To be read before The Royal Statistical Society at the Discussion meeting to be held at Broadway House, Westminster, London and online on 1st July 2026, Dr Rute Vieira in the Chair.

Randomisation in Clinical Trials: Is a Pragmatic Compromise between Randomisation- and Model-Based Inference Possible?

Stephen Senn

Consultant Statistician, Edinburgh

Address for correspondence: Stephen Senn, Consultant Statistician, 2 Merchiston Crescent, Edinburgh, EH10 5AJ, UK. Email: stephen@senns.uk

The defenders of randomisation point to the long run properties that randomisation underwrites. The critics argue that a long-run average is not relevant to the case in hand. Here I argue that both are right in a sense. Randomisation permits one to use the distribution in probability of the effects of covariates one has not seen but this distribution is not relevant for those one has. However, conditional inferences are embedded within marginal ones. Thus, although the latter should not be substituted for the former, the former are unlikely to be right if the latter are wrong. Randomisation is valuable for what we don't see and don't know. It should not be used as an excuse for ignoring what we do. I shall claim that randomisation does not solve all problems but doing better is harder than many suppose.

Keywords: marginal model, conditional model, analysis of covariance, randomisation distribution, linear model, , blinding.

1. Introduction

1.1. Scope and purpose

In this paper a view of randomisation in clinical trials will be presented from the perspective of one who has worked in the pharmaceutical industry and whose research has been mainly on problems of design and analysis in drug development. There is more to clinical trials than their application to drug development and there is more to experimental design as a whole than there is to clinical trials. Nevertheless, my view of pharmaceutical clinical trials has been much influenced by considering experimentation in other fields and I hope that what I have to say will have some relevance there also.

1.2. Marginal and conditional

To describe an effect or estimate as *marginal* can mean many different things. This can be illustrated by considering a clinical trial with sex of patient as a single potential covariate.

One might consider it possible that the treatment effect is different for men and for women. A marginal effect could then mean some sort of average effect for the sexes, either using the sample or some other proportions. A recent causal treatment of a binary outcome considers various approaches to analysis that are linear in the probability (Kuipers and Moffa 2022) and provides an example where the solution is complex, despite a simple model.

Now suppose that we take the case above but move from the probability scale to the log-odds ratio scale. It might be the case that the log-odds ratio for treatment is the same for men and for women. However, if sex is prognostic of outcome, then even if it is balanced by treatment, it will not be the case that the expected value of the treatment effect if sex is not in the model will be the same as if it were(Robinson and Jewell 1991). A marginal estimate might be the estimate not conditioning on sex or possibly one constructed by applying the conditional model to some target population(Lee and Nelder 2004).

Neither of these two cases will be considered here. Treatment by covariate interaction will be excluded by assumption and the development will be in terms of continuous outcomes. Where this is the case, if randomisation has been used, conventional estimates that do not condition on the covariate will be marginally unbiased (unbiased over all randomisations) and such an estimate will be referred to as *marginal* and those that condition on the covariates will be conditionally (and therefore also marginally) unbiased and referred to as *conditional*.

1.3. Some history

In his paper of 1926, RA Fisher set out the case for using randomisation as a crucial element of designed experiments (Fisher 1926). However, as pointed out by Nancy Hall, (2007), he was not the first to use randomisation, for example, Peirce and Jastrow had employed it in their work of 1884 (Peirce and Jastrow 1884). Nevertheless, he was the first to develop a theory of its use that allied the device to a general programme of statistical inference. An important contribution of Fisher's to this programme is his monograph of 1935, *The Design of Experiments* (Fisher 1935) and in his collected papers of 1950 (Fisher 1950), in the preface to the reprinted 1926 paper, he refers to the link between the two.

Fisher's theory on the analysis of randomised designed experiments was further expanded by many other statisticians, notably Frank Yates, his successor as head of statistics at Rothamsted, who developed, for example, its application to so-called *incomplete block* experiments (Yates 1936, Yates 1940) and by Yates's successor John Nelder, who constructed a general principled algorithmic approach to analysing a wide class of experiments satisfying what he called *general balance* (Nelder 1965a, b). (Although Nelder's work on general balance was carried out at Wellesbourne *before* he became head of statistics at Rothamsted.) Many other statisticians who worked on the theory of designed experiments worked at Rothamsted or were influenced by those who did. In addition to Fisher, Yates and Nelder, notable examples are Frank Anscombe, (1948), William Cochran, (1947) and David Finney, (1943), and more recently, Roger Payne and Graham Wilkinson(Payne and Wilkinson 1977) and Rosemary Bailey Bailey, (1983) and I think that one can reasonably refer to a *Rothamsted*

School in the design and analysis of experiments. Another important statistician who worked on randomisation and who was also at Rothamsted was Oscar Kempthorne (Kempthorne 1977) but his approach was rather different and I hesitate to include him in the school. On the other hand, one leading statistician who was never at Rothamsted but who could easily be added as regards his work on design of experiments, is David Cox, (1958). For each of these statisticians, I have cited one paper or monograph; as a set they may give a flavour of what was achieved by 1984, the 150th anniversary of the Royal Statistical Society, when David Cox gave a summary of the position on design of experiments. (Cox 1984).

However, this paper is not about randomisation in agriculture but about randomisation in clinical trials and a key figure here is not Fisher but Austin Bradford Hill (ABH), in his contribution to the design of the MRC trial of streptomycin in tuberculosis(Medical Research Council Streptomycin in Tuberculosis Trials Committee 1948). Writing of the trial over 40 years later, in 1990, ABH stated:

> I deliberately left out the words "randomization" and "random sampling numbers" at that time, because I was trying to persuade the doctors to come into controlled trials in the very simplest form and I might have scared them off. (Hill 1990, P77) .

Randomisation had been used in trials with human subjects before the MRC trial. The study by Peirce and Jastrow, (1884) already referenced is an example. A valuable summary of various early efforts to ensure fair comparisons of treatments in clinical trials is given in a book chapter by Iain Chalmers, the title of which declares, “Statistical Theory Was Not the Reason That Randomization Was Used in the British Medical Research Council's Clinical Trials of Streptomycin for Pulmonary Tuberculosis”. (Chalmers 2005, P309)

In discussing randomisation in clinical trials, one cannot escape its history in agriculture. Many issues are similar but there may be differences. Agricultural trials frequently have complicated treatment and block structures. This can make calculation of standard errors difficult but it can also make efficient estimation far from simple. Incomplete block designs are an example: an efficient estimate will combine information not only within but also between blocks. Drug trials will be rich in covariate information but might have a much simpler treatment structure (Cox 1984). Of course, there are exceptions and The Lanarkshire Milk Experiment (Leighton and McKinlay 1930, Senn 2023) provides an early example of a trial in humans with complicated treatment and block structure. For those who are interested in knowing more about the history of randomisation in clinical trials, the two part account by Robert Matthews is highly recommended (Matthews 2025a, b).

1.4. Design and analysis

A feature of the Rothamsted School is the deep relationship between design and analysis. This relationship can express itself in many ways, for example how a suitable parametric model can closely match what a test using the possible treatment permutations given the randomisation scheme would show, or perhaps in certain symmetry relationships in approaches to analysing data structures (Dawid 1988, Speed 1987).

Fisher’s work was in the context of field experiments in agriculture and he formulated a key requirement for valid error estimation as follows:

> The *estimate* of error afforded by the replicated trial depends on differences between plots treated alike. An estimate of error so derived will only be valid for its purpose if we make sure that, in the plot arrangements, pairs of plots treated alike are not nearer together, or further apart than, or in any other way, distinguishable from pairs of plots treated differently(Fisher 1926, P506).

The reference to pairs of plots can be explained by considering an alternative, in terms of all possible pairwise differences, to the usual formula for a variance. *Variance* is a measure and term that Fisher himself had proposed eight years earlier (Fisher 1918, P399) in a paper on Mendelian genetics. Given a sample of N observations $Y_1 \ldots Y_N$ and that the difference between two pairs of observations $Y_r - Y_s$ is d_{rs}, their corrected sums of squares, SS_T, can be written

$$SS_T = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{\sum_{r=1}^{N}\sum_{s=1}^{N}(d_{rs})^2}{2N}. \quad (1)$$

Here the middle term is the familiar analysis of variance form introduced by Fisher in *Statistical Methods for Research Workers (Fisher 1925)* and the right hand term is an expression in terms of the N^2 paired differences, including the N self-paired differences that must be zero. (There may, of course, be other differences that are zero, depending on the data set.) Note that in this formula, $d_{sr} = -d_{rs}$, $r \neq s$ and that, since squaring makes the sign irrelevant, each non-self pair contributes two identical squared values.

