STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:3729–3753
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2074

Added values

Controversies concerning randomization and additivity in clinical trials

Stephen Senn[1,2,*,†]

[1]*Department of Statistical Science, University College London, London WC1E 6BT, U.K.*
[2]*Department of Epidemiology and Public Health, University College London, London WC1E 6BT, U.K.*

SUMMARY

'As ye randomise so shall ye analyse', is one way of describing Fisher's defence of randomization. Yet, when it comes to clinical trials we nearly always randomize but we rarely analyse the way we randomize and Fisher himself was no exception. Two controversies involving Fisher in the 1930s are discussed: one with Neyman concerning additivity and the other with Student concerning randomization. Their relevance today is considered, as is whether randomization inference in clinical trials is dead and whether modelling rules the day, whether minimization is an acceptable procedure and to what extent trialists confuse experiments with surveys. It will be maintained that a number of different possible purposes of clinical trials have been confused because in the case of the general linear model, under strong additivity, they can all be satisfied by a single analysis. More generally, however, this is not the case. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: randomisation; additivity; unit-treatment interaction; history of statistics; clinical trails

1. INTRODUCTION

Randomization remains controversial. Most practising trialists and many applied statisticians value it; many even regard it as essential for valid inference. Yet it has been repeatedly criticized by some philosophers and statisticians as being irrelevant or even harmful. It seems to me to be a highly suitable topic to consider at a joint meeting of these two societies, which ought to have considerable (if rather different) interests in the subject. If randomization is truly necessary for valid inference, The society for clinical trials (SCT) needs no searching examination of its practices. On the other hand the International Society for Clinical Biostatistics must worry that, although much of what it does is valid, much else, being involved

*Correspondence to: S. Senn, Department of Statistics, 15 University Gardens, University of Glasgow, Glasgow, G12 8QQ, U.K.
†E-mail: stephen@senns.demon.co.uk

Received October 2003
Accepted September 2004

with epidemiology, is not. If, however, randomization is an irrelevant distraction, and possibly even harmful, then the SCT might have cause to worry that it has made a fetish of a piece of nonsense, whereas ISCB can gain solace from its wider interests in biostatistical modelling.

Of course, as might be expected from my willingness to raise the topic here, my personal beliefs are between the two extremes outlined above. I do not consider randomization is essential for valid inference but I do think that it is valuable in increasing the credibility of inferences drawn from clinical trials.

There is a related issue that is even wider, however, and this is to do with the inferential purpose of clinical trials. I shall discuss this in due course but simply raise here, in this introduction, the thought that an important and relevant statistical concept is that of additivity: the notion or the hope that there is some scale the statistician can find upon which the treatment effect makes a (near) constant difference. One of the most famous of such scales is the log-hazard scale used, of course, in survival analysis.

It is thus not entirely a coincidence that both of these themes, randomization and additivity, feature in books co-authored by the presidents of the two societies. John Lachin, is a co-author with William Rosenberger of a book, *Randomization in Clinical Trials* [1] and Maria Grazia Valsecchi with Ettore Marubini of *Analysing Survival Data from Clinical Trials and Observational Studies* [2]. I note, in passing the stress on clinical trials in the book by the president of SCT and the additional mention of observational studies in that of the president of ISCB and the fact that collaboration is involved in both these enterprises! However, more germane to my purpose here are the many references to randomization theory in the former and the fact that the latter has a whole chapter devoted to 'Validation of the proportional hazards models'. Proportional hazards are, of course, additive log-hazards.

These two related themes are so vast that I cannot possibly do them full justice here. There are two reasons. The first is that space and time would not permit; the second is that the task is beyond me. All that I do below is touch on a selection of issues. The outline of this paper is as follows. In Section 2, I consider two early controversies, one regarding randomization and the other additivity. In Section 3, I make some general remarks concerning randomization and additivity in clinical trials. In Section 4, I cover a series of controversial questions and then in the final section offer some tentative conclusions.

2. TWO EARLY CONTROVERSIES

Both of these related topics have formed the object of celebrated controversies and, if one looks for controversy in statistics, one name is frequently prominent, that of Fisher (1890–1962). A public dispute with Neyman (1894–1981) at a Royal Statistical Society (RSS) meeting in March 1935 over the analysis of Latin Squares [3] led to an irretrievable breakdown in relations between these two giants of statistical inference. (See Joan Fisher Box's biography of her father [4, pp. 262–265] and Constance Reid's biography of Neyman [5, pp. 122–124].) In March of the following year, again at the RSS, there was a similar public disagreement with Student (WS Gossett, 1876–1937) over the use of randomization in experiments. Fisher may have seen this as a spirited but amicable disagreement with a revered older colleague but there is evidence that Student, who died the following year, considered that he had been treated less than fairly by his younger friend. (See Pearson's biography of Gosset [6, p. 66

and Joan Fisher Box, pp. 268–270]). I shall look at both of these disputes briefly and then consider some more modern ones.

2.1. Fisher's dispute with Gossett

Fisher never convinced Student that randomization was valuable in designing experiments. They had been disagreeing in private on the subject for 10 years or so prior to the occasion of the 1936 RSS meeting. The published record of that meeting entitled, 'Co-operation in Large-Scale Experiments,' is of, 'A Discussion, opened by Mr W.S. Gosset,' [7] but Gosset is by far the major contributor, providing about a third of what was published, and the whole is rather like a modern RSS 'read paper' but without formal votes of thanks. Student's piece is extremely narrow in terms of examples, being concerned directly only with variety trials and in that connection, as might befit the interests of a brewer, only barley. However, the issues covered are wide-ranging and there are many that are relevant to medical statistics. For example, he describes how,

> 'in the 1880s and 1890s the Danes, working with comparatively large plots, with few replications, but at several co-operating stations and in a number of successive seasons, were able to establish that Prentice was the most suitable barley to grow in Denmark' [7, p. 115].

Presumably, a form of what we now call meta-analysis was being used. He also draws attention to what medical statisticians would now call *trial-by-treatment interaction* by citing another statistician, now famous to us:

> 'On the other hand, Mr Yates has pointed out that it is not uncommon, when using the most modern methods in manurial experiments, to obtain a significant result on one occasion but, on repeating the experiment in another year or in another field, to get an equally significant result in the opposite direction.' (p. 116)

Student also refers subsequently to, 'the further real variation due to the differential response of the varieties to soil, climate and farming technique,' and in several other places discusses and examines the difference between the local standard error for the result of an experiment and the error of the reproducibility of that result from experiment to experiment.

Neither of these points were ones that Fisher would dispute. However, on the first page of his contribution Gossett referred to Fisher as follows:

> '... about fifteen years ago, Professor Fisher introduced the principle of randomizing the position of the plots in the various systems of randomized blocks and Latin squares with which many of you are familiar. This enabled us to obtain a certainly valid estimate of the variability of our results, though usually at the expense of increasing that variability when compared with balanced arrangements.' (p. 115)

It was a position that he expounded in the appendix to the paper. He advocated a particular sort of balanced design, originally due to the barley breeder and maltster Edwin S. Beaven (1857–1941), a friend and business associate of Gosset's through his work for Guinness. This was used for trials in two varieties (say A and B) which would be sown in a systematic sequence ABBA. The analogous design of a clinical trial would be to have given every fourth patient in a given centre the same treatment, having started the first on A, the second on B, the

third on B and the fourth on A. Student preferred this design to the randomized one because it balanced for any linear trend in fertility, the sum of orders for treatment A being $1+4=5$ within blocks and hence being identical to that for treatment B, $2+3=5$. Student was aware that this did not balance for any local quadratic trend, within blocks of four, stating, 'Periodic fertility slopes may undoubtedly occur, but apart from those due to the works of man, they must be so rare as to add negligible risk' (p. 121). Also, in the context of variety trials, there was a practical advantage in that sowing of the field could proceed efficiently.

Fisher's is the first contribution to follow and he expresses the essence of his disagreement with Gosset as follows [8]:

> 'The serious fact is that the actual errors of the split-drill method are always unknown, and though the result of the trial may be ornamented by the addition of the standard error, estimated by some plausible process, such estimates can never be scientifically on the same level as are standard errors of known validity.' (p. 124)

Fisher did not leave it there. Shortly after the RSS meeting, in fact in the same year, Fisher, in the *Annals of Eugenics*, a journal he edited, published a paper [9] he had written with Stefan Barbacki (1903–1979), a Polish geneticist. Their paper was a re-analysis of a uniformity trial reported by the American agricultural scientist Gustav Wiebe (1891–1975) in 1935 [10]. This analysis is discussed in the appendix to this paper. It is sufficient here to note that *in this particular example* Barbacki and Fisher were able to show that Student was not correct and that schemes in which treatments were randomly allocated in one of two ways were more precise. The first scheme, 'randomized pairs' used blocks of two with the sequences AB or BA, chosen at random. The second used randomized sandwiches of the form ABBA or BAAB. Note that for either of these two schemes, the analysis that Fisher deemed appropriate would be to reduce the data to a summary measure first: for example, the difference within blocks in the yield of A and B. Thus, the degrees of freedom available for analysis are one fewer than the number of blocks. Barbacki and Fisher's analysis of the second of these schemes is discussed in more detail in Appendix A.1.

