2 Framing the question

In an apocryphal story told by Neil Postman of New York University, an epidemic disease struck a small community and killed many people, but some of the afflicted recovered. The victims lapsed into a deathlike coma and it was hard to know when and, indeed, if they had succumbed. The townspeople worried about burying the 'dead' too soon, and they were hard-pressed for a solution to the dilemma. It was suggested that coffins be well stocked with food and an air vent provided just in case the victim happened to be alive. Although this was expensive, it certainly seemed worth the effort. However, a second proposal was made that was both inexpensive and quite efficient. A twelve-inch stake was to be mounted on the inside of the coffin lid exactly at the level of the heart. When the coffin was closed, all uncertainty would end.

It is of interest that the two solutions were generated by two different questions. The first solution was an answer to the question, How can we make sure that we do not kill people who are alive? The second was a response to the question, How can we be sure that everyone we bury is dead? The point, Postman noted, is that the only answers we get are responses to questions. Although questions that refer to certain assumptions may not be evident, they design the form that our knowledge will take and; thus, determine the course of our actions.

The parable should come to mind as a prelude to the design of clinical studies and before reading the reports of past studies. 'What is the question?' takes precedence over all other considerations.

> 'God is the answer!
> But what is the question?'
> Gertrude Stein

GENUINE QUESTIONS

In order to carry out its directive work effectively, the express question to be addressed in a formal investigation must have the property of authenticity. A question should be considered genuine only if it refers to a hypothesis that can be overturned by defined events. A pseudo-question, on the other hand, is one in which the inferred supposition is at no risk since it cannot be contridicted by any conceivable event. For instance, the query 'Does Galen's treatment (p 2) work?' is a pseudo-question. The claim of infallibility is simply untestable; all treatment failures are ruled out by classifying these unfortunates as 'incurable'.

It is the property of refutability, philosopher Karl Popper has pointed out, that separates scientific questions from those in metaphysics. Moreover, the claims implied in explicit questions are more testable than those in non-specific statements. The former take greater risks of being overthrown and, as a result, are highly productive. What is envisioned is the Galilean interplay of question and experiment: step-by-step challenges of explicit claims with progressive narrowing of the area of uncertainty.

SEARCHING FOR QUESTIONS

Where do the questions come from? Traditionally in medicine they emerge from a background of observation, and I want to turn now to this concept-seeking function of descriptive information. (I will postpone a discussion of other aspects of observation, e.g. sense perception, observer behavior, and measurement until Chapters 6 and 7, which deal with outcome observations in experimental trials.)

Classification of question-seeking observations

Claude Bernard recognized two levels of observation:

> '... A spontaneous or passive observation which the physician makes by chance and without being led to it by any preconceived idea ... [secondly] an active observation ... made with a preconceived idea, with intention to verify the accuracy of a mental conception.'

Although the dichotomy is only relative (since the notion of unprejudiced observation is a myth), the classification is, nonetheless, very useful. It deserves a close look. Notice that the distinction between 'passive' and 'active' is made on the basis of the mind-set of the observer.

'Passive' observation A 'passive' observation is made by chance in the sense that it is unexpected according to the (unspoken) prevailing theory about the state of the world. The event or circumstance seems novel because

it was unpredicted and the observation may be considered 'passive' because the viewer did not prepare the physical act of perception with the mental acts of *new* theory formulation and forecast.

'Active' observation By contrast, the 'active' observation is made after some mental work has been performed. The point here is that a second-level observation is carried out in a newly defined framework of meaning specified in advance by the preconceived idea.

The hierarchal distinction is quite important because it takes notice of a progression in the pre-experimental screening of question-seeking observations. Since there is no end to the number of observables, we need to pick out the observations that are worth further exploration. The move from noting unpredicted incidents to the focus on prediction-confirmed events is just such a culling action; it narrows the search for challenging questions.

An epidemic of blindness

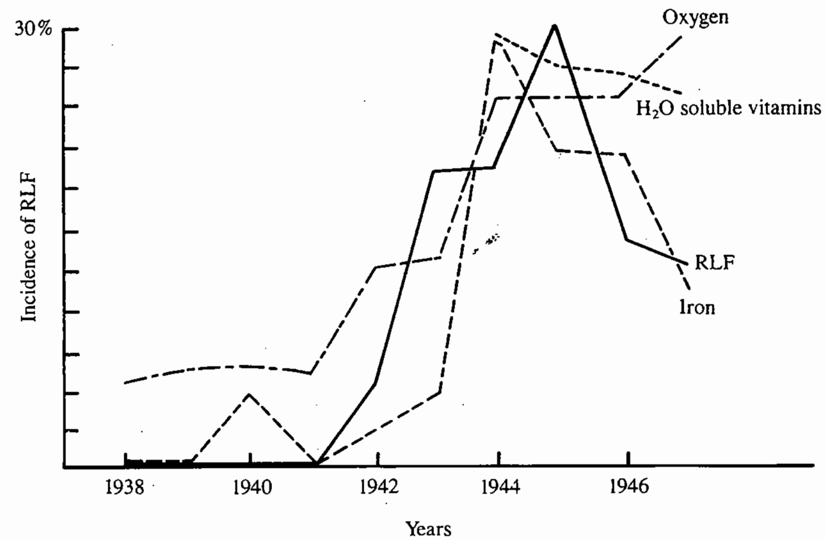
Concrete examples of the two levels of observation took place in 1949 and 1951 when an unexplained epidemic of blindness—retrolental fibroplasia, or RLF—affecting prematurely-born infants, raged throughout the United States and, to a lesser extent, in other countries throughout the world (see Appendix for a history of RLF).

