Commentary

THE USE OF STATISTICAL METHODS IN THE ANALYSIS OF CLINICAL STUDIES

DAVID SALSBURG

Central Research Division, Pfizer Inc., Groton, CT 06340, U.S.A.

(Received in revised form 17 August 1992)

Abstract—It is "standard" to analyze data from a clinical trial using a narrowly defined probabilistic mathematical model. This paper examines the ways in which mathematical models, in general, can be used in clinical research, the meaning of probability in the examination of clinical trials, and the philosophical flaws in the current "standard" method. An alternative formulation is proposed which is more flexible and which comes closer to meeting the needs of medical science. In this alternative formulation, significance tests are applied to the data from a study only as a first step to determine whether the data are worth further examination. After that, clinically relevant questions are answered with 50 and 95% confidence bounds. The initial significance test is tailored to be directed at a narrow class of hypotheses that, in turn, are dictated by clinical expectations.

Hypothesis tests Estimation Probability Cause-and-effect Clinical studies

I. INTRODUCTION

The purpose and direction of this paper is depicted in Fig. 1. The box entitled "Inductive Reasoning" represents what most of us do most of the time. We observe events and try to generalize, in order to anticipate future events, or to relate events to each other, or to "understand" what has happened. When we do "science", we engage in inductive reasoning carefully and with experimental checks on the conclusions. The figure shows three types of models that can be used in inductive reasoning. Sections of the box are not covered by these models, indicating that there are other ways of engaging in inductive reasoning. Some, such as leaps of intuition, often govern our ordinary day-to-day reasoning but are not usually considered adequate for science. The three models overlap in pairs, so some forms of probabilistic models are the same as some forms of deterministic models, and some forms of logical models are the same as some forms of deterministic models. But, logical and probabilistic models do not overlap—and there's the rub. This paper is concerned primarily with how the overlap and lack of overlap in Fig. 1 affect the interpretation of data from clinical trials, but the problem is much more general than that, and the reader can probe those generalities in the references.

The standard view of controlled clinical trials is that the design and the eventual analysis of the data are governed by statistical considerations, that protocols should include descriptions of null hypotheses that will be run and that the planned number of patients be justified by power calculations [1]. However, it may be useful to separate the scientific experiment from the statistical methodology. Horwitz *et al.* [2] point out that we can consider the randomized controlled clinical trial as a paradigm with aspects that can be adopted to other situations. Suppose we consider the complete paradigm as nothing more than a scientific experiment which allows for useful comparison of treatments *independent of how the data will be examined.* We can, then, step away from the standard statistically based approaches and ask a simple question:

What are useful ways of examining the data that arise from the trial?

For instance, we might identify patients who "respond" to treatment and describe the baseline characteristics of those patients, in an attempt to define a class of patients who might be best treated this way. Or, we might examine the measurements of disease over time in the control group in order to establish a "typical" pattern and look at individual patients on the experimental treatment to see how well they follow this "typical" pattern. One could think of other non-statistical summaries of the data that might be useful for predicting what the treatments will do in future patients. These may not produce very good predictions, and they will lack the ability of statistical methods to measure the degree of uncertainty associated with them. But, it is enlightening to consider them as possible approaches and to ask why we would use probabilistic models instead.

One answer is that probabilistic models are sophisticated mathematical models that have proven useful in the analysis of scientific data, allowing us to distinguish differences due to treatments and something else (termed "random noise" for purposes of modeling). However, it must be recognized that many branches of scientific research flourish without probabilistic modeling or with, at most, a minimal use of averages and standard deviations. These include molecular biology, mathematical biology, physiology, and physical chemistry.

The standard approach does more than impose probabilistic models on clinical trials, it imposes a very specific type of probabilistic model, which will be described in Section III. One result is a schizophrenic attitude towards the analysis of data from controlled trials that occurs when the authors examine the results and locate subsets of patients that seem to "respond" differently or unexpected "effects" seen in measurements that were not expected to change. This is seen in the MRFIT study [3], which discovers differential responses associated with different patterns of baseline variables and accompanies these "findings" with the statement: "It must be emphasized that this kind of analysis does not preserve the randomized controlled design of the MRFIT ..."

