STATISTICS IN MEDICINE, VOL. 10, 1679–1692 (1991)

FALSIFICATIONISM AND CLINICAL TRIALS

S. J. SENN

Medical Department, Pharmaceuticals Division, CIBA-GEIGY AG, 4002 Basle, Switzerland

SUMMARY

The relevance of the philosophy of Sir Karl Popper to the planning, conduct and analysis of clinical trials is examined. It is shown that blinding and randomization can only be regarded as valuable for the purpose of refuting universal hypotheses. The purpose of inclusion criteria is also examined. It is concluded that a misplaced belief in induction is responsible for many false notions regarding clinical trials.

Induction simply does not exist, and the opposite view is a straightforward mistake.

SIR KARL POPPER[1]

AN INTRODUCTION TO INDUCTION

Induction has been defined as 'a form of reasoning that usually involves generalization, i.e. the inference from an instance or repeated instances of some conjunction of characteristics that the conjunction obtains universally',[2] (p. 306) for example, concluding that *because* all swans so far observed have been white that *therefore* all swans are white. It is usually contrasted with *deduction* which in logic is 'the logical process by which one derives one proposition (the conclusion) from one or more others (the premisses), following certain general rules. A conclusion so obtained must be true if the premisses are'.[3] (p. 75). 'A deductive argument is thus distinguished from an inductive argument in which, however convincing it may be, the premisses could conceivably be true and the conclusions false'.[4] (p. 70). For this reason Induction is taken to be problematic in a way in which deduction is not. Furthermore, since 'some regard the process of scientific work as one of induction . . . going from particulars to universals in terms of more or less probable inferences',[3] (p. 263), the problematic nature of induction is held by some to carry over to science as well.

Clearly what is problematic for science must be problematic for medicine as science. Clinical trials are often regarded as being the perfect example of the scientific method applied to medicine. This essay is mainly concerned with the problem of induction as it affects clinical trials.

The man who is generally regarded as having first recognized the problem which induction poses and its importance is the Scottish philosopher, David Hume (1711–1766).[5,6] His paradox of induction, as it affects scientific inquiry, may be expressed by juxtaposition of three individually apparently plausible statements

First: Scientific knowledge is acquired by rational means only.
Second: Induction is necessary for acquiring scientific knowledge.
Third: Induction is not rational.

0277–6715/91/111679–14$07.00

(Here the word *rational* should be understood to mean 'based on reason' or 'capable of being defended by reason'.) These three statements taken together form a paradox which is generally considered one of the outstanding problems of epistemology (the theory of knowledge).

The heart of Hume's argument may be summed up in his own words thus:

> 'These two propositions are far from being the same. I have found that such an object has always been attended with such an effect, and I foresee, that other objects, which are, in appearance, similar, will be attended with similar effects.'[6] (p. 329).

The first of these two statements does not, of course, entail the second and yet this would seem necessary if induction is rational.

Since Hume, most philosophers have attempted to resolve the paradox in one of two ways. They have denied either the first proposition or the third. Hume himself opts for a partial mixture of these two courses. He accepts the logical limitations of induction but claims a psychological necessity for it and indeed certain of Hume's statements are of a kind we now call *Bayesian*. (Thomas Bayes, 1702–1761, was born before Hume but his 'An essay towards solving a problem in the doctrine of chances', appeared posthumously in 1763 whereas Hume's *A Treatise of Human Nature* appeared in 1739 and his *An Enquiry Concerning Human Understanding* in 1758. Bayes theorem is held by some to provide a probabilistic solution to the problem of induction.)

The denial of the first proposition is, at least for practising scientists, an unattractive way to resolve the paradox. The third proposition is, however, not easily denied. Its truth does not, for example, rest in our inability to prove that nature is not capricious. Even if we believe that the operation of natural laws has a pattern, the truth of whose form we would have to accept were it revealed to us, we have no rational guarantee of being able to discover these laws by induction, or indeed by any other means.

Sir Karl Popper, the Austro–British philosopher, has provided a solution to Hume's paradox, and that is to deny the second of the three propositions above. According to him induction is not necessary to the process of rational scientific enquiry. This essay is an attempt to show that many of Popper's ideas have practical relevance to the conduct and analysis of clinical trials. Before embarking on this exposition, however, I wish to make a few personal prefatory remarks.

I first encountered Popper's theories nearly 20 years ago as a first year undergraduate and have been interested in them ever since. I am not a philosopher, however, and am aware that Popper's theories have often been misrepresented. On the other hand I believe them to be extremely relevant to the theory of clinical trials I am about to outline. In outlining this theory I am attempting, therefore, to steer a nice course between the Scylla of misattribution and the Charybdis of plagiarism. A professional philosopher might well find what follows irritating, elementary and specious but then philosopheres themselves investigating clinical trials have not always escaped provoking similar judgements.[7] I have written this essay in the hope of attracting criticism, sharing as I do, Popper's belief that science progresses partly through the 'friendly-hostile co-operation of many scientists'[1] (p. 372).

