STATISTICS IN MEDICINE, VOL. 13, 217-230 (1994)

FISHER'S GAME WITH THE DEVIL*

STEPHEN SENN

Medicines and Clinical Development Department, Pharmaceuticals Division, CIBA Ltd., CH 4002 Basle, Switzerland

SUMMARY

The publication of Fisher's correspondence[1] on statistics has shed new light on his views on randomization. Quotations from this correspondence and from other works of Fisher are used to illustrate the role of randomization in clinical trials. It is concluded that Fisher's views not only are coherent but, despite having been developed over 60 years ago and with particular reference to agricultural experiments, are still relevant to the planning and analysis of clinical trials today.

INTRODUCTION

> . . . one has to consider the problem in an extreme form. Let the Devil choose the yields of the plots to his liking. . . If now I assign treatments to plots on any system which allows any two plots which may be treated alike an equal chance of being treated differently. . . then it can be shown both that the experiment is unbiased by the Devil's machinations, and that my test of significance is valid.
>
> R. A. Fisher (Reference 1, p. 269)

> . . . the essence of the minimax principle is to try and protect against the worst possible state of nature. The one situation in which this is clearly appropriate is when the state of nature is determined by an intelligent opponent who desires to maximize your loss.
>
> James O. Berger (Reference 2, p. 308)

> It would be most useful if the prior probability took account of previous information on human mendacity, but this has not, I think, been collected in a useful form!
>
> Harold Jeffreys (Reference 3, pp. 309–310)

The first quotation above is from a letter dated 30 May 1938 forming part of a fascinating correspondence between Fisher and Harold Jeffreys on the subject of randomization. These letters, as well as others by Fisher and his correspondents, have recently become available in an easily accessible form.[1] It is my purpose in this article to consider what light these letters and other writings of Fisher shed on the issue of randomization as it affects clinical trials.

The subject is worth discussing because randomization has long been an important element in the design and conduct of clinical trials but is coming under increasing attack from two quarters:

* A paper delivered at the PSI annual conference, Bristol, September 1991

0277-6715/94/030217-14$12·00

Received February 1994
Revised June 1993

first from Bayesian statisticians and philosophers, and secondly from proponents of 'minimization'. In discussing randomization I shall make liberal use of quotations from Fisher's work. This seems to me to be an interesting thing to do on many grounds, but first of all because Fisher was the prime promoter of randomization in experiments, and secondly because what he has to say is always worth considering. Furthermore, the practice of quoting Fisher may be justified by quoting Fisher! In a letter to Tukey of 27 April 1955 he writes:

> If you must write about someone else's work it is, I feel sure, worth taking even more than a little trouble to avoid misrepresenting him. One safeguard is to use actual quotations from his writings; better still a series of comparative quotations. (Reference 1, p. 221)

Nevertheless my object is to use Fisher's discussions to help put forward a defence of randomization in clinical trials rather than to represent his views on randomization in experiments in general: to find out what these are, the reader can do no better than consult Fisher's original writings.

I shall attempt two things in this article. First, I shall try to prove that there are certain types of trial, for which blinding is considered important, where even a medical statistician with a Bayesian philosophy should accept that randomization is *necessary*. Second, I shall claim that for a wider type of trial the Bayesian should accept that randomization is *harmless*. My ultimate purpose is to show that as regards *design*, randomization in clinical trials is not an issue which need divide Bayesians and classical statisticians, though of course they will have different views regarding analysis.

I draw attention here to two interesting papers which I have found useful. The first by Youden[4] expresses some reservations about showing too extreme an enthusiasm for randomization. The second by Kempthorne[5] is a careful discussion of the case for randomization.

THE DEVIL OR NATURE?

A criticism which could be made of Fisher's game, described in the first quotation in this paper, is as follows. It is not the Devil who chooses the yields of plots but Nature, and there is no rational basis for regarding Nature as a malevolent opponent; hence there is no need for randomization. Furthermore, since we may have some knowledge of Nature but she has no knowledge of us, our best moves involve us in using our knowledge to the best of our ability without fear of being out-thought. Hence, randomization is not only useless but may be harmful. Indeed Basu, who mistrusts randomization and finds randomization tests illogical, has questioned whether we may legitimately think that the 'scientist is engaged in something like a poker game against Mother Nature' (Reference 6, p. 594).

There are at least two good arguments which can be made for the experimenter to act as if faced with the Devil rather than Nature. The first applies to all experiments in which blinding is necessary. I shall claim below, using Fisher's game, that when this is the case not only is randomization indispensable but any form of analysis which does not use the distribution of the test statistic over all randomizations as its yardstick is potentially misleading. I believe this to be Fisher's view also, but it is worth mentioning that in the article which prompted the discussion cited above, Basu[6] claims that during the period 1935–56 Fisher's views on randomization underwent a major change. This is disputed by Kempthorne.[7] I shall show that this claim is not supported by Fisher's correspondence and that Fisher was still justifying randomization in 1955 in similar terms to those he had used 20 years earlier.