Similarly, about 60% of the paragraphs in the description of the LRC-CPPT Trial [4] deal with secondary end points although the methods section warns

"This method of determining significance was used for the primary end point of the study. Other statistical tests reported used the nominal level of significance. The reader is cautioned that interpretation of these nominal P values should include the possibility that some may be significant by chance \dots "

It seems reasonable to medical scientists to use the data from controlled trials to discover unexpected clinical effects. However, they are constrained by a community of orthodox statisticians whose best advice on dealing with multiple end-points and interesting subsets of patients is put forward by Pocock et al. [4], who warn that all comparisons must be described in advance. But, the purpose of scientific research is to discover things that are not known in advance. In fact, a case has been made [5] that is unethical to run a randomized controlled trial, which will assign patients to inferior treatment, when the only purpose of the study is to confirm what is already believed to be true.

Some statisticians, such as Abt [6], will throw a bone to the medical scientist, allowing for "Exploratory Data Analysis", provided the scientist penalizes all such explorations by bowing to the tyranny of the Bonferonni bounds and requiring extremely small p-values for "significance". As will be shown later, the basic problem here is that, even within the framework of probabilistic modeling, the use of p-values as an exploratory tool is inappropriate, so this is no solution.

When what is essential to a particular type of mathematical modeling appears inappropriate to the experimenters who are interested in scientific conclusions, there are serious philosophical problems at hand. Philosophy is not an abstract useless mental activity of "philosophers" to be buried in philosophical journals and never referred to by "real" people. Philosophy provides the logical and epistemological underpinnings of intellectual activity. When we engage in intellectual activity, such as planning and analyzing clinical trials, we rest our activity on certain philosophical positions, which are seldom stated and not always fully understood by those engaging in the activity.

Thus, this is a philosophical article designed to consider reasonable answers to the question posed at the beginning of this section, how best to analyze the data from a clinical trial. Section II is a conscious act of destruction. In it, the philosophical underpinning of the use of probabilistic models is examined. The timbers on which these methods rest is exposed. It is shown that what we wish to do with these models cannot be done, due to fundamental mathematical (or rather meta-mathematical) limitations. Section III continues the destruction, applying it to the methods of statistical analysis now widely used in clinical studies, showing that these methods are arbitrary and have nothing to do with the goals and methods of clinical research. Section IV salvages a few beams from the debris of Sections II and III and suggests alternative approaches to probabilistic modeling of clinical studies.

The topics of this paper have been discussed more thoroughly in the mathematical and philosophical literature. There is nothing new here, and in the framework of a short article, not all the subtleties that are involved can be covered, nor does this article provide a complete bibliography. Rather, this is an attempt to acquaint the readers of this journal with some of these problems. A more complete discussion along with a fully anotated bibliography can be found in Ref. [22].

The still ongoing discussion in the mathematical and statistical literature of these problems shows that what medical scientists would like to do with the data from these studies (such as examine results from subsets of patients) is not illegal, illogical, or sinful. If it clashes with the standard formulation of hypothesis testing, it may be that this activity is inappropriate or it may be that the mathematical model is inappropriate. This paper investigates that latter possibility.

One more disclaimer. This discussion of philosophical foundations is focused on the analysis of data from controlled randomized clinical trials. No attempt is made to extend the discussion to sample surveys or to the development of diagnostic tests, or any other aspect of modern science which makes use of probabilistic models.

II. PROBLEMS WITH PROBABILISTIC MODELS

Figure 1 displays the relationship between scientific research and mathematical models.

The box depicts what philosophers call inductive reasoning. This is the intellectual activity which uses observed data to construct descriptions of reality more general than the observations. The problems of whether inductive reasoning is possible or whether "reality" is real will not be examined. Let us start with inductive reasoning as an activity.

As indicated in Fig. 1, there are three types of mathematical models within which we can express the procedures and conclusions of inductive reasoning.

Probabilistic Models Deterministic Models Logical Models

Areas in the box are not covered by the regions of mathematical models, because we can engage in inductive reasoning without mathematical models. A child goes out in the rain and gets wet, so she concludes she will get wet whenever the sky darkens and water splashes against the ground. There was no planned experiment, no replication, and obviously no modeling. However, scientific methods developed over the past 300 years use mathematical models lead to useful conclusions. Figure 1 displays the three types of models as extending beyond the box representing inductive reasoning. The

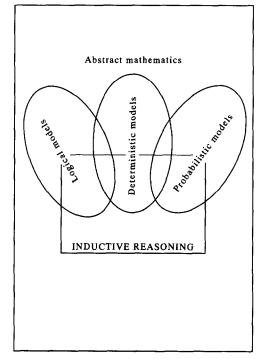


Fig. 1. The interplay of mathematical models, abstract mathematics, and inductive reasoning.

mathematics involved in these models can be investigated without reference to experiments or without attempting to describe "reality".

