Research Article

Statistics in Medicine

Received 22 November 2011, Accepted 27 November 2012 Published online 17 December 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5713

Seven myths of randomisation in clinical trials

Stephen Senn[*†]

I consider seven misunderstandings that may be encountered about the nature, purpose and properties of randomisation in clinical trials. Some concern the practical realities of clinical research on patients. Others are to do with the value and purpose of balance. Still others are to do with a confusion about the role of conditioning in valid statistical inference. I consider a simple game of chance involving two dice to illustrate some points about inference and then consider the seven misunderstandings in turn. I conclude that although one should not make a fetish of randomisation, when proposing alternatives to randomisation in clinical trials, one should be very careful to be precise about the exact nature of the alternative being considered if one is to avoid the danger of underestimating the advantages that randomisation can offer. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: randomisation; blinding; conditioning; covariates

1. Introduction

Attacks on randomisation in clinical trials have betrayed an ignorance as to how randomised clinical trials are run and what the value of randomisation is [1–5]. It may be that these misunderstandings are not fatal to the critical arguments that have been put. Nevertheless, it seems to be of some value to issue a correction. Either the misunderstandings are inessential to the overall argument, in which case their elimination should make the case against randomisation cleaner and clearer (because their continuing presence can only confuse and irritate those who are knowledgeable about clinical trials), or they are essential, in which case their clarification is an important task in the general refutation of the criticism of the role of randomisation in clinical trials.

The plan of this paper is as follows. I consider, first of all, by way of a general illustration of some points regarding probability, information and valid inference, a simple game of chance involving two dice that is played in three variants. I will then list and comment on various misunderstandings or myths of randomisation. These are divided into two unequal groups. Myths 1–5 are those I assume that no medical statisticians believe. Myths 6 and 7, however, are the ones that even (some) statisticians may believe. I finish with some discussion of unresolved issues.

Although I shall be critical of John Worrall's paper, 'What evidence is evidence based medicine?' [3], there is much in it with which I am in agreement. As I shall make clear in the conclusion, and indeed have stated elsewhere [6], I do not think that randomised clinical trials are sufficient for answering all our evidential needs. Worrall's paper provides many useful examples why this is so. Nevertheless, that paper and also the follow-up paper, 'Why there's no cause to randomize' [4], go too far in making this point and, in doing so, make statements about randomised clinical trials in need of rectification. Also, I shall be clearing up some misunderstandings that are not found in Worrall's papers and that cannot be ascribed to him.

2. Three variations of a game involving two dice

The game to be described mimics a clinical trial in that it looks at two possible sources of variation in results: prognostic information and treatment. Two dice, presumed fair, are involved: a red die whose

Competence Centre for Methodology and Statistics, CRP-Santé, L-1445 Strassen, Luxembourg
**Correspondence to: Stephen Senn, Competence Centre for Methodology and Statistics, CRP-Santé, L-1445 Strassen, Luxembourg.*
†E-mail: stephen.senn@crp-sante.lu

score provides an analogue to prognostic information and a black die that provides an analogue of adding information due to treatment. The game is played between a statistician (or other would-be forecaster) and a gambler. (See [7] for a fuller discussion.) The red die is to be rolled first and then the black die. A statistician is to state the probability, P, that the total score from the two dice is 10. The result is to be subject to a bet that will be placed by the gambler (with the statistician as the other party) using the implied odds, $P : (1 - P)$, from the statistician's statement of probability. The gambler may choose, however, whether to bet for or against the total score of 10. This device of allowing the gambler to choose is meant to encourage the statistician to forecast honestly. The three variants of the game are as follows.

1. Variant 1: the statistician states the probability of a total score of 10, and the two dice are then rolled together.
2. Variant 2: the red die is rolled so that the statistician (and gambler) can see it. The statistician then states the probability of a total score of 10, and the black die is then rolled.
3. Variant 3: the red die is rolled, and the score is known neither to the statistician nor to the gambler. The statistician calls the odds. The black die is then rolled.

What probabilities should the statistician issue? For variant 1, the statistician may argue as follows. There are 36 possible combinations of the six scores from the red die and the six from the black die, which I assume equally likely. Of these 36 combinations, three (four and six, five and five, and six and four) produce a total score of 10. Thus, the probability of a 10 is $3/36 = 1/12$.

For variant 2, the statistician will have information to condition on and should argue as follows. If the score on the red die is 3 or less, then it is impossible to obtain a 10 as the total score for both dice. The probability that the statistician should issue is 0. If on the other hand the score is 4 or more, then there is exactly one score of the six on the black die that will make the total up to 10. Thus, the probability is $1/6$. Thus, for the game as played in variant 2, the statistician *either* issues the probability 0 *or* $1/6$. Note that the statistician knows before the red die is rolled that there is half a chance that the forecast that will be made once the red die is seen will be 0 and half a chance that it will be 1/6. Hence, the average of the probabilities that will be issued is $(\frac{1}{2} \times 0) + (\frac{1}{2} \times 1/6) = 1/12$ as before. However, despite that, it would be inappropriate to use the value 1/12 as a forecast once the result of the red die is known.

Variant 3 is equivalent to variant 1. Although the red die has already been rolled, the score on it is known neither to the statistician nor to the gambler. Therefore, they should behave as if the game was played in variant 1.