Boston survey The experience with RLF in several US cities was surveyed in 1949 by V. Everett Kinsey and Leona Zacharias of Boston's Eye and Ear Infirmary, in the hope that some cause might be identified to account for the sudden increase of the previously rare disorder. Forty-seven factors relating to mothers and infants delivered in Boston between the years 1938 and 1948 were examined and these (for example, complications of pregnancy, delivery, and the newborn period; treatments administered to mother and baby; and so forth) were correlated with the occurrence of RLF.

The investigators had two goals in mind: to enumerate the conditions and treatments experienced by children affected and unaffected by RLF and to compare the trends of factors and outcomes over an interval before and after the sharp rise in frequency of RLF. But they had no single and *specific* hypothesis in mind. (The analytic procedure used has been termed 'data-dredging'; data obtained in the past are disintered, as with a dredge, to see what turns up.)

On completion of the survey, the observers discovered associations between increased RLF-blindness in premature infants and treatment with iron salts, the administration of water-miscible vitamins, and liberal use of oxygen therapy. Curves demonstrating time trends of the unpredicted correlations led the Boston authors to report that 'correlation between the rise

'Passive' associations in a Boston survey



Occurrence of RLF-blindness and the administration of three treatments to premature infants in a Boston nursery from 1938 to 1947. The vertical axis indicates the occurrence of RLF (range 0–30 per cent), the duration of oxygen treatment (range 0–15 days), drops of water-miscible vitamins (range 0–400 drops), and grains of iron sulfate (range 0–60 grains). The study population consisted of 53 RLF-blind and 298 sighted children born some years before the study began. (Redrawn from the data of Kinsey and Zacharias)

in incidence and [treatments] was less striking for oxygen than for water-miscible vitamin preparations and for iron.'

Rarity, interest and surprise Before going on, it would be worth pausing for a moment to consider the concepts of rarity, interest, and surprise as they relate to observations of events in medicine.

For example, the Boston analysts were quite justified in regarding the associations found in their survey as rarities: simple inspection of the distribution of the frequencies is convincing. (Even without statistical arithmetic, it is safe to say that co-events like these would be expected to occur rarely by chance alone.) Moreover, the associations were unquestionably interesting: the analysts (and the baffled medical community) were very much interested in the findings of the survey. What remains to be examined is the matter of surprise: Were the results surprising?

The requirements for justified astonishment were set out many years ago by mathematician Warren Weaver. He argued that an event should not be considered surprising merely because it is rare (i.e. the probability of occurrence is small in an *absolute sense*), but rather because its probability is quite small as compared with the probabilities of other possible alternatives.

In the Boston survey, a large number of 'alternative' correlates were examined (when the survey was undertaken, there was no theory concerning a mechanism of action of these factors). Thus, the likelihood of concurrence of any one of these forty-odd variants with RLF, simply as a meaningless coincidence, was of the same low order of probability as a fortuitous association with iron salts, vitamins, and oxygen. It was reasonable to conclude that the associations observed in this survey were rare and quite interesting. But they were *not* surprising. In fact, if the dredging process had continued long enough, other associations would have been found, as guaranteed by the definitions of improbable events.

The fundamental distinction between impossible events and improbable events is that the former cannot occur and the latter *must* occur if the observations continue indefinitely. (Following the Boston survey, other searches turned up haphazard correlations between the occurrence of RLF and cow's milk feedings, blood transfusions, low fluid intake, and rapid cessation of oxygen treatment.)

This, then, is the underlying weakness of 'passive' observations: incredibly rare events are occurring all around us; if we single out those which have already occurred because they are of interest to us, we cannot, with any confidence, attribute 'significance' or 'meaning' to their occurrence. We have no right to be surprised. And this is the basis for a fundamental tenet of the scientific method: hypotheses to be tested must be formulated before examining the data that are to be used to test them.

Melbourne prediction How does the situation differ in the case of 'active' observations? Here the relationship between the possible alternatives is changed by declaring the outcome-of-interest *before* examining the outcome-in-fact.