Fisher's opposition to Student's preference for 'balanced' (systematic) designs, as opposed to randomized ones, can be summarized as follows.

1. They may or may not be more precise than randomized designs.
2. If they are more precise than the randomized design we shall not *know* that this is the case; we can only suspect it.
3. If the balanced designs are more precise than the randomized designs, the standard errors we shall quote for them will actually be larger than if they had been randomized and so we will be misleadingly conservative.

This last point requires some explanation and will be covered below when discussing minimization. For the moment it is enough to note that Student regarded it as a virtue to err on the side of being conservative and Fisher did not. For further discussion of Fisher's views on randomization see the biography by his daughter [4] and also 'Fisher's Game with the Devil' [11].

Finally, it is only fair to Student to note that Fisher did not follow his own advice. Some years later he put his name to a paper with Atkins in which he calculated a highly significant difference between two groups of men as regards vitamin C intake [12]. This was attributed to the timing of dose of vitamin C (before or after breakfast), other explanations having been

discounted. The analysis has the following defects. 1.There was no randomization: two sections of men were compared. 2. Initially there were three groups but one was discarded after seeing the data. 3. The results were dichotomized after seeing the data. 4. Despite Fisher's describing it as an experiment it was no such thing: there was no initial intention to allocate men to receive vitamin C at different times with a view to studying absorption [13].

2.2. Fisher's dispute with Neyman

In his paper Neyman proposed to fill in some gaps he believed that Fisher had left in developing his theory of the analysis of variance as applied to randomized block designs and, in particular, Latin squares [3]. What he claimed to show was that the mean square error for treatments did not, using the combination of Fisher's analysis of variance and a Latin square design, have the same expectation as the mean square error; in fact it was slightly higher.

Fisher replied at great length in remarks that were extremely critical of Neyman himself. He maintained that the analysis of variance applied to Latin squares was valid. He also made a reference, that at the time must have meant nothing to most who were present, stating:

> '...it was only about a year since another academic mathematician from abroad had been as much excited about having proved that the Latin Square was mathematically exact, as Dr Neyman seemed to be at having proved it inaccurate.'

It now seems likely that the 'academic mathematician' concerned was Samuel Wilks (1906–1964), who in 1933 submitted a paper to the Royal Society, in which, using characteristic functions, he proved the validity of Fisher's z test. (See Box's account [4, pp. 266–267]). One of the referees for the Royal Society, however, was Fisher and he wrote directly to Wilks on 27 December, 1933, and again on 6 February, 1934 drawing attention to his own paper of 1925, 'Applications of 'Student's' distribution'. (These letters are included in pp. 299–304 in the volume of Fisher's statistical correspondence edited by Bennett [14].) Fisher did not convince Wilks that his proof, although original, was redundant. However, Wilks, although a Texan who had studied at the universities of Texas and Iowa, was at the time a fellow student of W.G. Cochran's at Cambridge, who was able to show him the very simple proof of the unbiased nature of the estimates from Latin squares [4].

The dispute between Fisher and Neyman can be seen to have its origins in the choice of model. It has been claimed that this was not obvious to Fisher [15] but I disagree. I think Fisher understood perfectly well what Neyman's argument was but rejected it as being foolish. In words, Fisher's null hypothesis can be described as being, 'all treatments are equal', whereas Neyman's is, 'on average all treatments are equal'. The first hypothesis necessarily implies the second but the converse is not true. Neyman developed a model in which on average over the field, the yields of different treatments could be the same (if the null hypothesis were true) but they could actually differ on given plots. Although it seems that this is more general than Fisher's null hypothesis it is, in fact, not sensible. Anyone who doubts this should imagine themselves faced with the following task: it is known that Fisher's null hypothesis is false and the treatments are not identical; find a field for which Neyman's hypothesis is true. A more modern related argument concerns type II and type III sums of squares. The latter, allowing for the possibility of interactions when main effects are absent, violate Nelder's marginality principles [16] and are similar in spirit to Neyman's approach.

Fisher's position is very similar to one he held in connection with the validity of Student's *t*-test when comparing two means, an application for which he himself was responsible. In connection with experimental work he wrote [17].

> 'It has been repeatedly stated, perhaps through a misreading of the last paragraph, 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, namely that the samples are drawn from the same normal population.' (pp.124–125)

We shall return to some of these issues when discussing additivity below. For the moment we note simply that Fisher regarded significance testing and estimation as two separate issues, whereas Neyman, who had developed confidence intervals from hypothesis testing, saw them as closely related. Under the null hypothesis it makes no sense to allow for treatment-by-unit interaction, hence it does not affect the validity of a Fisherian significance test. (This is not quite the same as saying that therefore it should never be allowed for since, to use a concept Fisher hated, it might affect the power of the test.) On the other hand, if the null hypothesis is false, then there are treatment main effects, and treatment-by-unit interactions are a distinct possibility that any estimation procedure perhaps ought to take it into account; it was thus perhaps not surprising that Neyman considered them. (Although this does not justify his particular approach as logical.)

For a careful examination of the difference between Fisher and Neyman's model see Cox [18].

3. RANDOMIZATION AND ADDITIVITY

In this section some general points will be made about randomization and additivity. These will be relevant to the issues covered in Section 4.

3.1. Clinical trials and randomization

An important difference between agricultural trials and clinical trials is that in the latter the experimental units are not identified in advance of the trial being designed. In an agricultural trial, such as those discussed by Gossett and Fisher, the field to be used could be identified, and the blocking structure likewise and the pattern of treatments then imposed on this. Chronologically, the situation is the reverse in a clinical trial. The treatment structure is defined first and the patients who will be treated are 'discovered' as the trial progresses. Of course, at least some, and sometimes all, of the centres in which the trial will take place are identified at the beginning but in standard clinical trials the units are not centres but patients, and one does not know with what frequency suitable patients will present, nor does one know how many will give consent. This means that highly organized prognostic blocking structures of the sort discussed by Gosset and Fisher are rare, (cross-over trials being a notable exception [19]).

Treatment allocation is usually implemented in the one of the two ways. Often, in double-blind trials carried out by the pharmaceutical industry, the randomization list is prepared in advance, the treatment packs, identical in appearance are numbered and provided in advance

of recruitment to the centres, and the pack with the lowest available number is given to any patient that is recruited. We can call this *pre-allocation*. Common elsewhere, and also increasingly used by the industry, is some form of *post-allocation*. The patient is entered onto the trial and then, possibly after covariate information has been taken into account and perhaps using some automated telephone or web-based system, the investigator is told the number of the pack that is to be given to the patient who has just been recruited.

When it comes to schemes for sequences of allocation, two are usual. Within the pharmaceutical industry the most common is the method of randomized blocks. A block size is chosen to be some multiple of the number of different treatments being given (unless unequal allocation is being used in which case it must be some multiple of the sum of the ratios). For a given block a suitable permutation is chosen at random. For example if the block size is four and there are two treatments A and B, then for a given block one chooses at random among AABB, ABBA, BBAA, ABAB, BABA and BAAB. (Had Student advocated this scheme rather than using only ABBA, Fisher would have had no objection.) This ensures that the treatments are allocated in the chosen proportions for each block. This general approach can use either pre-allocation or post-allocation, although the former is usual. It can only balance for covariates by stratification and employing a separate list for each stratum.

'Public' organizations such as the Medical Research Council (MRC) in the United Kingdom or the European Organization for Research and Treatment of Cancer (EORTC) have tended to rely instead on what Rosenberger and Lachin call 'covariate-adaptive randomization'. [1]. However, it is arguable as to whether all such schemes really involve randomization and we shall refer to this approach by the more general 'Allocation Based on covariates' (ABC). The patient must be entered first into the trial and values made known so that the central trial office can then announce the allocation, so that the method of post-allocation alone is suitable using either telephone or (nowadays) web-based allocation. A common form of ABC is 'minimization' or some variant of it. This is an attempt to balance trials by using prognostic information. The method of pre-agreed sequences cannot be used in connection with this approach.

The analysis of ABC trials will be considered in more detail below. For the moment we note simply that the common analysis of clinical trials randomized in blocks violates Fisher's prescription anyway. It is common to use sub-centre blocks, so that there are two or more blocks per centre. However in the pharmaceutical industry, in analysing the data, if a linear model is employed, it is usual to fit centre as a factor but unusual to fit block. If a non-parametric procedure were used, the van Elteren test [20] might be employed but would usually be stratified by centre and not block [21]. Outside the pharmaceutical industry, it is common practice to fit nothing apart from the treatment itself.