Fisher frequently dealt with experiments, common in agriculture, in which many treatments were being compared but I shall consider the most common case in clinical trials, that of comparing two treatments and welcome the simplification in discussion that this brings.

Suppose we have a parallel group trial with $N = 2n$ patients. Now partition the observations into two groups of equal size n. Define an indicator function I_{rs} such that $I_{rs} = 1$ if Y_r, Y_s are observations in the same treatment group but is otherwise equal to 0. We can then express the within sum of squares as

$$SS_W = \sum_{i=1}^{2}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{i.}\right)^2 = \frac{\sum_{r=1}^{N}\sum_{s=1}^{N} I_{rs}\left(d_{rs}\right)^2}{2n}. \quad (2)$$

Here $i = 1, 2$ is an indicator for group and $j = 1 \cdots n$ identified the patient in a given group. The RHS of expression(2) thus represents the within sum of squares in terms of paired differences between patients in the same group.

Since we have that SS_T reflects all pairwise differences, whether or not observations are in the same group, but SS_W only reflects pairwise differences for the same group, we have, from the familiar analysis of variance decomposition, $SS_T = SS_B + SS_W$ that the sum of squares between, SS_B must have a contribution from paired differences involving different groups. So, to adapt Fisher's requirement we have, "An estimate of error so derived will only be valid for its purpose if we make sure that, in the design, pairs of patients treated alike are not in any way, distinguishable from pairs of patients treated differently."

1.5. Putting my (shuffled) cards on the table

Fisher's espousal and promotion of randomisation has divided statisticians. Proponents tend to stress the support it brings to the validity of marginal inferences. Critics have pointed to the absurdity of regarding the long run properties of a statistical approach to analysis as an excuse for not producing an inference that applies to the case in hand. It is not hard to find eminent proponents of either extreme. Neyman, who was a bête noir of Fisher's, has this to say, "without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher and I consider it as one of the most valuable of Fisher's achievements." (Reid 1982, P45) On the other hand Basu, who was an admirer of Fisher's work, states, "we find no satisfactory answer to the question why randomize" (Basu 1980, P575). An example of this division in medical statistics is given by minimisation. This has been widely used by certain public organisations such the European Organisation for Research and Treatment of Cancer (Buyse and McEntegart 2004) but is regarded by many as not being a true randomisation method and has been criticised in drug regulatory circles (Day *et al.* 2005).

In this paper it will be argued that both the defenders and critics of randomisation are right in a sense and both are wrong in another. The claim is that the long run average property of randomisation cannot be used as an

excuse to ignore the particular conditional case at hand, whether or not a predictive factor that has been observed is balanced. However, a further claim is, that despite this, the long run property is valuable. The plan of the rest of this paper is as follows. In section 2, various general matters considering clinical trial questions and models and approaches to answering them are covered. In section 3, side issues will be considered. These are often mentioned when randomisation is discussed, although not essential and will be cleared out of the way. Section 4 uses a simple game of dice to illustrate how inferences should change when information is acquired, a matter that is central to the effect of observing covariates in randomised trials. Section 5 proposes a pragmatic compromise between viewing the long run property of randomised trials as being essential and the view that for the analysis of a given trial this is irrelevant. The last section offers some conclusions, of necessity tentative, but which, therefore, are suitable for debate.

What are not covered in any detail are technical matters to do with randomisation in clinical trials: its various forms and implementations. The reader who is interested in these and a deeper theory of randomisation in clinical trials should consult the classic text by Rosenberger and Lachin, (2016).

2. Statistical Frameworks

2.1. Clinical trial questions

There are at least five different types of question a clinical trial might address

"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)?"(Senn 2004)(p3738).

Note that it is a feature of these questions that if you can answer a later one you can generally answer an earlier one but not *vice versa*. For example, if you can establish what the effect was for different subgroups (Q4) and these are not all identical, then you can answer "no" to Q3. I take it as given that the answers to these questions involve probabilities, so, for example, to answer Q2 requires not only provision of an estimate but an estimate of its uncertainty.

Statisticians may fail to make it clear to clients that the statistical calculations they provide cannot provide an answer to Q5 except by making strong assumptions that are plausibly false. In the discussion here only Q1 and Q2 are addressed. The justification is that getting the answers to these questions right is the first step in understanding the effects of treatment and doing so is not easy. It is clear that the uncertainty that attaches to answering Q1 and Q2 is a gross underestimate of that attendant on any answers to Q5.

2.2. Approaches to modelling data accrual

In a clinical trial we have little control over the *presenting process*, which is to say all the many unknown factors that lead to patients being offered and accepting the chance to enter the clinical trials we run. On the other hand we have precise control of the *allocation algorithm,* that is to say how we decide which patients get which treatment(Senn 2004). Representative sampling does not take place(Magirr *et al.* 2025).

The fact that some working on clinical trials think that this is the case is only explicable in terms of a complete ignorance of sampling methodology. Inclusion criteria define a multivariate 'space' that cannot be exceeded by the sample of patients studied but there is no guarantee that the sample will fill this space in a representative way.

Fisher himself did make use of a sampling analogy but the 'population' is some ideal population that the sample obtained could be regarded as being taken from. This conceptual device has been much criticised but whether or not it is reasonable, it is not the same as an actual physical population which would, in any case, be very difficult to define. The logic of clinical trials is comparative and randomisation is designed to promote reasonable answers being given to Q1 and Q2.

Nevertheless, there is a relevant connection to sampling theory. The debate as regards the relative merits of design versus model based inference that one finds in design of experiments is mirrored by a corresponding debate in the world of sampling. Dev Basu, VP Godambe, (1955) and Richard Royall, (1970) have been important contributors to this debate. See Little, (2004) for a very insightful review.

2.3. Approaches to modelling effects

In what follows, unless stated to the contrary, it will be assumed that a single continuous outcome is being modelled. At a later point what changes, if this is not the case will be considered.

Pharmaceutical statisticians make frequent use, often implicitly, of three frameworks when analysing clinical trials. In discussing these frameworks, it will be assumed that a parallel group trial is involved.

The most common framework is that of the linear model. In its simplest form this has treatment as a main effect and a single random error term. Main effects of centres are often included and usually as fixed effects. If covariates are added, they are added as fixed main effects. The primary analysis will usually not include treatment-by-covariate interactions and regulatory advice to the pharmaceutical industry encourages the examination of such as a secondary matter and discourages putting it in the primary analysis(International Conference on Harmonisation 1999). Randomness is assumed to arise via notional sampling from a fictional hyper-population. This will be referred to as the LM approach.

Adding covariates to a model explains a source of variation that is otherwise treated as random. This would imply that any model that does not include the 'correct' covariates is incorrectly specified. However, the whole point of randomisation is to be able to ignore factors that might affect the outcome but have not been seen. This suggests, instead, treating outcomes and covariates in a truly multivariate model, where the effect of treatment can only affect the former and not the latter and the covariates are conditioned on to isolate the effect of treatment. This will be referred to as the MV approach. See Aldrich, (2005) for a discussion as to how Fisher developed the linear model from the then dominant multivariate approach.

A third approach is to regard the outcomes as fixed and that random variation is induced by the act of random allocation itself. By permuting the treatment labels in a way that reflects the randomisation one can obtain an empirical randomisation distribution (RD) for the treatment estimate under the null hypothesis and hence a significance test(Zhang and Zhao 2023). For a development of this approach for multi-centre trials see Zheng and Zelen, (2008). This will be referred to as the RD approach.

The RD approach tends to be rather restrictive. However, for simple analyses, and as regards calibrating expectations as to what is reasonable, it is useful. Consider, for example, the case of a small trial with just ten patients randomised equally into the two groups. There are just $10!/(5!5!) = 252$ possible allocations. One does not have to be wedded to the randomisation distribution framework of analysis to see that any parametric two-sided P-value that is less than $2/252 \approx 0.8\%$ is suspect. This is particularly so if blinding is regarded as essential(Senn 1994). The LM approach is commonly used for analysis but the MV approach appears to provide a justification for allowing covariates that have been omitted to contribute to the random variation in response, given that the model is unlikely to be perfectly specified.