The second applies when the experimenter himself is required to prove that he cannot be responsible for any bias in allocation of treatments to replicates. In such cases the randomization may be regarded not so much as the protection which the experimenter provides for himself but one which he provides for the scientific community. In such a case randomization must be regarded not as the justification for ignoring prognostic information but as a guarantee that the investigator is not using hidden covariates to improve experimental results.

BLINDING IN EXPERIMENTS

The purpose of blinding in clinical trials, as I understand it, is to reduce the danger that an observed link between treatment and outcome is due to any features of the treatment which we regard as inessential. For example, if we judge that there is a difference between the effect of an active treatment and a placebo, we would like to have eliminated, as far as possible, the danger that this is due to prejudice regarding the effect of the treatment and knowledge of the treatment given, leaving the conclusion that it is directly due to pharmacology. The important point here is not that the patient (or doctor) should remain in ignorance of the treatment throughout the trial but that, in any trial of efficacy, if any unblinding occurs it occurs as a result of the patient recognizing the treatment as a consequence of its efficacy and not (say) as a result of its side-effects. The practical difficulties are usually formidable and it is doubtful whether they are ever perfectly overcome. It should also be recognized that the closer one gets to studying treatment as used in practice the less relevant blinding becomes. For example, to blind a comparison of a twice daily treatment to one taken four times daily, two daily placebo treatments would have to be added to the twice daily schedule. If this were done there would be no point in asking patients to rate the treatments as regards their convenience in administration. An excellent discussion of many of the issues concerning the use of placebos has been given by Joyce.[8]

Nevertheless, in much of drug development, it is considered desirable to run blind trials. In what follows I shall take it for granted that the physical and practical problems of blinding (matching for taste, colour, size and so on) have been overcome, in order to see what difficulties still remain. A distinction is usually made between single-blind trials, in which the patient is in doubt as to which of the treatments he is receiving but the physician is not, and double-blind trials, in which the physician is also in ignorance. On the whole I regard the blinding of the physician as being the more essential feature since he reports on many patients and has no direct experience of the drug himself and also because he allocates patients to treatment. Exceptions, perhaps, are *n*-of-1 trials, where a single patient reports on many episodes of treatment.[9] I shall not be making any explicit reference to these distinctions but I implicitly assume that it is essential to achieve the highest degree of blinding possible.

To the extent that the experimenter regards blinding as necessary he fears deception by an intelligence, and to the extent that he takes the possibility of deception seriously he must regard that intelligence as that of a malevolent genius: the Devil of Fisher's game. This statement is extreme and requires some justification. One might argue, for example, that in most cases we do not think that the patient or the investigator is trying to cheat, merely that he may be subject to subconscious bias. I would reply that, if this is the case, it is his subconscious which causes the problem and, to adopt the language of Freudian psychoanalysis, however highly we think of his ego we have to regard our problem as being that of dealing with an uncooperative id. Again one might claim that even if we accept that a subject's subconscious is trying to thwart us, there is no need to expect that it will be particularly clever at doing this. I shall give reasons below as to why this argument is very unsafe.

That Fisher would regard randomization as an essential part of blinding can be supported through quotation. Consider this, for example, from the letter to Jeffreys I have already quoted:

> It [randomization] is as it seems to me, a tribute to our ignorance of the nature of the errors to which our results will be liable. Thus, if I want to test the capacity of the human race for telepathically perceiving a playing card, I might choose the Queen of Diamonds, and get thousands of radio listeners to send in guesses. I should then find that considerably more than one in 52 guessed the card right. . . Experimentally this sort of thing arises because we are in the habit of making tacit hypotheses, e.g. 'Good guesses are at random except for a possible telepathic influence.' But in reality it appears that red cards are always guessed more frequently than black. (Reference 1, pp. 268–269)

The telepathic experiment provides the perfect paradigm of blinding since, just as the rate at which pure guesswork might be successful provides the yardstick by which we judge the existence of telepathy, so in a blind experiment we likewise need to know what is the degree of success which the combination of ignorance regarding treatment allocation and guesswork as to allocation might produce. Interestingly, the first example which Fisher considers in *The Design of Experiments*,[10] the famous tea-tasting experiment, is just such a case where it is essential to establish the background rate of success to which guesswork would lead. Fisher shows that this may be done by randomization but that it is equally important that the subject should know just how rich the randomization is. If these conditions are fulfilled the probability calculation can be made.