3.2. Clinical trial questions and additivity

Many standard statistical models are additive in the sense that the *simplest* assumption consistent with them is that, on the scale chosen, the effect of treatment is to induce a constant shift in the response. (This is not to say that this is always the *necessary* interpretation.) Consider, for example, a parallel group trial with k treatment groups and (rather unrealistically and purely for the sake of simplicity) exactly n patients per group. If the response, Y is continuous, we may model some transformation of it Y' (most simply the identity transformation and very commonly a log transformation [22]) using the general linear model. Thus we might

write something like

$$Y'_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

for the response for patient j in treatment group i. Here τ_i, is the effect of treatment i. However such individual treatment effects are not estimable and attention is focussed on estimable contrasts usually in terms of the difference between pairs of treatments and hence of the general form $\tau_h - \tau_k$, $h \neq k$. The ε_{ij} terms are disturbance terms and will reflect general differences between subjects and also what one might call measurement error, as well as sometimes differences from occasion to occasion within subjects, since, although we are only assuming that we measure a given subject once, this is often only one of many occasions on which we *might* have measured the subject. Such disturbance terms may also reflect subject-by-treatment interaction, a feature that Neyman clearly allowed for in his model (although his simultaneous constraint that the treatment differences were zero was then rather bizarre). However, without further measurements in the form of covariates, we do not in general have any way of telling what the relative contribution of these various elements is to the overall variability [23].

However, as Fisher pointed out in 1938 in a letter to Henry Daniels [14], a clue to the presence of interactive effects can sometimes be gained by studying variances. As he put it:

> '...although on the data it could not be said that the means were different...supposing the test were made between two varieties of the plant, the fact of a real difference in the variances shows that in some circumstances one variety is the better and in other circumstances that it is the worse'. (p. 64).

In the clinical context, where response to treatment differs from patient to patient, the variance may be expected to increase. Indeed, in a placebo-controlled trial if such an interaction is suspected, a more powerful test of the null hypothesis can be used by basing it on the variance from the placebo group only [24]. Sometimes a transformation can restore additivity. Figure 1 is a dot-plot from a pre-clinical study in which inhibition of thromboxane B2 (TXB2) is being studied in six groups of rats. There are four experimental treatments and one positive control (marketed product) as well as one negative control (vehicle). There is a clear effect of treatment and clear evidence of heteroscedasticity. Bartlett's test rejects homoscedasticity ($\chi^2_5 = 50.9,\ p < 0.0001$). However, a log-transformation seems to cure the problem as is shown by Figure 2 ($\chi^2_5 = 8.95,\ P = 0.111$).

More complex data sets may permit us to identify non-additivity at some level or other. For example, suppose that we have recorded the sex of patients. Then by comparing the treatment effect in women to that in men, we can see whether the response is the same for the two sexes. This is a specific example of what might be referred to as block-by-treatment interaction. Note that such interactions are qualitatively different from treatment-by-treatment interactions, as might be studied in factorial experiments, since they involve factors, the block structures, that the trialist does *not* allocate. For example, we might later discover that in this trial all the males are young and all the females are old so that what had been interpreted as sex-by-treatment interaction might equally well be regarded as age-by-treatment interaction.

Sometimes an apparent lack of additivity may have a more subtle explanation. Suppose that we have a parallel group trial in epilepsy and decide that the number of seizures should follow a Poisson distribution. We note, however, that the variance of the number of seizures within groups is larger than the mean number of seizures. This might be a reflection of

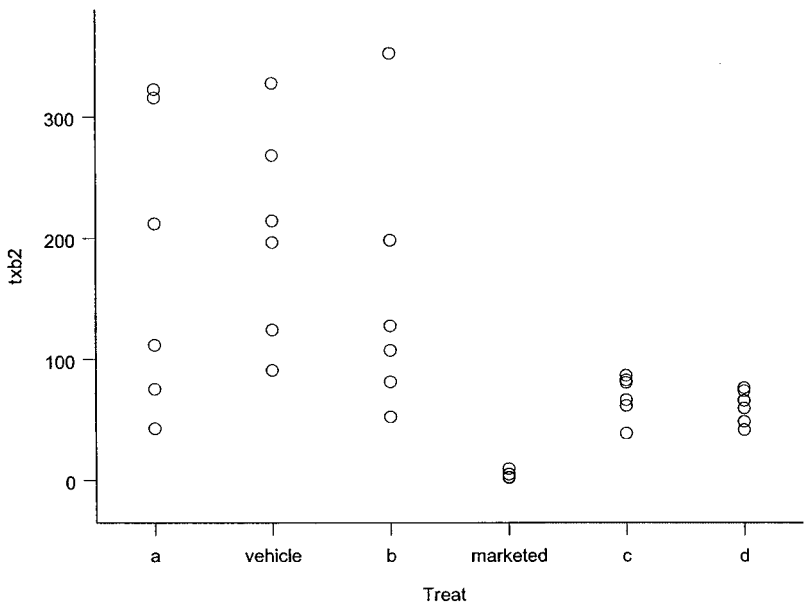

Figure 1. Experiment comparing four experimental treatments to a positive and negative control as regards degree of thromboxane B2 inhibition.

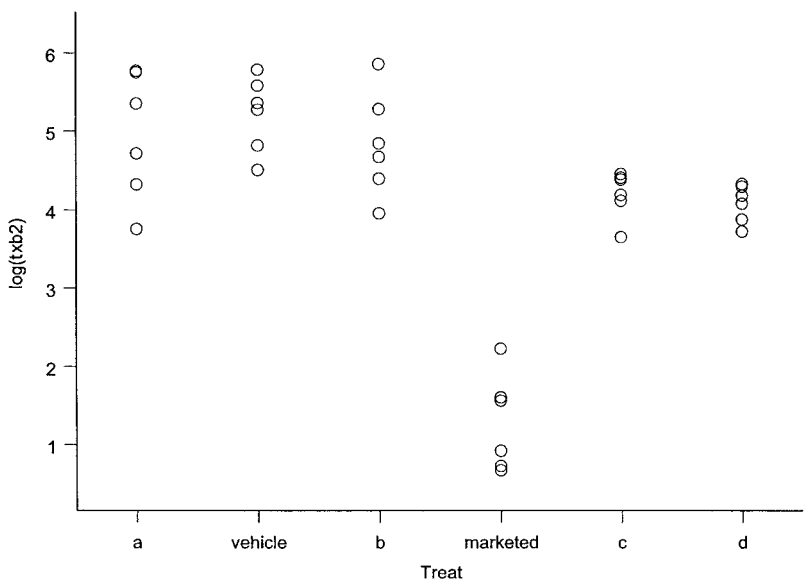

Figure 2. The same data as in Figure 1 but log-transformed.

patient-by-treatment interaction but equally it might simply be a reflection that the patients are not homogenous. For a cross-over trial in which each patient could provide his or her own local control there might be no extra-Poisson variation [19].

However, I am not so interested here in technical aspects of additivity, such as whether or not it is important to assume it, when and how it can be detected and what to do if it does not apply. Rather I am interested in how different assumptions regarding it can have an impact on the interpretation of clinical trials and the questions we might wish to answer with them. To make matters simple, suppose that we have a two-aimed trial comparing an active treatment or verum to a placebo. The effect of verum compared to placebo will be referred to as the effect of treatment. These are a number of questions we might want to answer in such a trial.

Q1. Was there an effect of treatment in this trial?
Q2. What was the average effect of treatment in this trial?
Q3. Was the treatment effect identical for all patients in the trial?
Q4. What was the effect of treatment for different subgroups of patients?
Q5. What will be the effect of treatment when used more generally (outside of the trial)?

For expert discussion of the sorts of questions that may arise in experiments in general see Cox [18] and also Dawid [25].

Given an assumption of what might be called local (or weak) additivity, that is to say that the effect of treatment was identical for all patients in the trial (in other words that the answer to Q3 is 'yes'), then Q1, Q2, & Q4 can all be answered using the same analysis: a confidence interval or posterior distribution for the mean effect of treatment says it all. The effect on each patient is the average effect Q2 and is hence the effect in every subgroup Q4 and if it is implausible that this effect is zero, then the treatment has an effect Q1. Given a further assumption of universal (or strong) additivity, this observed effect is the effect to every patient to whom it might be applied; this also provides an answer to Q5.

Now, in my view, nobody believes literally in local, let alone universal additivity. However, there are circumstances under which there is no point in worrying about it, primarily when there is nothing much that can be done about the possible lack of it: for example, if we have run a simple randomized trial in which we have failed to collect any covariate information and if ethical considerations prevent us from running further trials. Suppose that we have shown that on average a new treatment is (highly) effective in the patients we have studied. It may be that this effect is an average of exceptional benefit for some patients and none at all for others but unless we can identify the sort of patient for whom it works there really is no choice but to use the average to inform our decisions. Of course, if the data permit, then some attempt should be made to find evidence regarding Q3. However, in some examples I consider below, I shall show that considerable harm has been done by looking naively for interactions.

