Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments

Iain Chalmers

Histories of clinical trials have recorded and analysed the development of quantification in therapeutic evaluation, the emergence of probabilistic thinking, the application of statistical methods and theory, and the sociology, ethics and politics of clinical trials; but it is surprising that they only rarely identify as a distinct theme the development of efforts to control biases. An exception is Kaptchuk's recent account of the history of blinding and placebos for reducing observer biases. In this complementary paper I introduce and discuss some milestones between 1662 and 1948 in the development of methods to control selection biases when assembling therapeutic comparison groups, to ensure, as far as possible, that 'like is compared with like'.

In the paper I note (i) that treatment allocation based on strict alternation abolishes selection bias as effectively as treatment allocation based on strict random allocation; (ii) that use of schedules based on random numbers is more likely to prevent foreknowledge of allocation schedules, and thus the risk of introducing selection bias at the point of recruitment to trials; (iii) that a concern to conceal allocation schedules was the rationale for using schedules based on random numbers in the Medical Research Council trials of vaccination for whooping cough and streptomycin for pulmonary tuberculosis; and (iv) that the introduction of allocation concealment more than half a century ago remains the most recent substantive milestone in the history of efforts to control selection biases in therapeutic experiments.

Accepted 10 January 2001

Control of bias: insufficiently explored in existing histories of controlled trials

The results of different studies addressing similar questions about the effects of treatments often vary. One cause of this variation is the play of chance, and there are ways of assessing the likelihood that this is the explanation. Another cause of variation is the extent to which biases of various kinds have been avoided. Although it is more difficult to be confident that this is an explanation for variation, it is widely accepted that, when possible, steps should be taken to avoid biases. In particular, therapeutic studies should be designed to control biases when comparison groups are being assembled (control of selection bias), and when outcomes are being assessed (control of observer bias).

Analyses comparing therapeutic studies that have addressed similar questions, but which have varied in the extent to which biases are likely to have been controlled, often (but not always) show differences in the results. Crucially, however, it is not possible to predict when differences attributable to uncontrolled biases will emerge, let alone to predict the direction and magnitude of any differences that are found. Kunz and Oxman[1] have referred to this as 'the unpredictability paradox'. For example, Schulz *et al.*[2] compared the results of studies in which rigorous steps had been taken to control bias with the results of those revealing evidence that inadequate methodological precautions had been taken. On average, the methodologically

UK Cochrane Centre, NHS Research and Development Programme, Summertown Pavilion, Middle Way, Oxford OX2 7LG, UK. E-mail: ichalmers@cochrane.co.uk

This paper is dedicated to Professor Jerry Morris on the occasion of his 90th birthday, with affection and gratitude for having given me the most stimulating and refreshing year of formal education of my life, and with special thanks for encouraging me to challenge authority.

inferior studies yielded estimates of treatment effects that were exaggerated by 40%; but 6 of the 33 component analyses suggested that biases had led to estimates distorted in the opposite direction. Further research may clarify when estimates of treatment effects derived from studies varying in design features intended to control biases can be confidently predicted to be similar. For the time being, however, it remains important to design therapeutic studies in ways that can be expected to minimize biases, particularly because the costs to patients of biased estimates of the effects of treatments can be substantial.

Given the practical implications of therapeutic studies for patients, it is not surprising that historians have shown interest in their evolution. Histories of clinical trials have recorded and analysed the development of quantification in therapeutic evaluation, the emergence of probabilistic thinking, the application of statistical methods and theory, and the sociology, ethics and politics of clinical trials. All these dimensions of the history of clinical trials are important and interesting. It is surprising, however, that existing histories only rarely identify—as a distinct theme—the development of efforts to control biases.

An important exception is Kaptchuk's impressive account of the history of blinding and placebos—the method used to reduce the biases in assessing treatment outcome which can result from knowledge of the identity of the treatments being assessed.[3] Early examples of efforts to reduce these 'observer biases' include late 18th century blinded assessments of the purported therapeutic properties of 'magnetism'. Thus 'Mesmerism' was assessed by a French royal commission headed by Benjamin Franklin[4,5] and 'Perkinism' by John Haygarth,[6] a physician in Bath, England.