In my view Fisher regarded randomization as being essential in all experiments in the same way that he regarded it as being essential in telepathic and psychophysical experiments: the estimate of error followed exactly from the richness of the randomization. Although this argument is more difficult to accept where blinding is not essential, what is astonishing is how little it is accepted even when 'blinding' is carried out. In particular the essential role of randomization in blinding in clinical trials is poorly understood. How else can one explain the widespread belief that blind run-in periods in which all patients are treated with placebo are possible? This notion rests entirely on a presumption that the patient is ignorant, what one might term 'the argument from the stupidity of others'. But suppose patients know or guess that doctors are in the habit of starting trials with placebo run-ins. The real blindness in these trials is not the patient's blindness regarding treatment but the trialist's blindness regarding human nature. In my opinion the correct scientific and ethical approach to blinding is to offer the patient the chance to read the protocol and, whether or not the offer is accepted, to regard nothing that might be known from reading the protocol as secret.

Consider also these quotations from Urbach, a very harsh critic of randomization in clinical trials: 'For example, the people in both groups may be led to believe they are receiving an effective treatment. This can be achieved by use of a placebo. . . a fastidiously conducted trial will ensure that the doctor does not know whether he or she is administering the drug or placebo', and 'there is no advantage to be gained from allocating patients to test and control groups in a random fashion' (Reference 11, pp. 269, 270). This is wrong on two counts. First, the object of the placebo is not to deceive the patient into believing he is receiving a verum but to leave him in doubt as to what he is receiving (again the standard of the open protocol applies). Second, the blinding is imperfect without randomization, as consideration of Fisher's game will show.

FISHER'S GAME

I am interested in showing how randomization is required to support blinding, but it is worth noting that the converse may also be the case. Chalmers *et al.*,[12] in an extremely interesting paper comparing blind randomized trials, unblinded 'randomized' trials and non-randomized trials, found some evidence that the treatment allocation was biased unless the trials were randomized and blind, the implication being that knowledge of the treatment to be allocated had interfered with the randomization process. They also found that on average non-random trials had a general imbalance of prognostic factors in favour of the experimental group, an extremely important finding which is not the issue here, but which will be considered later.

I shall now consider Fisher's game as applied to clinical trials. I consider an example of an n-of-1 trial[9] which has the purpose of investigating a new therapy for asthma. The patient will receive, on a set number of occasions (say 8), either a single dose of a new therapy or a control therapy. On each of these 8 occasions the forced expiratory volume in litres per second is measured after 5 minutes, and this may be assumed to be (approximately) normally distributed. There is an adequate washout between treatments. The patient's response is measured by a doctor, and the sponsor or experimenter is worried that the doctor may prejudice the results. Both treatment and control are indistinguishable in appearance, taste, smell and so on. To consider the experiment in an extreme form, as Fisher would invite us to do, the doctor is the Devil. (If you prefer a more practical context, assume that the doctor is arranging a demonstration of the effect of a homeopathic remedy for a sceptical scientist.) A losing strategy for the experimenter is one which allows the Devil an appreciable probability of making the experimenter *strongly* believe (or alternatively, decide at a low level of significance) that the treatment is effective when it isn't.

The rules are as follows:

1. The experimenter chooses the comparator (for example a placebo or another active treatment) and the method of allocation.
2. The 'Devil' may or may not choose the result for each experimental unit. (In this case, an experimental unit is an episode of treatment.)

We shall follow four games.

Game 1

The experimenter chooses an active comparator known to be effective and wishes to test for equivalence with the new treatment.

The treatment is in fact ineffective, but the Devil plays (I show the FEV_1 readings he produces)

3·5 3·5 3·5 3·5 3·5 3·5 3·5 3·5 litres

and wins against any experimental allocation, since the experimenter, given any reasonable prior, *must* now conclude that treatments are equivalent and hence that the ineffective treatment is effective. (Or at least, it may be argued that no stronger evidence of equivalence is possible and that however many repeats of the experiment were organized the Devil could continue with this ploy.)

Since we defined a losing strategy as being one which allowed the Devil an appreciable probability of making the experimenter strongly believe that the treatment was effective when it wasn't, the experimenter has chosen a losing strategy. This example shows that blinding is irrelevant to the interpretation of equivalence.

Game 2

The experimenter informs the Devil that he will run a placebo-controlled study and choose on the flip of an unbiased coin either the double-sandwich sequence VPPVPVVP or else PVVPVPPV (where V stands for verum and P for placebo). Student's t-statistic (in its two independent sample form) will be used to evaluate the result.

The Devil plays

3·5 2·5 2·5 3·5 2·5 3·5 3·5 2·5 litres

The value of the t-statistic is $-\infty$ with probability 1/2 or ∞ with probability 1/2.

Again the experimenter loses. It is true that the statistic he has chosen has given that his randomization strategy is unbiased, but this is not protection enough. For adequate protection he needs a richer randomization. This example shows that there is no such thing as a truly blind analysis, for even knowing which results belong together is a degree of unblinding. For example, a statistician analysing a placebo-controlled clinical trial and knowing which results were obtained under treatment A and which under treatment B, but not knowing which of A and B was the placebo and which was the verum, could decide at random to favour A. It is true that A might turn out to be placebo but there is an even chance that it is the verum, and this strategy will have a type I error rate and power of 50 per cent.