For example, in 1951, an Australian physician, Kate Campbell, heard a rumor that RLF frequency increased in Britain when oxygen use was liberalized (after inception of the National Health Service made increased funds available for the purchase of medical equipment). She proceeded to examine the experience of babies in three Melbourne hospitals with the preconceived idea that RLF occurrence might be related to the use of oxygen. And she found that the disorder was most frequent in the institution which used oxygen freely.

The association found by this 'active' observation was more credible than the correlations noted earlier in Boston. Note, however, that the improved quality of the new evidence was not based on the numbers of babies in-

RLF in Melbourne

Years	*Free use of oxygen*[a] (*Institution I*)		*Conservative use of oxygen*[b] (*Institutions II and III*)	
1948–50	No. of Infants 123	No. with RLF 23	No. of Infants 58	No. with RLF 4
RLF%	19%		7%	

a. In Institution I oxygen was piped into the ward and was given in an oxygen cot: the percentage of oxygen ranged between 40 and 60 per cent. Oxygen was given before symptoms appeared as well as during periods of obvious need (blue spells).
b Institutions II and III oxygen was administered sparingly (by a catheter placed in the nose or by a funnel placed over the face, sometimes by tent or closed cot).
(Taken from the observations of Campbell)

volved in the two sets of observations nor on any calculation of the comparative rarity-by-chance of the Boston and Melbourne experiences. Both sets of observations were certainly of considerable interest. But the distinctive qualities of the 'active' observation were: 1) the Melbourne thesis satisfied Popper's requirement—it risked failure by predicting a high frequency of RLF only in hospitals using oxygen liberally, and 2) it satisfied Weaver's requirement for surprise—the association was unusual as *compared with all other associations considered together*.

It is as if I declare that only a bridge hand of thirteen spades interests me and lump together all other hands as imperfect. The probability of drawing any specified bridge hand on a single deal is vanishingly small (1 divided by 635 013 559 600) and none are surprising since each one of the thousands of millions of possibilities is equally rare. My proclaimed hand becomes surprising only when I compare it with the sum of the probabilities of the alternatives (the likelihood of receiving an imperfect hand approaches certainty). Moreover, if I am dealt a hand of thirteen spades, I may become

A misunderstanding about chance

Horace C. Levinson, the mathematician, illustrated a common misunderstanding about *a priori* probability by relating the story of an embattled sailor who put his head through the first hole made in the side of his ship by an enemy cannon ball. The man reasoned that it was highly improbable that another ball would come through the same hole.

Although the sailor was correct in believing it was highly unlikely that two balls would hit the ship at the same spot, he was entirely mistaken in his belief that after the first strike the likelihood of an identical repeat is smaller than the chance it would hit any other spot *designated in advance*. Before the engagement began, the betting odds were huge against two balls landing on the same spot, but once half the 'miracle' had been accomplished, the betting odds were immediately reduced to the odds that any indicated point would be safe.

suspicious about the dealing process or the dealer, but the happening is invested with considerably more 'meaning' if I declare my suspicion *before* the deal. Again, the improbability of the event as reckoned before its occurrence is the same in both instances, but rarity of the hand is not as important as the level of observation.

The mere fact that a probability is low should not in itself lead to amazement. I am not justified in declaring, 'This event *must* have some meaning because the probability of occurrence by pure chance is incredibly remote—surely nothing as improbable as this could ever occur.' An evening of bridge is convincing evidence that the occurrence of events of fantastically small probabilities is, in fact, inevitable.

DATA-DREDGING PROCEDURES

Quite often in medicine, we are faced with baffling disorders or phenomena and we are forced to begin the search for meaning with very weak questions (as in the 1949 survey: What can account for the rise in RLF?). In the absence of any reasonable specific theory, it is entirely reasonable to undertake an epidemiologic survey of unplanned events which have already occurred, but it is prudent to take some precautions.

The problem has been likened, by economists Hanan C. Selvin and Alan Stuart, to a hunter stalking an unknown quarry through an unfamiliar landscape with an arsenal of complex weapons. The evocative names they have given to two of several data-dredging procedures in survey analysis are 'fishing' and 'hunting'.

'Fishing'

When pre-search questions are vague, the principal motive for embarking on a survey is to provide the observational material from which a precise theory may be formed. This process of angling for a model is quite literally a 'fishing expedition', using the observed data to choose which of a number of candidate determinants to include in an explanatory thesis.

Once a 'fish' is caught, the survey analyst must have a different body of data to evaluate the explanatory model developed in the first search. This need not involve a new set of observations if the initial collection was divided (with suitable precautions to insure that the sub-sets are not too dissimilar) into two parts for just this purpose.