The analogy that can be made with the game involving the two dice is that of a clinical trial in which covariate (that is to say, prognostic) information may or may not be available at baseline. In variant 1, there is no information. In variant 2, it is observed at baseline. In variant 3, it might in principle be available, but nobody has seen it. I shall make use of this example in discussing some of the myths of randomised clinical trials. I now proceed to list and discuss these.

3. Seven myths of randomised clinical trials

I list and discuss in this section seven myths regarding the running and interpretation of randomised clinical trials, in particular as regards the role of randomisation. Many of these myths are held as truths by those who are not involved with clinical trials, but myths 6 and 7, and in particular the last of these, are also commonly encountered amongst trialists.

3.1. Myth 1. Patients are treated simultaneously in clinical trials

This myth is astonishingly persistent with critics of randomisation. However, patients are generally treated in clinical trials when (or soon after) they 'present'. This follows on at an interval after they fall ill. In consequence, patients are entered sequentially onto a clinical trial, and the recruitment period (the period in which patients are entered onto a trial) will, in many indications, be longer than the follow-up period (the period for which patients are observed). In consequence, it is not uncommon for some patients to have completed a clinical trial before others have started. There are occasional exceptions. Many phase I studies use 'banks' of healthy volunteers. Also, in most cluster randomised trials, the clusters are identified before treatment is started. However, for classical clinical trials in patients, it is extremely rare that the patients can be treated simultaneously.

Thus, it is rarely possible to match patients according to baseline characteristics because the trial must be started with *some* patients well before the last patients are recruited. It is thus almost always impossible to do what Worrall proposes [4] when he writes

> If, having created groups matched with respect to those 'known' factors, one then goes on to decide which will be the experimental and which the control group by some random process—in the simplest case by tossing a fair coin—then one can do no epistemic harm, though one also does no further epistemic good. (p. 463)

Nor is it possible to do what Urbach [2] proposes when he writes

> For example, one could arrange for the matching to be performed by a panel of doctors representing a spectrum of opinion on the likely value of the drugs and whose criteria of selection have been made explicit. (p. 272)

because this would clearly require that all patients to be entered into the trial were known and measured before the start of the trial.

It is also not possible to re-randomise. So when Worrall states [4], for example,

> Those involved in clinical trials usually talk about this as 'checking for baseline imbalances.' Everyone agrees, as the Bayesian points out, that if there *is* a clear 'baseline imbalance' one should not proceed to draw any conclusion from the trial. The classical statistician (rather quixotically) insists that one should then re-randomize (if necessary again and again) until we see no reason to think the division unbalanced. (p. 463)

he is not describing what happens in practice, because trialists know re-randomisation to be impossible (see also [1, p. 151]).

3.2. Myth 2. Balance of prognostic factors is necessary for valid inference

To discuss this issue, it is necessary to make a distinction between observed and unobserved covariates. In understanding the relevance of this distinction, it is useful to have in mind the example of the two dice. An observed covariate corresponds to the situation where the score for the red die is revealed first and the unobserved covariate corresponds to the case where the score is not revealed.

Consider the case where an important prognostic covariate is revealed and the values of the covariate correspond to two strata in a clinical trial. To give a concrete example, one might suppose one has patients already on steroids and patients not on steroids in a trial in asthma comparing treatment with a beta-agonist with placebo. Suppose that the two arms of the trial are not balanced as regards use of steroids. This is equivalent to saying that if one formed two sub-trials, one for patients on steroids and one for patients who were not, then the numbers in each trial would be unbalanced. Thus, in this instance, imbalance in covariates is imbalance in numbers. However, each stratum constitutes a valid clinical trial, and it is only necessary to compare proportions (or rates or means, or some other statistic standardised by the number of subjects) to make a valid comparison within each stratum. The valid comparisons within strata can then be combined appropriately across strata. This technique, which is well known to trialists, is called *post-stratification*. An analogous technique called analysis of covariance can be used to adjust for imbalances in continuous prognostic variables [8, 9]. It thus follows that imbalance of an observed prognostic covariate is not a problem.

So Worrall is quite wrong when he states (in the passage already quoted) that, 'everyone agrees that if there is a clear "baseline imbalance" one should not proceed to draw any conclusion from the trial'. There is no requirement for baseline balance for valid inference.

Now consider unobserved covariates. Borgerson quoting Worrall [3] with approval claims it is necessary for these to be balanced. (See in particular [5, pp. 222–223].) However, such balance is not necessary. It is sufficient to know what their distribution in probability may be. This is true even though one knows that if their actual distribution were revealed, then the calculation that is based on their distribution in probability would be irrelevant. The analogy is to variant 3 of our game with two dice. This can be treated as if it were variant 1 although it is known that if the information regarding the red die were revealed, then it would have to be treated as variant 2.

Thus, Worrall's point that

> There is no reason to think that any actual randomized trial reflects the 'limiting average'. [4, p. 465]

is not relevant. There is no reason to think that the actual unobserved score on the red die is 3.5 although this is the average score. In fact, there is every reason to believe it is not 3.5! It is quite sufficient for the purpose of calculating the probability that the total of the two dice will be 10 to know that each of the six scores for the red die are equally likely.