2.4. Principles

One can think of the analysis of clinical trials in terms of five principles.

1. A statistical inference is an uncertainty statement (for example in terms of likelihood, confidence or a fiducial or posterior interval).
2. Clinical trials are experiments and their goal is to make valid and efficient causal inferences.
3. Conditional inference trumps marginal inference.
4. If an inference is invalid, the conditional inferences to which it is marginal are plausibly also invalid.
5. Experimentation is a public exercise in which trust is important.

The first of these means that concentrating on point estimates is not enough. A good explanation as to why is given by Cox and Hinkley, (1974) (p250). The second justifies prioritising answering Q1 and Q2 above. The third

means that randomisation cannot be used as an excuse for ignoring relevant information. The fourth suggests that despite conditional inferences being the goal, it is still worth considering the conditions under which marginal inferences are valid. The fifth is a reminder that the data that are collected in a clinical trial may be used by many parties who may not necessarily share the beliefs or background knowledge of the trial's designers.

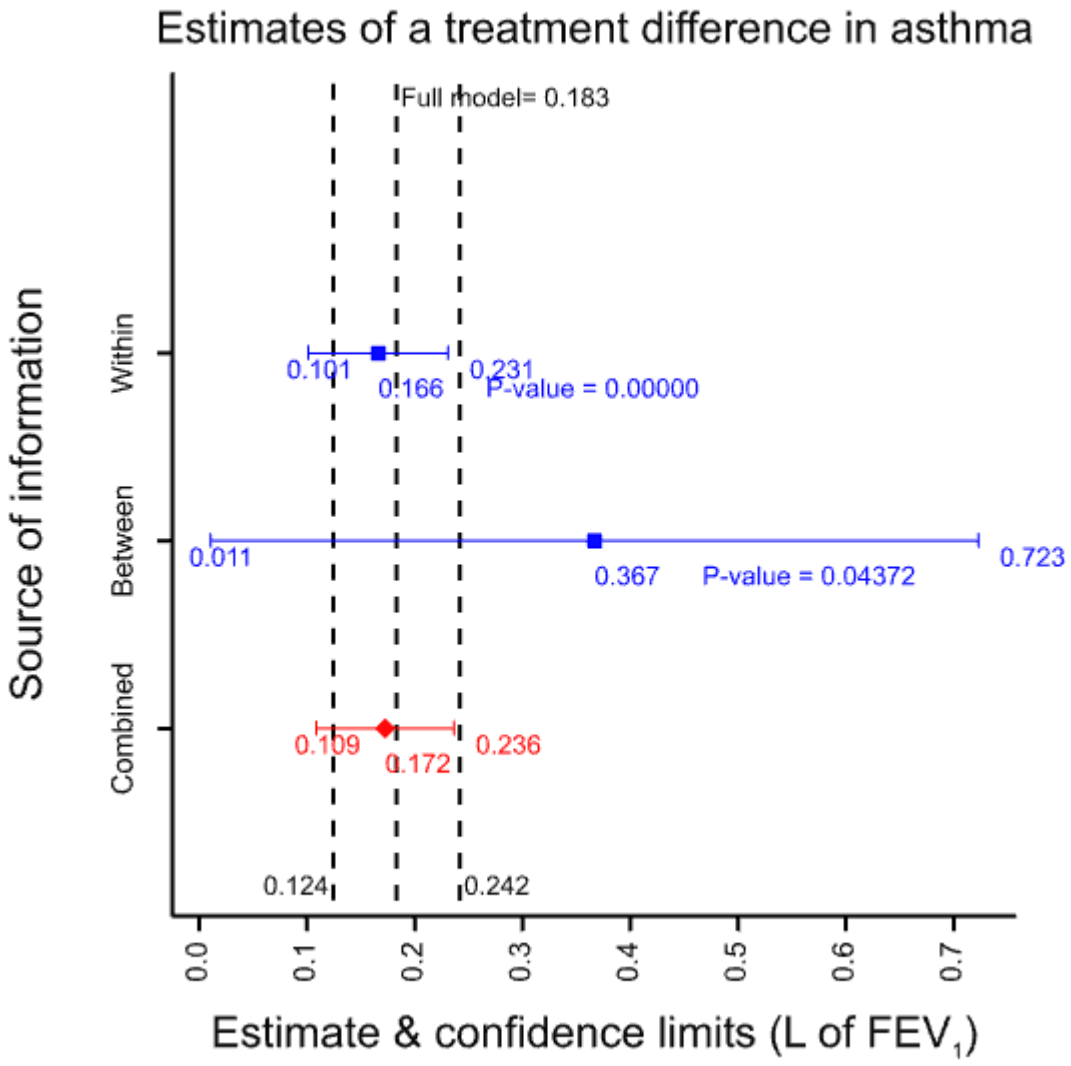

Figure 1 Various 95% confidence intervals from an incomplete blocks cross-over. The outcome is litres of forced expiratory volume in one second (FEV_1) measured 12 hours after treatment.

The first of these is worth illustrating, since failure to understand this point is the source of much confusion (Krauss 2018). Figure 1 shows various analyses of a parallel assay in asthma(Senn 2022, Senn *et al.* 1997) designed to show that two formulations were equipotent. It failed to do this. In order to have high precision, it was desirable to have a within-patient study but seven treatments were compared (three doses of a new formulation, three doses of a reference formulation and placebo)and only five periods could be used. As a solution an incomplete blocks cross-over with twenty-one sequences was used and any given patient only received five of the seven treatments. Here only two of the treatments are compared: the highest dose of the existing formulation, ISF24 and the lowest dose of the new formulation, MTA6. The figure shows point

estimates and 95 % confidence intervals for various analyses. The measurement is litres of forced expiratory volume in one second measured 12 hours after treatment.

The analysis labelled *Between* compares results for 37 patients given ISF24 but not MTA6 and 37 patients given MTA6 but not ISF24. The analysis labelled *Within* uses the results on 71 patients given both treatments on different occasions. These are two statistically independent sources of information from the same trial. The analysis labelled *Combined* is a weighted combination of the two. Also shown as a vertical dashed line is what the full mixed model analysis, using all data from all treatments, would show. This will have a little additional indirect information of the double contrast form but is scarcely different from *Combined* which is largely dominated by *Within* anyway.

Any critic of randomisation who thinks in terms of its ability to balance factors can only regard the *Between* information as being sadly deficient compared to *Within*. The latter balances for tens of thousands of genes and all life history until the start of the trial and the former for none of these. Sure enough, the between-patient point estimate is considerably different from that given by the other methods and plausibly this is because some influential factors are unbalanced. However, this just illustrates the fallacy of regarding a point estimate alone as constituting a valid inference.

A common criticism of randomisation is that because it cannot possibly balance for all hidden factors (Borgerson 2009) it cannot help with dealing with them. Consider, for example, this criticism of Worrall's, "…given that there are indefinitely many possible confounding factors, then it would seem to follow that the probability that there is some factor on which the two groups are unbalanced (when remember randomly constructed) might for all anyone knows be high." (Worrall 2002, P s324). There are three errors in this. First, given that it is the joint effect of all factors that matters, it concentrates on the 'indefinitely many' rather than their sum. But it is a trivial mathematical fact that the sum of an infinite series can be bounded (Senn 2013a) and it is a practical one that the combined effect of all factors has a finite variance that can be estimated. Second, it fails to note that these factors affect variation not only *between* but also *within* groups and that the basis of the Fisherian analysis is to establish a valid probability statement that compares both. Third, it does not address that the statistical inference includes a statement of uncertainty. This is illustrated by Figure 1. The between-patient analysis is affected by the myriad factors that the within-patient analysis eliminates. Of course, it is also based on fewer observations, there being $37+37=74$ as opposed to $2\times 71=142$ but this is only part of the explanation. For both these reasons, but in particular the first, the between-patient confidence interval is much wider and the difference between the two inferences is that one is more precise than the other, not that it is more valid. In other words, *if we knew all potential confounding factors were balanced in a randomised clinical trial, the conventional analysis would be wrong*.