For example, the Boston survey might have employed this 'double-pond' approach by using half of the observations for 'fishing'. When the three 'fish' (iron, vitamins, and oxygen) were landed, the remaining half of the experience could have been used to make 'active' observations about the associations in the unexamined pool of information held in reserve.

'Hunting'

'Fishing' uses a limited number of candidate variables, but 'hunting' involves no restrictions. The data are widely probed for information in the area of interest. The practice can take many forms. Basically it involves searching through a body of data and examining many relations in order to find rare co-occurrences. Or one can examine a single hypothesis seeking confirmations in many different bodies of data.

The practice leads to a pernicious problem when only 'significant' results are published, but I will return to this in a later chapter. Suffice it to say here that 'hunting' offers the maximum scope for the data stalker since there are no ground rules.

A 'hunting accident' The history of medicine is rich with stories of 'hunting accidents'. One such event occurred in 1949 when the results of treating 208 newborn infants with exchange transfusion were examined. This technique

A 'hunting accident'
Sex of blood donor and survival of babies after exchange transfusion*

Sex of blood donor	*Outcome of exchange transfusion* *No. treated*	*No. survived*	*Per cent survived*
Male	137	110	*80%***
Female	42	42	*100%***

* Newborn infants suffering from anemia and jaundice (erythroblastosis fetalis).
** These proportions were processed with statistical arithmetic and the difference was declared 'significant'; that is, the likelihood of obtaining a similar difference in survival by chance alone is less than 1 in 100 repeated trials involving comparable numbers of patients.
(Taken from the data of F.H. Allen, Jr. and others)

of replacing blood had been introduced the year before as a method for preventing death and brain damage in infants who were born with erythroblastosis fetalis, a severe form of anemia and jaundice caused by Rh-blood-group incompatability between mother and fetus. The analysts wrote that they were surprised to note that although mortality was over 15 per cent in the whole group, there were no deaths in a group of 42 babies who happened to receive blood from female blood donors exclusively.

The unexpected observation prompted a searching analysis in the *same* collection of data (10 other state-of-the-mother-and-infant variables were examined) and the authors concluded that 'it appears certain that exchange transfusion using blood from a female donor is the treatment of choice for babies with erythroblastosis fetalis. In the past two months we have given this type of blood by exchange transfusion to 13 babies with erythroblastosis fetalis. All have recovered.' A report of the female-donor benefit appeared as the lead article in a highly-regarded medical journal and attracted widespread interest. However, the authors had wisely advised that others examine their own experiences for evidence of a similar beneficial effect. It was quickly found that the hopeful results could not be duplicated in subsequent studies.

PROBLEMS INDUCED BY VARIABILITY

I have noted that questions and bold new theories in medicine arise from a number of sources. Informal, yet accurate, on-going personal observations of events by practising physicians (case studies) have been fundamental to progress in the past and they remain indispensable as a source of original hypotheses. But the shift of emphasis from the individual patient to groups of patients (populations) as the unit of study reflects a change in thinking about causality in medicine.

A simple view of causal relationships, a single cause resulting in a single effect, is gradually being replaced by a concept of multiple causes and variable consequences.

Variation in medical events

The change in outlook emphasizes the difference between the type of uniformity that can be contrived in the laboratory or seen in repetitions of phenomena that give the same result over and over (such as the time required for a ball to fall from a fixed height) and the case to case irregularity that characterizes medical incidents (for example, the course of illness with or without treatment). Invariable outcomes, such as death or disability, are rare, and infallible treatments are virtually non-existent in medicine.

It is the variability in clinical events that leads to the need for *collections* of observations to generate explicit (number-specific) questions and the need for some method of untangling the strands in a web of relationships so the questions can be tested rigorously.

Oxygen and RLF Many of the problems of variability were encountered during the early attempts to explain the rise of RLF. I said that the preconceived Melbourne idea concerning a causal relationship between oxygen treatment and RLF received substantial support from the observations in compared hospitals (p 18). Now I must point out the weakness in that evidence.

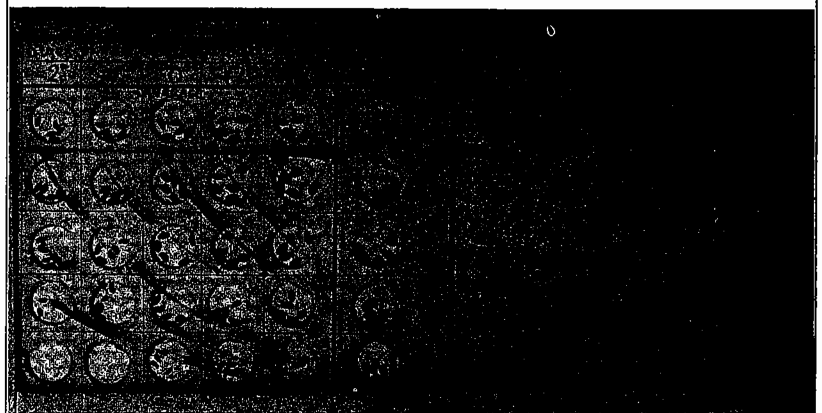
First, the majority of infants who received liberal oxygen treatment did not develop RLF; only one-fifth of the babies so treated were affected. Clearly, other risk determinants exerted a considerable influence on outcome. And there was no assurance that these unfavorable factors were distributed equally among the groups of babies in the surveyed hospitals.