Nor is Worrall's criticism that

> In particular, randomizing cannot deliver us from the possibility that the two groups are—of course (by definition) unbeknown to us—relevantly different with respect to some factor that we have not yet thought about. [4, p. 463]

of any relevance. It is not necessary for the groups to be balanced. In fact, the probability calculation applied to a clinical trial automatically *makes an allowance for the fact that groups will almost certainly be unbalanced*, and if one knew that they were balanced, then the calculation that is usually performed would not be correct. Every statistician knows that you should not analyse a matched pairs design as if it were a completely randomised design. In the former, there is deliberate balancing by the device of matching; in the latter, there is not. In other words, not only *despite* but also *because* we know that in practice many hidden confounders will be unbalanced, the conventional analysis of randomised trials is valid. If we knew these factors to be balanced, then the conventional analysis would be invalid.

In fact, when Borgerson claims that 'in order to begin to address this problem of confounding factors, the randomization would have to be repeated an indefinite number of times' [5, p. 223], she confuses a probability statement regarding the possible effects of possible imbalances (which is what with the usual statistical calculations provide) with a requirement for perfect balance (which does not exist). In variant 1 (or 3) of the game with two dice, one can make a bet on a single throw using as a guiding rule for the appropriate probability one's belief that this is a fair die, that is to say *one for which the relative frequency would be* $1/6$ *given that the experiment were 'repeated an indefinite number of times'*. One knows full well that given a single throw, one of the numbers will have an observed relative frequency of 1 and the others of 0.

3.3. Myth 3. Blinding can be carried out effectively without randomisation

In Fisher's classic description of a lady tasting tea to see whether the cups she has been given to taste have had milk added first or last, he writes

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

I have discussed this example extensively elsewhere [11, 12], but briefly, the following points are important. (1) There is a careful description as to how allocation is to proceed. (2) Randomisation implies that every possible sequence of cups is equally likely. Because there are $8!/(4!4!) = 70$ possible sequences, the probability of any given sequence is $1/70$. (3) Because the lady has been told how the allocation will be made, double guessing is not necessary; that is to say, in calculating the probability that she would guess the given sequence of cups correctly, it is sufficient to use the average probability over all randomisations. In consequence, the experiment satisfies the principle of what might be called mutually agreed deception. The random element of the experiment is agreed between subject and experimenter, and this simplifies the probabilistic modelling of the process of guessing.

Instead of this careful description, consider the following from Lindley, a distinguished Bayesian critic of frequentist approaches. He describes some possible experiments involving a lady tasting tea.

> In each of the experiments the obvious randomization is supposed to have taken place. Actually no physical act of randomization is needed: all that is required is that the lady is reasonably entitled to make the assumption of exchangeability required below. For this purpose a haphazard arrangement is all that is required. [13, pp. 456–457]

The problem with this is that as a method of assignment, *haphazard arrangement* is not rigorously described. In what does it consist? How would it apply to the lady tasting tea? Is it a method that chooses one of the 70 sequences with equal probability? In that case, it seems to be nothing more or less than randomisation. If it is not randomisation, then the 70 sequences are presumably not chosen with equal probability. Some of them are more or less probable, and one can only assume that this feature is somehow tied up with Lindley's thought processes, which now become (unfortunately, because this greatly complicates the matter) part of the essential description of the experiment.

Fisher, in a letter to Jeffreys, explained the dangers of using a haphazard method, thus

> ... if I want to test the capacity of the human race for telepathically perceiving a playing card, I might choose the Queen of Diamonds, and get thousands of radio listeners to send in guesses. I should then find that considerably more than one in 52 guessed the card right. Experimentally this sort of thing arises because we are in the habit of making tacit hypotheses, e.g. 'Good guesses are at random except for a possible telepathic influence.' But in reality it appears that red cards are always guessed more frequently than black. [14, pp. 268–269]

So when, for example, Worrall states

> ...if the trial was, and remained, double-blind then randomization could play no further role in this respect. [4, p. 454]

he misses the point. Full blinding is achieved with the help of randomisation, which assures maximum unguessability of any sequence. Nobody running an experiment where blinding really mattered, say to check claims about extra sensory perception [15] or about the efficacy of homeopathic medicines, would accept anything less.

3.4. Myth 4. Randomisation is inefficient

There is a trivial sense in which this is not a myth but a fact. Randomisation is not generally fully efficient. For example, if you wished to balance for a linear trend, then some method of alternation might be superior. For example, if you allocated all patients in a continuing predictable 'sandwich' sequence ABBAABBA, and so on and recruited $4n$ patients, where n was an integer, then you could guarantee that the total of the sequence number of the patients given A was the same as those given B: under A, we would have patients 1, 4, 5, 8, and so on and under B patients 2, 3, 6, 7, and so on. Thus, for any set of four patients, the total of the position numbers under each treatment is five with an average of 2.5 for either treatment. So, the average difference in position between A and B is 0. On the other hand, if you randomised in pairs, then there could be some slight imbalance on average in position. On the other hand, strict alteration, ABABAB... would be worse than randomisation in pairs because the average difference between A and B would be one place. We can sum up these schemes in terms of mean square error for the difference (B-A) in position as in Table I.