One might quibble that these three models can be subsumed in a greater abstraction, but it is important to keep the differences among these models in mind when considering how to analyze the data from a clinical trial. It is possible to use deterministic models to describe biological phenomenon. This is done, for instance, when large arrays of differential equations are used to model and simulate patterns of arrhythmia as defined by electrical activity across the surface of the heart muscle [7].

Similarly, logical models have also been used in inductive scientific reasoning. Most of the advances in computer techniques involve logical structures. The theory of neural networks [8] is based on Boolean algebra (a tool of mathematical logic). Some techniques of image processing on computers use statistical models, but most of the successful ones use simple iterations of logical or deterministic equations [8].

Thus, the first tearing up of foundations is the recognition that one can legitimately engage in inductive reasoning without invoking probabilistic models.

Figure 1 shows an overlapping of logical and deterministic models and of deterministic and probabilistic models within the box. However, there is a synapse between probabilistic models and logical models. This displays a well-known (among philosophers) theorem [9] which often escapes those who use statistical methods. When we engage in inductive reasoning we would like to describe "cause" and "effect". We would like to say that treatment A "causes" patients to be cured of disease X. But, cause and effect is a slippery set of concepts, and the philosophical literature is filled with the paradoxes that emerge. Mathematical logic finesses this problem by defining something which acts like "cause" and "effect" in most of the ways we want to use those ideas. This is the concept of material implication. We say A implies X,

$A \rightarrow X$.

In mathematical logic, material implication is equivalent to the counter positive

not $X \rightarrow \text{not } A$.

There is something like this in probabilistic models. This is the concept of conditional probability. We can talk about the probability of X, given A,

$\operatorname{Prob}\{X|A\},\$

and it is tempting to say that, if $\operatorname{Prob}\{X|A\}$ is very close to 1.0, we can conclude that $A \to X$. However, the conditional probability that mimics the counter-positive is

$$Prob\{not \ A \mid not \ X\},\$$

and it is very easy to create situations where $Prob\{X|A\}$ is very high but $Prob\{not A | not X\}$ is arbitrarily small.

Thus, the second rotten timber is exposed. It is impossible to use probabilistic models alone to come to the type of conclusions we would like to make in inductive reasoning.

Keynes [14] proposed that we need to associate a set of weights with units of experimental evidence, that are independent of the calculated probabilities, in order to come to useful conclusions. This is what happens, for instance, when the medical scientist views some studies as "more reliable" than others and reaches conclusions that may be based on interpretation of only some of the evidence.

Suppose, however, we plunge ahead and use probabilistic models to aid us in inductive reasoning (recognizing that we will need some additional ideas or information in order to convert the conclusions into "cause" and "effect"). The mathematical concept of probability is well defined. Any graduate student in mathematics can manipulate the symbols and theorems of mathematical probability. But, when we come to applying probability to "real" life, what does probability mean? For instance, suppose we compute a 95% confidence bound on the odds ratio for probability of death due to MI comparing two treatments. What is that 95% of? There is an answer from within the orthodoxy of hypothesis testing, but let us rip up a few more floor boards before accepting it at face value.

Figure 2 shows the region of overlap between probabilistic models and inductive reasoning in greater detail, with different views of probability. Mathematically, probability is a measure of elements of an abstract space. So, we need to identify the space which is being measured. We can think of it as a set of events that can be observed or as a set of propositions about reality. When he proposed the Latin word we now call "probability", in the **Ars Conjectandi**, Bernoulli made it clear that he considered probability to be a measure of the validity of a

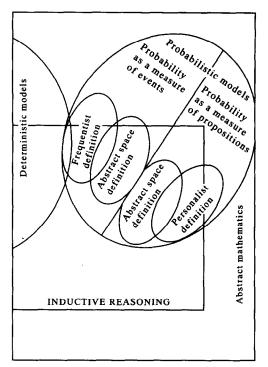


Fig. 2. Probabilistic models further refined.

proposition. That is, to Bernoulli, the only use of probability was to make statements like:

It is XX% probable that this treatment is efficacious.

The abstract space is, in this view, the set of statements or hypotheses we can make about reality. The major problem with this view is that it is not clear how one might calculate probabilities on this space.

There are two general interpretations of probability as a measure of propositions. One views the probabilities as existing independent of the observer, estimable from the data and other external sources. The other view is that probability is a measure of the personal belief of an individual scientist. Thus, in the first view, we make a statement like

The probability that the odds ratio lies between 1.03 and 2.25 is 95%.

In the second view, we make a statement like

I am 95% sure that the odds ratio lies between 1.03 and 2.25.