Further development of this argument will be left to sections 4 & 5 . In the next section some side issues will be considered.

3. Side issues

In this section various issues are considered that are important to randomised clinical trials but either not unique to them or not a universal feature of them or not central to the justification of randomising patients to treatment.

3.1. Blinding

Not all randomised clinical trials are double-blind but no clinical trials that are not randomised are credibly double blind. The degree to which any sequence may be guessed is governed by the randomisation employed. In my view it is best to be explicit as regards the scheme used. Fisher's famous tea-tasting experiment is a case in point. He used a scheme in which each of the $8!/(4!4!)=70$ sequences of milk in first and tea in first is equally likely and insists that the subject must be "told in advance of what the test will consist"(Fisher 1990, P11). Based on correspondence with ABH and subsequent correspondence with his son David Hill, reported by Iain Chalmers, (2005) (pp320-321), it seems that ABH thought this was a mistake of Fisher's who should not

have shared these details with the subject. However, this raises a problem: suppose that the woman reasons the same way as the experimenter. She will then guess that there will be four cups of each kind, in which case her probability of guessing the sequence correctly is indeed 1 in 70. By not telling her, the experimenter does not know whether 1 in 70 or some lower probability is appropriate(Senn 1994).

Similarly, suppose fortuitously that the sequence employed, as a result of the randomisation, has the types of teacup arranged in pairs. Of the 70 possible sequences, 16 have this property. Having noticed this, the experimenter must worry that the subject wrongly assumed that this would have to be the case. Given the wrong assumption the subject has a 1 in 16 chance of guessing correctly. Of course, unconditionally, employing this wrong assumption, the probability is $\left(16/70\times 1/16\right)+\left(54/70\times 0\right)=1/70$. However, by telling the subject what will happen, Fisher does not have to use this marginal probability over all randomisations. He can use a conditional probability of $1/70$.

What Fisher is doing is eschewing any attempts at further complication that rely on an *argument from the stupidity of others*. This is an appropriate ethical and scientific attitude to have: open protocol but hidden random allocation. If so, the case for randomisation as a device for helping the blinding of clinical trials is obvious: it works via mutually agreed deception.

3.2. Randomisation is not possible for all factors

Howson and Urbach's exposition of the Bayesian approach to scientific reasoning(Howson and Urbach 1989) makes some rather confusing criticisms of randomisation in clinical trials. They suggest, for example, that patients could be divided into two similar groups (p151& 253). However, for reasons explained in section 3.4 this is usually impossible. Nevertheless, they make the valid point that we do not randomise everything (pp150-151).

A common example can be given. For clinical trials in which two active treatments are compared, blinding is often carried out using the double dummy approach. In a parallel group trial comparing treatments A & B, patients will be randomised either to receive A and placebo to B or B and placebo to A. We do not commonly randomise the order in which patients are instructed to take these. We implicitly assume it does not matter and where pills are involved, they may, of course, be taken together. However, I once helped design a clinical trial in asthma comparing formoterol to salbutamol, both delivered by inhaler, as regards onset of action. Inhalers cannot be used simultaneously; they have to be used sequentially even if the gap between them is short. Onset of action can be detected by patients within one minute of taking the medication. To address this, it was necessary to use four inhalers: formoterol active (F_A), formoterol placebo, (F_P), salbutamol active (S_A) salbutamol placebo (S_P). Suppose that on a given occasion a patient is given formoterol they will either take F_A followed by S_P within a few seconds, which may be written F_A S_P or the reverse order of administration S_P F_A. Thus, irrespective of any other design considerations, for example whether the trial is run as a parallel group trial, or, as was actually the case here, a cross-over trial, we need four different treatment sets as illustrated in Table 1.

Set	Treatment	First Treatment	Brief Interval	Second Treatment
$F_A S_P$	**Formoterol**	Formoterol Active		Salbutamol Placebo
$S_P F_A$		Salbutamol Placebo		Formoterol Active
$F_P S_A$	**Salbutamol**	Formoterol Placebo	A few seconds	Salbutamol Active
$S_A F_P$		Salbutamol Active		Formoterol Placebo

Table 1 Treatment sets for a double-blind trial in asthma in which the double dummy technique is used and the order of active and placebo treatment is balanced.

Note that patients cannot distinguish between (say) $F_A S_P$ or $F_P S_A$ in appearance, although the treatments are not identical**,** but they can distinguish between (say) $F_A S_P$ or $S_P F_A$, although the treatments are the same. Unfortunately, due to a communication failure, although treatments were randomised, order within treatments was not. However, that both should be randomised was our intention. Most clinical trials do not address this issue.

If, however, we consider factors as being divisible into three sorts: negligible, important but identifiable, important and unidentifiable, then the first sort can be ignored, the third sort can be blocked (if possible) and fitted, whether blocked or not and randomisation can be used to deal with the second sort.

Here *patient* is such a potentially important factor that it must be treated as random in a parallel group trial, or no reasonable analysis is possible and since it must be treated as random, there seems to be no harm in allocating at random.

3.3. Modelling

In my opinion, randomisation is valuable but cannot be used as an excuse for ignoring observed prognostic factors. A pre-specified model is a principled way to incorporate further valuable covariate information. Since more information cannot be worse than less, it would seem that nothing ought to be lost by fitting a more elaborate model. Unfortunately, this is not quite true, at least if ordinary least squares are used. The variance of the estimated treatment effect $\hat{\theta}$ for a model fitting k prognostic covariates in a two-armed clinical trials with N patients in total is

$$V\left(\hat{\theta}\right) = q_k \frac{4}{N}\sigma_k^2,$$

where σ_k^2 is the unexplained variance of the outcome given the k fitted covariates and $q_k \geq 1$ is a reflection of the degree of imbalance. For example, even if no covariates are fitted and $k = 0, q_0$ will be greater than 1 unless the number of subjects on each arm is the same. There are three consequences of fitting an extra prognostic covariate as a main effect when estimating a treatment effect(Senn 2000, Siegfried *et al.* 2023):

1. If the covariate has additional prognostic value, $\sigma_{k+1}^2 < \sigma_k^2$.
2. To the extent that the new covariate is imbalanced then $q_{k+1} > q_k$.
3. The degrees of freedom for estimating the error variance are reduced.

For a given covariate with partial correlation ρ (given other covariates in the model) with the outcome variable, the first consequence of fitting the covariate is that the expected mean square error becomes $\left(1-\rho^2\right)\sigma_0^2$, where σ_0^2 is the expected value for the unadjusted mean square error. To judge the second consequence, we may use that

$$E[q_k] = (\nu_0 - 1)/(\nu_0 - k - 1)$$
$$E[q_{k+1}] = (\nu_0 - 1)/(\nu_0 - k - 2)$$
$$E[q_{k+1}]/E[q_k] = \frac{\nu_0 - k - 1}{\nu_0 - k - 2}, \nu_0 \geq k+3$$

where ν_0 is the number of degrees of freedom for the model without any covariates(Cox and McCullagh 1982, Senn *et al.* 2024). The combined effect of the first consequences is multiplicative. The effect of the third , which might be called second-order precision may be considered in terms of the variance of the t-distribution. This is $\nu/(\nu-2)$, where ν is the degrees of freedom. Not fitting a covariate, we have $\nu = N - 2$. Fitting k covariates, we have $\nu = N - k - 2$. However, combining the effect of this with the other two is controversial and was the subject of a rather heated correspondence between Fisher and Nelder in November and December of 1958 (Bennett 1990)(pp280-283). See also (Gilmour and Trinca 2012).

Since covariates may be regarded as stochastic under the MV approach of section 2.3, the Gauss-Markov theorem does not apply unless one is interested in conditional unbiasedness. Note, however, that the situation here is somewhat different from the one considered by (Popper Shaffer 1991) who finds that even when predictor variables are stochastic, the Gauss-Markov theorem does apply in the MV case. Here, the covariate regression parameters are only of interest to the extent that they help in estimating the treatment effect. One might do better by not fitting the covariate if it is only weakly predictive(Cox and McCullagh 1982). There is an analogy here with incomplete block designs. An efficient analysis will recover between-block information by treating the blocks as random rather than fixed.