By contrast with the attention that Kaptchuk has given to the history of control of observer biases, the history of methods to control selection bias in assembling comparison groups in prospective therapeutic experiments has been less satisfactory. This seems to have resulted partly from a tendency to ascribe a special status to the use of random allocation in creating comparison groups, and, as a result, a tendency to suggest that the influence of the statistical theories of Ronald Fisher in the middle of the 20th century were of seminal importance. I believe such accounts to be mistaken.[7] The history of efforts to make fair treatment comparisons in medicine predates Fisher by centuries. In fact, until relatively recently, examples of steps taken to generate unbiased comparison groups in therapeutic experiments reflected—not a preoccupation with the statistics of randomness—but a concern to make fair, unbiased comparisons. As Richard Doll noted recently in respect of the Medical Research Council randomized trials of vaccination for whooping cough[8] and streptomycin for pulmonary tuberculosis,[9] 'Randomization was introduced to control selection biases, not for any esoteric statistical reason.'[10]

Another problem in previous histories of clinical trials is that there has been insufficient acknowledgement that there are two essential steps in the methods required to ensure that like will be compared with like.[2] The first step involves deciding how the allocation schedule will be generated. Unbiased comparison groups can result either from using schedules based on *random processes* (coin tosses, selection of different coloured beads from an urn, reference to random number tables, and so on), or by using *unbiased systematic processes*, such as strict alternation or rotation of patients in a consecutive series to one of two or more comparison groups.[11,12] As Peter Armitage has noted (personal communication, 31 July 2000):

> 'In principle, alternation can go wrong if the successive responses are not statistically independent, not so much in producing a bias but rather in giving the wrong error variance and thus invalidating the tests, etc. If, for instance, there is a trend in responses, the true error variance will be smaller than that given by the usual formulae, which assume randomization: treatment groups are in fact more alike than would be the case with randomization. If, on the other hand, there is negative correlation between successive responses, the error variance will be underestimated—an error in the opposite direction. I think it is reasonable to ignore the second of these possibilities in clinical trials, although the first may be present in some cases. I doubt whether the effect would be important.'

Whichever method has been used to generate the allocation schedule—strict randomization or strict alternation or rotation—a further essential step in creating unbiased comparison groups relates to the application of the schedule, in practice. Regardless of whether the schedule has been based on a random process or on alternation, any departure from strict adherence to it is likely to introduce bias.[2] For this reason, it is important to conceal the allocation schedule from those (including patients) involved in assessing eligibility to participate in a trial and, thereby, the group to which eligible people will be allocated. In practice, allocation schedules based on random numbers (particularly if fairly complex schedules have been prepared) are more likely to remain concealed than schedules based on alternation, so the former are almost always to be preferred.