The math stat literature has extensive discussion of these two views. I, for one, find myself convinced by the arguments of L. J. Savage [10] that only the personalistic view avoids serious problems of definition. Savage proposed that each person carries an internal probability measure about propositions, which is "coherent". It has to meet a small number of conditions, in order for this personal probability to obey the rules of mathematical probability.

Unfortunately, most psychological experiments that have been run to elicit this coherent set of probabilities from individuals have failed to do so [11]. It would appear that most people are not coherent when dealing with probabilities ranging from 0 to 1. The only personal probability statements that people are consistent about are statements about very sure probabilities and 50:50. A probability "space" which consists of only things that are only very sure or 50:50 is not an interesting one, but it is coherent. Thus, the experiments lead us to conclude that personal probability can be used only when dealing with these two types of events.

In Fig. 2, there is a second way of describing a probability space in terms of "reality". This defines probability as measuring sets of observable events. Fisher and "Student" [12] proposed that the experimental result we observed is one of a large set of possible results. Probability is measured on the set of all possible experimental results. Later, Fisher refined this idea and proposed that we could think of the act of random assignment as generating the probability space [13]. In this case, we note that the observed results for individual patients are fixed and compute probabilities in terms of the set of all possible permutations of treatments among those patients.

Another view is that the probability of an observable event is the percentage of times that event will occur in a long run of identical or nearly identical experiments. This is the frequentist definition. It lies at the heart of the standard hypothesis testing formulation. Unless one accepts the frequentist interpretation, the reasons for treating the standard formulation as an optimal procedure disappear.

At about the same time that Neyman and Pearson [17–19] were resting their formulation of hypothesis testing (which became the standard one) on a frequentist definition, Keynes had already shown that the frequentist definition is not a well-defined mathematical concept and that use of this definition leads to serious inconsistencies [14].

Thus, the next set of rotten beams:

The link between probability calculations and "reality" has only two philosophically solid forms. (1) We can think of probability as a measure of personal belief. (2) We can compute probabilities based on all possible permutations of treatment assignment. In the first case, we can make statements about propositions, but we have to recognize that those statements will be internally consistent for most people only if we restrict them to statements about almost sure probabilities and probabilities of 50%. In the second case, probability is very well defined, but calculations can only be made where patients have been randomly assigned to treatment before the start of the study.

To summarize this section:

- (A) Probabilistic models are not the only types of mathematical models that can be applied to the data from clinical studies.
- (B) There is no link between a conditional probability statement and material implication (much less between probability and "cause" and "effect"). To go from a probabilistic statement to a logical implication requires the use of ideas external to the probabilistic model.
- (C) The only solid definitions of probability in terms of "reality" are personal probability statements about propositions (and then restricted to "very sure" and 50:50) and probabilities based on permutations of random assignments.

A thorough discussion of these problems can be found in Cohen [15].

III. HYPOTHESIS TESTING VERSUS SIGNIFICANCE TESTING

Popper has pointed out that a fundamental characteristic of a scientific theory (as opposed to other descriptions of reality) is that a scientific theory is falsifiable [15]. One can construct an experiment or describe an observation that would show the theory to be false, if such-andsuch an event were observed. There are times, however, when what can be observed is highly variable and the chances of observing a clearly falsifying event are very small. Suppose, for instance, that a new NSAID drug is accused of causing a higher than expected rate of bleeding ulcers in elderly patients. Suppose we have good enough records to estimate the incidence of this event in patients 65 or older treated with other similar drugs at less than 3%. We observe 50 patients of this age on the new NSAID and none of them develops a bleeding ulcer. This, by itself is not a falsifying event. After all, if we had an urn filled with a large number of balls, 3% of which were red and the rest white, it is perfectly possible to draw 50 white ones in a row.

Now, suppose that we continue to observe patients until we have 300 of them, none of whom develops the event. It is still not a falsifying event, since it is possible to draw 300 consecutive white balls from the urn. However, it is highly improbable (less than 1 in 1000). Thus, if we continue to observe patients and none develop the event, we have to conclude that, if the theory is correct, we have observed a highly improbable event.

Since we don't often observe improbable events, we can take this as a falsifying event.

R. A. Fisher refined this method for controlled experiments by defining a straw man "hypothesis" which he knew in advance was probably not true [16]. This straw man null hypothesis was that the different treatments had no effect on the outcome. He then produced a single number, a test statistic, calculated in such a way that he would know its theoretical probability distribution under the null hypothesis. If the test statistic produced a number t_0 and if

Prob{Test Statistic $\geq t_0$ |null hypothesis}

was very small, Fisher would take this as falsifying the null hypothesis. In effect, this told him that the data in the study were capable of showing a differences among treatments, and he would estimate those differences and comment upon them. Fisher denounced the idea that a large probability would imply that the null hypothesis is true [16]. He insisted that the experiment was not designed to allow for acceptance of the null hypothesis. The null hypothesis was nothing but a straw man designed to determine if the data were sufficiently precise to allow for further analysis.