I have, however, made it easy for the reader to check my interpretations by restricting my quotations and references to one easily accessible selection of Popper's writings, namely *A Pocket Popper*, edited by David Miller and published by Fontana.[1] (Unfortunately this has just gone out of print but the U.S. edition, *Popper Selections*, is still available.[8])

Finally, I should point out that I cannot claim that this is the first attempt to apply Popper's ideas to the theory of medicine. Indeed there is a journal, *Medical Hypotheses*, which has Popper himself as a member of the editorial board and which is devoted to the purpose (which may be described as Popperian) of publishing hypotheses prior to testing them with data. Popper's ideas

have also aroused considerable interest among epidemiologists.[9] Furthermore, Briskman[10] has proposed that Popper's criterion of demarcation (see below) can 'explain the rationality of our preference for Western medicine over the superstitious practices of, for example, the witchdoctor' (p. 1109). I believe, however, that there is a great deal of modern medical theory, which cannot be defended on the basis of the criterion of demarcation, that some of it at least has been 'supported' through the irrational use of clinical trials and that application of Popper's theories on scientific investigation in general to the particular field of clinical trials would be extremely beneficial.

A THUMBNAIL SKETCH OF SOME OF POPPER'S THEORIES

Popper's particular views on rational enquiry and the role of criticism and deduction in it have been applied by him in many fields: for example, not only to the philosophy of science but also to politics. I shall only be concerned here with those elements of his theories which I consider of relevance to the planning, analysis and interpretation of clinical trials.

At the heart of Popper's views on scientific method lies his rejection of induction. He denies both that it can be logically justified and that it is psychologically necessary. He maintains instead, that, 'What we do use is a method of trial and the elimination of error'[1] (p. 104). Not only are we unable to prove that induction is justified but we know it to be invalid as a mode of inference: 'induction is invalid in *every sense*, and therefore *unjustifiable*'[1] (p. 103). He cites the example of the Newtonian system as one which despite the apparent mass of evidence for it had eventually to be replaced by relativity and quantum mechanics when new evidence came to light which was not consistent with it.

Popper maintains it is the business of the scientist to devise severe and ingenious ways of putting theories to the test. He has proposed an acid test by which the scientific may be divided from the non-scientific. This 'criterion of demarcation' between scientific and non-scientific theories is that scientific theories should be falsifiable: capable in principle of being proved false. It is important to recognize, however, the difference between principle and practice in this respect. Universal statements[11] (that is to say statements of the form 'all As are Bs' or 'no As are Bs') are not in principle capable of being proved true. 'Only the falsity of the theory can be inferred from empirical evidence, and this inference is a purely deductive one'[1] (p. 102). On the other hand true universal statements are only capable of being proved false in principle; it is only false statements that are capable of being proved false in practice. In practice this may, however, be difficult to do. It is crucial, therefore, that when we consider the body of conjectures called science, as Miller has put it, 'no hypothesis may be admitted unless there is some way in which, if necessary, it may be expelled'[12] (p. 23).

Popper has this to say:

> 'This then, is roughly the *methodological form* of (D), of the criterion of demarcation. Propose theories which can be criticized. Think about possible decisive falsifying experiments – crucial experiments. But do not give up your theories too easily – not, at any rate, before you have critically examined your criticism.'[1] (p. 126).

Popper has been misrepresented on this point and has been accused of failing to note that scientific theories never fit the facts except to a given degree and that the precision to which measurement is possible has to be taken into account when judging them. Popper is in fact perfectly aware of this: 'We must clearly distinguish between falsifiability and falsification'[1] (p. 150). What many critics of Popper have failed to note is that although probability may make the business of deductive rejection of theories difficult it does not therefore provide a justification

for induction and even if the phenomena we observe are not only measured with perfect precision but also behave with exemplary regularity the problem with induction remains.

What Popper believes is that theories which withstand severe attempts to refute them are 'corroborated'. They are not confirmed as being true, they are not even made more probably true (and in fact they may well be false) but they are shown to be fruitful – capable of making successful predictions in inherently difficult circumstances.

Popper regards scientific development as an evolutionary process but the term 'evolutionary' is not used lightly. He means evolutionary in the Darwinian sense, using only instruction from within and not for example in the Lamarckian sense of instruction from without. Medawar and Smith's entry under Genetic Code in the *Fontana Dictionary of Modern Thought*[13] (p. 258) explains the importance of the distinction: 'The central dogma of molecular biology is that coded information can pass only from DNA to protein and never the other way about.' Similarly our theories given certain initial conditions entail certain observations. The information, however, does not pass the other way. The observations do not entail the theory. Theories are eventually found wanting and replaced by fitter theories.

Popper's evolutionary analogy, however, provides a model of the process by which we learn about our theories. It is not implied that theories necessarily arise at random like mutations[1] (pp. 81–83), simply that their survival is governed in a similar way to the survival of species. The way in which scientific theories arise Popper regards as being more legitimately the province of the psychologist rather than the philosopher.

> 'The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it. The question how it happens that a new idea occurs to man – whether it is a musical theme, a dramatic conflict, or a scientific theory – may be of great interest to empirical psychology; but it is irrelevant to the logical analysis of scientific knowledge.'[1] (p. 133).