Secondly, the prediction concerning an association between treatment and outcome failed to specify *in advance* the size of difference in RLF occurrence that would be accepted as decisive. Since the range of possible differences was large in this highly variable disorder (a wide range of occurrence was reported in groups of babies who were presumed to be treated uniformly, see p 35), a single confirmation of the numberless prediction was not entirely convincing.

Francis Galton's contribution to the study of cause and effect

Francis Galton's introduction of the idea of correlation in 1889 opened the way to a deeper study of the problems of cause and effect. Up to that time, only simple causal relations could be described in quantitative terms. His concept provided a means for mathematical analysis of multiple causes: 'the degree of relationship, or of partial causality, between the different variables of our ever-changing universe' could be represented by a system of numbers.

Galton began by comparing the size of two generations of seeds from cress plants (the size of seeds sown and the size of seeds produced by their progeny). He attempted to formalize the associations by sorting the mother seeds into five size categories and similarly ranking daughter seeds produced in each of the seeds-sown categories. Thirty groups (5 sown and 25 produced; each size grade in small, round containers) were set out in a tabular array:



The sixth vertical column of Galton's 'table' (labeled 'size of seeds sown') ranked the five containers of parent seed by size (from the largest '+2°' to the smallest '−2°'); each of the horizontal rows displayed the seeds of progeny in similar rank order. The array was covered with a sheet of glass and a series of contour lines were drawn connecting seed ranks of similar size (Galton termed these 'isograms'; the lines are smudged and almost obliterated).

The rough observations contained the germ of an idea for measuring the intensity of resemblance between characteristics exhibiting a range of variability due to 'different and equally probable combinations of a multitude of small independent causes.' Galton went on to develop a mathematical approach to associations between many kinds of measurements; the dimensions were expressed in statistical terms and relationships were summarized as coefficients of correlation. His insight revolutionized the study of natural phenomena.

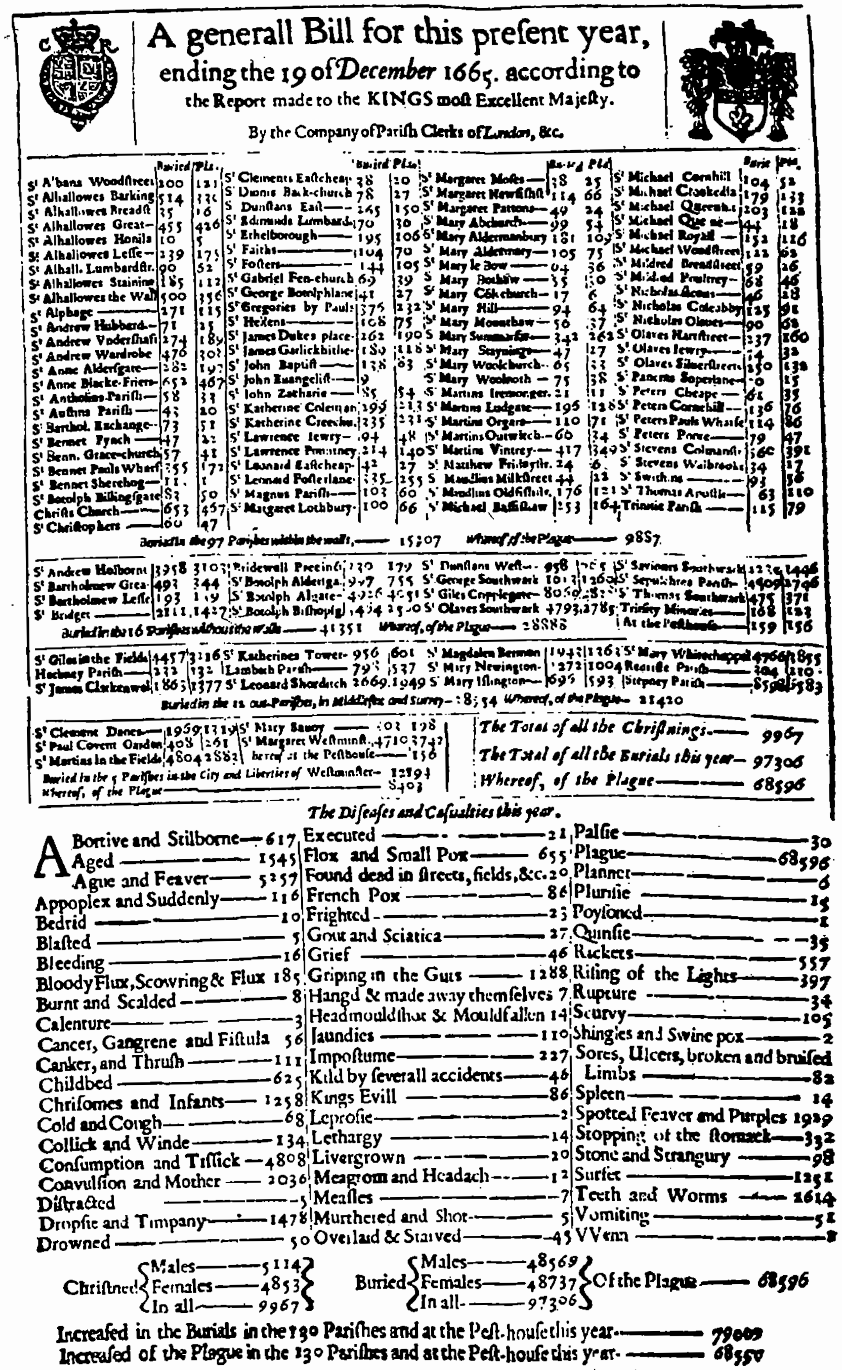
Discordances between an isolated 'cause' and an isolated effect in medicine
(For Example: Supplemental Oxygen and RLF)