At the same time, Karl Pearson was using this idea to determine whether a set of data belonged to a specific probability distribution. He would assemble data on, say, cranial capacities of skulls from a Roman burial ground, and try to fit it to one of a set of distributions based on four parameters. His basic tool was the chi square goodness of fit test (which he had invented) [12]. If the fit between the data and the hypothesized distribution produced an improbable chi square, he would take another hypothesized distribution, until he found one with a chi square goodness of fit test that was small enough to be explained as resulting from chance variation.

In the late 1920s, his son, Egon Pearson, approached the Polish mathematician, Jerzy Neyman, with a question. When we use a goodness of fit test to compare data to a postulated distribution, how can we be sure that the distribution is "true", if the goodness of fit test is not significant? Neyman and Pearson examined this problem over several summers and then during one year while Neyman was on sabbatical in England. Their correspondence and the sequence of papers [17-20] that resulted provide an excellent picture of mathematical work-in-progress, showing how the problem evolved as they considered it. However, this is not a paper on the sociology of science, so we skip to the final result.

They concluded that the problem of hypothesis testing consisted of considering two descriptions of "reality", a null hypothesis and an alternative hypothesis. The data in the experiment allow the scientist to decide that either the null hypothesis or the alternative hypothesis is true. The problem, stated this way, revolves around three numbers:

- $\alpha = \operatorname{Prob}\{\operatorname{reject} \text{ the null when the null is true}\}$
- $\beta = \text{Prob}\{\text{reject the alternative when the alternative is true}\}$
- δ = The "distance" between the null and the alternative.

Every scientific discipline has "standard" approaches to problems, and mathematics is no different. If you have a problem where there is a choice of methods and a set of numbers that describe the event, you "solve" the problem by finding that choice which is "optimum" in terms of the set of resulting numbers. To reach an "optimum" solution here, you want to reduce the level of α and β as far as possible. The role that δ plays in the optimization depends upon how you view the distance between the two hypotheses. If you want a method that works best when the null and alternative are close together, you want a method that optimized for small values of δ . If you want a method that works best when they are widely separated, then vou optimize for large values of δ . Another possibility is to optimize over all possible values of δ . How about minimizing α and β ? We could minimize some function of the two such as

$$\begin{array}{l} \alpha + \beta \\ [\alpha/(1-\alpha)][(1-\beta)/\beta] \\ \alpha\beta \\ \text{etc.} \end{array}$$

Or, we could fix one and minimize the other.

The point made here is that (1) the act of optimization is part of the culture of mathematics and may not be appropriate to clinical research and (2) what is being optimized is arbitrary and one choice may be more appropriate for clinical research than another.

The Neyman-Pearson formulation consists of fixing α and minimizing β across the range of values that δ might be expected to occur. A procedure that minimizes β across all alternatives is a "uniformly most powerful (UMP) test". However, it turns out that uniformly most powerful tests seldom exist. In fact, UMP tests do not exist for the comparison of two proportions, or of two normally distributed means with unknown variance, situations that frequently occur in clinical studies.

This "solution" of Neyman's requires that we fix α , the probability of falsely rejecting the null. He chose to fix α because he could then provide a frequentist interpretation to the solution. The problem he faced was that the p-value we calculate from the test statistic is a random variable whose value depends upon the random fall of the data in the study. A slight change in a few patients leads to a different *p*-value. This *p*-value has no frequentist interpretation, since there is no way of constructing a sequence of future trials which will have the same *p*-values. However, if we use a fixed cut-off (say 0.05) and reject the null whenever the *p*-value is less than that cut-off, the event of rejecting the null becomes one with a frequentist interpretation.

This means, among other things, that there is no difference between a *p*-value of 0.049 and one of 0.00001. Both provide the same degree of evidence against the null. It means nothing, within the framework of the Neyman-Pearson formulation, to talk about "very" significant or "highly" significant. A result is either significant $(p \le 0.05)$ or not. Keifer and Arrow [21] have shown that it is impossible to make any other distinction, as long as we use the frequentist definition of probability.

If we do not use the Neyman-Pearson formulation, there is no reason to consider a predetermined cut-off value as important. We protect the alpha-level of a study only because the purely arbitrary formulation is based upon fixing α and minimizing β .

