
'True-negative' ratio $= d/b+d$

The proportion of patients correctly classified as 'positive', (a), among a group who are affected by a disorder, ($a+c$), reflects the *sensitivity* of a diagnostic test. Similarly, the fraction of 'true-negatives', (d), among those free of the disorder, ($b+d$), is a measure of the *specificity* of the test.

These estimates are relevant only to the particular experience reported since the values are dependent upon the proportion of abnormals in a given sample. Moreover, the labels 'sensitivity' and 'specificity' determined in this way should not be regarded as inherent properties of the test or observation. (The estimates of predictive accuracy, i.e. $a/a+b$ and $d/c+d$, must be interpreted with considerable caution.)

* Confirmation by some independent criterion.

Evaluation of accuracy When there is an independent (and unequivocal) method of confirming the occurrence of an event of interest, the calculation of 'true-positive' and 'true-negative' ratios of observations provides a limited basis for describing their accuracy. These estimates cannot be extended (with any assurance) beyond the particular experience under study, but they do provide a method of comparing different diagnostic criteria for the same event and between-observer differences in diagnoses.

Simple criteria of outcome

Finally, the requirement of simplicity of differential criteria takes into account the practical matters of patient comfort, time, and expense in carrying out studies involving suffering human beings. Careful consideration must be given to the cost, in these practical terms, whenever elaborate tests and observations are proposed in exchange for relatively small gains in discrimination.

At every step in the planning of studies that require people to enroll in a regimented program, we must return to the questions concerning external relevance: To whom are the results of the study meant to apply? At the event-of-interest step we must ask, Are the diagnostic criteria practical for everyday application? The pros and cons as seen from a community-wide perspective must be weighed before deciding on the observations and tests that are to be used in a bedside trial.

TARGET EVENTS

How many target events may be 'lined up in a row' in a focused clinical trial? Common sense and good citizenship require that we try to obtain as much firm evidence as possible when we undertake an exercise that places so many demands on the participants and on the community's resources. But how can these stipulations be met, given the actions of the 'inconstant, dangerous, and delicately balanced' goddess of fortune? (As the number of end-points increases, the likelihood of chance associations *must* rise.)

The dilemma can be resolved if we make a clear distinction between three kinds of targets: a 'called shot', several 'practice targets', and lastly, unexpected 'hits'. If we are to use the logic of chance as a guide to the interpretation of the occurrence of observed events, we are forced to return to the aim of the study: the results must be examined in the context of the pre-trial questions and the details of the experimental design.

Primary event of interest

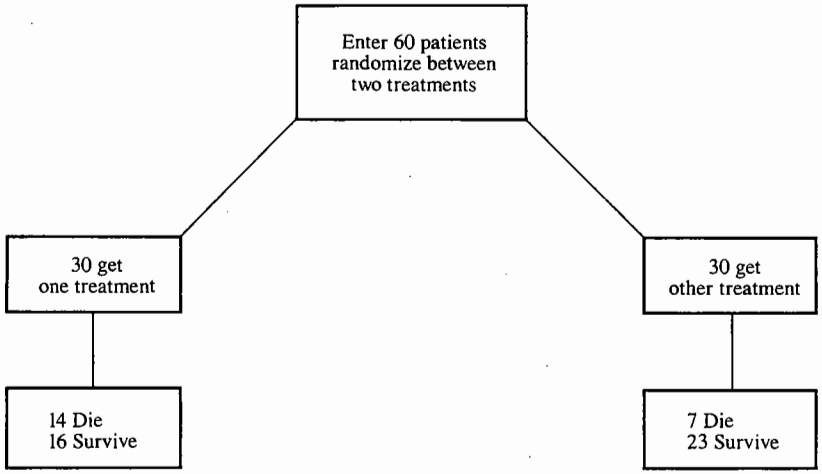
The primary target event is the outcome that is defined in terms which relate to the specific question posed before the trial begins. (In the national RLF trial, the primary outcome of interest was the appearance of scarring

eye changes during a trial time that concluded when each infant reached the age of 2½ months.) It is the number of such outcome events that determines the dimensions of the trial, and it is the prediction about these numbers that is put to a severe test. (The proportional proposition that risked refutation in the RLF trial stated that curtailed oxygen management was expected to reduce the frequency of the scarring form of RLF in enrolled babies from a little over 10 per cent to about 2 per cent or less).