If the block effects are small enough, one could do better by ignoring them altogether rather than treating them as fixed. For example, suppose we have centres of size three in a two-armed trial and consider two centres one with two treatments and one placebo patient and the other with one treated patient and two placebo patients. Let us denote the design by T,T,P/P,P,T, with the order within centres being immaterial. Estimating the T-P contrast, ignoring the blocks , the weights are $\frac{1}{3},\frac{1}{3},-\frac{1}{3}/-\frac{1}{3},-\frac{1}{3},\frac{1}{3}$. From this we calculate the sum of the squared weights as $q_0 = \frac{2}{3}$. On the other hand, eliminating fixed centre effects, the weights for the linear combination are $\frac{1}{4},\frac{1}{4},-\frac{1}{2}/-\frac{1}{4},-\frac{1}{4},\frac{1}{2}$ and the sum of their squares is $q_1 = \frac{3}{4}$. The difference between the two is thus $q_1 - q_0 = \frac{1}{12}$ Of course, the two residual variance terms are not the same and we have $\sigma_1^2 \leq \sigma_0^2$. Nevertheless, it is possible that $q_0\sigma_0^2 < q_1\sigma_1^2$.

Trying to do better than either faces the practical problem that recovering information is only possible by having enough blocks to generate the necessary degrees of freedom for estimating the variance components and by regarding them as exchangeable. Exactly how to deal with the fact that a nuisance parameter (the ratio of between block to within block variances) is not known is a long- running issue in fitting REML models(Kenward and Roger 1997, Patterson and Thompson 1971, Senn 2015). It is not obvious to me, when fitting covariates, how to deal with this. In short, there must be something better than ordinary least squares, when adjusting for covariates.

However, this can be regarded as a difficulty of estimation that has nothing to with randomisation *per se*. It is a practical matter that should be considered when choosing models.

3.4. Sequential recruitment

A number of critics of randomisation have implied that it is common practice in clinical trials to examine random allocations and reject any that are unsatisfactory in terms of covariate balance. These critics have not just included philosophers of science(Howson and Urbach 1989, Worrall 2002, 2007) but also statisticians(Lindley 1982). This, of course, doesn't happen, being usually impossible, since for most clinical trials, patients cannot be simultaneously allocated, so that, for example, we might have a randomisation list in

which it was dictated (say) that the 89th patient would be given treatment B but we would have no idea who this patient would be and hence know nothing of their covariate measurements.

This argument is a 'litmus test' as to whether suggestions for improvement should be taken seriously. Nevertheless, it is not essential to any criticism of randomisation and in any case (projected) time of recruitment is one covariate that could be noticed ahead of running the trial . For example, we could in principle notice that the randomisation had an unfortunate temporal structure, say most patients under the new treatment to be recruited early and most under the comparator late, and it is partly to deal with this that the method of permuted blocks is used(Pocock 1983, pp77-79) (Rosenberger and Lachin 2016, pp46-47). Such temporal trends are commonly ignored but this is not necessarily wise(Altman and Royston 1988).

For balancing covariates dynamically, techniques such as minimisation(Pocock and Simon 1975, Taves 1974) or Atkinson's algorithm(Atkinson 1982, Senn *et al.* 2010) can be used. However, in the pragmatic compromise proposed here, all such 'balance' issues can be seen to contribute nothing to improving validity. They may improve efficiency but compared to randomisation, the improvements may be expected to be very small(Burman 1996, Senn, et al. 2024) and it is possible that in doing so they harm validity.

3.5. Beyond the general linear model

The argument that is made here works best for the linear model. Other generalised linear models raise further issues. The Normal distribution is a two-parameter model but many other models that statisticians use in regression are based on single parameter probability distributions. Such models are often not robust. A good example is Poisson regression. A given Poisson regression model implies that two patients with identical covariates could only differ due to the bed-rock variability of the Poisson distribution itself. This in turn suggests that there are no unincluded covariates and therefore that the model is perfectly specified, which no medical statistician would believe. In practice, over-dispersion is to be expected(McCullagh and Nelder 1989, section 6.2.3). Medical statisticians tend to deal with over-dispersion on count data, such as, for example, exacerbations in respiratory diseases, either by re-scaling the standard error by using the residual deviance(Liu and Menjoge 2008) or by using a two-parameter distribution such as the negative binomial(Hilbe 2007, Keene *et al.* 2007). Another possibility is the Poisson lognormal.

A further issue is that since single parameter models cannot sweep up the effect of unobserved and unfitted covariates in a variance term, there will be an effect on parameter estimation. In a linear (Normal) model the average estimate over all randomisation of a treatment effect will be the same whether or not a prognostic covariate is fitted but this is not so for other models(Beach and Meier 1989, Ford *et al.* 1995, Gail *et al.* 1984, Robinson and Jewell 1991). However, such apparent incompatibilities can usually be resolved by looking at prediction space(Lane and Nelder 1982, Lee and Nelder 2004). In any case, despite the fact that standard errors will not reduce by fitting predictors, but may be expected to increase, fitting prognostic covariates does not adversely effect the power(Harrell 2015, Hernández *et al.* 2004, Steyerberg *et al.* 2000). The reason is that when there is a treatment effect, the estimate of it ignoring a prognostic covariate may be expected to be lower than the estimate fitting it and thus the ratio of estimate to standard error can be expected to be larger when fitting than when not fitting the covariate.

4. A game with two dice and its lessons for clinical trials

In this section an elaboration of a game described elsewhere(Senn 2013b) is used to illustrate conditioning and the implications for the analysis of clinical trials.

4.1. The game

It is supposed that a statistician is involved in a game in which two fair dice, a red die and a black die, are rolled once each, the total score being noted. The object is for the statistician to make a reasonable call as to what the probability is of obtaining a total score of 10. There is another party to the exercise, the gambler, who can place bets using the prices determined by the statistician's call. The whole is overseen by a banker.

There are four variants of the game.

Variant 1: both dice are rolled together and the probability must be called before they are rolled.

Variant 2. It is agreed that the dice will be rolled sequentially: first the red die and then the black die. The statistician and the gambler will be shown the result of rolling the red die before the black die is rolled and must then call the probability.

Variant 3. It is agreed that the dice will be rolled sequentially: first the red die and then the black die. Although the banker will see the result of rolling the die, neither the gambler nor the statistician will be shown the result. However, once the red die is rolled and the fact has been announced, the statistician must call the probability.

Variant 4. The red die will be rolled first. It is not established beforehand whether or not the statistician will be shown the result. Whether or not the statistician is shown the result, they must call the probability after it has been announced that the roll has taken place.

Note that there are thirty-six possible combinations of scores for the two dice and three (4&6, 5&5, 6&4) produce a score of ten.

For variant 1 the statistician can argue that each combination is equally likely and three out of thirty-six give a score of ten, so the probability is 1/12. For variant 2, if the statistician sees a score of 1,2 or 3 for the red die, they must recognise that a score of ten is now impossible so the probability is 0. If, however, the red die score is 4,5 or 6, then the probability of getting a black die score that will produce a total of ten is 1/6. Variant 3 is obviously the same as variant 1 but is included because I have encountered arguments that hidden confounders may cause problems in randomised clinical trials or even random changes once the trial has started. If the statistician feels the need to speculate on what the unseen red die score is, they can argue that there is half a chance it will make it impossible to achieve a score of ten and conversely there is half a chance that the probability will be 1/6. Thus, they have $\left(\frac{1}{2}\times 0\right)+\left(\frac{1}{2}\times\frac{1}{6}\right)=\frac{1}{12}$, which is the same end result as in variant 1. For variant 4, it is not clear how the statistician should argue. Suppose that the die has not been shown, then perhaps the situation is like variant 3, so the probability is 1/12. However, suppose that the statistician would be shown the red die if the score were 4,5 or 6. Then the fact that the die has not been shown is indicative that the red die score is 1,2 or 3, so the probability of a total of ten is actually 0.

4.2. Lessons for clinical trials

We can regard the red die here as being analogous to covariate information that may or may not be available at the beginning of a trial with the black die providing the further outcome information. Variants 1 and 3 of the game tell us that we can use distributions in probability when actual values have not been observed but variant 2 tells us that the actual value trumps the distribution in probability if and when it has been observed.