For these three schemes, if balance for *linear trend* is the only object, then scheme 1 is the best and scheme 3 is the worst. (Of course, as regards blinding, the only one of these three worth considering is scheme 2.) So, to a limited extent, this is not a myth but a fact. However, there is also a common belief that one can do *much* better in terms of efficiency of inference if one can balance treatment groups by prognostic factors rather than randomising, and this *is* a myth. Certainly, there is some gain in balancing numbers because, in a two-group parallel trial, for any total number of patients N allocated in such a way that fN are allocated to one group and $(1-f)N$ are allocated to the other, $0 < f < 1$, the variance of the treatment contrast is proportional to the sum of the reciprocals of the numbers of patients in each group and hence to

$$\frac{1}{fN}+\frac{1}{(1-f)N}=\frac{1}{Nf\,(1-f)}=\frac{1}{N\left[\frac{1}{4}-\left(f-\frac{1}{2}\right)^2\right]}=\frac{4}{N}\times\frac{1}{1-4\left(f-\frac{1}{2}\right)^2},$$

which by inspection clearly reaches a minimum of $4/N$ when $f = {}^1/_2 = (1-f)$, that is to say, when the two groups are balanced. This is the case that applies when there are no covariates. The reciprocal of the factor $4/N$, that is, $N/4$, is the upper bound for the efficiency of the design (which one may rescale to equal 100%), and it turns out that it also applies in the more general case where there are covariates, although conditioning on them will (to the extent that they are prognostic) bring a dividend in terms of

Table I. Statistics for the mean difference in 'place' between two treatments using three simple allocation schemes.

Scheme	Mean	Variance	Mean square error
1. Sandwich	0	0	0
2. Randomised pairs	0	$1/(2n)$	$1/(2n)$
3. Alternation	1	0	1

reduced conditional variance. However, any deterioration in balance leads to a reduction in the efficiency attained by any given trial. See [16, 17] for a technical discussion.

It may seem that this formula is not relevant here as it refers to imbalance in totals and not in covariates. However, as was pointed out in the discussion of myth 2, the two are closely related. Consider two trials comparing the same treatment but in different populations. One randomises patient A to patient B in the ratio 1:2 and the other in the ratio 2:1. It is clear that each trial is inefficient compared with one that used an equal division of patients. It is also clear that if the two trials are combined naively, then there is a bias because the difference between treatments is partly confounded with the difference between populations. However, it is also clear that a meta-analysis stratifying by trial (and hence population) would (a) not be biased but (b) inherit the inefficiency that the unequal randomisation of the trials caused in the first place. Thus, imbalance in a covariate (in this case population), *to the extent that it is dealt with*, is very closely related as a phenomenon to that of imbalance in total numbers.

The argument here is in terms of numbers of a binary covariate (in this case, whether a patient belongs to trial 1 or trial 2), but an analogous argument applies to other sorts of covariates. For example, for a continuous covariate, such as a baseline measurement of some continuous outcome variable, full efficiency is only achieved if the difference in groups in the mean baseline is zero. (See *Statistical Issues in Drug Development* [18, Chapter 7] for a technical discussion.) Thus, a randomised trial, because it will not balance covariates perfectly, may experience some small loss in efficiency compared with a perfectly balanced trial. However, it is also known that the expected loss in efficiency for a randomised design compared with a perfectly balanced design (which is in practice unobtainable) is equal to one patient for every covariate fitted in the model [19–22].

Extensive theoretical analysis and simulation of clinical trials show that the loss involved in randomising is not generally important in clinical trials even if one takes the (unrealistic) case that is least favourable to randomisation; that is, there is no doubt as to which factors are relevant, and one could in principle balance them perfectly [17, 19].

So, it is known that the disadvantage of randomisation compared with perfect balance is minimal [22, 23]. However, it is much better than some alternatives that have been proposed. Consider, for example, the proposal of Urbach

> ....or one could simply permit the subjects to choose their own groups, always ensuring of course that they have not been informed of which treatment is to be applied to which group. [2, p. 271]

There is no guarantee that this will provide even approximately equal numbers per group. Suppose that the groups in this bizarre scheme are labelled A and B. The assumption (amongst many) must be that human beings have no prejudice in favour of one letter or the other as say for red cards in favour of black ones. Any common prejudice will tend to lead to unbalanced groups. Note also that for this scheme to be believed adequate as a means of *efficiently* conducting a randomised clinical trials, it is not *sufficient* for the Bayesian to have a subjective expectation that the probability θ of a randomly chosen patient choosing A would be $1/2$. For example, such a condition is satisfied by a uniform prior distribution on θ. However, such a uniform prior distribution makes every possible *proportion* of patients allocated to A equally likely [12], and this is far less peaked around the optimal value 0.5 than is a completely randomly allocated scheme and hence may be expected to be less efficient. Of course, a uniform prior distribution is perhaps somewhat pessimistic. However, to understand what Urbach is up against when facing randomisation, one has to appreciate that randomisation is equivalent to having a completely informative Bayesian prior distribution that any given patient will receive A with probability 0.5. This means that only stochastic (and not epistemic) uncertainty applies. As the sample size increases, the distribution collapses around 0.5, and the variance of the proportion allocated goes to 0. This does not happen with Urbach's scheme for which the variance of the prior distribution (which admittedly may be smaller than the value of $1/12$ that applies to the uniform distribution but in practice will be greater than the value of 0 that applies to randomisation) provides a lower bound for the variance of the predictive distribution of the proportion of patients allocated.