Popper distinguishes between the logical and methodological aspects of scientific discovery. The first are concerned with the logical process of deduction. It is a matter of logic that however many white swans are observed does not prove that all swans are white whereas the observation of a single black swan disproves the theory. It is a matter of methodology, for example, that we choose to look further and further afield in the hope of finding a non-white swan.

Central also to Popper's theories is a belief in objective knowledge.

> '. . . without taking the words "world" or "universe" too seriously we may distinguish the following three worlds or universes: first, the world of physical objects or of physical states; secondly, the world of states of consciousness, or of mental states, or perhaps of behavioural dispositions to act; and thirdly, the world of objective contents of thought, especially of scientific and poetic thoughts and of works of art.'[1] (p. 58).

I shall not attempt to summarize this extremely controversial and difficult part of Popper's theories except to speculate that it is closely related to a distinction between beliefs (world 2) and assumptions (which in my opinion belong to world 3) which may be drawn in certain theories of statistical inference.

Finally it is important to mention Popper's rejection of what he has termed the 'bucket theory of the mind'[1] (p. 105) by which 'Our knowledge . . . consists of an accumulation, a digest, or perhaps a synthesis of the elements offered us by our senses'. Popper's rejection of induction applies not only to scientific enquiry but also to psychology (Popper thus rejects Hume's claim of psychological necessity for induction). Perception is impossible without theories and the theories

we hold as human beings are tested and corroborated or rejected according to the same logic by which we examine theories as scientists.

Readers who are interested in learning more about Popper's philosophy will find useful introductions in *Popper*[14] by Magee or *A Pocket Popper*.[1] A clear and comprehensive defense of falsificationism against some common criticisms will be found in Miller's, 'Conjectural knowledge: Popper's solution of the problem of induction'.[12]

FALSIFICATION AND STATISTICAL INFERENCE

It is tempting to claim immediately that a Popperian view of scientific enquiry leads logically to the system of statistical inference associated with classical hypothesis testing and is incompatible with Bayesian inference. I think that such a demonstration would not be easy and I do not consider myself qualified to attempt it.

I do, however, believe the following. First, that Popperian philosophy, whether or not it can be made compatible with the mechanics of Bayesian statistics, is incompatible with any absolute inductivist interpretation of the results. Popper and Miller have debated the issue in *Nature* with a leading Bayesian[15-17] and extended the argument in an important paper published by the Royal Society.[18] Secondly that it may be possible that Popperian philosophy is consistent with statistical systems which adopt *operational* axioms of regularity (that is, acting *as if* the future will resemble the past) but that the minimum requirement for this to be so is that it is accepted that (a) the accumulation of data does nothing to establish inductively the truth of the operational axiom itself (that is, we have 'induction', if at all, only under the axiom and in fact given the axiom this 'induction' is effectively a deduction) and (b) that the axiom is operational and nothing more and will eventually be modified. There are many such operational axioms which are adopted by statisticians (Bayesian and classical): for example, exchangeability, postulated random sampling, additivity, assuming that all alternative hypotheses are known, and so on. I exclude here genuine random sampling (which can only ever take place from a finite population) and randomization (which has to do with a finite number of possible arrangements).

This position is not too different from Popper's and Miller's, if I have understood them correctly.[18] What they show is that if there is any support provided for a hypothesis by evidence, this support is not inductive (although it may sometimes be called 'inductive'). Their conclusions are:

> '. . . there is nothing in this world that can be settled beyond doubt. Thus the duty to go on testing a hypothesis severely, with the use of all our imagination and intelligence in our search for loopholes, is not discharged even if the hypothesis has been established by some process called induction. (Aspirin has been tested countless millions of times, yet there may be a significant, not yet suspected, side effect.) . . . no amount of testing can provide a hypothesis with anything like inductive support. And no amount of probabilistic support, or even inductive support, can release us from the responsibility to impose further tests.'[18] (p. 585).

Thirdly, I believe that within the framework of classical hypothesis testing a Popperian philosophy requires us to recognize a crucial distinction between deciding that the null hypothesis is false and deciding that the alternative hypothesis is true if this alternative hypothesis is anything other than the strict complement of the null hypothesis. We may postulate a plausible alternative hypothesis when devising a test which can lead to the rejection of the current theory. The rejection of the current theory does not imply that the alternative hypothesis is true. (It is

frequently held that such an alternative hypothesis is an essential component of any hypothesis test,[19] but as Fisher points out,[20] (p. 246), any preference for one test statistic over another is more logically justified by claiming good results with this test statistic in the past, rather than by stating that the good results obtained with this test statistic lead to a preference among the possible alternative hypotheses of that alternative hypothesis which makes this particular test statistic the optimal choice.)

In fact I shall not be much concerned with probability in this essay. Most of the problems I shall discuss are pre-statistical. They are problems of inference which would arise even if clinical trials were experiments which exhibited the utmost regularity (which is obviously not the case). What I shall attempt to show is that in planning and interpreting clinical trials we should think deductively rather than inductively, we should distinguish between the fundamental logic of the trial and the methodological aspects of applying that logic and that wherever we set out to prove that any hypothesis we wish to uphold is true we end up in severe logical and methodological difficulties. If the relevance of Popper's theories to what I have to say is not at first obvious I beg the reader's patience and claim that the connection will become clear by the end.