The lesson for the analysis of clinical trials is that inferences should be made conditional on prognostic covariates if and when they are observed. As regards this point, I agree with the views of Basu and Lindley(Basu 1980, Lindley 1982). The inference using the distribution over all randomisations is not directly relevant and certainly cannot be substituted for the conditional inference once a prognostic covariate has been observed. Where I disagree is in arguing that *therefore* randomisation has no value. The various conditional inferences are probabilistically embedded within the marginal one. Variant 4 of the game shows us that we cannot necessarily condition on what we have observed. Randomisation is valuable for what we have not observed.

Note that it is observed covariates we have to deal with in randomised clinical trials and the fact that they are inevitably measured with some error and hence that the regression of outcome on covariate is subject to *regression dilution bias*(Hutcheon *et al.* 2010) is not a problem. There are two arguments that justify this. The first is a simple and obvious intuitive one that I prefer: since one is dealing with an imbalance with an observed covariate, it is the regression of the outcome on the observed covariate that has to be used, not the regression on the true covariate, which is unobserved.

The second argument goes like this. Suppose that we have

$$x = X + \varepsilon$$

Where X is a 'true' covariate, ε is a random error term and x is an observed covariate and Y is the outcome value. The variance-covariance matrix of these four random variables, in order, X, ε, x, Y will be given by

$$\begin{array}{c} \\ X \\ \varepsilon \\ x \\ Y \end{array} \begin{array}{c} \begin{array}{cccc} X & \varepsilon & x & Y \end{array} \\ \begin{pmatrix} \sigma_X^2 & 0 & \sigma_X^2 & \sigma_{XY} \\ 0 & \sigma_\varepsilon^2 & \sigma_\varepsilon^2 & 0 \\ \sigma_X^2 & \sigma_\varepsilon^2 & \sigma_X^2 + \sigma_\varepsilon^2 & \sigma_{XY} \\ \sigma_{XY} & 0 & \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \end{array}. \tag{3}$$

From this we see that the regressions are

$$\beta_{YX} = \frac{\sigma_{XY}}{\sigma_X^2}, \quad \beta_{Yx} = \frac{\sigma_{XY}}{\sigma_X^2 + \sigma_\varepsilon^2}, \quad \beta_{xX} = \frac{\sigma_X^2}{\sigma_X^2} = 1, \quad \beta_{Xx} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\varepsilon^2}, \tag{4}$$

where the first subscript term indicates the predicted variable and the second one the predictor. A conceptual mistake here is to assume that because in (4) $\beta_{Yx} < \beta_{YX}$ (which is the regression dilution), adjusting for an observed imbalance baseline in the covariate using β_{Yx} will under-correct for the imbalance. However, if one is tempted to use some estimate of the undiluted regression, β_{YX}, this cannot be used with the observed imbalance, it has to be used with the true imbalance, which is unknown and must therefore be estimated. However, expression (4) shows that although the regression of observed on true is 1, the regression of true on observed is less than 1, being given by the fourth term in (4) and not the third. Putting the two regressions together (outcome on true and true on observed) we would get

$$\frac{\sigma_{XY}}{\sigma_X^2} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\varepsilon^2} = \frac{\sigma_{XY}}{\sigma_X^2 + \sigma_\varepsilon^2} = \beta_{Yx}. \tag{5}$$

In other words, we end up with the intuitive and obvious solution of adjusting for what you have seen(Senn 1995). The conceptual error involved can be described as follows. It is assumed that although the true individual covariate values cannot be observed, the true mean difference between groups can.

What this implies is that, although you cannot use randomisation as an excuse for ignoring what you have measured, you are entitled to adjust for what you have observed and not worry about 'true' values. The above argument easily extends to unobserved covariates. However, if you have not randomised you will have to use some other argument to justify that the regression terms cancel out in this way. In observational studies it is a worry and not dealt with so easily and residual confounding is a serious problem(Fewell *et al.* 2007).

5. Proposal for a pragmatic compromise

The proposed approach to thinking about randomisation involves accepting the following points.

1. There are two general purposes in designing and analysing clinical trials and both are crucial:
    a. To provide an estimate of the effect of a treatment that is as precise as possible.
    b. To provide a valid estimate as to how precise that estimate is.
2. Statistical approaches must be judged by their ability to promote these two objectives.
3. It is the joint effect of covariates on the outcome that is relevant when attempting to satisfy the objectives above.
4. Covariates should be allowed for as follows:
    a. The joint effect of observed covariates that are prognostic should be modelled using their actual values, whether or not the covariates are imbalanced.

    b. The joint effect of unobserved prognostic covariates should be allowed for in the estimate of precision through the residual variance.
    c. Covariates that are not prognostic should be ignored, whether or not they are imbalanced.
5. The role of design includes balancing covariates to the extent that this is possible, in order to minimise the variance inflation factor.
6. The purpose of randomisation is:
    a. To permit valid estimation of the effect of unobserved covariates by allowing for their effect in probability since the relevant data are unobserved.
    b. To promote confidence in the results.
    c. To facilitate blinding where this is desirable and practical.
7. Randomisation is relevant for that which has not been observed and is not an excuse for ignoring that which has.

This may seem obvious, but it explains much that is regularly ignored.

For instance, we should stop justifying randomisation in terms of what happens as the sample size goes to infinity. This is unconvincing to non-statisticians and not what randomisation and the associated calculation relies on. Instead we should stress what Harry Marks referred to as, 'rigorous uncertainty'(Marks 2003) and the role of randomisation and the associated statistical analysis in honest reflection of uncertainty.

It also explains why a randomised block design should not be analysed like a completely randomised design, even though the point estimate would be the same: the estimate of precision differs. Since differences between blocks do not contribute to uncertainty in the estimate, they should not contribute to the estimate of that uncertainty. However, it is not the design *per se* that makes the difference; it is that the design only makes sense if *block* is considered prognostic and since block is prognostic, it is relevant to the calculation of the standard error. If statistical analysis of clinical trials were taught this way, it would save many a student confusion as to whether a matched-pairs t or a two-sample t test was needed.

Again, this may be regarded as being trivial. However, since the propensity score is identical for randomised block and completely randomised designs, a closely related consequence is that the propensity score (PS) has no place in the analysis of randomised clinical trials(Guo and Dawid 2010, King and Nielsen 2019, Senn *et al.* 2007). Construction of the PS concentrates on that which is predictive of assignment. This maximises the effect of the variance inflation factor and omits factors that are predictive of outcome only. All attempts to rescue the PS from this deficiency make explicit or implicit appeal to ANCOVA. Once this is understood the PS can be seen to be irrelevant.

It also shows that many common habits are illogical. For example, to the extent that baselines are believed to be prognostic (as they nearly always are) and have been measured (as they usually are), they should be in the model. Thus, simple analysis of change from baseline would be generally invalid, since the baselines must have been measured. (The only exception would be very small trials where the correlation was believed strong.) The simple analysis of raw outcomes could only be valid in the rare cases where the baseline had not been measured. (The only exception would be small trials in which the correlation was believed weak.)

In the Rothamsted approach to the analysis of designed experiments, attention was paid not only to the treatment structure but also to the block structure, which defines the variation in the experimental material and hence the null analysis of variance (Anscombe 1948) against which the full analysis of variance will be judged. The design matrix describes the mapping of treatment structure onto block structure and these three things together define how the analysis will proceed(Nelder 1965a, b, Payne and Wilkinson 1977). There will be a set of exchangeable allocations that would produce the same form of analysis and randomisation will choose one at random.

In its first version, the form of the analysis followed from block structure, treatment structure and the design matrix. This particular approach is extremely powerful in proposing efficient valid analyses of designs that might otherwise be overlooked. It also seems to follow Fisher's dictum 'as you randomise so shall you analyse.'

However, subsequent development also permitted the fitting of covariates(Payne and Tobias 1992) and since these are not part of the design except in that the experimenter must have chosen to measure them, this implies that the matter is not so rigid. It is, however, compatible with a reversal of the usual dictum, 'as you analyse so shall you randomise'. Here the idea would be that the experimenter has in mind factors known and unknowable that might affect the outcome. A decision is made, bearing in mind that measurement is not free, as to which of the former should be measured in order for them to be put in the model for analysis. Since there is a penalty to be paid for non-orthogonality, what can be blocked easily should be blocked but what is predictive, cannot be blocked but can be measured, should be put in the model for analysis. For very small trials it may be necessary to reduce the covariates to a single prognostic score by relying on historical data(Holzhauer and Adewuyi 2023).