Game 3

The experimenter again chooses a placebo-controlled trial but a sequence of 4 verum and 4 placebo treatments at random. Student's t-statistic is again used to evaluate the result.

The Devil plays

3·29 3·45 3·51 2·47 2·44 3·56 2·43 2·54 litres

and wins this game also. To see this we need only note that there are only $8!/(4!4!) = 70$ possible allocations. If it turns out that all the values in excess of 3 are obtained under verum and all those less than 3 are obtained under placebo (that is to say that the actual sequence was VVVPPVPP) then the t-statistic will be 15·6 on 6 degrees of freedom. The probability of a t-value as extreme as this, given that there is no difference between treatments and given the usual normality assumption, is minute. This point is, of course, that the data are not normally distributed, or at least that the errors do not come from a single normal distribution. In constructing this example I generated four observations from each of two normal distributions, each with variance 0·01 litres^2 but one with mean 3·5 litres and the other with mean 2·5 litres, and then distributed the observations at random amongst the 8 possible periods in the sequence. A given observation is in fact nothing but 'noise' with expectation 3 litres, variance 0·26 litres^2 and a distribution which is far from normal. There is, however, a 1 in 70 chance that the observations will divide amongst the treatments in a way which favours verum and which is perfectly consistent with their coming from two normal populations with quite different means. If this happens there is no way in which the experimenter can detect that it has happened. Note that the precision of the observations here is a red herring; I chose two decimal places but could have used any number desired. A similar point applies to games 1 and 2. To make the examples look more natural I could easily have added a little 'noise'. Although this would have made the t-statistics less impressive they could, given a small enough standard deviation, still have had very extreme values.

Thus, there is a small but not minute probability of the Devil causing the experimenter to conclude extremely strongly that the verum is effective when it is not.

Game 4

As game 3 but a randomization test is used.

This time the experimenter wins. The minimum '*P*-value' which the experiment can produce is 1/70 (one tailed) which, if the issue of blinding is a serious one, I maintain is just about right for an experiment of this richness.

BLINDING, CONDITIONING AND RANDOMIZATION

The Bayesian objection to randomization is that once the experiment is finished, the data are fixed and what might have happened is of no interest. It is the experiment you ran that you have to base your inferences on, not the experiments you might have run but didn't.

This is not to say that Bayesians are not prepared to allow a certain role for randomization in design. Consider an extract from Lindley's comment on Basu's[6] paper:

> The value of randomization in design may then be illustrated by an experiment to test the efficacy of treatment T in aiding the recovery R of a patient. We require the probability of a patient's recovery were the patient to be given a treatment, $p(R/T, D)$, using data from a planned experiment. This may differ from $p(R/T, D, A)$, where A is some factor unrecognized by us. . . In order to make reasonably sure that our design does not confound the effects of T and A, we may assign treatments at random, that is independent of A. (Reference 13, p. 590)

The problem with experiments which require blinding is that the numbers recorded are the results not just of the experiment but also of the mental processes of their recorder. The numbers themselves are not *sufficient* for inference, and to condition on them as if they were is to mistake science for mathematics. If the likelihood is to be defined exactly, these mental processes must be known, and if they are capable of being elucidated at all, it is only by the person who is being blinded. If that person can be trusted to report fully and honestly then there is no need for blinding.

Fisher's solution to this is the open protocol and the blind randomized experiment. The tea-tasting experiment is defined as follows:[10]

> Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order. The subject has been told in advance of what the test will consist, namely she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance. (Reference 10, p. 11)

There are two extremely important points in this: first, that the lady knows that there will be four cups of each kind; and secondly, that she knows that they will be presented to her in random order. Given that she understands what this means, this permits the experimenter to calculate the probability with which she could correctly identify all the cups. If he hadn't told her that there will be four of each, she might guess that there will be. As a result, if she guesses correctly in every case he does not know whether this represents a chance of 1 in 70, determined by the 70 possible sequences of two types of cups of tea replicated four times, or whether it represents the more impressive success of 1 in 256, determined by the number of sequences obtainable if the cups are allocated randomly without restriction. Of course the experimenter could always incorporate his prior belief regarding the lady's ability to guess that he had used a restricted randomization into the calculation. No doubt tastes will differ as to the desirability of this, but the fact that the lady understands what is to happen makes it unnecessary. The fact that she knows that the cups will be allocated in random fashion means that the experimenter need not be concerned if some obvious pattern arises as a result of the randomization. Suppose the sequence ABABABBA arises. This is a scheme which could have arisen if the cups were allocated at random within pairs. If the lady has not been told that, subject to the restriction of the four/four split, the sequence will be completely randomized, she might assume that the cups will be allocated at random within pairs. Given this false assumption regarding the experimenter's intention and the lucky fall of the randomization, the chance of guessing correctly is only 1 in 16. Of course if the experimenter is an extreme Neymanite he could argue that over all randomizations the probability is still 1/70, being composed of the conditional probability 1/16 with probability 16/70 and the conditional probability 0 with probability 54/70. Again tastes will differ as to the degree to which this argument is found compelling, but again it is unnecessary provided experimenter and subject agree exactly on the richness of the randomization process to be used.