PROOF WITHIN THE TRIAL

The basic purpose of a clinical trial is to provide deductive inference about treatment, 'proof within the trial'.[21] It is necessary to insist on this trivial point because it is in practice frequently overlooked and the typical clinical trial report consists of a mass of statistics and commentaries presented in such a way that what is truly valuable is scarcely distinguished from what is irrelevant.[22] The strongest inference which it is possible to make from a clinical trial is that the experimental conditions being compared differ as regards some outcome. In philosophical terms we may claim that the objective of the trial is to disprove a universal statement of the form 'experimental condition A always produces the same effects as experimental condition B'. Such a statement is scientific in the sense that it satisfies Popper's criterion of demarcation since it is in principle capable of being disproved.

It should be noted that to disprove this statement it is sufficient to show that it is *sometimes* false and we may do this (to the extent that we accept the frequentist approaches to statistical analysis) for a given group of patients by showing that the mean response to A differed from the mean response to B by an amount we prefer not to explain by chance. The Fisherian significance test, for example, can be understood in terms of this requirement. If, for example, we have a treatment which is effective in some patients but no different from a placebo in others it may be the case that none, all or some patients given this treatment in a given placebo controlled clinical trial will respond. If all respond equally then for continuous measurements a two-sample t-test might well be a particularly effective way of disproving the statement. If only some respond then the variance in the treated group will be increased and it will probably be less effective but the validity of the procedure is uncompromised for as Fisher says, 'It has been repeatedly stated . . . that our method involves the 'assumption' that the two variances are equal. This is an incorrect form of statement; the equality of the variances is a necessary part of the hypothesis to be tested.'[23] (p. 124).

If the universal statement above is disproved it then follows as a matter of deductive logic that it may be replaced by its complement, 'experimental condition A does not always produce the same effects as experimental condition B'. This second statement is not scientific, however, since it does not satisfy the criterion of demarcation. At a later stage I shall consider the difficult problem of how further scientific statements may be generated from the results of the trial.

For the moment, I shall only be concerned with the first deductive step. The methodological devices which are used to support that deduction are control, randomization and blinding (in order of importance) and it is only statistical deductions which are made regarding variation in the experimental dimension which have true clinical trial validity.

The experimental dimension is the one in which the trialist can and does vary the initial conditions among the patients. Usually it is only the treatments that vary and when this is the case only comparisons of treatments have clinical trial validity and the deduction which results (if it results) has strictly limited scope.

I shall use the example of a trial to investigate the protective effect of beta agonists on the late-phase reaction in allergic asthma as measured by its effect on forced expiratory volume in one second (FEV_1) to illustrate this point. Some allergy sufferers suffer not only an acute reaction on exposure to allergen but, having recovered from this reaction, may suffer further effects. Such effects are called late-phase reactions. The late-phase reaction is, however, notoriously difficult to reproduce. Beta-agonists are a class of rapid acting bronchodilators which act on the β_2 receptors in the lung causing the bronchial passages to widen. Their efficacy in terms of improved FEV_1 may be demonstrated quite easily in all sorts of asthma sufferers and indeed even in healthy volunteers. A common trial design in such cases is a two-period placebo controlled crossover in which patients are selected because they are believed to have exhibited late-phase reactions in the past, are pre-treated (double-blind) with the trial drug or placebo and are then given an allergen challenge. A significant difference in favour of active substance compared to placebo during the period in which the late-phase reaction may be expected is then taken as proof that the drug protects against the late-phase reaction.

This conclusion is, of course, a deductive nonsense.[21] There is no deductive proof within the trial that a late-phase reaction took place at all and indeed, since beta-agonists are bronchodilators, the theories we hold about them predict that we should, without allergen challenge, see a significant advantage over placebo when we employ them. The fact that we find such an advantage tells us nothing about the effect of the allergen challenge. The design is inadequate and will not provide any deductive proof regarding the role of allergen and the effect of treatment in modifying it, or at least not a proof of the standard we expect from clinical trials unless we apply control, randomization and blinding to the challenge itself. We need to run a factorial experiment with dummy and active challenge as levels of one factor and active treatment and placebo as levels of another.[24]

THE DEDUCTIVE FUNCTION OF CONTROL, RANDOMIZATION AND BLINDING

I am well aware that the control treatments we are able to find in clinical trials are not always ideal, that randomization is not uncontroversial and that for a host of practical reasons it is extremely difficult if not impossible to ensure perfect blinding. I wish to make a Popperian distinction here between methodology on the one hand and logical purpose on the other and in fact to assert that on the whole in discussing these topics far too much attention has been given to the methodological problems associated with these three scientific 'devices' and not nearly enough to their logical purpose. I shall therefore assume in the discussion which follows that where a study is described as having been randomized we may trust that the investigator has indeed randomized correctly and that where it is described as being blind the blinding is absolute, in order to investigate what logical difficulties remain even if the methodological difficulties are overcome.

Blinding

The easiest to consider is blinding. There is a very important simple point which is regularly overlooked about blinding, namely that it can only strengthen the conclusion we obtain from a trial if that conclusion is that the treatments are different. Thus to the extent that active control equivalence studies (ACES)[25] (the purpose of which is to demonstrate equivalence of treatments) are successful, whether they were blind or not is of no interest in interpreting the results. The reason is made clear if we consider (a) the nature of deductive inference in clinical trials and (b) the logical function of blinding in supporting this inference.