I note also, by the by, that proposals to do better than randomisation often do worse as regards improving inferences and dealing with bias. Consider for example, this proposal:

"The control group could be formed from past records and the new treatment applied to a fresh set of patients who have been carefully matched with those in the artificially constructed control group…"(Howson and Urbach 1989, P153).

Quite apart from any other difficulties of matching, if the patients are matched as regards age they cannot be matched as regards birth cohort and *vice versa* and these are both factors that can be important and, whatever they are matched on, they cannot be matched on era of treatment. Furthermore, we have ample evidence that variation of results for control patients from study to study is important. When this is formally taken into account, the information content of such historical controls is found to be much less than one might naively suppose(Schmidli *et al.* 2014). There is a tendency to regard all non-concurrent information as if it had been collected in an unbalanced parallel group trial whose deficiencies can be restored by matching or modelling. However, cluster allocation might be a better analogy.(Collignon *et al.* 2020)

6. Conclusion

A weak justification of randomisation can be given as follows. Not conditioning on observed prognostic variables is a failure of judgement and poor practice. However, if the decision not to use the prognostic covariates is not based on the results, the information contained in the covariates is missing completely at random(Rubin 1976). If so, a series of trials analysed in a way that respects the randomisation allows a third party to make a valid inference. The same is not true of a combination of trials that have not been randomised.

Nevertheless, there is no need and indeed no justification in using randomisation to disparage modelling. It is clear that our decisions as to how to design experiments and what to measure must reflect what we consider is relevant to interpreting the results and hence what we should put in our models. The marginal inference that would apply in ignorance cannot be used to trump the conditional inference which knowledge indicates, just because on average it would be right. There is value in adjusting analyses for prognostic variables and good practice requires these to be prespecified in the analysis plan(Raab *et al.* 2000).

However, this does not mean that randomisation is not valuable. Quite apart from its role in double-blind studies, randomisation provides a template of trust. It enables us to use results produced by others and it permits the use of the distribution in probability of those confounders for which the actual distribution is not observed. Once a model for analysis has been chosen, then it may suggest a feasible blocking structure to which randomisation can be applied. Trying to do better than this will at best bring very small gains in efficiency but risks damaging trust and increasing uncertainty.

Acknowledgements

I am grateful to Rosemary Bailey, Chris Brien, Iain Chalmers, Simon Day, Anthony Edwards, Frank Harrell, Robert Matthews, Roger Payne, Hans-Peter Piepho and Qingyuan Zhao for helpful comments and discussions and to the referees and the editors for comments on earlier versions .

References

Aldrich, J. (2005). Fisher and regression. *Statistical Science, 20(4),* 401-417. <Go to ISI>://000235104600006

Altman, D. G. and Royston, J. P. (1988). The Hidden Effect of Time. *Statistics in Medicine, 7(6),* 629-637. <Go to ISI>://A1988N738600001

Anscombe, F. J. (1948). The validity of comparative experiments. *Journal of the Royal Statistical Society. Series A (General), 111(3),* 181-211.

Atkinson, A. C. (1982). Optimum Biased Coin Designs for Sequential Clinical-Trials with Prognostic Factors. *Biometrika, 69(1),* 61-67.

Bailey, R. A. (1983). Restricted Randomization. *Biometrika, 70(1),* 183-198. https://doi.org/10.1093/biomet/70.1.183

Basu, D. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test. *Journal of the American Statistical Association, 75(371),* 575-582.

Beach, M. L. and Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials, 10(4 Suppl),* 161S-175S., http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=2605965

Bennett, J. H. (1990). *Statistical Inference and Analysis: Selected Correspondence of R.A. Fisher*, Oxford: Oxford University Press.

Borgerson, K. (2009). Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine, 52(2),* 218-233. S1529879509200056 [pii]

10.1353/pbm.0.0086

Burman, C.-F. (1996). On Sequential Treatment Allocations in Clinical Trials In *Department of Mathematics*, Gothenburg: Chalmers University of Technology.

Buyse, M. and McEntegart, D. (2004). Achieving balance in clinical trials. *Applied Clinical Trials, 13(5),* 36-40.

Chalmers, I. (2005). Statistical Theory Was Not the Reason That Randomization Was Used in the British Medical Research Council's Clinical Trials of Streptomycin for Pulmonary Tuberculosis In *Body counts: medical quantification in historical*

and sociological perspectives eds G. Jorland, O. A and W. G), Montreal: McGill-Queens University Press.

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics, 3(1),* 22-38.

Collignon, O., Schritz, A., Senn, S. J. and Spezia, R. (2020). Clustered allocation as a way of understanding historical controls: Components of variation and regulatory considerations. *Statistical Methods in Medical Research, 29(7),* 1960-1971. 10.1177/0962280219880213

Cox, D. R. (1958). *Planning of Experiments*, New York: John Wiley.

Cox, D. R. (1984). Present Position and Potential Developments: Some Personal Views: Design of Experiments and Regression Present Position and Potential Developments: Some Personal Views: Design of Experiments and Regression. *Journal of the Royal Statistical Society, Series A, 147(2),* 306-315.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, London: Chapman and Hall.

Cox, D. R. and McCullagh, P. (1982). Some Aspects of Analysis of Covariance. *Biometrics, 38(3),* 541-554.

Dawid, A. P. (1988). Symmetry Models and Hypotheses for Structured Data Layouts. *Journal of the Royal Statistical Society Series B-Methodological, 50(1),* 1-34.

Day, S., Groulin, J.-M. and Lewis, J. A. (2005). Achieving balance in clinical trials. *Applied Clinical Trials, 14(1),* 24-26.

Fewell, Z., Davey Smith, G. and Sterne, J. A. C. (2007). The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology, 166(6),* 646-655. 10.1093/aje/kwm165

Finney, D. (1943). The fractional replication of factorial arrangements. *Annals of Eugenics, 12(1),* 291-301.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52(*339-433.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33(6),* 503-513.

Fisher, R. A. (1935). *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Fisher, R. A. (1950). *Contributions to Mathematical Statistics*, New York: Wiley.

Fisher, R. A. (1990). The Design of Experiments In *Statistical Methods, Experimental Design and Scientific Inference* (ed J. H. Bennett), Oxford: Oxford.

Ford, I., Norrie, J. and Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine, 14(*735-746.

Gail, M. H., Wiand, S. and Piantadosi, S. (1984). Biased estimates of treatment effects in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika, 71(*431-444.

Gilmour, S. G. and Trinca, L. A. (2012). Optimum design of experiments for statistical inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 61(3),* 345-401.

Godambe, V. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological), 17(2),* 269-278.

Guo, H. and Dawid, A. P. (2010). Sufficient covariates and linear propensity analysis In *International Conference on Artificial Intelligence and Statistics*, pp. 281-288.

Hall, N. S. (2007). R.A. Fisher and his advocacy of randomization. *Journal of the History of Biology, 40(2),* 295-325.

Harrell, F. (2015). *Regression Modeling Strategies, With Applications to Linear Models,Logistic and Ordinal Regression,and Survival Analysis*, New York: Springer.

Hernández, A. V., Steyerberg, E. W. and Habbema, J. D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces

sample size requirements. *Journal of clinical epidemiology, 57(5),* 454-460.
Hilbe, J. M. (2007). *Negative binomial regression*: Cambridge.
Hill, A. B. (1990). Suspended judgment. Memories of the British Streptomycin Trial in Tuberculosis. The first randomized clinical trial. *Control Clin Trials, 11(2),* 77-79. 10.1016/0197-2456(90)90001-i
Holzhauer, B. and Adewuyi, E. T. (2023). "Super-covariates": Using predicted control group outcome as a covariate in randomized clinical trials. *Pharmaceutical Statistics, 22(6),* 1062-1075.
Howson, C. and Urbach, P. (1989). *Scientific Reasoning: the Bayesian Approach*, La Salle: Open Court.
Hutcheon, J. A., Chiolero, A. and Hanley, J. A. (2010). Random measurement error and regression dilution bias. *BMJ, 340(*c2289. 10.1136/bmj.c2289
International Conference on Harmonisation (1999). Statistical principles for clinical trials (ICH E9). *Statistics in Medicine, 18(*1905-1942.
Keene, O. N., Jones, M. R., Lane, P. W. and Anderson, J. (2007). Analysis of exacerbation rates in asthma and chronic obstructive pulmonary disease: example from the TRISTAN study. *Pharmaceutical Statistics, 6(2),* 89-97. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17230434
Kempthorne, O. (1977). Why Randomize. *Journal of Statistical Planning and Inference, 1(1),* 1-25. <Go to ISI>://A1977DY09800001
Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53(3),* 983-997.
King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis, 27(4),* 435-454.
Krauss, A. (2018). Why all randomised controlled trials produce biased results. *Ann Med, 50(4),* 312-322. 10.1080/07853890.2018.1453233

Kuipers, J. and Moffa, G. (2022). The variance of causal effect estimators for binary v-structures. *Journal of Causal Inference, 10(1),* 90-105.