A further point of interest is Fisher's insistence that 'determination arbitrarily by human choice' is not random enough. I take this as yet another instance of Fisher's deep understanding of experimental realities and careful attention to detail, although it does raise one difficulty. Consider this passage from a review article outlining pitfalls in research into the paranormal, a field where blinding is essential:

> Another well-known error lies in 'subjective random generation'. Put simply, most people have no idea how random numbers behave. When they are asked to generate a string of random numbers many people avoid repeating the same digit twice – it is as though they think that this would not be random. (Reference 14, p. 64)

So it is unwise to rely on subjective random number generation since sequences generated in this way are not random. The problem it raises in connection with the tea-tasting experiment is that the lady may have an imperfect notion, *despite explanation*, of what constitutes a random sequence. For example, she might consider that sequences MMMMTTTT and TTTTMMMM are excluded (where, following Lindley,[15] M is milk in first and T is tea in first). This increases her chance of guessing correctly, given that those sequences have not occurred, from 1/70 to 1/68. A possible solution to this difficulty would be to train the lady in the theory of random sequences. A Bayesian solution might be to try and model our belief as to what sort of a guess she might make. I think that most practical scientists would call it a day here and say that once the nature of the experiment had been explained to the subject then we may, provided we randomize, calculate our probabilities as if she were guessing at random so that no distinction need be made between unconditional probabilities (over all randomizations) and conditional probabilities given a particular sequence chosen at random.

In an extremely subtle and interesting paper, Lindley has given a discussion of Bayesian approaches to analysing the tea-tasting including an investigation of two alternative designs. He writes:

> In each of the experiments the obvious randomization is supposed to have taken place. Actually no physical act of randomization is needed: all that is required is that the lady is reasonably entitled to make the assumption of exchangeability required below. For this purpose a haphazard arrangement. . . is all that is required. (Reference 15, pp. 456–457)

As the argument above shows, however, physical randomization *is* needed. For an arrangement in itself to have the property of haphazardness would imply either that all arrangements have this property, in which case there is no good reason for not choosing one at random, or that some arrangements (perhaps, for example, MMMMTTTT and TTTTMMMM?) are excluded. But if some arrangements are excluded and the lady knows this then we run into the difficulties outlined above, and to assume that she doesn't know that some arrangements are excluded makes appeal to the argument from the stupidity of others. As regards the possibility of choosing a sequence at random, either such a process is no different from physical randomization or, as our quotation above suggested, people do indeed have a preference for certain sequences, in which case we cannot rule out correlation between preferences. In short, randomization is the only safe course.

To sum up, my claim here is that when an experiment is run which requires blinding it may be understood in terms of a game. In Fisher's tea-tasting the 'opponent' is the lady. In a double-blind clinical trial run by the pharmaceutical industry, for example, the sponsoring firm takes Fisher's role and the investigator that of the lady. (Once the results are presented to the regulatory authorities then the regulator may be regarded as taking Fisher's role and the sponsor that of the lady.) Randomization provides the protection against deception.

It is of interest to note that Fisher himself specifically made the connection between games and randomization, as may be shown by quoting the opening paragraph from his paper discussing the game of 'le Her':

> The process of randomisation has in recent years come to play such a central part in experimental design that it is of some interest to find that it affords a means of resolving one of the oldest paradoxes which arose in discussions of gaming. (Reference 16, p. 294)

As Barnard (Reference 17, p. 163) and later Savage (Reference 18, p. 452) have pointed out, Fisher seems to have independently developed the idea of a minimax randomized strategy. There is no space here to consider this particular paper of Fisher's, and it is only mentioned because of the specific connection between randomization in experiments and in games which Fisher makes. An interesting discussion of the problem and history of its solution is given by Hald (Reference 19, pp. 314–322) who points out that Fisher's solution to this game was anticipated by Waldegrave in 1714.