I find it helpful here to consider an example in which blinding was considered absolutely essential for deductive reasoning. The example I have in mind is that of the *in vitro* experiment conducted by Jaque Benveniste and his colleagues under the supervision of John Maddox, the editor of *Nature*,[26] to investigate the possible degranulating properties of anti-IgE anti-serum at dilutions so high, that effectively no molecules were present in the solution. Elaborate steps to ensure blinding were insisted upon by Maddox and his colleagues as they investigated the claims made originally by Benveniste.[27] Whatever the interpretation of the outcome of the experiment,[28,29] one thing is clear: no blinding, however elaborate, could have supported the claims made for a homeopathic effect if the experiment had been run with an active control only and the objective had been to show that there was no difference between high dilution (homeopathic) and high concentration anti-IgE anti-serum. In such a case all a prejudiced observer would have to do would be to record similar values for high concentration and high dilution.

In fact only when a difference is observed does the issue of the adequacy of blinding arise at all. Thus since Benveniste had observed that, '. . . 7 control tubes and 3 tubes containing a dilution previously determined as active (1×10^{34}) were counted blind: basophil degranulation was $7{\cdot}7 \pm 1{\cdot}4$ per cent for the controls, and 44·8 per cent, 42·8 per cent and 45·7 per cent for the tubes containing diluted anti-IgE'[27] (p. 817), the adequacy of the blinding became an issue. It is only the difference observed which makes the blinding of interest.

In general it is only unblinding through efficacy which is of interest in clinical trials and hence it is irrelevant to the scientific community at large whether an ACES which failed to find a difference between the new and standard therapy was blinded or not. Any competent statistician could have faked equivalence between the treatments to any degree desired without knowing the randomization codes.[21] The reason this point is regularly overlooked is that statisticians have a lamentable tendency to think about trials inductively rather than deductively and hence have missed the point that the purpose of blinding is to strengthen the only deductive inference which is possible from a trial, namely that the treatments are not equivalent.

Randomization

A similar point may be made about randomization. The topic is, of course, one of notorious controversy in statistics and the device itself is disliked by many statisticians: not surprisingly chiefly by those who favour induction. Here I think that statisticians have missed an important point by concentrating on the probabilistic aspects of randomization rather than on a more fundamental experimental one, namely, the demonstration of causality.

Suppose that a trialist tells me 'I found a highly significant and clinically relevant difference between the two treatment groups *despite the fact that the patients were allocated at random to treatment*'. Now whether I belong to the statistical tribe whose members are happy to calculate probabilities over all randomizations or whether I feel that the particular distribution of covariates actually produced by the randomization must be reflected in the analysis, I can still

acknowledge the potential value of the qualification in italics. If the statement is true it protects me against possible deception on the part of the trialist. To look at the problem the other way round, if the trialist were to insist on one and only one particular allocation of the patients to treatment in order to demonstrate a treatment effect we would rightly be suspicious. We ought to require him to be prepared to accept a sufficiently rich number of alternative allocations of patients to treatment for his proof to be acceptable. If we wish to protect our interpretations from the possibility of 'bad' randomizations then we may use restricted randomization or some analytic protection such as analysis of covariance.[30] (It ought to go without saying that the form of such an analysis is to be determined *a priori*.) Since, however, the trialist not only treats but allocates we must have some way of satisfying ourselves that it is the treatment and not the allocation which brings about the effect. On this interpretation randomization is a necessary but not sufficient guarantee for probabilistic calculations.

Suppose, however, that a trialist who has run an active control trial with the purpose of proving equivalence between a new and existing therapy says 'I found no evidence of any difference between the two treatment groups *despite having allocated patients at random to treatment*'. The value of the qualification in italics is now unclear. It is not intuitively obvious that it makes the demonstration of equivalence stronger. The extent to which the trialist could make equivalence more likely by manipulating allocations is limited. (This is perhaps reassuring in view of the widespread use of the allocation method known as 'minimization' in fields where spectacular differences are rarely found and not always sought.)

To see why this is so consider the quantities involved for a two-treatment parallel bio-equivalence study with n patients in each group. (Many such studies are in fact carried out as crossovers but to consider a crossover study would only complicate the statistical argument without adding anything of benefit to the philosophical essence of the problem.) Suppose the trialist decides to calculate a two-sided confidence interval using the approach of Kirkwood[31] to show that this lies entirely within the range of effective clinical equivalence (see O'Quigley and Baudoin for a review of this and alternative methods[32]). Now, if we take the randomization argument, in the well known identity for the total sum of squares (SST), the between sum of squares (SSB) and the within sum of squares (SSW) for the analysis of variance

$$\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW} \tag{1}$$

the left hand side may be regarded as a constant and the terms SSB and SSW vary according to the allocation.

Now let the difference between the two treatment means be Δ and the estimated error variance σ^2. Then,

$$\mathrm{SSB} = n\Delta^2/2$$

and

$$\mathrm{SSW} = 2(n-1)\sigma^2.$$

The objective of the trialist is to show that

$$\omega = |\Delta| + t\sigma/(2/n)^{1/2} < k. \tag{2}$$

where t is an appropriate value of Student's t distribution and k is some value reflecting generally received opinion as to what constitutes a clinically relevant difference.