4. MODERN ISSUES AND CONTROVERSIES CONCERNING RANDOMIZATION AND ADDITIVITY

In this Section I, consider some modern issues regarding randomization and additivity. They are modern in the sense that they are still live, but they are not necessarily new and indeed

some reflect the disputes covered in Section 2. Except where explicitly stated otherwise, I shall assume that design and analysis of a two-armed parallel group trial is being discussed.

4.1. Does randomization balance covariates?

There is a trivial sense in which the answer is *no* and another trivial sense in which the answer is *yes*. A given randomization is hardly ever perfectly balanced (never as regards continuous covariates, purely fortuitously for discrete covariates) but averaged over all randomizations the groups are balanced. This is one reason why testing baseline balance using significance tests as part of a strategy of statistical analysis is a complete waste of time: it is irrelevant both as regards the trial run and the infinity one might have run [26–30]. If it has any purpose at all it is purely as a matter of quality control: as part of a more general strategy in order to check whether the trial was randomized. It should be noted here, however, that despite the fact that a trial may be intended to be randomized various subtle biases could subvert such intentions. See Berger and Bear for a discussion [31].

It is sometimes maintained that balance improves with increasing sample size, but this depends what is meant. Consider a concrete example: mean baseline diastolic blood pressure at baseline in a two group trial in hypertension with no blocking of any sort apart from in terms of total numbers of patients and with n patients per arm and hence $v = 2n - 2$ degrees of freedom for error. The expected absolute mean difference between groups will decrease as sample size increases; on the other hand, the expected absolute sum of differences will increase. Between the middle of these two extremes we have standardized differences such as the t-statistic for the difference between groups. As soon as more than a very few patients are recruited, the variance of this statistic hardly changes with increasing sample size being given by

$$\mathrm{var}(t) = v/(v-2) = (2n-2)/(2n-4)$$

and, of course, if we go one step further and make the P-value transformation, then we cannot say that a value of $P \leqslant 0.10$ is more or less likely for larger sample sizes compared to smaller ones.

Furthermore, standard (frequentist) probability statements such as, for example, confidence intervals, reflect sample size increases in terms of increased precision rather than increased coverage probability; hence, as regards this we are no better off in large trials than in small ones [32]. (A similar point has been made in connection with sample surveys by Cumberland and Royall [33]) Whether this matters or not is an issue for debate; I am simply pointing out here that a great deal of nonsense is talked about this when discussing clinical trials. A 95% confidence interval from a large trial is likely to be narrower than a corresponding interval from a small trial. It is not less vulnerable to chance covariate imbalance. This point is discussed further in Appendix A.2.

4.2. Minimization and randomization

In an editorial together with Tom Treasure, the late Ken MacRae, one of the greatest of communicators of statistical ideas to the medical profession, claimed that minimization had become the platinum standard of clinical trials [34]. Minimization is a method of dynamic allocation first proposed by Taves [35] and subsequently by Pocock and Simon [36]; it works by using a total score based on adding patient characteristics.

As long as it is only the total number of patients one wishes to balance on trials, two simple schemes are as follows; one is analogous to Student's approach and the other to Fisher's. Let each patient be identified by a unique number from 1 to $2n$ indicating order of recruitment into the trial. The analogy to Student's approach would be to give all odd numbered patients one treatment and all the even numbered patients the other, only the first patient's treatment being chosen at random, if at all. (Although unlike the Student/Beaven sandwich arrangement this would be vulnerable to linear trend.) Fisher's approach would be to randomize in blocks of size 2. For block number j, $j=1,\ldots,n$, it would be decided at random which treatment the first patient in the block, that is to say patient $i=1+2(j-1)$ of the trial, would get. The second patient in the block would then get the other treatment. Both Student and Fisher would then analyse the trial in terms of the treatment difference. This is equivalent to forming an analysis of variance table with degrees of freedom (DF) as follows:

Source	DF
Blocks	n-1
Treatments	1
Error	n-1
Total	2n-1

Corresponding to this table there are two possible allocations for Student's scheme and $2''$ possible allocations for Fisher's.

In this context, unlike the agricultural one where sowing in complex patterns may be difficult, there is no advantage to Student's scheme compared to Fisher's. However both suffer from the disadvantage that in an open trial the allocation of every even-numbered patient is known (and in Student's case every odd-numbered patient except the first) and by deciding whether or not to enter patients in a trial the allocation process could be biased. In drug regulation very few pivotal trials are open, so that this is not a serious disadvantage for Fisher's approach. However, where open trials are used, an alternative approach, a so-called 'biased coin' design described by Efron may have some advantages [37]. This works as follows. At the point at which patient i is recruited, let $\tilde{D}$ be the excess of patients on experimental treatment (A) compared to control (B). Let $\frac{1}{2}\leqslant p\leqslant 1$ be the 'coin bias'. Then, if $\tilde{D}<0$ assign A with probability p, if $\tilde{D}=0$ assign A with probability $\frac{1}{2}$ and if $\tilde{D}>0$ assign A with probability $1-p$. If $p=1$ we have Student's design and if $p=\frac{1}{2}$ we have a completely randomized trial.

If the purpose is to balance for covariates, then Efron proposes that this procedure be carried out separately with appropriate strata. It has been claimed that there is a problem in implementing this, however, and that is if k covariates are involved the allocation must be carried out in 2^k strata, some of which may be poorly represented or empty. (Although, actually, it is not empty strata that cause a problem, since they are balanced, but rather sparsely represented strata, which may be unbalanced.)

As an alternative, Taves [35] and subsequently Pocock and Simon [36], independently proposed a strategy for achieving a form of total balance. An excellent simple account is given by Pocock in his classic text [38] and a more technical description by Matthews [39]. Suppose that we wish to decide how to allocate patient i and that we are 'minimizing' by k factors. Let $\tilde{D}_{h(i)}$ be the 'excess' (which may be negative) of patients in group A compared to group B

having the same level of factor h as patient i. Then let

$$\tilde{D}_i = \sum_{h=1}^{k} \tilde{D}_{h(i)}$$

This is then a score that can be used in the same way as Efron's.

The problem with this score is that it has no theoretical justification as being a good way to balance. The question raised by considering this matter further is, 'what is balance for?'. Achieving greater balance has nothing to do with increasing the validity of inference with respect to confounding. This can be done by fitting the prognostic factors in an analysis of covariance (ANCOVA). Discussion of this option is sometimes misleading. Although stratified allocation requires use of 2^k strata, this achieves balance for all main effects and all interactions. Minimization does not use interactions at all and so the analogous control using ANCOVA is to fit main effects only: in other words k, not 2^k prognostic variables are fitted. What is also known is, that when this is done, the expected loss in efficiency of inference for a randomized design compared to the optimal (which minimization will not reach) is equivalent to a loss of k patients [40].

Atkinson has suggested an alternative to minimization based on sound design principles [41, 42]. This uses the fact that for a general linear model the variance of a vector of regression coefficients, $\boldsymbol{\beta}$, associated with a design matrix, $\mathbf{X}$ is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \tag{1}$$

The right hand side of (1) is the product of two factors, the first of which, $(\mathbf{X}'\mathbf{X})^{-1}$ depends only on the allocation of patients and not on the response. One of the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ will be that corresponding to the estimate of the effect of treatment. Let us call this element M (for multiplier). In a trial with $2n$ patients the minimum value that this could achieve would be $M = 2/n$. In practice this will not be achieved but given knowledge of the covariates of the patients recruited into the trial up to and including patient i, who is about to be allocated, we can calculate the value M will have if the patient is allocated to A as M_{iA} and if allocated to B as M_{iB}. Then, if we define $\tilde{D}_i = M_{iB} - M_{iA}$ we can use this in Efron's scheme. Atkinson also considers alternative ways of using this general design principle to 'bias the coin'.

I do not much like ABC schemes, principally because their advantages have been over-stressed and they have some disadvantages (to be discussed below). However, in terms of the old saying, 'I would rather be hanged for a sheep than a lamb'; if I were to use such a method I would use one of Atkinson's schemes. There is no place, in my view, for classical minimization.

If an investigator uses such ABC schemes, she or he is honour bound, in my opinion, as a very minimum, to adjust for the factors used to balance, since the fact that they are being used to balance is an implicit declaration that they have prognostic value. In the case of a linear model the standard error quoted will generally be too small if this is not done. More generally, the estimates themselves will be biased towards zero [27, 43, 44]. The point of view is sometimes defended that analyses that ignore covariates are superior because they are simpler. I do not accept this. A value of $\pi = 3$ is a simple one and accurate to one significant figure, and even, as a referee has reminded me, with perhaps some biblical justification (see I Kings, 7: 23–26 and II Chronicles 4:2–5). However, very few would seriously maintain that

it should therefore be generally adopted by engineers. It is a sad reflection on the standards we believe should apply to medical research if this is our attitude.