3.5. Myth 5. Randomisation precludes balancing covariates

In a discussion of randomisation, Urbach states

> ...it seems to me, even in these circumstances randomizing is not the best way of getting matched groups. [2, p. 266]

as if matching were an *alternative* to randomisation. However, the standard view expressed in dozens of books on experimental design is that randomisation is an *adjunct* to matching, although the practical reality of clinical trials dictates that matching is difficult. Certainly, matched pairs designs are almost never possible unless the matching is within patients, for example, when treating both eyes for cataracts, each being given a different treatment. On the other hand, stratified randomisation using a limited number of strata is possible and is frequently carried out, in particular if centres form the strata [24]. It is then attempted to balance the numbers on each treatment arm within the strata, most commonly using permuted blocks. A nice treatment of experimental design that does not rely heavily on mathematics is the classic by Cox written over 50 years ago [25]. For a more modern and heavily mathematical treatment of randomisation in clinical trials, see [26].

In fact, the Fisherian prescription for experiments might be summed up as 'balance what you can and randomise what you can't'. The theory of experimental design makes extensive use of this, starting with the simplest example that of randomised blocks and continuing via more complicated designs such as split-plot designs with control and randomisation at different levels. John Nelder's theory of general balance [27–30] was developed to produce analytic algorithms that corresponded to the randomisation used. For example, in the package GenStat® that he developed, you can declare the blocking structure and the treatment structure and provide the design matrix and the result of the experiment, and the analysis then follows automatically.

Elaborate treatment and blocking structures are much more common in agricultural and industrial experiments than in clinical trials, largely for the reasons outlined in the discussion of myth 1: patients have to be allocated to treatment when they present. Nevertheless, some blocking in clinical trials is common, and attempts are frequently made to balance numbers on the arms of a trial within centres, for example, [24]. Such blocking is also possible in a cross-over design. This is one in which patients are allocated to sequences of treatment for the purposes of comparing individual treatments [31]. The units of inference then become episodes of treatment, and the blocks are the patients. This does not mean that randomisation is not employed because the order in which patients receive treatments will be randomised.

It will be helpful to consider an actual example, because this also illuminates some of the points made in discussing the previous myths. I choose a famous trial described by Hills and Armitage [32] in which 29 patients suffering from enuresis were treated in separate periods of 14 days with a placebo and a verum (a treatment under investigation). The number of dry nights were noted. In what follows, in order to simplify the discussion, I shall assume that period and carry-over effects can be ignored. In that case, an appropriate analysis is a so-called *matched pairs t-test*. The observed mean difference in numbers of dry nights (verum–placebo) is 2.172, and the one-sided p-value for the matched pairs t-test is 0.00074. If you do not like the parametric test because the normality assumption seems less than ideal, then you can do a permutation test as an alternative. (These were proposed by Fisher [33] and Pitman [34] in the 1930s. For a simple explanation, see Colquhoun's book [35].) This leaves the actual results as they are paired by subject but then randomly switches the labels as to which period is under verum and as to which is under placebo. When I did this, using 10 000 simulations, the proportion of mean differences ($\pm$SE) that are greater than 2.172 was found to be 0.0008 ($\pm$0.0003), which then becomes the permutation p-value. If you do not like this, then you can always do a Bayesian analysis for which, given a standard vague prior and repeating the dubious assumption of normality, the probability that the treatment effect is greater than 0 is 0.001 [36, 37]. In other words, it really makes no difference which inferential framework, parametric frequentist, randomisation inference or Bayesian you use.

However, if you repeat the analysis ignoring the pairing, then you obtain very different results. The parametric t-test now gives a p-value of 0.0141 and the permutation test of 0.013($\pm$0.001). In other words, ignoring the blocking factor leads to a p-value that is much less impressive (nearly 20 times as large as previously).

Of course, this second analysis is not reasonable because it reflects neither the way the trial was carried out nor what we believe about the data (because we expect observations to be correlated by patient), but it serves to make a point, namely that an allowance is made in statistical analyses for covariates that are not measured. In a parallel group trial in enuresis, the results would not be matched in the way that they are in the cross-over trial. The random variability from treatment group to treatment group would be much greater. But the analysis will reflect this. Although the point estimate of 2.172 is the same for the two analyses, the 95% confidence intervals are (0.91, 3.43) for the (parametric) matched pairs analysis and (0.24, 4.10) for the (parametric) analysis ignoring matching. The much wider confidence interval for the latter reflects the fact that *if* the data come from two randomly allocated groups of patients rather than

being matched, then in that case, we are much less certain about the result. But this is not a problem for the expression of uncertainty itself. It seems to me that this point is consistently overlooked by the critics of randomised trials and their analysis, and so it is worth re-stating here. The analysis of randomised trials makes an allowance for unmeasured covariates. When these covariates are measured and adjusted for, the uncertainty in the estimate is (usually) reduced.

In fact, confidence intervals for parameter estimates appear to be being misinterpreted by the critics of randomisation. These are as wide and imprecise as they are *because* they make an allowance for imbalance. An analogy can be made with engineering here. It is rather missing the point to claim that engineering calculations are useless because they depend on mathematical idealisations that cannot take all factors into account whilst overlooking the fact that precisely *because* this is so engineers build allowance and tolerance factors into their calculations and specifications. Conventional statistical calculations have tolerance built into them. They use the fact that patients will differ randomly not only between groups but also within groups. They use the disparity in results within groups as a means of scaling the uncertainty expressed about the between-group difference. Note that this does not require one to measure an infinity of factors (which may or may not be affecting the within-group and between-group results); it only requires that one has observed the outcome that these factors have affected and to have a theory that relates within-group variation to between-group variation. It is this theory that randomisation helps support.