KNOWLEDGE, ESTIMATION AND RANDOMIZATION

I think Fisher regarded these three things as being intimately linked. Consider this extract from a letter to Yates dated 2 November 1955, which shows, amongst other things, that Basu's speculation regarding the change in Fisher's views is wrong:

> . . . with all the great advantages of the Knut Vik square, if the results of using it were reduced by an analysis of variance, or one of the cruder techniques that preceded it, the probability statements obtained in the z test would be erroneous, whereas if proper randomization were applied, as I think you and Eden once demonstrated experimentally, the z test was made to be reliable.
>
> Of course if a method were available to give a reliable test of significance for the use of the Knut Vik square, there would be no advantage in wider randomization. In the analogous case of eliminating blocks in a randomized block arrangement, or rows and columns in a Latin square, we do, and I think you will agree, properly and inevitably consider an experiment laid out in randomized blocks as one of a population subject to

> this restriction, and not as one of the larger population, to which it also belongs. (Reference 1, pp. 242–243)

Then, explaining why he regards it as wrong of an investigator to force a feature into design which he does not take account of in analysis, he writes:

> In my view it would be simply erroneous in exactly the same way as a rain maker who claimed significant success by comparing the frequency of rain following his experiments with that of the annual frequency observable in his neighbourhood, although it is within his knowledge that the frequency of rain is greater than the annual frequency in that part of the year during which his experiment were carried out. Of course we do not know a probability unless we know it, and it is only when it is within our knowledge that it is erroneous to substitute for it a less appropriate probability. It is when we lack this knowledge, that randomization provides the safeguard. (Reference 1, p. 243)

This is an extremely important statement of Fisher's views, and I propose to consider it in some detail. To understand it, I think it is useful to reverse the usual order of looking at randomization and analysis. Rather than asking 'can the randomization justify the analysis?', we should ask 'is the scheme of allocation justified in view of the analysis?'

Suppose, in a placebo-controlled trial of an active treatment, with n patients in each group, with an outcome measure Y, we analyse the data using the standard two-sample t-test or, equivalently, a one-way analysis of variance. The partition of the degrees of freedom and sums of squares is as follows:

Source	*DF*	*SS*
Between treatments	1	SSB
Within treatments	$2n - 2$	SSW
Total	$2n - 1$	SST

Now, under the null hypothesis of strict equality of the treatments, and under any alternative by which the treatment effect is perfectly additive, SST is unaffected by any split of the patients between treatment groups. Of course, it is *not* unaffected by treatment, but given any constant additive effect of treatment it makes no difference to the value of SST how we allocate patients to treatment groups. Thus under these circumstances an important point applies: the trialist cannot reduce (or increase) *the total sum of squares by allocation of the patients.* (He may, of course, be able to affect it by appropriate *selection* of patients to the trial.)

For SSW, a different point applies: we cannot maintain that this is not affected by allocation of patients, for it is. Under the additivity assumption, however, it is not affected by treatment.

We see that the treatment sum of squares, SSB is affected both by the allocation and by the treatment.

If we define mean squares MSB = SSB and MSW = SSW/($2n - 2$), then we shall find that under the null hypothesis of strict equality of treatments the expected values of these two mean squares are identical, so that E(MSB) = E(MSW). This expectation may be defined in at least two different ways, for example as an average over all randomizations, or as the average given repeated random selection of patients from some population. (But note that in the latter case, although the expectation property applies, the total sum of squares is not a constant.)

Since the value of MSW is unaffected by treatment it provides a natural yardstick by which to judge the effect of treatment, which is captured by MSB. Under the null hypothesis of equivalence of the treatments, the expected values of the two sums of squares are identical. Where a treatment effect is present, the expected value of MSB is greater than that of MSW. These are important properties central to Fisher's development of the analysis of variance.

Now, of course, it may be argued that the identity in terms of expectations of these two mean squares under the null hypothesis is a rather weak property and is not adequate as the basis for inference. In any case in practice we need to be able to specify the distribution of the two mean squares as well, and we usually do this in one of two ways: either by assuming that 'errors' are normally distributed, or by using the randomization distribution. Furthermore, we might well argue that the expectation property, following as it does from an averaging over all the allocations that did not occur as well as the one that did, is irrelevant, particularly if the individual allocation can be recognized as being different in some way which is likely to affect outcome. Indeed, Fisher himself stressed the concept of recognizable subsets. Referring to a gambler rolling a die, he writes:

> Before the limiting ratio of the whole set can be accepted as applicable to a particular throw, a second condition must be satisfied, namely that before the die is cast no such subset can be *recognized*. . . On this condition we may think of a particular throw, or a succession of throws, as a random sample from the aggregate, which is in this sense subjectively homogeneous and without recognizable stratification. (Reference 20, p. 35)

What is completely incoherent, however, is to analyse the data as if the expectation property applied but to insist on allocating the patients as if it did not. For example, the statistician who insists that he must balance the allocation by sex so that exactly the same number of females and males appear in the verum group as in the placebo group, but who then excludes sex from his statistical model, is claiming at one point to consider the factor as being essential and at another as being irrelevant. His behaviour at allocation shows that he does not (or ought not to, if he thinks it through) consider the simple mean square within as the basis for an adequate estimate of the variance of the treatment effect. His behaviour is thus like the rain maker in Fisher's letter. Of course, his inferences can be shown to be conservative in that the true standard error will be smaller than the one he reports but, curiously enough, it is precisely in Bayesian modes of inference that such conservative inferences are most difficult to justify.