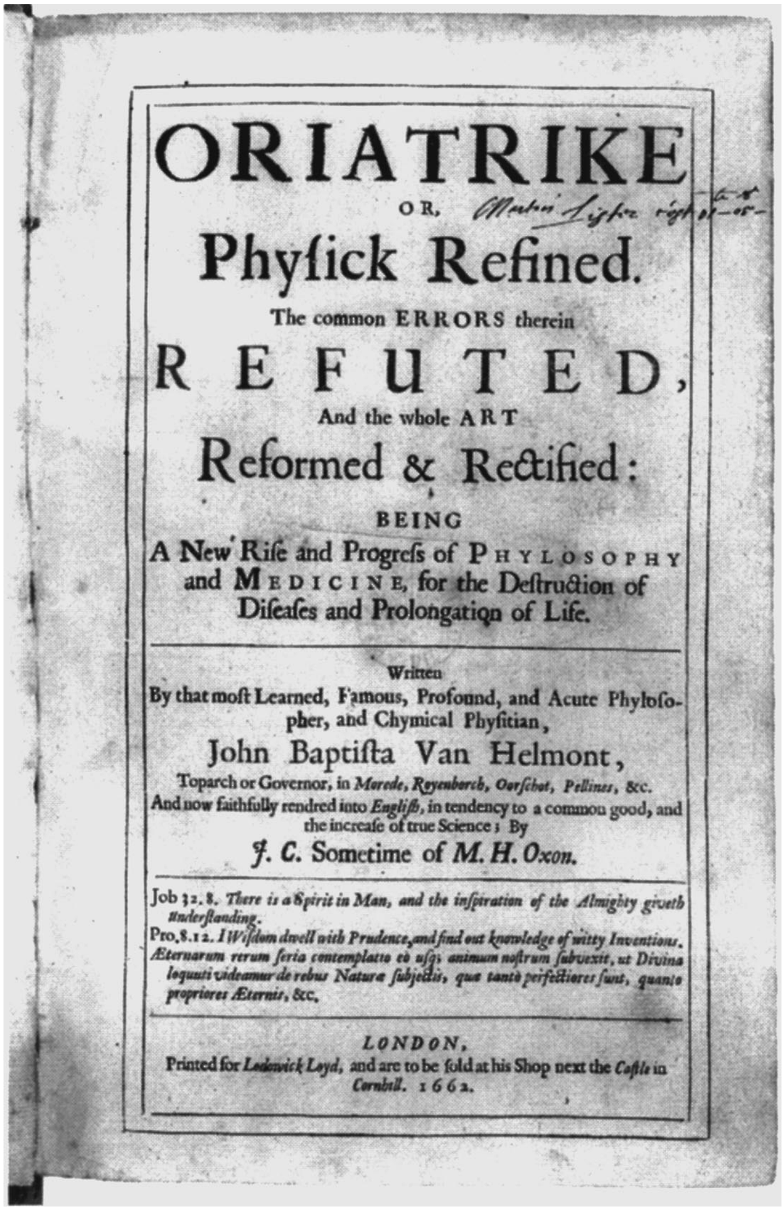
The remainder of this paper documents some milestones in the evolution of these two essential steps in creating unbiased comparison groups in therapeutic experiments.

Milestones in the use of potentially unbiased allocation schedules

The concept of making fair comparisons by taking steps to ensure that like is compared with like goes back a long way. In the 17th century the Flemish physician Jean Baptiste Van Helmont wrote as follows:

> 'If ye speak truth, Oh ye Schools, that ye can cure any kind of Fevers without evacuation, but will not fear of a worse relapse; come down to the contest ye Humorists: Let us take out of the Hospitals, out of the Camps, or from elsewhere, 200, or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them in Halfes, let us cast lots, that one half of them may fall to my share and the other to yours; I will cure them without bloodletting and sensible evacuation; but do you do as ye know (for neither do I tye you up to the boasting, or of Phlebotomy, or the abstinence from a solutive Medicine) we shall see how many Funerals both of us shall have: But let the reward of the contention or wager, be 300 Florens, deposited on both sides: Here your business is decided.'[13]

We shall probably never know whether practitioners of mainstream 17th century medicine accepted Van Helmont's challenge



ORIATRIKE
OR, Martin Lister [illegible]
Phyſick Refined.
The common ERRORS therein
REFUTED,
And the whole ART
Reformed & Rectified:
BEING
A New Riſe and Progreſs of PHYLOSOPHY
and MEDICINE, for the Deſtruction of
Diſeaſes and Prolongation of Life.

Written
By that moſt Learned, Famous, Profound, and Acute Phyloſopher, and Chymical Phyſitian,
John Baptiſta Van Helmont,
Toparch or Governor, in *Merode, Royenborch, Oorſchot, Pellines,* &c.
And now faithfully rendred into *Engliſh,* in tendency to a common good, and the increaſe of true Science; By
J. C. Sometime of *M. H. Oxon.*

Job 32. 8. *There is a Spirit in Man, and the inſpiration of the Almighty giveth Underſtanding.*
Pro. 8. 12. *I Wiſdom dwell with Prudence, and find out knowledge of witty Inventions.*
Æternarum rerum ſeria contemplatio eò uſq; animum noſtrum ſubvexit, ut Divina loquuti videamur de rebus Naturæ ſubjectis, quæ tantò perfectiores ſunt, quanto propriores Æternis, &c.