Lane, P. W. and Nelder, J. A. (1982). Analysis of Covariance and Standardization as Instances of Prediction. *Biometrics, 38(3),* 613-621.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science, 19(2),* 219-228. <Go to ISI>://000227042800001

Leighton, G. R. and McKinlay, P. L. (1930). Milk consumption and the growth of school children, Edinburgh and London: Department of Health for Scotland.

Lindley, D. V. (1982). The role of randomization in inference In *PSA: Proceedings of the Biennial meeting of the philosophy of science association*, pp. 431-446: Cambridge University Press.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association, 99(466),* 546-556.

Liu, D. and Menjoge, S. (2008). Statistical analysis of chronic obstructive pulmonary disease (COPD) exacerbations. *European Respiratory Journal, 32(5),* 1422-1423; author reply 1423. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18978152

Magirr, D., Wang, C., Przybylski, A. and Baillie, M. (2025). Estimating the Variance of Covariate-Adjusted Estimators of Average Treatment Effects in Clinical Trials With Binary Endpoints. *Pharm Stat, 24(4),* e70021. 10.1002/pst.70021

Marks, H. M. (2003). Rigorous uncertainty: why RA Fisher is important. *The International Journal of Epidemiology, 32(6),* 932-937. doi: 10.1093/ije/dyg288

Matthews, R. A. J. (2025a). The problematic history of randomised controlled trials Part 1: presumption and confusion on the road to randomisation. *Journal of the Royal Society of Medicine*,

Matthews, R. A. J. (2025b). The problematic history of randomised controlled trials Part 2: Hill's "pragmatic" view of randomisation and its consequences. *Journal of the Royal Society of Medicine*,
McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, London: Chapman and Hall.
Medical Research Council Streptomycin in Tuberculosis Trials Committee (1948). Streptomycin treatment for pulmonary tuberculosis. *British Medical Journal, ii(*769-782.
Nelder, J. A. (1965a). The analysis of randomised experiments with orthogonal block structure I. Block structure and the null analysis of variance. *Proceedings of the Royal Society of London. Series A, 283(*147-162.
Nelder, J. A. (1965b). The analysis of randomised experiments with orthogonal block structure II. Treatment structure and the general analysis of variance. *Proceedings of the Royal Society of London. Series A, 283(*163-178.
Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58(3),* 545-554.
Payne, R. and Tobias, R. (1992). General balance, combination of information and the analysis of covariance. *Scandinavian Journal of Statistics,* 3-23.
Payne, R. and Wilkinson, G. (1977). A general algorithm for analysis of variance. *Applied Statistics,* 251-260.
Peirce, C. S. and Jastrow, J. (1884). On small differences in sensation.
Pocock, S. J. (1983). *Clinical trials, A Practical Approach*, Chichester: Wiley.
Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics, 31(1),* 103-115.
Popper Shaffer, J. (1991). The Gauss-Markov theorem and random regressors. *The American Statistician, 45(*269-272.
Raab, G. M., Day, S. and Sales, J. (2000). How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials, 21(4),* 330-342.

Reid, C. (1982). *Neyman, from Life*, New York: Springer.
Robinson, L. D. and Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review, 58(*227-240.
Rosenberger, W. F. and Lachin, J. M. (2016). *Randomization in clinical trials: theory and practice*, Hoboken: John Wiley & Sons.
Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika, 57(2),* 377-387.
Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63(3),* 581-590. <Go to ISI>://A1976CP66700021
Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics, 70(4),* 1023-1032. 10.1111/biom.12242
Senn, S. J. (1994). Fisher's game with the devil. *Statistics in Medicine, 13(3),* 217-230. http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=0008202650
Senn, S. J. (1995). In defence of analysis of covariance: a reply to Chambless and Roeback [letter; comment]. *Statistics in Medicine, 14(20),* 2283-2285. http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=0008552904
Senn, S. J. (2000). Consensus and controversy in pharmaceutical statistics (with discussion). *Journal of the Royal Statistical Society Series D, The Statistician, 49(*135-176.
Senn, S. J. (2004). Added Values: Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine, 23(24),* 3729-3753. doi: 10.1002/sim.2074
Senn, S. J. (2013a). A Brief Note Regarding Randomization. *Perspectives in biology and medicine, 56(3),* 452-453.
Senn, S. J. (2013b). Seven myths of randomisation in clinical trials. *Statistics in Medicine, 32(9),* 1439-1450. 10.1002/sim.5713
Senn, S. J. (2015). Various varying variances: The challenge of nuisance parameters to the practising biostatistician. *Statistical Methods*

in Medical Research, 24(4), 403-419. 10.1177/0962280214520728

Senn, S. J. (2022). Empirical studies of balance do not justify a requirement for 1,000 patients per trial. *J Clin Epidemiol, 148(*184-188. 10.1016/j.jclinepi.2022.02.010

Senn, S. J. (2023). Student and the Lanarkshire milk experiment. *Eur J Epidemiol, 38(1),* 1-10. 10.1007/s10654-022-00941-x

Senn, S. J., Anisimov, V. V. and Fedorov, V. V. (2010). Comparisons of minimization and Atkinson's algorithm. *Statistics in Medicine, 29(7-8),* 721-730. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20213718

Senn, S. J., Graf, E. and Caputo, A. (2007). Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment on exposure. *Statistics in Medicine, 26(30),* 5529-5544. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18058851

Senn, S. J., König, F. and Posch, M. (2024). Stratification in Randomised Clinical Trials and Analysis of Covariance: Some Simple Theory and Recommendations. *arXiv preprint arXiv:2408.06760*,

Senn, S. J., Lillienthal, J., Patalano, F. and Till, M. D. (1997). An incomplete blocks cross-over in asthma: a case study in collaboration In *Cross-over Clinical Trials* eds J. Vollmar and L. A. Hothorn), pp. 3-26, Stuttgart: Fischer.

Siegfried, S., Senn, S. and Hothorn, T. (2023). On the relevance of prognostic information for clinical trials: A theoretical quantification. *Biom J, 65(1),* 10.1002/bimj.202100349

Speed, T. P. (1987). What is an analysis of variance. *The Annals of Statistics, 15(3),* 885-910.

Steyerberg, E. W., Bossuyt, P. M. M. and Lee, K. L. (2000). Clinical trials in acute myocardial infarction: Should we adjust for baseline characteristics? *American Heart Journal, 139(5),* 745-751.

Taves, D. R. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics, 15(5),* 443-453. doi: 10.1002/cpt1974155443

Worrall, J. (2002). What evidence is evidence based medicine. *Philosophy of Science, 69(S3),* S316-S330.

Worrall, J. (2007). Why There's No Cause to Randomize. *British Journal for the Philosophy of Science, 58(*451–488.

Yates, F. (1936). Incomplete randomized blocks. *Annals of Eugenics, 7(2),* 121-140. doi: 10.1111/j.1469-1809.1936.tb02134.x

Yates, F. (1940). The recovery of inter-block information in balanced incomplete block designs. *Annals of Eugenics, 10(1),* 317-325. doi: 10.1111/j.1469-1809.1940.tb02257.x

Zhang, Y. and Zhao, Q. (2023). What is a randomization test? *Journal of the American Statistical Association, just-accepted),* 1-29.

Zheng, L. and Zelen, M. (2008). MULTI-CENTER CLINICAL TRIALS: RANDOMIZATION AND ANCILLARY STATISTICS. *Annals of Applied Statistics, 2(2),* 582-600. 10.1214/07-aoas151