It is also sometimes maintained that the fact that inference based on balancing but not conditioning is conservative is perfectly acceptable but such an argument severely undermines the other usual argument in favour of minimization. Suppose that we have a parallel group trial in asthma and are balancing by dichotomizing baseline forced expiratory volume in 1 s (FEV_1) at 2.75 l. Doing this would at best lead to an economy of one patient in the trial compared to a randomized design *if the baseline value were fitted in ANCOVA*. The expected residual variance having analysed the data using ANCOVA would be approximately $(1-\rho^2)$ of what it would be if ANCOVA were not used and $(1+\rho)/2$ of what it would be if a so-called change score analysis were used, the latter option only being a possibility in the unique case of baseline as covariate [45]. Neither of these factors can be greater than 1. In practice the correlation might be about 0.7 leading to a 15% saving in patients compared to the change score and a 50% saving compared to using raw outcomes. Fitting is more important than balancing.

Furthermore, as Fisher pointed out in commenting on Student, if we balance by a predictive covariate but do not fit the covariate in the model, not only do we not exploit the covariate, we actually increase the expected declared standard error. This is because of the analysis of variance identity

$$\text{Total sum of squares} = \text{Treatment sum of squares} + \text{Error sum of squares}$$

Minimization tends to reduce the treatment sum of squares but under the null hypothesis we can regard the total sum of squares as a quantity which no allocation scheme can possibly affect. Therefore, under the null hypothesis, a reduction in the treatment sum of squares must lead to an increase in the error sum of squares.

A simple example may make this clear. Imagine a trial with 100 females and 100 males. Assume that we will not adjust for sex but that this factor has prognostic value as would, say be the case in asthma where the FEV_1 of males is higher on average than females. Clearly we need to split males and females 50:50 in each treatment group if we wish to eliminate the effect of sex on the treatment estimate. Consider, however, what allocation would eliminate the effect of sex on the estimate of the variance. We would have to allocate all the males to one group and all the females to the other. In other words the allocation that minimizes the effect of sex on the variance is that which maximizes its effect on the treatment estimate and *vice versa*.

Of course, we do not need to renounce our ability to put sex in the model and this is Fisher's point. We can block by sex, and eliminate it from the estimate of error. Fisher was opposed to restrictions that go beyond what is indicated in the analysis. A Bayesian justification would be that what you put in your model indicates what you think is important. Your behaviour is incoherent if you restrict what is not important.

Do we satisfy Fisher if we use an ABC approach and fit the covariates in the model? Not necessarily. Such methods not only produce a degree of balance by the end of the trial but at all stages throughout it. Consider even the simplest variant, Efron's biased coin design, in which the only factor being balanced for is the number of patients. Suppose that there is a strong trend in prognosis throughout the trial. The trial will be of necessity very closely balanced by prognosis whereas a randomized trial that balances by number could, theoretically,

allocate the first n patients to one group and the last n to the other and hence be affected by this evolution of prognosis. Efron's design would give a more precise answer.

But it would also claim to be less precise and this raises a problem for combining the results of experiments which Student appears not to have considered even though this is what, ostensibly, his paper was about. In carrying out a meta-analysis we wish to weight trials by their precision and thus we need to know how precise they are. To claim that minimized trials are the platinum standard and randomized trials are only the gold standard but then, other things being equal (in this case the number of patients recruited), to give platinum less weight than gold, would be illogical [21]. One solution may be to fit for minimized trials, in addition to covariates used, a linear trend over the trial as a whole. This may have some advantages for randomized trials also [21].

4.3. Fixed and random effect meta-analysis

In his RSS discussion Student noted that the variation of differences in yields from trial to trial was much greater than would be suggested by the individual standard errors. This would imply that modern fixed and random effects meta-analyses might yield different conclusions and certainly that the latter would have a wider confidence interval. However, the more modern debate over the appropriate way to perform a meta-analysis also has echoes of the Neyman versus Fisher controversy. It is often stated that the key to whether a fixed or a random effects analysis should be performed is whether there is trial-by-treatment interaction. This is rather misleading. The absence of trial-by-treatment interaction means that point estimates and confidence intervals are very similar whichever approach is used. To test the null hypothesis of no treatment effect, a fixed-effects analysis would always be a reasonable thing to do, since the hypothesis of no interaction belongs to the hypothesis of no treatment effect. It is analogous to Fisher's test. As soon as we accept that there is a treatment effect then Neyman's model becomes interesting (although not his constraint under the null hypothesis that the treatment effect may be zero). It then becomes reasonable to accept that there may be trial-by-treatment interaction and, indeed, if our purpose is then to answer the analogous question to Q3 of Section 3.2, the estimation of the relevant component of variation using perhaps either the method of DerSimonian and Laird [46] or Hardy and Thompson [47] is indicated. This may also help in answering Q5 but claims for this should not be over-stressed. Our assessment of trial-by-treatment interaction is limited to those trials in the meta-analysis and neither the patients nor centres are a random sample of those that might use the treatment being compared, *nor could they ever be* [48].

4.4. Randomization versus random sampling

Although I have no great enthusiasm for pure randomization-based inference, permutation tests and so forth, I do believe that at the very least the analyses of randomized trials should be comparative: they should stress contrasts rather than group means. It is a curious vice, of which even statisticians are frequently guilty, to calculate standard errors of group means from clinical trials. Such standard errors are more or less ritually calculated by taking the standard deviation and dividing by the square root of the sample size. This formula, which everybody learns in 'Stat 1' at university, is justified by simple random sampling, a thing which *never* happens in connection with clinical trials *and never could*, as pointed out in the preceding section. A statistician who worked in survey work would be horrified by this cavalier attitude.

The cure, in my view, is not to try and carry out representative clinical trials. (But see Longford, 1999, for the contrary view [49].) Even the limited aim of recruiting typical centres would be incredibly difficult and, of course, it is a logical impossibility to sample from the future. The cure is to recognize that trials must stress comparisons, that such comparisons require suitable scales, and that their results may require careful implementation when used for prediction. (This point is taken up in Section 4.10 below.)

Personally, I should like to see standard errors for group means abolished [50].

4.5. Main effects of trials as random and unequal randomization

There is another kind of random effect that we can have in connection with meta-analysis of trials, although it does not appear to have been discussed much. We could have the main-effect of trials as random. In the case where the trials have used unequal allocation, this might permit the recovery of inter-trial information. However, such recovery violates the principal of concurrent control and may not be advisable since different allocation ratios may have been used for different treatments over time thus introducing confounding with any secular trends. Its justification would be purely model based and would be not based on any randomization as actually carried out. However, an apparently analogous situation in multi-centre trials would be to treat centre effects as random and recover inter-centre information. This ought not to be affected by the same biases, and might even have a randomization justification, but appears to be rarely employed [48].

4.6. Choice of variance in trials with three or more arms

Pair-wise treatment contrasts are nearly always of far greater interest in clinical trials than global comparisons (via an F-test) between *all* treatments. Of course in a two-armed trial there is no distinction between these. A common habit, however, in analysing trials with three or more arms is to pool the variances from all arms when calculating the standard error of a given contrast.

In my view, this is a curious practice, or at least, it is curious that it is so little questioned. It relies on an assumption of additivity of *all* treatments when comparing only *two*. Here, unlike Fisher's defence of the two sample t-test, equality of variances (since *all* variances have to be assumed equal) is *not* implicit in the hypothesis being tested. When comparing two active treatments in a trial that also includes a placebo it could be liberal. It seems to be a habit inherited from the analysis of agricultural field trials where degrees of freedom for error were scarce: a five by five Latin square would have 12 degrees of freedom for error; however a phase III parallel group trial, typically has well in excess of 100 available for any pair-wise contrast.

A Bayesian, of course, by putting a prior on the extent to which variances can differ from treatment to treatment would be able to adopt a compromise analysis between the two extremes of pooling and not pooling. For the frequentist it would seem that where degrees of freedom are not scarce there is potentially more to be lost than gained when pooling, since a classical t-test is robust to heteroscedasticity provided that sample sizes are equal in the groups being compared and that the variance estimate is internal to these two groups but is not robust where an external estimate is being used.

4.7. Type III versus type II sums of squares

There is a long-running controversy regarding the analysis of multi-centre trials (and also more generally) regarding the appropriate sums of squares to fit for looking at the effect of treatment as to whether type II or type III sums of squares are most appropriate for this [16, 51, 52]. By adjusting main effects for interactions a type III analysis is similarly illogical to Neyman's hypothesis test. It violates Nelder's marginality principle [16]. It is to be hoped that this controversy is at long last being resolved in favour of type II sums of squares [53, 54].

4.8. Individual response to treatment

As already discussed in Section 3.2, a worry in clinical trials is that the treatment effect may vary from patient to patient. There is also a danger, however, that we may underestimate the extent to which apparent differences between patients do not reflect true differences in response but either measurement error or some other variation within patients. Recent papers have claimed to be able to provide novel ways of detecting interactive effects. These claims are false.