'Cause' and effect associations	*The 'cause' is necessary*	*The 'cause' is sufficient*
A	+	+
B	+	−
C	−	+
D	−	−

A. The 'cause' is necessary and it is sufficient to result in the effect. Such a straightforward relationship is referred to as *deterministic* causality. For example, if exposure to supplemental oxygen and nothing but such exposure was needed to cause RLF, it would be logical to consider this agent as the sole necessary and sufficient cause of RLF. In late 1954 when such a relationship was 'established', research activities into matters related to the disorder virtually ceased. Over the next 20 years, it slowly became evident that this simplistic causal model was inadequate.
B. The 'cause' is necessary but it is not sufficient to result in the effect. Here there is a producer-product relationship, and the term *probabilistic* (nondeterministic) causality may be applied. Early in the investigation of the association between supplemental oxygen and RLF, it was noted that exposure to supplemental oxygen did not result in RLF in all newborn infants. An additional factor, immature development of the blood vessels of the eye, was found to be a necessary determinant. Moreover, only a minority of immaturely developed infants exposed to the treatment developed RLF. Slowly it was realized that there must be other determinants not yet identified.
C. The 'cause' is not necessary but it is sufficient to result in the effect. The association is referred to as *correlation* (the purported 'cause' and the effect tend to be present or absent together); the association may not involve causality at all. For instance, it became evident that RLF occurred in a few infants who had never been exposed to supplemental oxygen. This supported the suspicion that there were multiple causes of RLF—or that the incompletely developed eye could be adversely influenced by the oxygen (21 per cent) in atmospheric air.
D. The 'cause' is neither necessary nor is it sufficient to result in the effect. For instance, RLF was found in some stillborn infants. Here the mechanism appeared to be related to severe oxygen *lack*. Recently there has been reason to suspect that exposure to oxygen after birth may be only a contributory determinant; the causal role may have been exaggerated when the gas was administered in high concentrations for prolonged periods of time (in the years before 1955).
(Adapted from Mervyn Susser's arguments)

The numerical method and the idea of formal comparisons in medicine

The concept of formal enumeration in medicine had its origin in the London Bills of Mortality which were kept regularly beginning with publication of the bill dated December 29, 1603. (Earlier Bills, which furnished the number of deaths caused by the plague compared with all other fatal sickness, began in 1532, but these appeared only sporadically.) The London Bills described medical events in the parishes; they were published on the Thursday before Christmas Day. In the year 1665, London experienced the last of many epidemics of plague:


A generall Bill for this present year, ending the 19 of *December* 1665. according to the Report made to the KINGS most Excellent Majesty.
By the Company of Parish Clerks of *London*, &c.

Buried in the 97 Parishes within the walls,—— 15207 Whereof, of the Plague—— 9887

Buried in the 16 Parishes without the Walls —— 41351 Whereof, of the Plague—— 28888

St Giles in the Fields	4457	3216	St Katherines Tower	956	601	St Magdalen Bermon	1943	1363	St Mary Whitechappel	4766	3855
Hackney Parish	232	132	Lambeth Parish	798	537	St Mary Newington	1272	1004	Redriffe Parish	304	210
St James Clarkenwel	1863	1377	St Leonard Shoreditch	2669	1949	St Mary Islington	696	593	Stepney Parish	8598	6583

Buried in the 12 out-Parishes, in Middlesex and Surrey — 18554 Whereof, of the Plague— 21420

St Clement Danes	1969	1319	St Mary Savoy	303	198
St Paul Covent Garden	408	261	St Margaret Westminst.	4710	3742
St Martins in the Fields	4804	2883	hereof at the Pesthouse		156

Buried in the 5 Parishes in the City and Liberties of Westminster— 12194
Whereof, of the Plague—— 8403

The Total of all the Christnings—— 9967
The Total of all the Burials this year— 97306
Whereof, of the Plague—— 68596

The Diseases and Casualties this year.