> Since the principle of life in animals is a force which is ever active, which is constantly endeavoring to overcome obstacles, and since nature when left to its own devices cures many diseases by itself, it follows that when a remedy is applied, it is infinitely difficult to determine what effects are due to nature and what to the remedy. The result presents itself to the wise man merely as a greater or lesser probability, and that probability can be converted into certainty only by a large number of facts of the same kind.
>
> Lavoisier (1784)

Total number of primary events Comparative trials are relatively insensitive to fairly substantial true differences between treatments because chance variations in outcomes between groups of patients tend to be quite large. The fluctuations in the usual small-scale trial may either obscure true differences or excite false interest in a new intervention.

A chance difference in mortality between two groups of patients



In a report to the Medical Research Council (Britain), it was noted that results at least as extreme as these occur in about 10 per cent of small clinical trials which compare *equivalent* treatments. Variations of this magnitude can be expected as the result of chance allocation of patients at relatively high risk to one treatment group.

The ability of a planned test to distinguish between the merits of two treatments depends on how many patients suffer a relevant 'event' rather than how many patients are enrolled. A committee of Britain's Medical Research Council has emphasized, for instance, that a study with 100 patients, 50 of whom die, is about as sensitive as a study with 1000 patients, 50 of whom succumb.

The discriminating power of a trial also depends on the magnitude of the difference between treatments. The study of a few dozen patients can, in most cases, detect an unusually effective treatment which prevents two thirds of deaths, but more realistic effects, such as preventing about one third of deaths, requires well over 100 patients if the difference is to be detected.

The essence of performing a successful clinical trial, then, is to enroll a sufficient number of *at-risk* patients. I will postpone a discussion of how this 'sufficient number' may be estimated—particularly when there are multiple, and opposing, end-points of interest—until the chapters on The Stopping Rule and on Inferential Decision. The reasoning must be coupled with that used in making inferences from proportional propositions.

Additional 'active' observations

There are almost always a number of outcomes of interest in a formal study that are quite properly classified as 'active' observations, or practice targets. The opportunity to make preliminary observations about the occurrence of such expected events is an important part of every randomized clinical trial. There is, however, a fundamental difference between the two kinds of defined targets. The primary outcome is under critical test, but, if no clearly specified pre-trial predictions have been made about the additional 'active' observations, we can hardly claim that they have been in any great danger of refutation.

It is imperative that the additional observations be made, but it is equally necessary that they be clearly labeled as results of a pilot exercise, for it is the 'range' of these latter targets that is under investigation. Further testing is needed to explore the limits of applicability of the newly formed proportional propositions.

Oxygen treatment and survival The relationship between oxygen treatment and survival was a prominent pre-trial concern of the planners of the national controlled RLF trial, but the question was not framed in numerical terms that could be addressed in a formal way. The deaths of infants assigned to two oxygen treatment regimens were monitored week by week during the first three months of the study to determine if there was a systematic difference that could be made out. These *pilot* observations indicated only a small disparity in favor of routine (uncurtailed) oxygen

treatment. We can now see, with the wisdom of hindsight, how unfortunate it was that these preliminary observations were not followed by focused studies of the survival question.

Intensive treatment and major handicap The Australian planners of the alternate-assignments trial of intensive care (p 87) stated that they wished to determine if the innovative techniques of 'unproven benefit might even have detrimental effects.' Here the end points of the trial, death and major handicap, were given equal weight, but neither were primary outcome events in the strict sense required for a severe test. Like survival outcome in the RLF trial, the questions in the Australian pilot experience were not number-specific. Twelve years of study (four years of patient intake and eight years of follow-up) provided only an estimate of the range of differences that would be expected in rigorous tests of the *two* questions.

Unpredicted outcomes

The third kind of target event is the most difficult to deal with. On the one hand, the role of serendipity in investigation must not be dismissed. (The princes, in the fairy story *The Three Princes of Serendip*, were always making discoveries, by accidents and sagacity, of things they were not in quest of.) The opportunity to 'dredge' the hard-to-obtain data made available in a well-conducted large clinical trial simply must not be missed. Most commonly, this takes the form of searching for unpredicted associations in many subclasses of enrolled patients. And yet, as probability theorists never tire of warning, we must keep in mind the Fallacy of the Enumeration of Favorable Circumstances—if enough independent phenomena are studied and correlations sought, some will, of course, be found. At the time of an unexpected 'win', it would be well to remember the human drama that takes place in the gambling casino at Monte Carlo. When a patron rakes in a huge pile of chips after an *en plein* bet at the roulette table (thirty five times the amount placed on a single number), a tremble of excitement passes around the room. The 'house' merely yawns, and says, in effect, keep playing.