For instance, Horwitz *et al.* took a 31 centre trial comparing propranolol to placebo as regards its effect on myocardial infarction [55]. In ten of the centres the patients appeared to better under placebo. The treatment effect was then compared in these ten centres to the 21 centres in which the propranolol group had the better response using the Gail–Simon test [56] for qualitative interaction. The difference was highly 'significant'. They opined that searches for differential effect in subgroups were legitimate,

> 'provided that (1) a clear clinical or biological explanation is available to support the numerical results, and (2) a formal statistical test for qualitative interaction is performed that shows a statistical significant effect that excludes chance as an explanation for the divergent results.'

When, however, Senn and Harrel pointed out that the Gail–Simon test had been used in a way that was specifically forbidden by its authors and that in fact a random effects analysis of the 31 centres showed, that far from chance being excluded as an explanation, there was no excess variation in the treatment effect compared to chance [57, 58], Horwitz *et al.* changed their opinion as to the importance of formal analysis, cheerfully admitting, 'we indeed violated Gail and Simon's admonition' [59] but failing to explain what in that case the meaning of the p-value they had calculated was. In a follow up letter, they stated,

> 'Only when the investigator has explored a thorough set of factors that could cause the observed differences should chance be accepted as a plausible explanation.' [60]

So by now, in fact, they believed the reverse of what they had previously stated as their point (2); instead of having to exclude chance first, one now had to exclude every other possible explanation before chance could be considered. If this is how interactions are to be judged why not main effects also? The number of cures we shall find will rise dramatically.

As another example of the genre, consider the claim by Guyatt *et al.* to have found a 'method for estimating the proportion of patients who benefit from a treatment when the outcome is a continuous variable' [61]. In fact, they had done no such thing. For most trials conventionally run, the proportion is unidentifiable and it was so in all the examples they

used. For example in a cross-over trial in asthma comparing quality of life for salmeterol to salbutamol, they calculated the proportion of patients for whom the *observed* quality of life differed by 0.5 points between the two treatments. On this basis a proportion of 0.32 were 'better off' on salmeterol and 0.1 'better' on salbutamol. This was interpreted as a net benefit of 0.3 which was, almost inevitably for a *British Medical Journal* paper these days, translated into that easy to interpret measure the number needed to treat (NNT), the value in this case being 3.3. As the authors put it,

> 'Once investigators have excluded chance as an explanation for differences between groups they can examine the proportions of patients who have deteriorated, remained the same, or improved as an aid in interpreting the importance of the results.'

However, unfortunately they forgot that chance can explain not only averages but individual results also [62]. There is nothing in the data which excludes the possibility that, were patients to be given salmeterol over two periods, the chance difference between periods would be greater than 0.5 for some. We should then have the absurd position, using this definition of preference, that some patients preferred salmeterol to salmeterol, whereas others preferred salmeterol to salmeterol. I want to make it quite clear here that I have no objection in principle to the use of observed preferences in analysing cross-over trials (except that this is often inefficient), such as for example leads to Fisher's exact test [19]. I am simply objecting to over-interpretation of such preferences.

In fact, conventional cross-over trials do not permit one to resolve residual error into patient-by-treatment interaction and pure within-patient variability, the former being analogous to the term on which Neyman placed so much stress in his analysis of Latin squares [63]. To do that one needs a cross-over design in which patients are repeatedly treated with the same treatment. The difficulty with conventional cross-over trials applies a *fortiori* to parallel group trials [23]. The paper by Guyatt *et al*. is only one of a growing number that illustrate the dangers of NNT-induced dichotomania. The ease of interpretation of these measures is a delusion [64–66].

It is to be regretted that we are likely to see more such unjustified inferences. When I checked the Web of Science on 29 June 2003, the paper by Horwitz *et al.* had been cited 28 times and that by Guyatt *et al.* had been cited 79 times. The letters pointing out the fallacies had been cited only 8 and 5 times respectively.

4.9. Non-linear models

As soon as one moves away from the linear model, even to the rather confusingly (and in some ways misleadingly) named generalized linear models [67, 68], then some of the re-assurances provided by randomization and balanced designs disappear. In the Normal case balanced allocations produce unbiased estimates but invalid estimates of standard errors. Randomized designs produce estimates that are unbiased over all randomizations and produce valid estimates of standard errors. For other cases this is not generally correct and an ignored prognostic covariate will produce biased estimates [27, 43, 44, 69]. Thus, for example, if we pool heterogeneous strata, the odds ratio of the treatment effect will be different from that in every stratum, even if from stratum to stratum it does not vary [43].

Of course, one might argue that this does not matter. If covariates have not been identified in the trial, then they may not be identified in future. Hence it is the odds-ratio of the pooled

data that is relevant. However, against this argument, we may note that the patients in a clinical trial are not a representative, let alone a random, sample of the target population and the degree of heterogeneity will vary considerably from trial to trial and will cause problems in any meta-analysis. This is one reason, why measures that attempt to quantify treatment effects by the degree of overlap between distributions are far from the ideal solution to our inferential difficulties that is sometimes claimed. These measures will differ from trial to trial even if the effect is constant, simply because standard deviations will change [70]. Such measures cannot be correctly interpreted, *pace* Stine and Heyse [71], as, 'a clinically relevant measure of similarity by using individual patient responses,' (p. 231) for reasons, amongst others, discussed in Section 4.8.

Part of the problem with Poisson, proportional hazard and logistic regression approaches is that they use a single parameter, the linear predictor, with no equivalent of the variance parameter in the Normal case. This means that lack of fit impacts on the estimate of the predictor. One way to deal with this is to allow for the random effect of unobserved covariates as for example in frailty models.

The issue this raises, however, is what is the value of randomization if, in all except the Normal case, we cannot guarantee to have unbiased estimates. My view, which I believe was Fisher's later view, whether or not it was his earlier view, was that the form of analysis envisaged (that is to say, which factors and covariates would be fitted) justified the allocation and *not vice versa*. With this philosophy we do not need to make a fetish of randomization but it nevertheless becomes relevant to ask ourselves if, a wide set of possible allocations are equally or nearly equally efficient given a particular model, whether we are not prepared to choose an allocation at random and if not why not.

4.10. Prediction in practice

Scales of measurement that show additivity may not be the most clinically relevant. There may thus be a conflict between measuring what a treatment does and how well it does what we should like it to do. For example, the log-odds scale would be a better default choice of additive measure than the probability scale but the latter is relevant to practical decision-making. This would seem to provide a conflict but a possible resolution is to use the additive measure at the point of analysis and transform to the relevant scale at the point of implementation [72, 73]. This transformation at the point of medical decision-making will require auxiliary information on the level of background risk of the patient.

4.11. Binary equivalence

The above approach might be a way to handle binary equivalence trials [21]. Here the problem is that the equivalence margin, or more usually the non-inferiority margin, may be set in terms of probabilities but the control group probability will not be known until the trial has been run and even then may not be a reliable estimate of the probability in the population. This is complicated by the fact that clinical trials do not study representative patients, as discussed in Section 4.4. Consequently relevant uncertainties attach not only to the results of the trial but to the background information. The trial analysis will almost certainly have to be supplemented by an extensive background summary of the properties of the control treatment, perhaps in the form of a meta-analysis. This, together with an analysis in the trial on the additive scale,

can then be used to estimate the probability difference when used in practice, as discussed in Section 4.10.

However, this alone may not be enough and it may be necessary in addition to use the sort of formal decision analysis that has been proposed by Lindley [74, 75].

5. CONCLUSION

I hope I have succeeded in showing that some issues that formed the subject of considerable controversies nearly 70 years ago remain live and relevant today. It would be arrogant of me to suppose that I can resolve them. All I propose to do here is make some optimistic remarks as to how we might think about them and their implications for clinical trials.

The dispute between Fisher and Student is a fundamental one philosophically but perhaps not as important practically for us as it appeared to them. If I tend on the whole to side with Fisher, this is not to say that Student does not have many good arguments. There is no space here to cover the full debate but I can recommend Student's posthumous article in *Biometrika* [76] as both a spirited reply to Fisher and as a testament to the keen insight and practical sense of a statistician we perhaps tend to underrate.

There are reasons to randomize in clinical trials that do not apply in agriculture (principally, the need to blind trials) and there are reasons that applied in agriculture that do not apply in medicine (the correlation structures we face in trials are usually weaker). I do not believe that we should make a fetish of randomization. I do believe, however, that claims to do better than randomization should be critically examined. I also believe, that we should make use of prognostic information where we have it. We are much better off in our ability to model the effects of this than we were 70 years ago and our ability is increasing all the time. We are also much more knowledgeable about which factors are prognostic than we were 55 years ago at the dawn of the new age of controlled clinical trials. Researchers should remember that we may often be studying new therapies but we more rarely find ourselves studying new diseases [21]. The decision to fit prognostic factors has a far more dramatic effect on the precision of our inferences than the choice of an ABC or randomization approach and one of my chief objections to the ABC approach is that trialists have tended to use the fact that they have balanced as an excuse for not fitting. This is a grave mistake.