And so, the Neyman-Pearson formulation lays in rubble at our feet. It is an arbitrary construction with no apparent relationship to the needs of clinical research. It rests on the rotten beam of frequentist probability. The basic optimization that it attempts is impossible in most clinical studies. And, it does not allow us to make relative judgments about two studies, one of which shows a major difference and one of which shows a barely "significant" difference.

In the previous section, we managed to recover two battered but usable beams from the destruction, personal probability and permutation tests. Can we recover anything from the ruins of Neyman–Pearson? Yes, we should give careful consideration to Neyman's basic insight that it makes no sense to test a null hypothesis without having an alternative against which to test it.

Neyman's later career exploited that insight. He realized that the power of a statistical test, its ability to detect a difference in treatments, depends very heavily upon how you define the class of alternatives. This led to the concept of a restricted test. A fuller discussion of restricted tests for clinical studies can be found in Ref. [22]. It is sufficient to point out, here, that the power of a statistical test can be increased by increasing the number of patients or by narrowing the class of alternatives, and that very powerful tests can be constructed for small studies, if the class of alternatives is sufficiently narrow.

IV. PROPOSALS FOR THE ANALYSIS OF CLINICAL DATA

After the last two sections we stand on a windy plain with data from clinical studies raining down upon us, holding a few battered beams. Can we construct anything from these? There are situations where it makes more sense to go into deterministic models. If we have a known infective agent, which can be cultured from the patient's wound and where the disease is there only when the live agent is there, a deterministic model is adequate to investigate whether a new agent will kill the infective agent and "cure" the patient.

If the disease and patterns of change in that disease involve variability, and if there is a good chance that any differences in effect will be masked by some of that variability, or if we plan to apply the treatment to a heterogenous group of patients and wish to extrapolate to a larger population, then probabilistic models are called for. This means that, at the very least, we should have a means of distinguishing between treatment differences that are "true" and those that might be the result of random noise. It would also be useful to be able to measure the uncertainty associated with more specific conclusions that are reached.

In order to make use of the few beams I have salvaged from the destruction of the previous section, I need to distinguish between the use of a probabilistic model to determine whether there is any signal in the midst of the noise and the use of a probabilistic model to provide a range of reasonable answers to questions about the possible effects of an experimental treatment on future patients. Traditionally, the statistical literature distinguishes between tests of significance and estimation theory.

Testing whether there is an effect

Logically, one should first establish that the clinical trial was adequate in design and execution to show that there was a difference in effect between the treatments being compared. That is, we need to start with some sort of a significance test. If we cannot show that there is a difference in effect between treatments, it makes no sense to attempt to estimate the difference that might be there. To construct a significance test, we need to start with a probability space. Yet, as shown before, there are only two viable probability spaces that we can use in a randomized controlled clinical trial. One is the set of possible hypotheses and probabilities that measure the scientist's personal belief in their validity. The other is the set of all possible permutations of treatment assignments, conditional on the observed set of responses. The personal probability space does not provide us with an easy and universally accepted method for determining when there is signal in the noise. The permutation space does. We need only compute a test statistic, some number which will be large if there is a treatment difference, and compute the tail probability of that test statistic in the set of all possible permutations of treatment. If that probability is very small, we conclude that we can reject the hypothesis that there is no difference between treatments, since we have seen a falsifying event.

But, we don't choose just any test statistic. Neyman's insights tell us that we can do better if we consider as narrow a class of alternatives as possible. In addition to defining the null hypothesis to be tested, the authors of a protocol should consider what might happen to individual patients if the treatments differ in effect. They could start with a set of typical scenarios. Examination of those scenarios should lead them to identify a set of measurements that will be taken and to consider how those measurements might appear among patients with treatment related responses. For instance, suppose we want to test a drug which may improve the lipid profile against a placebo. We don't want to test whether there is a difference in mean change in triglycerides, then in total cholesterol, then in HDL cholesterol, etc. Instead we might construct a set of change patterns across all the lipids, and order the patterns in terms of what is desired. Then, each patient might be categorized into a specific pattern, and the test statistic would compare the relative proportion of patients as a function of increasing category.

The point of this is that we can use Neyman's insights and Fisher's definition of probability in terms of permutations to construct a reasonably powerful significance test, which can be expected to detect treatment effects if they are discernible. The exact way in which the method is invoked is unimportant in this paper. Specific methods are described in Ref. [22]. It suffices to suggest that powerful permutation tests can be constructed, provided the class of alternatives is clearly established and narrowly defined.