3.6. Myth 6. Observed covariates may be ignored because one has randomised

Unfortunately, this myth is one some statisticians appear to believe.

To ignore observed prognostic covariates is to treat variant 2 of the game as if it were game 1. This is not logical. In fact, it is not logical to ignore prognostic covariates even if they are perfectly balanced once they have been observed. Valid inference depends on conditioning on what is known. It is only when the actual distribution of covariates has not been observed that it is acceptable to substitute their distribution in probability. In the case of the game involving two dice, it is relevant, indeed necessary, to consider with what probability each of the six possible scores of the red die could arrive if the die cannot be observed before betting. However, once one has the actual score, this probability is no longer relevant.

Of course, there is the practical problem of knowing what is prognostic. The solution commonly favoured in drug regulatory circles of selecting a limited set of known prognostic covariates to be used in the model is a sensible pragmatic compromise. Obviously, what has not been measured cannot be put in the model, and it is here that randomisation proves valuable. However, if it was chosen to measure something because it was prognostic, then it is not reasonable to behave as if one had not seen it. This issue is discussed again at the end of the paper.

3.7. Myth 7. Large trials are more balanced than small ones

As already explained, conventional analyses of randomised trials make an allowance for the distribution of unmeasured confounders. They do this by judging the probability with which the groups can differ from each other by looking at the way in which results differ within groups. Unmeasured confounders make a contribution to both of these measures of variation (between and within group), and the comparison of the two is the cornerstone of the technique of analysis of variance developed by RA Fisher in the 1920s [38].

When sample sizes increase, it is certainly the case that the expected random difference between two groups will reduce, and this reflects, amongst other things, the greater expected balance in proportionate terms between groups. In this sense, the belief that larger trials are more balanced than smaller ones is not a myth. However, by the same token, the standard error of the treatment effect will be smaller and the confidence interval will be narrower, and for any given observed difference at outcome, the p-value will be smaller. Thus, the effect of increasing sample size is *consumed by conventional analyses in terms of increased precision.* There is no further benefit in terms of increased reliability [6] (except the rather limited one that increasing sample size makes the difference between permutation analyses and those based on the normal model smaller). A similar point has been made in connection with sampling schemes by Cumberland and Royall [39, 40].

Another way of saying this is that when comparing a large clinical trial with a small one, other things being equal, a smaller difference at outcome will be significant. Hence, smaller differences at baseline become important.

4. Some technical matters

To sum up, my position is that randomisation entitles you to ignore unmeasured covariates but that you must condition on what you observe. The technical problem that this raises is that, eventually as you include more covariates in the model, the precision of your inferences decreases if you use conventional frequentist approaches. In consequence, if the partial correlation coefficient of a predictor given other predictors in the model is weak, then the net result of fitting a covariate can be to increase the variance of the treatment effect [41]. A related fact is that the Gauss–Markov theorem does not apply to stochastic regressors [42, 43](which is what covariates in a randomised trial are) so that you can do better than a naïve least squares analysis, but it also raises the issue as to how Bayesians ought to handle models with infinitely many predictors (see [44] for a suggestion).

There are deep issues here to which I do not pretend to have solutions except to say that the pragmatic compromise of agreeing beforehand a relatively limited set of covariates to use in a model seems reasonable. However, a formal investigation of this would be valuable, and I regard this as an area where further research would be interesting and useful.

It would, however, be a mistake to think that this is a fatal difficulty for randomisation. First, to the extent that it is a difficulty, it is a difficulty of modelling. Any non-randomised study relies at least as heavily, and in practice more so, than a randomised clinical trial on modelling. Second, it is a mistake to think that there are an infinite number of factors whose possible imbalance fatally undermines randomisation. From one point of view, we may consider a single factor. This is the observed outcome of the 'response' that we choose to measure effects by. We know that this outcome could vary between groups even when the effects of the treatments being compared are the same. We know this because we can see that within the groups of patients treated identically, this outcome is different from patient to patient. We use these observed within-group differences to judge how much the between-group difference could be even if the treatments were identical. The point is that observed covariates may lead us to refine our judgement based on the within-group difference as to what the between-group difference might be. However, when Worrall states

> Even if there is only a small probability that an individual factor is unbalanced, 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.

it seems to me he fails to realise how a coherent Bayesian who had been shown the classical Fisherian analysis would have to predict what the effect of these unmeasured covariates would be. It is not enough for such a Bayesian to suppose the existence of such rogue factors; he or she has to consider how likely it is that they are highly prognostic and as always must not forget that it is a *probability statement* that has been issued by the analysis of variance. It is the validity of the probability statement that has to be attacked, and this is quite hard. If there are indefinitely many covariates, then we might measure that these can vary *within* groups as well as *between* them and the probability statement issued from an analysis of variance is not just based on the variation between groups but also on the variation within. So, what Worrall has overlooked is that inference is based on the *ratio* of between to within variability. For his argument to follow through, he has to show that the infinity of factors will have a greater effect on the numerator than the denominator.

Another way of putting this is to say that what really matters is differences in outcome. Differences in covariates are only relevant to the extent that they help us predict outcomes we would have seen between groups in the absence of treatment. If we knew what these differences would be, then the differences in covariates would be irrelevant. Thus, we do not need to concentrate on the indefinitely many varying covariates. Their relevance is bounded by outcome and if we have randomised the variation within groups is related to the variation between in a way that can be described probabilistically by the Fisherian machinery.