Abortive and Stilborne	617	Executed	21	Palsie	30
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in streets, fields, &c.	20	Planner	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	15
Bedrid	10	Frighted	23	Poysoned	1
Blasted	5	Gout and Sciatica	27	Quinsie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Rising of the Lights	397
Burnt and Scalded	8	Hangd & made away themselves	7	Rupture	34
Calenture	3	Headmouldshot & Mouldfallen	14	Scurvy	105
Cancer, Gangrene and Fistula	56	Jaundies	110	Shingles and Swine pox	2
Canker, and Thrush	111	Impostume	227	Sores, Ulcers, broken and bruised Limbs	82
Childbed	625	Kild by severall accidents	46	Spleen	14
Chrisomes and Infants	1258	Kings Evill	86	Spotted Feaver and Purples	1929
Cold and Cough	68	Leprosie	2	Stopping of the stomack	332
Collick and Winde	134	Lethargy	14	Stone and Strangury	98
Consumption and Tissick	4808	Livergrown	20	Surfet	1251
Convulsion and Mother	2036	Meagrom and Headach	12	Teeth and Worms	2614
Distracted	5	Measles	7	Vomiting	51
Dropsie and Timpany	1478	Murthered and Shot	5	VVenn	8
Drowned	50	Overlaid & Starved	45		

Christned { Males—— 5114, Females—— 4853, In all—— 9967 }
Buried { Males—— 48569, Females—— 48737, In all— 97306 } Of the Plague —— 68596

> For many years, the death rolls were used merely to warn the sovereign of the need to move to clean air. Major Greenwood, the British statistician, reviewed this period in the history of medical statistics. He came upon some novel questions for the Bills to answer in the papers and correspondence of a sceptical physician, William Petty:
>
> 'Whether of 1000 patients to the best physicians, aged of any decade, there do not die as many as out of the inhabitants of places where there dwell no physicians. Whether of 100 sick of acute diseases who use physicians, as many die and in misery, as where no art is used, or only chance.'
>
> Although these particular analyses were never carried out, the seed of the idea of formal comparisons was planted. Out of the casual correspondence between Petty and his friend, John Graunt, in the mid 17th Century, a new method of scientific investigation germinated and grew slowly.

It became important to know if the observed discrepancy in RLF risk was reproducible. Experiences in other parts of the world were quickly examined but the issues remained unsettled. Some observers found an increased risk with oxygen treatment, others found no association; one found a decreased risk (the early changes of RLF improved when infants were placed in oxygen-enriched incubators), and RLF occurred in some babies who were never treated with oxygen.

The descriptive studies served to sharpen the questions that needed to be asked about oxygen treatment, such as: What gas concentration and duration of treatment is used? What groups of premature infants are susceptible? What specific eye changes are diagnostic of RLF? What range of difference in risk of RLF is expected in treated and non-treated groups? But a firm link between cause and effect was not established by the observational methods.

NUMERICAL APPROACH TO VARIABILITY

The complexities introduced by variability in illness outcomes were recognized by William Petty, a 17th century English physician, who proposed that groups of patients be compared in order to distinguish between 'art' and 'chance'. Ironically, this fertile idea grew more rapidly in other fields of biology than in medicine.

I have already noted that the class of problems in which exact outcomes are not predictable was recognized by experimenters in agriculture. The techniques of numerical comparison permitted them to work with the evidence as they found it and to measure an effect against the background of fluctuations. They did not try to idealize an experiment, and, instead, accepted the reality of variability 'caused' by multiple influences.

The move from descriptive study to a statistical kind of experimentation which made it possible to approach the problem of uncontrollable con-

founding factors in a real-world setting was a major step in the biological sciences. Mathematician and philosopher Jacob Bronowski said of this revolutionary shift to statistical methodology: 'It replaces the concept of *inevitible effect* by that of the *probable trend*.' (It is interesting that when the investigation of physical phenomena reached subatomic dimensions, physicists encountered uncertainties of the kind so familiar in biology. For guidance, they employed the statistical outlook which had already been developed in biological research.)

Statistical reasoning

A central question needs to be considered when statistical reasoning is used in human experimentation: How do we decide whether or not a pattern of observed events (or measurements) is to be attributed to chance (random occurrences) or to systematic influences (that is, to a planned intervention or to unplanned forces that we classify as bias)?