Suppose, then, that our significance test provides a very small p-value. How small? one might ask. The answer to that is determined by how much effort the analyst wants to put into a study that may not yield useful information. The smaller the *p*-value the greater the chance that something useful will be found. Suppose, on the one hand, that we have managed to compare AIDS patients on placebo and a new anti-infective agent for some rare opportunistic infection. Since such a study is unlikely to be repeated, a p-value of 0.50 might be sufficient to suggest that it is useful to examine the data more thoroughly. On the other hand, a crossover study of single doses of an established bronchodilator that fails to show a significant difference between treatments is a study that is easily repeated, and it might make more sense to expend resources on planning a better study than in analyzing non-significant results.

Estimating effects

I noted that there are two types of questions the probabilistic models can answer. One has been dealt with above, how do we know that the apparent treatment effects are not due to random noise? If the results of the permutation test are "significant", then we want to answer questions about the treatment differences that have emerged. I think that those questions should be medically meaningful ones. What do I mean by "medically meaningful"? Suppose the new treatment is an antihypertensive. The significance test may be based on comparing the average change in diastolic blood pressure between the two treatments. However, knowing the average change in blood pressure is of no value to the practicing physician, who will be using the new treatment. Average change combines both patients who respond and those who do not. If the new treatment "works" but only in some patients, then the physician who plans to use it will want to know whether it will work on a given patient. One way is to follow the patient until the blood pressure drops. But, how long should the patient be followed before deciding that the new treatment will not work on this patient? That is a medically meaningful question.

But, even if we can find an answer to our medically meaningful question, the exact answer we calculate from the data will most likely not be true. These data refer to the results from a small set of patients. They will be used to predict results for a larger and more varied set of patients on whom the treatment will be used in the future. We need to apply probability theory to compute a range of answers. Thus, if we decide that almost all patients who responded to the new treatment did so within two weeks, we need to find bounds on that answer so we can say, for instance, that we are reasonably sure that future patients who are responders will respond within 12–26 days.

How can we put this "smear" of uncertainty about the answers we compute from the data?

Let us dig up a couple of the timbers we've salvaged in Section II. We can consider personal probabilities associated with such statements. However, keeping with the human limitations on the interpretation of probabilities, we should restrict the calculations to 95–99% regions and 50% regions. Thus, the end result of the analysis should be statements like

I am 50% sure that any patient who will respond will do so within 18 days.

I am 95% sure that, if the patient has not responded by 30 days, that patient is not a responder.

We can compute these probabilities with the formal use of Bayesian statistics. However, we

are only looking for rough answers. The true probability associated with a 50% statement does not have to be exactly 50%. As long as it it closer to 50% than to 95% the statement is adequate. We can use standard confidence intervals to compute such rough informal "smears" of uncertainty. These have the advantage that the width of the interval reflects both the number of patients used and the variability of the measure within those patients.

Suppose, for instance, that we find that left handed females under the age of 33 appear to have a different response. We can compute 95 and 50% probability intervals about the estimate of that response. The smaller number of patients in this subset means that the intervals will tend to be wide. However, if all of the patients in this subset produce the same or nearly the same result, the interval will be narrow. And, this is as it should be. If we have some small subset of patients whose responses are uniformly the same, this should mean something, even if that subset was identified as a result of examination of data from the study.

Note that I have slid into an aspect of data analysis that is an anthema to the standard orthodox procedures. I have allowed the data to identify subsets of patients with "interesting" responses and then used statistical methods to estimate the degree of response.

When would we use 50% confidence intervals? We would use them under circumstances where action is reasonable even if there is only a 50:50 chance that we are correct. Suppose a treatment compared to placebo produces 50 and 95% confidence interval on the odds ratio as follows:

50%:[1.55, 1.98] 95%:[0.93, 2.55]

We are 50% sure that the treatment will improve the chances of cure by more than half, but we cannot be sure that it does any good at all. If the treatment is relatively inocuous and not very costly, the 50% confidence interval will be enough to suggest that it is useful. If, on the other hand, the treatment is invasive, or costly, or carries serious risk of injury, the fact that a 95% confidence interval includes 1.0 means that we cannot use the data from this study to justify use of the treatment.

V. CONCLUSIONS

When we examine the results of a controlled clinical trial, we are engaging in inductive reasoning, and there are many modes of inductive reasoning. In the end, we will know if a particular mode was useful only if the predictions it makes about future events come true. Until we have such verification, there is no objective reason to prefer one mode over another. Given a particular situation, we can often reject some modes as inappropriate because it is not clear how to use them to answer the questions that have been posed. However, no one of the remaining modes has any claim of *a priori* "legitimacy" over the others.