LONDON,
Printed for *Lodowick Loyd,* and are to be ſold at his Shop next the *Caſtle* in
Cornhill. 1662.

Figure 1 Title page from *Oriatrike or Physick Refined* ... by J B Van Helmont, 1662.
Courtesy of the Bodleian Library, University of Oxford. Reference (shelfmark) Lister D.46

to find out whether they were killing some of their patients with bleeding and purging; but the gauntlet he threw down to the 'Humorists' contains two points that remain as important today as they were then. First, well-intentioned practitioners sometimes inadvertently do their patients more harm than good, so they have a professional responsibility to identify which treatments are likely to help patients. Second, when comparing alternative treatments, steps must be taken to ensure that the comparisons are fair—in this case, by casting lots to decide which patients were to be treated with bloodletting and purging, and which with less drastic methods. Indeed, Van Helmont's proposal to cast lots reflected a much older tradition in many cultures and circumstances to use this device to make fair decisions.[14] It is not clear whether Van Helmont was proposing casting lots to decide how to assign individual patients or one of two similar groups (clusters, in contemporary jargon) to treatment with or without bloodletting and purging. What is important is his recognition that lottery would result in a fair therapeutic contest—in which like would be compared with like.

An awareness of the need to compare like with like in therapeutic comparisons appears to have first gathered momentum in 18th century Britain.[15] The wording in James Lind's famous account of his comparison of six alternative treatments for scurvy is significant:

> 'On the 20th *May* 1747, I took twelve patients in the scurvy, on board the *Salisbury* at sea. *Their cases were as similar as I could have them* (my emphasis). They all in general had putrid gums, the spots and lassitude, with weakness of their knees.