Now, with perfect knowledge of the experimental material and no difference between treatments, we can manipulate allocations so that $\Delta = 0$ in which case

$$\omega = ([\mathrm{SST}/\{2(n-1)\}]\,(2/n))^{1/2}. \tag{3}$$

Thus, given perfect knowledge of the experimental material but in the absence of any difference between treatments (3) represents the limit of what we can achieve. Of course if there is no difference between treatments we need hardly be concerned about dishonestly 'proving' that there is not.

On the other hand, if there is a difference between treatments, then the investigator has the further problem of having to know what this is in order to be able to do as well as the limit given by (3). In short, the investigator's ability to manipulate allocations so as fraudulantly to 'prove' equivalence of treatment is far more restricted than his ability to manipulate allocations so as dishonestly to demonstrate a difference.

Control

Control is the most important feature of clinical trials. Without it blinding and randomization are impossible but this is not its only purpose. By using controls we are able to use simple experimental means to deal with phenomena such as time trends, regression to the mean and the particularity of the patient population selected[33] and with minimal resort to the more complicated and less certain help of mathematical modelling. Just as with blinding and randomization it turns out that control only reveals its full power as a device where we seek and find differences between treatments.[21,34] The reason that this is so is because under such circumstances a judgement of the competence of the experiment may be delivered.

In my view, the trialist is faced with the objective of designing fair and competent experiments. Devices by which to ensure fairness are, for example, blinding, randomization and allocation generally and also analysis. The fairness of the trial, however, is decided on external grounds. We cannot decide by examining the results that the trial was blind or that it was randomized. It is essentially a question of scientific trustworthiness and this is one reason why it is reasonable to expect high standards in handling these logical devices. Sometimes, as for example in the Benveniste investigation,[26–29] elaborate steps are taken to provide the external guarantee.

When it comes to competence, however, this cannot be entirely established on external grounds. By competence of a clinical trial I mean the ability of the trial to detect a difference in treatments if it exists. Of course, as part of the contribution to designing a trial the statistician is expected to make a sample size determination on grounds of power or precision. This power calculation is a mathematical solution to determining the competence of the experiment and not an experimental one and this is still the case if the power calculation is retrospective.

Competence of the experiment may, however, be determined if a difference between treatments is found. It is then a (trivial) matter of deductive logic that the experiment was capable of detecting a difference between the treatments in question since the difference has been found. This then constitutes the basic problem of an equivalence trial. If the trial is successful the experimenter can provide no demonstration that it was competent.

BEYOND REFUTATION: USING THE RESULTS FROM CLINICAL TRIALS

As I have noted above the null hypothesis of equality of treatments satisfies the criterion of demarcation but its complement does not. This is potentially worrying because it might suggest that, having successfully brought to conclusion a clinical trial and found a treatment effect, we nevertheless can do nothing with the results since we are left with a theory, 'the treatments are not always equal', which we can neither defend (since the treatments might in fact be equivalent for any given group of patients we wish to treat) nor replace (since such a statement could never be refuted). I shall now attempt to provide a solution to this problem.

The role of alternative hypotheses

It is obvious that in designing clinical trials we must have alternative hypotheses in mind even if these are only imprecisely defined. In practice we have to make a vast number of assumptions to determine the way in which we are to collect observations and also how we are to analyse those observations. It is also generally held that the optimal design and test depend on the choice of alternative hypotheses and indeed that without having these alternatives in mind no test is possible at all[19] (but note Fisher's point of view referred to above[20]). Recently Salsburg has suggested that some of the alternative hypotheses[35] which we routinely use could profitably be replaced, thus leading to more powerful tests. (I have no objection to this recommendation in principle but do not like the implication in Salsburg's paper that the choice of alternative hypothesis may be determined after seeing the data.)

A common assumption which is made in analysing clinical trials is to assume additivity of the treatment effect within a given frame of reference. This is equivalent to assuming that the effect of treatment within the frame of reference is the same whoever it is applied to. Just what the appropriate frame of reference should be in experiments for which additivity is assumed has been a matter of considerable controversy. Should it for example be the patients recruited to the trial or should it be all patients satisfying the inclusion criteria for the trial?

Progress may fruitfully be made here by recognizing the crucial difference between the absolute rejection of the null hypothesis and the tentative acceptance of the alternative hypothesis which has been used to propose the test of the null hypothesis. Acceptance of Popper's realization that we can never prove that the theories we hold are true frees us to look for ways of using and criticizing them.

I shall return to this point once I have disposed of various misconceptions regarding inclusion criteria.

The purpose of inclusion criteria

A view of clinical trials is sometimes proposed whereby patients are regarded as a random sample (and implicitly a simple random sample) of the population satisfying the inclusion criteria for the trial.