A further controversial matter is what to do about situations where partial balance is possible in a way that cannot be reconciled easily with randomisation inference. For example, minimisation as proposed by Taves [45] and Pocock and Simon [46] is very popular in many public-sector trials although generally avoided in a regulatory setting. It continues to attract both defenders [47–50] and critics [6, 23, 51]. The advantage of minimisation compared with randomisation is a potential gain in efficiency. However, it is also known that the gain is very small and dependent on a number of assumptions.

The disadvantages include an increase in predictability with its attendant disadvantages (loss of blinding vulnerability to manipulation) [11, 52] and increased uncertainty about precision itself [6]. My judgement is that this particular game is not worth the candle, and I strongly prefer randomisation to minimisation.

5. Conclusion

I cannot end without expressing my exasperation with some of the critics of randomisation. There may be good reasons to doubt the value of randomisation, but one should not underestimate what Fisher provided. He not only produced a method of allocating treatments to experimental units but also developed an approach to analysing the data, the analysis of variance, that matched analysis to design. If you do not like this and want to propose something better, then you should make its details clear. A word such as *haphazard* is not unambiguous . Instead, provide a precise description of how allocation would proceed with a precise description of the corresponding algorithm for the analysis of resulting datasets and a justification of their use. In this sense, it is almost impossible to argue with many critics of randomisation because it is impossible to establish what they are proposing. Precise recipes and not vague statements are needed if we are to evaluate the counter proposals to randomisation.

Note that the claim I am making is not that randomisation is necessary for all inference but that randomisation makes inference more precise by, in particular, reducing the extent to which we have to model the behaviour of the experimenter as part of the analysis of the experiment.

I do not make a fetish of randomisation, and in fact, there are many points in favour of alternative designs in Worrall's paper on evidence-based medicine [3] with which I agree. One important point is that randomised trials will never be enough to answer all the questions we need answering, and we should not let the fact that we cannot be as precise as we would like to deter us from collecting and using information. Nevertheless, the attacks on randomised clinical trials from certain quarters betray many misunderstandings. As regards the relationship between randomisation and modelling, my philosophy is that the model determines the design rather than *vice versa* [6, 7, 11, 53]. That is to say, I consider that reasonable belief about what is prognostic (and can be observed) guides our choice of intended model. Once we have chosen this model, a number of possible allocations of patients to treatment will be equally efficient. I would be very suspicious of anybody who was not prepared to choose between such allocations at random where this is easy to manage. Such a choice is what we mean by randomisation. Of course, this is a cyclic procedure, and one can iterate several times before settling on a model/design combination. See [54] for a very nice statement from the frequentist point of view.

In short, I see randomisation as being valuable but not essential. It is a means of increasing the precision of our inferences because, amongst other advantages, it reduces the extent to which we have to model the brains of others as part of our model of the world. I personally find this simplification welcome. Where one can randomise, one should. If others do not agree, then I want to see precise descriptions of how they would proceed.

I will finish by giving the last word to Fisher in a passage I have previously quoted elsewhere [11]

> . . . randomisation was never intended from the first moment it was advocated to exclude the elimination from the error of components which could be completely eliminated it only requires that these components shall equally be eliminated from the estimation of error. I often put this by saying that it is only the components which contribute to the actual error of the experiment which need to be randomised to provide an estimate of that error. [14, p. 271] (Letter to Harold Jeffreys 26 September 1938)

Acknowledgements

I am grateful to Alexander Bird, Iain Chalmers, David Colquhoun, Simon Day, Ben Djulbegovic, Frank Harrell, Jeremy Howick, Adam La Caze and anonymous referees for their helpful comments on earlier drafts and to Stephen Evans for helpful discussions. The initial impetus for work on this paper was provided by hearing John Worrall talk at the conference on philosophy and statistics organised by Deborah Mayo and Aris Spanos at the London School of Economics in 2010, and I am grateful to the speaker and organisers. A version of the paper was presented at the MRC clinical trials hub network conference in Bristol in October 2011, and I am grateful to the organisers for the invitation to present this work. Finally, I thank my previous employer the University of Glasgow and my current employer CRP-Santé for the support.