On the other hand, the statistician who claims that sex is totally irrelevant to outcome and that therefore all allocations are equivalent, chooses one of the possible allocations at random and ignores sex in the analysis is perfectly consistent in his behaviour (Reference 21, pp. 25–28). Furthermore, if the sex of the patients is not recorded, then another statistician who did regard sex as important, whilst regretting the loss in precision which has resulted by failing to record sex, could nevertheless accept as valid the analysis of the trial. For although he does not know what the distribution of sex was, he knows what it was *in probability*, and this effect of sex in probability is expressed in the standard error of the treatment effect. Thus the analysis represents a valid expression of his ignorance regarding the distribution of sex in the experiment.

Now, if the actual distribution of sex is made known to him and he regards this as important then, in my opinion, he is justified in regarding the distribution of sex in probability as irrelevant. It will be the case then that neither the unconditional (with respect to sex) treatment estimate nor its standard error will reflect what he now knows (or believes he knows) about the experiment. But the fact that the trial was randomized is no bar to his carrying out a valid analysis. He can still stratify by sex. Of course, he may complain that sex is not perfectly orthogonal to treatment and that this has led to some loss in efficiency, but this is the only complaint he can make. This illustrates a particular point about balance in clinical trials which is often misunderstood: actual balance has nothing to do with validity of statistical inference; it is an issue of efficiency only. Balance of factors we have measured is not necessary for valid inference since we can correct for known distributions of covariates in analysis. What we cannot correct for in analysis is the distribution of unknown or unseen covariates. By randomizing we can make do with the second-best solution of ensuring that the contribution in probability of these unknown factors to the treatment estimate and its standard error is such that our probability statements remain valid.

This, then, is the true value of randomization. The trialist uses an allocation method which means that others, despite lacking information regarding factors which might be important, can nevertheless accept the analysis of the experiment provided. As the research of Chalmers *et al.*[12] shows, they would be foolish to accept such analyses where randomization has not been carried out.

SOME PROBLEMS

Urbach has pointed out that there are far more factors affecting a clinical trial than can be dealt with by randomization.[11,22] Unfortunately he makes the statement in a passage which shows a general unfamiliarity with clinical trials, so that the point has been misunderstood. He refers, for example, to the 'nurse assigned to the test group' (there is, of course, no test *group* in this sense in a clinical trial), and also claims, 'After forming the control group by assigning patients randomly, they check to ensure that the resulting groups are well matched. If they are not then the experimenters randomise again' (Reference 22, pp. 53–54). Of course, this is not done in clinical trials, for the simple reason that patients are recruited sequentially and most if not all of the patients have been treated by the time the baseline distribution has been found. It is worth drawing attention here to a particular feature of clinical trials which has not been understood by some critics of randomization who do not work in the field. Contrary to what is sometimes claimed, randomization is not a nuisance in clinical trials: from the practical point of view it is one of the easiest allocation procedures to implement.

In general, factors which are associated with patients at allocation are randomized by random allocation of treatment. In a blinded trial, differences which arise subsequently might also be regarded as being either at random or due to treatment (but I do not claim that there are no difficulties with this). There are, however, other factors which are not randomized. One example occurs in any trial in which dummy loading is employed. For example, patients may be allocated treatment A and placebo to B, or treatment B and placebo to A. Usually they will take these treatments approximately simultaneously without instruction as to the order in which they are to be taken. It would, of course, be possible to allocate patients at random (and preferably blocked in equal numbers) to an order of administration, but this is not usually done, and in such a case the inference regarding the effects of treatments would have to be conditional on the assumption either that natural randomization had taken place or that order of administration was unimportant. To admit either of these points, however, would seem to cast some doubt on the value of deliberate randomization. (Note that there are cases where the assumption that order of administration is not important is unreasonable, and dealing with it explicitly by blocking and randomization would be a useful step, for example in a trial in asthma to measure onset of action of bronchodilators.)

Urbach makes the point, which I accept, that in practice we have to make a judgement as to what is important. He then goes on to state, however, 'that a surer way of balancing the conditions in the two groups would be to control for these significant factors' (that is to say, those factors that cannot be ignored) (Reference 22, p. 54). This recommendation is not workable, however, and an example where it is unworkable is precisely the allocation of patients to treatment. The statement presupposes that we can only believe that there are important differences between experimental units where we have identified what these differences are. I may, however, make the perfectly plausible statement, 'patients differ from each other in important ways', and believe it to be true without having a reasonable way of classifying patients according to these differences. You may have balanced your trial by sex because you believe this to be important, and I may also believe it to be true, but I may still believe that 'female patients differ from each other in important ways'. You may agree with me but argue that there are no other important factors that you can identify. However if, having said this, you then refuse to consider as equally valid all possible allocations of the patients to treatment consistent with balancing by sex, then your behaviour is inconsistent.