In my opinion the specific dispute between Neyman and Fisher has been resolved in Fisher's favour. Others may not agree. Nevertheless, the interest in unit-treatment interaction that Neyman showed, is one that is finding many echoes in modern approaches to modelling data from clinical trials. We should welcome the ability that we have to examine such interactions where the structure of data sets permit; we should also guard against over-interpretation where it does not.

APPENDIX A

A.1. Barbacki and Fisher's analysis

Barbacki and Fisher [9] (B&F) re-analysed data on a uniformity trial in wheat that had been carried out in Idaho in the summer of 1927 and that had been reported by Wiebe [10] in

Table A1. Wiebe's data as reported and summarized by Barbacki and Fisher.

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)	(xii)
A	4410	4035	3865	3640	3650	3985	3490	3330	3358	3712	3487	3781
B	3950	3865	3295	2960	2925	3685	3400	3040'	2889	3195	3496	3576
B	4185	4075	3325	2860	2965	3770	3240	2735	2764	3460	3273	3442
A	3785	3515	3255	2815	2630	3295	2875	2630	2775	3040	2940	3152
A	3870	3780	3660	2980	2650	3250	2925	2915	2933	3277	3042	3363
B	3910	3690	3705	3050	2910	3630	2985	3130	2986	3040	2778	3123
B	3890	3695	3720	2990	2970	3315	2910	2985	2851	2635	2906	3081
A	4190	3970	4335	3350	3325	3870	3120	3015	3097	2909	2936	3628
A	4170	4070	4455	3610	3365	**3460**	2970	2855	2877	2834	3020	3632
B	4015	4480	4730	3805	3375	3545	3080	2810	2794	2974	2770	3805
B	4150	4755	**5065**	4125	3550	3740	3425	2690	2789	2810	2895	3695
A	4190	4740	5265	4415	3675	3965	3685	3030	2782	2904	3080	**2798**
A	4095	5075	5495	4270	3760	4010	3695	3255	2759	3118	**3287**	3547
B	3805	4360	4415	3870	3585	3785	4025	3300	3199	3407	3473	3572
B	4005	4225	3840	3800	3780	3780	4025	3710	3564	3616	3539	3853
A	3700	4325	3550	3455	3540	3660	3980	3705	3577	3759	3558	3673

The columns of the field created by averaging are labelled (i)–(xii). The rows have been labelled A and B according to the supposed treatment pattern that might apply if the Beaven split drill approach had been used.

1935. In total 1500 rows of wheat were sown. The rows were 15 ft long and 12 in apart. They can be regarded as having been arranged in 12 columns, each column consisting of 125 rows. Generally, the columns had a space between them but column 2 ran on to column 3 and column 4 ran on to column 5. This complication was ignored by B&F.

They then impose an imaginary structure on the field as follows. (1) They supposed that six adjacent rows in every column starting with row one would be given the same treatment but that one row would be sacrificed wherever the treatment was changed. (2) They supposed that a regular scheme of four sandwiches of the form ABBA ran down each column. (3) Although they did not state this explicitly, the net consequence was that they totalled yields from six rows of wheat in each of Wiebe's columns starting with rows 1, 8, 14, 21, 27, 34, 40, 47, 53, 60, 66, 73, 79, 86, 92, 99. (These are either at intervals of six or seven rows depending on whether the treatment is changed or not.) (4) They thus ignored the last 20 rows in each column of Wiebe's data, although again they did not mention this explicitly.

Table A1 gives the yields in grams as reported by B&F. Some comments are appropriate. (1) The figures given in bold are incorrectly reported by Barbacki and Fisher. Working with Wiebe's data I find values of 5060 not 5065, of 3465 not 3460, of 3288 not 3287 and of 3798 not 2798. (2) The last of these appear to be a simple misprint and is not reflected in their calculation. The others may be errors in calculation. (In what follows B&F's calculation is followed using the data as they reported them except for the value of 3798 rather than 2798.) (3) It is noticeable that the precision in the first eight columns is to the nearest 5 g but that in the last four columns is to the nearest gram [77]. This is traceable to differences in precision in Wiebe's table, and attracts comment neither from him, nor from B&F, nor subsequently from Student.

B&F then proceed to calculate as follows. (Some details are omitted) First they report the total yield for A as 339 535 g and for B is 333 660. The difference in favour of A is 5875 g

Table A2. Analysis of variance table of Barbacki and Fisher.

	Degrees of freedom	Sums of squares	Mean square
Varieties	1	719 076	719 076
Estimated error	47	4 819 203	102 356
Total	48	5 538 249	

and since the average of the two figures is 336 598 the percentage error is 1.745%. For each of the $4 \times 12 = 48$ sandwiches they calculate the difference A−B. B&F calculate the sum of uncorrected squares for the 48 differences as 5 538 279 (the correct value is 5 538 249). They point out that this is the sampling variance of the difference between the treatment totals. This follows because (a) on average over all randomizations the effect must be zero so we can pool the one degree of freedom with the other 47(b) we are talking about the variance of the total and this must have a variance equal to the sum of the individual variances.

They then calculate an Analysis of variance table as given in Table A2. (The correct value for the error sum of squares, taking the values of the data as Barbacki and Fisher used them, should be 4 819 173. The varieties sum of squares is correct.). Next they state without explanation that the standard error of the difference between total yields as estimated from the experiment is 2218.50 or 0.659% of the mean. What they have done is equivalent to having taken the mean square error and multiplied it by 48 and then taken the square root.

B&F point out that in this case the actual error of the experiment is nearly 2.5 times what it should be (1.745% as opposed to 0.699%). They also point out that the declared standard error is smaller than it would be for a randomized design (0.659% as opposed to 0.699%).

A.2. Balance, conditional type I error rates and sample size

We suppose that we have two statistics of interest, both continuous on the real line, W, the statistic of primary interest used to test the null hypothesis, and R, a statistic that describes balance as regards some covariate. We suppose that nuisance parameters are known and that the joint distribution $f(W,R)$ may be specified. We transform W and R, so that they are replaced by their one-sided P-values, $\alpha = U_1(W)$ and $\gamma = U_2(R)$. The second of these has a uniform distribution by randomization and the first under the null hypothesis. If these are monotonic transformations, then the joint distribution under the null hypothesis is a form of bivariate uniform

$$g(\alpha,\gamma) = h(U_1^{-1}(\alpha), U_2^{-1}(\gamma)) \left| \frac{\mathrm{d}U_1^{-1}(\alpha)}{\mathrm{d}\alpha} \right| \left| \frac{\mathrm{d}U_2^{-1}(\gamma)}{\mathrm{d}\lambda} \right| \tag{A1}$$

The marginal distributions are identical and may be written as:

$$h(\alpha) = 1, \quad 0 \leqslant \alpha \leqslant 1, \quad h(\gamma) = 1, \quad 0 \leqslant \gamma \leqslant 1$$

Clearly these do not depend on the sample size.

As a specific example consider a two-group clinical trial with sample sizes n_1, n_2 and let $q = (1/n_1 + 1/n_2)$. An efficacy variable $Y \sim n(\mu_\gamma, \sigma_\gamma^2)$ and a covariate $X \sim n(\mu_X, \sigma_X^2)$ are measured. We may write $W = \bar{Y}_2 - \bar{Y}_1$, $R = \bar{X}_1 - \bar{X}_2$. Under the hypothesis of no treatment effect,

the joint distribution, of W, R is

$$f(W,R) = n(0,0,q\sigma_\gamma^2, q\sigma_X^2, \rho) \tag{A2}$$

As a first step to calculating α and γ we standardize W and R. This transformation leaves the correlation between them unchanged and we thus see instantly that in this case (A.1) cannot depend on the sample size. In fact by obtaining first of all the density of a bivariate Normal for the standardise variables W^* and R^* and then performing the transformations $\alpha = 1 - \Phi(W^*)$ and $\gamma = 1 - \Phi(R^*)$, where $\Phi(.)$ is the distribution function of the standard Normal, we obtain

$$g(\alpha,\gamma) = \frac{\exp(-[\{\rho\Phi^{-1}(1-\alpha)\}^2 - 2\rho\Phi^{-1}(1-\alpha)\Phi^{-1}(1-\gamma) + \{\rho\Phi^{-1}(1-\alpha)\}^2]/\{2(1-\rho^2)\})}{(1-\rho^2)} \tag{A3}$$

This confirms the lack of dependence on q, the function of the sample sizes.

More generally, some slight dependence on sample size is possible but the fact that the margins of the distribution given by (A.1) do not depend on sample size limits this dependence.

ACKNOWLEDGEMENTS

I thank three referees for helpful comments.