References

1. Howson C, Urbach P. *Scientific Reasoning: The Bayesian Approach*. Open Court: La Salle, 1989.
2. Urbach P. Randomization and the design of experiments. *Philosophy of Science* 1985; **52**:256–273.
3. Worrall J. What evidence is evidence based medicine. *Philosophy of Science* 2002; **69**:S316–S330.
4. Worrall J. Why there's no cause to randomize. *British Journal for the Philosophy of Science* 2007; **58**:451–488.
5. Borgerson K. Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine* 2009; **52**:218–233.
6. Senn SJ. Added values: controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine* 2004; **23**:3729–3753.
7. Senn SJ. Baseline balance and valid statistical analyses: common misunderstandings. *Applied Clinical Trials* 2005; **14**:24–27.
8. Senn SJ. Covariate imbalance and random allocation in clinical trials [see comments]. *Statistics in Medicine* 1989; **8**:467–475.
9. Senn SJ. Testing for baseline balance in clinical trials. *Statistics in Medicine* 1994; **13**:1715–1726.
10. Fisher RA. The design of experiments. In *The Design of Experiments*, Bennett H (ed.). Oxford: Oxford City; 1990.
11. Senn SJ. Fisher's game with the devil. *Statistics in Medicine* 1994; **13**:217–230.
12. Senn SJ. *Dicing with Death*. Cambridge University Press: Cambridge, 2003.
13. Lindley DV. A Bayesian lady tasting tea. In *Statistics: An Appraisal*, David HA, David HT (eds). Iowa State Press: Ames, 1984; 455–485.
14. Bennett JH. *Statistical Inference and Analysis Selected Correspondence of R.A. Fisher*. Oxford University Press: Oxford, 1990.
15. Blackmore S. The lure of the paranormal. *New Scientist* 1990; **22**:62–65.
16. Lesaffre E, Senn S. A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine* 2003; **22**:3583–3596.
17. Senn SJ, Anisimov VV, Fedorov VV. Comparisons of minimization and Atkinson's algorithm. *Statistics in Medicine* 2010; **29**:721–730.
18. Senn SJ. *Statistical Issues in Drug Development*. Wiley: Hoboken, 2007.
19. Atkinson AC. Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in Medicine* 1999; **18**:1741–1752.
20. Burman C-F. *On Sequential Treatment Allocations in Clinical Trials*. Chalmers University of Technology: Gothenburg, 1996.
21. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* 1976; **34**:585–612.
22. Senn S. Modelling in drug development. In *Simplicity, Complexity and Modelling*, Christie M, Cliffe A, Dawid AP, Senn S (eds). Wiley: Chichester, 2011; 35–49.
23. Senn S, Anisimov VV, Fedorov VV. Comparisons of minimization and Atkinson's algorithm. *Statistics in Medicine* 2010; **29**:721–730.
24. Senn SJ. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* 1998; **17**:1753–1765; discussion 1799–1800.
25. Cox DR. *Planning of Experiments*. John Wiley: New York, 1958.
26. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice*. Wiley: New York, 2002.
27. Nelder JA. 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* 1965; **283**:147–162.
28. Nelder JA. 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* 1965; **283**:163–178.
29. Senn SJ, Nelder J. From general balance to generalised models (both linear and hierarchical). In *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS*, Adams NM, Crowder MJ, Hand DJ, Stephens DA (eds). Imperial College Press: London, 2004; 1–11.
30. Bailey RA. Principles of designed experiments in J. A. Nelder's papers. In *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS*, Adams N, Crowder M, Hand D, Stephens D (eds). Imperial College Press: London, 2004; 171–194.
31. Senn SJ. *Cross-over Trials in Clinical Research*, 2nd ed. Wiley: Chichester, 2002.
32. Hills M, Armitage P. The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology* 1979; **8**:7–20.
33. Fisher RA. *The Design of Experiments*. Oliver and Boyd: Edinburgh, 1935.
34. Pitman EJG. Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society, Supplement* 1937; **4**:119–130.
35. Colquhoun D. *Lectures on Biostatistics*. Clarendon Press: Oxford, 1971.
36. Grieve AP. A Bayesian analysis of the two-period crossover design for clinical trials. *Biometrics* 1985; **41**:979–990.
37. Teather D, Morrey GH. Bayesian methods in cross-over trials. *Byocybernetics and Biomedical Engineering* 1995; **15**:41–52.
38. Fisher RA. Statistical methods for research workers. In *Statistical Methods, Experimental Design and Scientific Inference*, Bennet JH (ed.). Oxford University: Oxford, 1925.
39. Cumberland WG, Royall RM. Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society Series B- Methodological* 1988; **50**:118–124.
40. Royall RM, Cumberland WG. Conditional coverage properties of finite population confidence-intervals. *Journal of the American Statistical Association* 1985; **80**:355–359.

41. Cox DR, McCullagh P. Some aspects of analysis of covariance. *Biometrics* 1982; **38**:541–554.
42. Popper SJ. The Gauss-Markov theorem and random regressors. *The American Statistician* 1991; **45**:269–272.
43. Senn SJ, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine* 2007; **26**:5529–5544.
44. Fang BQ, Dawid AP. Nonconjugate Bayesian regression on many variables. *Journal of Statistical Planning and Inference* 2002; **103**:245–261.
45. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics* 1974; **15**:443–453.
46. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; **31**:103–115.
47. McEntegart D. The pursuit of balance using stratified and dynamic randomization techniques: an overview. *Drug Information Journal* 2003; **37**:293–308.
48. Buyse M, McEntegart D. Achieving balance in clinical trials. *Applied Clinical Trials* 2004; **13**:36–40.
49. Taves DR. Rank-minimization with a two-step analysis should replace randomization in clinical trials. *Journal of Clinical Epidemiology* 2012; **65**:3–6.
50. Rosenberger WF, Sverdlov O, Hu F. Adaptive randomization for clinical trials. *Journal of Biopharmaceutical Statistics* 2012; **22**:719–736.
51. Day S, Groulin J-M, Lewis JA. Achieving balance in clinical trials. *Applied Clinical Trials* 2005; **14**:24–26.
52. Atkinson AC. The comparison of designs for sequential clinical trials with covariate information. *Journal of the Royal Statistical Society Series A-Statistics in Society* 2002; **165**:349–373.
53. Senn SJ. *Statistical Issues in Drug Development*, 2nd ed. Wiley: Hoboken, 2007.
54. Bailey RA. A unified approach to design of experiments. *Journal of the Royal Statistical Society Series A-Statistics in Society* 1981; **144**:214–223.