In other words, the statistician who adopts a maxim that he will block for factors whose effects are known and ignore those which are unimportant does not have a strategy to cover all cases. There may be cases where important differences between units are strongly believed to exist but no prospect of matching is available.

CONCLUSION

In my view Fisher's prescription for analysis and randomization is consistent and practical and may be expressed as follows:

1. Determine what is to be allowed for in analysis (that is eliminated from both treatment and error estimates).
2. If practical block for these factors.
3. Eliminate through analysis the factors which have been blocked and any further important measured factors.
4. Do not block for what is not to be allowed for in analysis but randomize instead.

It is best, however, to finish by letting Fisher put it in his own words:

> . . . randomization was never intended from the first moment it was advocated to exclude the elimination from the error of components which could be completely eliminated. . . it only requires that these components shall equally be eliminated from the estimation of error. . . I often put this by saying that it is only the components which contribute to the actual error of the experiment which need to be randomized to provide an estimate of that error. (Reference 1, p. 271)

And then elsewhere:

> There are, however, many factors relevant to the precision of our comparisons, which, while they cannot be equalized, can be measured, and for which we may reasonably attempt to make due allowance. (Reference 23, p. 274)

ACKNOWLEDGEMENTS

I am grateful to Professor Bennett and the Oxford University Press for kind permission to quote from Fisher's correspondence and other works. I thank the referees for helpful comments on an earlier draft.

REFERENCES

1. Fisher, R. A., in Bennett, J. H. (ed.), *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*, Oxford University Press, Oxford, 1990.
2. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn., Springer, New York, 1985.
3. Jeffreys, H. *Theory of Probability*, paperback edn., Oxford University Press, Oxford, 1983.

4. Youden, W. J. 'Randomization and experimentation', *Technometrics*, **14**, 13–22 (1972).
5. Kempthorne, O. 'Why randomize?', *Journal of Statistical Planning and Inference*, **1**, 1–25 (1977).
6. Basu, D. 'Randomization analysis of experimental data: the Fisher randomization test', *Journal of the American Statistical Association*, **75**, 575–582 (1980).
7. Kempthorne, O. 'Comment on Basu', *Journal of the American Statistical Association*, **75**, 584–587 (1980).
8. Joyce, C. R. B. 'Placebos and other comparative treatments', in de Saintonge, D. M. C. and Vere, D. W. (eds.), *Current Problems in Clinical Trials*, Blackwell, Oxford, 1984.
9. Senn, S. J. 'Suspended judgement: *n*-of-1 trials', *Controlled Clinical Trials*, **14**, 1–5 (1993).
10. Fisher, R. A. *The Design of Experiments*, reprinted in Bennett, J. H. (ed.), *Statistical Methods, Experimental Design and Scientific Inference*, Oxford University Press, Oxford, 1990.
11. Urbach, P. 'Randomization and the design of experiments', *Philosophy of Science*, **52**, 256–273 (1985).
12. Chalmers, T. C., Celano, P., Sacks, H. S. and Smith, H. 'Bias in treatment assignment in controlled clinical trials', *New England Journal of Medicine*, **309**, 1358–1361 (1983).
13. Lindley, D. V. 'Comment on Basu', *Journal of the American Statistical Association*, **75**, 589–590 (1980).
14. Blackmore, S. 'The lure of the paranormal', *New Scientist*, 22 September, 62–65 (1990).
15. Lindley, D. V. 'A Bayesian lady tasting tea', in David, H. A. and David, H. T. (eds.), *Statistics: an Appraisal*, Iowa State University Press, Ames, 1984.
16. Fisher, R. A. 'Randomisation and an old enigma of card play', *Mathematical Gazette*, **18**, 294–297 (1934).
17. Barnard, G. A. 'Fisher's contribution to mathematical statistics', *Journal of the Royal Statistical Society, Series A*, **126**, 162–166 (1963).
18. Savage, J. 'On rereading R. A. Fisher', *The Annals of Statistics*, **4**, 441–500 (1976).
19. Hald, A. *A History of Probability and Statistics and their Applications before 1750*, Wiley, Chichester and New York, 1990.
20. Fisher, R. A. *Statistical Methods and Scientific Inference*, reprinted in Bennett, J. H. (ed.), *Statistical Methods, Experimental Design and Scientific Inference*, Oxford University Press, Oxford, 1990.
21. Senn, S. J. *Cross-over Trials in Clinical Research*, Wiley, Chichester and New York, 1993.
22. Urbach, P. 'Clinical trial and random error', *New Scientist*, 22 October, 52–55 (1987).
23. Fisher, R. A. *Statistical Methods for Research Workers*, reprinted in Bennett, J. H. (ed.), *Statistical Methods, Experimental Design and Scientific Inference*, Oxford University Press, Oxford, 1990.