I cannot show any sympathy for this view. An example may help to explain why. Suppose I go to an investigator telling him that I wish to run a trial in asthmatics, aged 18 to 65 and having a forced expiratory value in one second (FEV_1) between 60 per cent and 80 per cent of predicted. I ask him if he can provide me with 20 patients. He looks through his records and discovers that he has 50 patients aged 24 to 62 with FEV_1s between 65 per cent and 78 per cent of predicted, feels confident of being able to persuade 20 to co-operate with a study and agrees to participate in the trial. It is clear, however, that using the very same patients the investigator could have participated in many different trials with quite different inclusion criteria: for example, specifying patients aged 18 to 75 or aged 24 to 62. (Of course if an age range of 18 to 75 had been specified he might well have included further patients but there is no guarantee that he would have done.) The patients entered in a clinical trial cannot be regarded as a simple random sample of the population satisfying the inclusion criteria, or even as a substitute for one.[21]

Miettinen has argued forcefully that the notion of representativeness of subjects studied in clinical trials (or, for that matter, in epidemiology) is quite misleading[36,37] and that it is experimental convenience (which in my opinion would include ethical and practical considerations) which does and should determine the nature of the study. In my opinion the true purpose of inclusion criteria is to provide the experimenter freedom to allocate. He must not include patients

in the trial unless on both practical and ethical grounds he is prepared to allocate them to either treatment.

We cannot, therefore, use the inclusion criteria as a means of generalizing the results from a clinical trial. We cannot use them as a means of replacing the singular statement 'the treatments were not equivalent in this trial' (which in any case does not preclude the possibility that the treatments would have had identical effects for certain patients in the trial nor for that matter that a minority responded differently to the majority) with a universal statement 'the treatments will be different for all patients satisfying the following qualifications'.

What can we then use? Can we use the characteristics of the patients studied themselves? Again the answer is no. We should have, for a start, to restrict generalization to patients who were prepared to volunteer for clinical trials but of course the restrictions would never end. To return to the example above the trial may upon further inspection turn out to have an absence of individuals with a height between 5'9" and 5'11". Does this mean that the results may not be applied to such individuals?

Generating universal statements from clinical trials

Universal statements from clinical trials can only be produced using theories and thus, like the theories which produce them, they are at best tentative. We may use results obtained by studying a particular group of patients for a quite different group of patients only if the theories we hold allow us to do so. All such assumptions of regularity may, however, when subject to further study be shown to be false.

Thus, for example, in dossiers submitted for registration many authorities expect to see separate studies for children and adults. This is because the theory that children and adults should react similarly has been shown to be false in the past. However, it is not a requirement, for example, in studying beta-agonists to have separate trials for the AB blood group. If at some future date, however, a researcher succeeds in demonstrating that patients of the AB blood group respond quite differently to a given beta-agonist then it will become a requirement to study such patients separately and furthermore, of course, there will be considerable interest in studying all beta-agonists currently in use with regard to this sub-group.

Predictions are thus seen to be (a) a matter of theory and (b) always in principle capable of being false. It is also my belief that probabilistic limits on predictions from clinical trials, if they can be established at all, can only be established within a framework of assumed regularity and that the framework itself is incapable of being the subject of a probability statement.

Thus to sum up. We may and do use the results from clinical trials according to the other theories which we hold. These theories may have to be modified in the light of further experience. Patients in a trial are not to be regarded as a random sample from some population satisfying the inclusion criteria.

POPPER AND CLINICAL TRIALS

In summary I should like to underline the points at which Popper's theories are relevant to the argument above.

First there is the recognition by Popper that the problem with induction is a genuine one. That is, unlike many philosophers who have attempted to demonstrate that Hume's objection to induction is persuasive sophistry, Popper has recognized that Hume's objection is not only true but practically relevant. The consequence for clinical trials is that we need to recognize the importance of deduction.

Secondly there is a crucial distinction in Popperian terms between refuting a theory (something which is in principle decisive however difficult it may be in practice) and corroborating a theory (which is always at best tentative). What I have tried to show above is that this distinction is not just philosophical but practical. The clinical trial is designed with refutation in mind; its ability to corroborate is of necessity more limited since blinding, randomization and control only reveal their value fully in the case of a refutation. This is especially so in equivalence trials in which it is hoped that nothing will be found and for which this none-discovery is judged a success.

Thirdly there is the distinction which Popper makes between the logic of scientific discovery and its methodology. I find this distinction extremely important and in particular consider that blinding, randomization and control can only be adequately understood by making it.

Finally there is a point related to the first. We need not only to recognize the importance of deduction in clinical trials but the irrelevance of any sort of inductive argument for the purpose of using the results of our deductions. We use the results of trials according to theories we hold. These theories may be examined deductively.

ACKNOWLEDGEMENTS

I thank David Miller of Warwick University for many helpful comments on an early draft of this paper and the referees for their comments on a later draft. The errors which remain are my own.