This article tries to show that there are serious flaws in the use of probabilistic models and that the rigid formulation of hypothesis testing due to Newman and Pearson makes use of arbitrary constructs that are probably inappropriate to most clinical research. However, the observed variability is so great in most clinical studies that we can analyze the results best within the framework of a probabilistic model. Also, a sequence of steps of analysis has been proposed that involves probabilistic models but avoid most of the philosophical problems. The new house has been constructed from solid, if battered, beams, but it is a flimsy structure, held together by appeals to common sense. The scientist who analyzes her data in this mode cannot appeal to the "authority" of "correct" procedures, and the conclusions could very well be wrong—but they might also be right.

The referees who considered an earlier version of this paper pointed out that generally accepted principles of cause and effect in science require that each study cannot be considered in a vacuum. Unexpected findings are unexpected because either previous studies or biological theory have not suggested them. In the end, scientific cause and effect has to be demonstrated with a combination of biological replication and reasonable theory to support it. I do not disagree. However, this paper has confined itself to the problem of analyzing data from a single controlled clinical trial. What we see in such a trial might be unexpected. If so, we should be allowed to note it, since this might be the first in a sequence of biological replications.

REFERENCES

- Donner A. Approaches to sample size estimation in the design of clinical trials—A review (with discussion). Stat Med 1984; 3: 199-214.
- Horwitz RI, Viscoli CM, Clemens JD, Sadock RT. Developing improved observational methods for evaluating therapeutic effectiveness. Am J Med 1990; 89: 630-638.

- 3. The Multiple Risk Factor Intervention Trial (MR-FIT). JAMA 1982; 248: 1465-1470.
- 4. The Lipid Research Clinics Coronary Primary Prevention Trial Results (LRC-CPPT). JAMA 1984; 251: 351-364.
- 5. Ware JH. Investigating therapies of potentially great benefit: ECMO (with discussion). Stat Sci 1989; 4: 337 - 340
- 6. Abt K. Planning clinical trials on the basis of descriptive data analysis. Stat Med 1991; 10: 777-795.
- 7. Pogwizd SM, Corr PB. Mechanisms underlying ventricular tachycardia and fibrillation in the ischemic heart: relation to nonlinear dynamics, mathematical approaches to cardiac arrhythmias. Ann NY Acad Sci 1990; 591: 278-300.
- 8. Jacyna GM, Marlet ER. Classification performance of a Hopfield neural network based on a Hebbian-like learning rule. IEEE Trans Inf Theory 1989; 35: 263-280.
- 9. Carnap R. Inductive logic and probability. In: Carnap R, Jeffrey C., Eds. Studies in Inductive Logic and Probability. Berkeley, CA: University of California Press; 1971.
- 10. Savage LJ. The Foundations of Statistics. New York: John Wiley; 1954.
- 11. Kahnemann D, Slovic P, Tversky A. Judgement under Uncertainty: Heuristics and Biases. Cambridge University Press; 1982.

- 12. Fisher RA. Professor Karl Pearson and the method of moments. Ann Eugenics 1937; 7: 303-318.
- 13. Fisher RA. The Design of Experiments. New York: Hafner; Section 9, pp. 17-21. 14. Keynes JM. A Treatise on Probability. London:
- Macmillan; 1921.
- 15. Cohen LJ. An Introduction to the Philosophy of Induction and Probability. Oxford: Clarendon Press; 1989.
- Fisher RA. Statistical Methods and Scientific Inference. 16. Oxford: Oliver and Boyd; 1959.
- 17. Neyman J, Pearson ES. On the use and interpretation of certain test criteria. Biometrika 1928; 20a, Part I: 175-240; Part II: 263-294.
- 18. Neyman J, Pearson ES. The testing of statistical hypotheses in relation to probabilities a priori. Proc Cambridge Philos Soc 1933; 29: 492-510.
- 19. Neyman J, Pearson ES. On the problem of the most efficient of statistical hypotheses. Philos Trans A 1933; 231: 289-337.
- Neyman J. Sur la verification des hypothese statis-20. tiques composees. Bull Soc Math Fr 1935; 63: 246-266.
- Berger J. The frequentist viewpoint and conditioning. 21. In: Le Carn LM, Olshen RA, Eds. Proc Berkeley Conf in Honor of Jerzy Neyman and Jack Kiefer. Monterey, CA: Wadworth; 1983: Vol. 1.
- Salsburg DS. The Use of Restricted Significance Tests 22. in Clinical Trials. New York: Springer-Verlag; 1992.