It is very important to understand that our ability to nullify biasing influences rests entirely on precautions taken in the design of an experimental plan. (Most of the space in this book will be taken up with discussions of the control of bias). Statistical methods do not offer a formula to distinguish between planned and unplanned *systematic* influences. On the other hand, numerical analysis provides a tool that allows us (with reasonable assurance) to differentiate random from non-random patterns of outcome.

Forecasting in gambling It is useful to compare the happenings we encounter in medicine with the events in games of chance. For the reasoning of the gambler (not devil-may-care as he would have us believe) leads the way to a practical approach to the problems which bedevil physicians.

In the simple game of 'heads or tails', for example, the results in successive tosses of a coin may be thought of as a series. The separate outcomes (like the single events in medicine) seem to occur erratically when we confine our attention to a few tosses at a time. But when the results in a long succession are examined, a pattern emerges. Finally the regularities of the workings of chance are quite distinct.

In a game involving hundreds of throws of a coin, the proportion of 'heads' or 'tails' is likely to be very close to one-half. And, in a very large experience, runs of successive 'heads' and of 'tails' also approach fixed proportions of the total. The point here is that in examining a long series of events (both in coin tossing and in medicine) order gradually emerges out of disorder.

Although there are enormous differences between the complex events in medicine and the straightforward occurrences in coin tossing, both provide the same kind of information needed to make predictions about repeated series of similar observations. However, a gambler has the very practical advantage of being able to calculate theoretical probabilities of outcomes *before* making any real world observations.

Several reasonable assumptions are made: 1. the physical forces that may influence the outcome of each toss operate haphazardly—they do not align themselves in favor of either 'heads' or 'tails'; 2. since the coin has no memory, the outcome of each toss is not influenced by preceding results; and 3. all of the possible and equally likely outcomes are obvious by inspecting the coin. Once these are accepted, the probabilities of various outcomes that will be approached in an up-coming game can be computed with equations based on the laws of chance.

If the results in a number of games do not agree with those predicted by theory, the gambler may be led to frame a new hypothesis, one which specifies the expected behavior of the unusual coin in hand.

Forecasting in medicine The doctor is unable to calculate the 'expected' proportions of outcomes in advance; the phrase 'all possible and equally likely events', so obvious by looking at a coin, has no clear meaning in complex problems. In medicine we are obliged to appeal to experience for statistical probabilities.

The variation, for example, in the occurrence of blindness begins to approach a fixed proportion in a long series of observations of infants treated with oxygen. After this information is in hand, we can make predictions with the same confidence as in a game of coin-tossing (or in weather forecasts, insurance risks, highway-accident projections ... and predictions made about many events in the natural world about which we have little or no precise 'causal' knowledge). The regularities in the aggregate make it possible to make inferences from what John Venn, the probability theorist, called 'proportional propositions'.

Uncomfortable as it is to dwell on the analogy, the resemblance between the gambler and the physician cannot be denied. Both must frame their questions about the state of the world in numerical terms. The gambler's hope for 'doctored' coins that will defy *a priori* calculations of outcome are exactly like the hopes of physicians for favorable treatments. Both dream of winning by consistently upsetting usual outcome probabilities, but both are forced to test their fantasies in real world experiments.

MEANINGFUL QUESTIONS

Now that I have argued for asking number-specific questions when planning clinical trials (and I will develop the details of this issue in Chapter 9), I must emphasize that reality imposes an additional demand. It must not be forgotten that medical events, unlike those encountered in the casino or in the laboratory, are complex social phenomena with wide ramifications.

Research questions, it has been said, are like blinkers on a horse; they resist distraction, but they limit the possibilities of perception by removing the context in which meaning is embedded. The horse must remember to turn his head from time to time.

The questions asked in human experiments cannot be framed in a vacuum; there must be some kind of value judgment—a community consensus about 'meaning'—that validates the inquiries. For example, clinical trials are undertaken to increase the store of medical knowledge and to provide practical information that can be applied in the practice of medicine. The relative emphasis given to the two objectives in a specific trial requires a decision heavily influenced by the public interest.

Step-by-step questions

The growth of applied knowledge in medicine can be envisaged as a continuous process in which each cycle begins with good questions and ends with better questions. I can appreciate how disconcerting it is to read these words about a body of knowledge in which we all have a vested interest, a system of information that is concerned with our well-being and our lives. Common sense demands a search for better *answers*. But a little reflection should make it obvious that the 'answers' in bedside medicine are ephemeral. (The 'half-life' of medical knowledge is constantly shrinking.) In the presence of uncertainties, it is the questions, as illustrated in the parable at the beginning of this chapter, that are doing their work. *They* determine the course of our actions.

> 'We have learned the answers, all the answers:
> It is the question that we do not know.'
> Archibald MacLeish