REFERENCES

1. Popper, K. *A Pocket Popper*, edited by Miller, D., Fontana/Collins, Glasgow, 1983.
2. Quinton, A. 'Induction' *in* Bullock, A. and Stallybrass, O. (eds), *The Fontana Dictionary of Modern Thought*, Fontana/Collins, Glasgow, 1977.
3. Vesey, G. and Foulkes, P. *Dictionary of Philosophy*, Collins, Glasgow, 1990.
4. Urmson, J. O. 'Deduction' *in* Urmson, J. O. and Rée, J. (eds), *The Concise Encyclopedia of Western Philosophy and Philosophers*, Unwin Hyman, London, 1989.
5. Hume, D. *A Treatise of Human Nature*, edited by Mossner, E. C. Penguin, London, 1984.
6. Hume, D. 'An enquiry concerning human understanding', in *The Empiricists*, Doubleday, New York, 1974.
7. Miké, V. 'Philosophers assess randomized clinical trials: the need for dialogue', *Controlled Clinical Trials*, **10**, 244–253 (1989).
8. Popper, K. *Popper Selections*, edited by Miller, D., Princeton University Press, 1985.
9. Rothman, K. J. (ed) *Causal Inference*, Epidemiological Resources Inc, Massachusetts, 1988.
10. Briskman, L. 'Doctors and witchdoctors: which doctors are which?', *British Medical Journal*, **295**, 1108–1110 (1987).
11. Basson, A. H. and O'Connor, D. J. *Introduction to Symbolic Logic*, University Tutorial Press, Foxton, 1953.
12. Miller, D. 'Conjectural knowledge: Popper's solution of the problem of induction' in *In Pursuit of Truth*, Humanities Press and Harvester Press, 1982.
13. Medawar, P. and Smith, J. M. 'Genetic code' *in* Bullock, A. and Stallybrass, O. (eds), *The Fontana Dictionary of Modern Thought*, Fontana/Collins, Glasgow, 1977.
14. Magee, B. *Popper*, Fontana/Collins, Glasgow, 1975.
15. Popper, K. and Miller, D. 'A proof of the impossibility of inductive probability', *Nature*, **302**, 687–688 (1983).
16. Good, I. J. (letter), *Nature*, **310**, 434 (1984).
17. Popper, K. and Miller, D. (letter), *Nature*, **310**, 434 (1984).
18. Popper, K. and Miller, D. 'Why probabilistic support is not inductive', *Philosophical Transactions of the Royal Society of London A*, **321**, 569–591 (1987).
19. Bahadur, R. R. and Savage, L. J. 'The non-existence of certain statistical procedures in non-parametric problems', *Annals of Mathematical Statistics*, **27**, 1115–1122 (1956).

20. Fisher, R. A. 'Fisher to C. I. Bliss: October 1938' *in* Bennet, J. H. (ed), *Statistical Inference and Analysis, Selected Correspondence of R. A. Fisher*, Oxford University Press, Oxford, 1990.
21. Senn, S. J. 'Clinical trials and epidemiology', *Journal of Clinical Epidemiology*, **43**, 627–632 (1990).
22. Senn, S. J. and Auclair, P. 'The graphical representation of clinical trials', *Statistics in Medicine*, **9**, 1287–1302 (1990).
23. Fisher, R. A. 'Statistical methods for research workers', in *Statistical Methods, Experimental Design and Scientific Inference*, Oxford University Press, Oxford, 1990.
24. Senn, S. J. 'The use of baselines in clinical trials of bronchodilators', *Statistics in Medicine*, **8**, 1339–1350 (1989).
25. Makuch, R. and Johnson, M. 'Issues in planning and interpreting active control equivalence studies', *Journal of Clinical Epidemiology*, **42**, 503–511 (1989).
26. Maddox, J. '"High-dilution" experiments a delusion', *Nature*, **334**, 287–290 (1988).
27. Davenas, E., Beauvais, F., Amara, J., Oberbaum, M., Robinzon, B., Miadonna, A., Tedeschi, A., Pomeranz, B., Fortner, P., Belon, P., Sainte-Laudy, J., Poitevin, B. and Benveniste, J. 'Human basophil degranulation triggered by very dilute antiserum against IgE', *Nature*, **333**, 816–818 (1988).
28. Benveniste, J. 'Dr Jacques Benveniste replies', *Nature*, **334**, 291 (1988).
29. Benveniste, J. 'Benveniste on *Nature* investigation', *Science*, **241**, 1028 (1988).
30. Senn, S. J. 'Covariate imbalance and random allocation in clinical trials', *Statistics in Medicine*, **8**, 467–476 (1989).
31. Kirkwood, T. B. L. 'Bioequivalence testing: a need to rethink (reader reaction), *Biometrics*, **37**, 589–591 (1981).
32. O'Quigley, J. and Baudoin, C. 'General approaches to the problem of bioequivalence', *The Statistician*, **37**, 51–58 (1988).
33. Pocock, S. J. *Clinical Trials a Practical Approach*, Wiley, New York, 1983.
34. Temple, R. 'Government viewpoint of clinical trials', *Drug Information Journal*, **16**, 10–17 (1982).
35. Salsburg, D. 'Use of restricted significance tests in clinical trials: beyond the one versus two-tailed controversy', *Controlled Clinical Trials*, **10**, 71–82 (1989).
36. Miettinen, O. S. 'The clinical trial as a paradigm for epidemiologic research', *Journal of Clinical Epidemiology*, **42**, 491–496 (1989).
37. Miettinen, O. S. 'Unlearned lessons from clinical trials: a duality of outlook', *Journal of Clinical Epidemiology*, **42**, 499–502 (1989).