REFERENCES

1. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials*: *Theory and Practice*. Wiley: New York, 2002.
2. Marubini E, Valsecchi MG. *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley: Chichester, New York, 1995.
3. Neyman J. Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society (Suppl.)* 1935; **2**:107–154.
4. Box JF, Fisher RA. *The Life of a Scientist*. Wiley: New York, 1978.
5. Reid C. *Neyman*, *from Life*. Springer: New York, 1982.
6. Pearson ES. In *Student A Statistical Biography of William Sealy Gosset* (Edited and augmented by Plackett RL with the assistance of Barnard GA). Clarendon Press: Oxford, 1990.
7. Gosset WS. Co-operation in large-scale experiments. *Journal of the Royal Statistical Society* (*Suppl.*) 1936; **III**:115–122.
8. Fisher RA. Co-operation in large-scale experiments. *Journal of the Royal Statistical Society* (*Suppl.*) 1936; **3**:122–124.
9. Barbacki S, Fisher RA. A test of the supposed precision of systematic arrangements. *Annals of Eugenics* 1936; **7**:189–193.
10. Wiebe GA. Variation and correlation in grain yield among 1500 wheat nursery plots. *Journal of Agricultural Research* 1935; **50**:331–357.
11. Senn SJ. Fisher's game with the devil. *Statistics in Medicine* 1994; **13**:217–230.
12. Atkins WRG, Fisher RA. The therapeutic use of vitamin C. *Journal of the Royal Army Medical Corps* 1944; **83**:251–252.
13. Senn SJ. *Dicing with Death*. Cambridge University Press: Cambridge, 2003.
14. Bennett JH. *Statistical Inference and Analysis Selected Correspondence of R.A. Fisher*. Oxford University Press: Oxford, 1990.

15. Holschuh N. Randomization and design: I. In *RA Fisher*: *An Appreciation*, Fienberg SE, Hinkley DV (eds). Springer: Heidelberg, 1980; 35–45.
16. Nelder JA. A reformulation of linear models. *Journal of the Royal Statistical Society A* 1977; **140**:48–77.
17. Fisher RA. Statistical Methods for Research Workers. In *Statistical Methods*, *Experimental Design and Scientific Inference*, Bennet JH (ed.). Oxford University: Oxford, 1925.
18. Cox DR. The interpretation of the effects of non-additivity in the Latin square. *Biometrika* 1958; **45**:69–72.
19. Senn SJ. *Cross-over Trials in Clinical Research*. Wiley: Chichester, 2002.
20. van Elteren PH. On the combination of independent two-sample tests of Wilcoxon. *Bulletin de l'Institut International de Statistique* 1960; **37**:351–361.
21. Senn SJ. Consensus and controversy in pharmaceutical statistics (with discussion). *The Statistician* 2000; **49**:135–176.
22. Keene ON. The log transformation is special. *Statistics in Medicine* 1995; **14**:811–819.
23. Senn SJ. Author's reply to Walter and Guyatt. *Drug Information Journal* 2003; **37**:7–10.
24. Conover WJ, Salsburg DS. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to respond to treatment. *Biometrics* 1988; **44**:189–196.
25. Dawid AP. Symmetry models and hypotheses for structured data layouts. *Journal of the Royal Statistical Society, Series B, Methodological* 1988; **50**:1–34.
26. Altman DG. Comparability of randomized groups. *Statistician* 1985; **34**:125–136.
27. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials* 1989; **10**:161S–175S.
28. Canner PL. Covariate Adjustment of Treatment Effects in Clinical-Trials. *Controlled Clinical Trials* 1991; **12**:359–366.
29. Senn SJ. Testing for baseline balance in clinical trials. *Statistics in Medicine* 1994; **13**:1715–1726.
30. Senn SJ. Base logic: baseline balance in randomized clinical trials. *Clinical Research and Regulatory Affairs* 1995; **12**:171–182.
31. Berger VW, Bears JD. When can a clinical trial be called 'randomized'? *Vaccine* 2003; **21**:468–472.
32. Senn SJ. Covariate imbalance and random allocation in clinical trials (see comments). *Statistics in Medicine* 1989; **8**:467–475.
33. Cumberland WG, Royall RM. Does simple random sampling provide adequate balance. *Journal of the Royal Statistical Society, Series B, Methodological* 1988; **50**:118–124.
34. Treasure T, MacRae KD. Minimization: the platinum standard for trials? Randomization doesn't guarantee similarity of groups; minimization does. *British Medical Journal* 1998; **317**:362–363.
35. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics* 1974; **15**:443–453.
36. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; **31**:103–115.
37. Efron B. Forcing a sequential experiment to be balanced. *Biometrika* 1971; **58**:403–417.
38. Pocock SJ. *Clinical Trials. A Practical Approach*. Wiley: Chichester, 1983.
39. Matthews JNS. *An Introduction to Randomized Clinical Trials*. Arnold: London, 2000.
40. Burman C-F. On sequential treatment allocations in clinical trials. *Ph.D. Thesis*, Chalmers University of Technology, Gothenburg, 1996.
41. Atkinson AC. Optimum biased coin designs for sequential clinical-trials with prognostic factors. *Biometrika* 1982; **69**:61–67.
42. Atkinson AC. Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in Medicine* 1999; **18**:1741–1752.
43. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991; **58**:227–240.
44. Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine* 1995; **14**:735–746.
45. Senn SJ. *Statistical Issues in Drug Development*. Wiley: Chichester, 1997.
46. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
47. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619–629.
48. Senn SJ. The many modes of meta. *Drug Information Journal* 2000; **34**:535–549.
49. Longford NT. Selection bias and treatment heterogeneity in clinical trials. *Statistics in Medicine* 1999; **18**:1467–1474.
50. Senn SJ, Auclair P. The graphical representation of clinical trials with particular reference to measurements over time. *Statistics in Medicine* 1990; **9**:1287–1302 (published erratum appears in Statistics in Medicine 1991;**10**(3):487).
51. Nelder JA. The great mixed-model muddle is alive and flourishing, alas! *Food Quality and Preference* 1998; **9**:157–159.

52. Speed FM, Hocking RR, Hackney OP. Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association* 1978; **73**:105–112.
53. Senn SJ. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* 1998; **17**:1753–1765; discussion 1799–1800.
54. Gallo P. Center-weighting issues in multicenter clinical trials. *Journal of Biopharmaceutical Statistics* 2001; **10**:145–163.
55. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *Journal of Clinical Epidemiology* 1996; **49**:395–400 (see comments).
56. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41**:361–372.
57. Senn SJ, Harrell Jr FE. On subgroups and grouping for significance. *Journal of Clinical Epidemiology* 1998; **51**:1367–1368 (letter; comment).
58. Senn SJ, Harrell F. On wisdom after the event. *Journal of Clinical Epidemiology* 1997; **50**:749–751 (see comments).
59. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. On reaching the tunnel at the end of the light. *Journal of Clinical Epidemiology* 1997; **50**:753–755 (see comments).
60. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Clinical versus statistical considerations in the design and analysis of clinical research. *Journal of Clinical Epidemiology* 1998; **51**:305–307.
61. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *British Medical Journal* 1998; **316**:690–693 (see comments).
62. Senn SJ. Applying results of randomised trials to patients. N of 1 trials are needed. *British Medical Journal* 1998; **317**:537–538 (letter; comment).
63. Senn SJ. Individual therapy: new dawn or false dawn. *Drug Information Journal* 2001; **35**:1479–1494.
64. Grieve AP. The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes? *Pharmaceutical Statistics* 2003; **2**:87–102.
65. Hutton JL. Numbers needed to treat: properties and problems. *Journal of the Royal Statistical Society A* 2000; **163**:403–419 (with comments).
66. Senn SJ. Odds ratios revisited. *Evidence-Based Medicine* 1998; **3**:71.
67. Nelder JA, Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society A* 1972; **132**:107–120.
68. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall: London, 1989.
69. Gail MH, Wiand S, Piantadosi S. Biased estimates of treatment effects in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; **71**:431–444.
70. Senn SJ. Testing for individual and population equivalence based on the proportion of similar responses. *Statistics in Medicine* 1997; **16**:1303–1306 (letter; comment).
71. Stine RA, Heyse JF. Non-parametric estimates of overlap. *Statistics in Medicine* 2001; **20**:215–236.
72. Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics* 1982; **38**:613–621.
73. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *British Medical Journal* 1995; **311**:1356–1359.
74. Lindley DV. Decision analysis and bioequivalence trials. *Statistical Science* 1998; **13**:136–141.
75. Senn SJ. Statistical issues in bioequivalance. *Statistics in Medicine* 2001; **20**:2785–2799.
76. Gossett WS. Comparison between balanced and random arrangements of field plots. *Biometrika* 1938; **29**:363–378.
77. Preece DA. Distribution of final digits in data. *The Statistician* 1981; **30**:31–60.