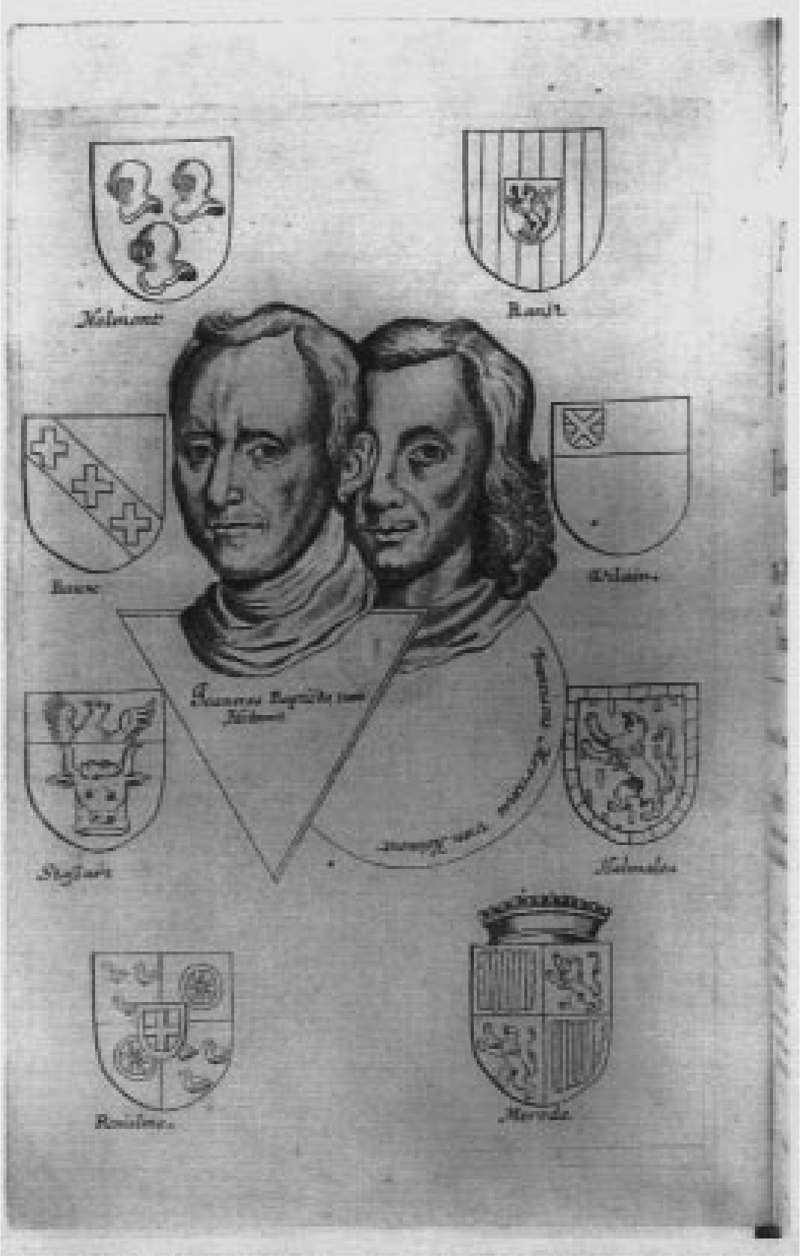


Figure 2 JB and FM Van Helmont 1662. From *Oriatrike or Physick Refined* ... by J B Van Helmont, 1662. Courtesy of the Bodleian Library, University of Oxford. Reference (shelfmark) Lister D.46

> They lay together in one place, being a proper apartment of the sick in the fore-hold; and had one diet common to all, *viz.* water-gruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times puddings, boiled biscuit with sugar, *etc.*; and for supper, barley and raisins, rice and currants, sago and wine, or the like.'[16]

Having thus attempted to ensure that he had assembled comparable patients who were being cared for in similar circumstances, Lind assigned two patients to each of six different 'therapies' (the two sailors who had been prescribed oranges and limes recovered much more quickly than the others).

Lind was a Scottish naval surgeon, and he and others contributed to the growth of quantified approaches to therapeutic evaluation in Britain during the second half of the 18th century and the early years of the 19th century.[15] Over this period, there was increasing recognition of the need for quantitative data to describe progress following treatment, the need to report disappointing as well as heartening results, the inadequacy of small samples, and the need to organize prospective, concurrent comparisons of alternative therapeutic strategies. The growth of hospitals and public dispensaries and the adoption of quantitative approaches within the armed forces had made it more possible to implement many of these steps towards obtaining less biased, empirically based evidence to guide clinical practice. At the height of their activities around 1780, the pioneers in this movement stressed the novelty of these methodological steps and the need to adopt what they referred to as 'medical arithmetic and experimentation'.[15] After about 1800, many civilian and military doctors tended to take the methods for granted as 'standard' techniques, albeit often not always in their entirety.[15]

One consequence of these developments was an increase in the application of ideas about therapeutic evaluation in the British Army and Navy during the early 19th century. James McGrigor, the Duke of Wellington's surgeon-general in the Peninsular War, had recommended that military surgeons use hospital cases for clinical trials.[17] This injunction was reflected in the earliest account of alternate allocation to comparison groups of which I am aware. In his 1816 doctoral thesis at the University of Edinburgh, Alexander Hamilton describes how he and two other army surgeons assessed the effects of bloodletting in an evaluation involving 366 sick soldiers in the Peninsular War:

> 'It had been so arranged, that this number was admitted, alternately, in such a manner that each of us had one third of the whole. The sick were indiscriminately received, and were attended as nearly as possible with the same care and accommodated with the same comforts. One third of the whole were soldiers of the 61st Regiment, the remainder of my own (the 42nd Regiment). Neither Mr. Anderson nor I ever once employed the lancet. He lost two, I four cases; whilst out of the other third (treated with bloodletting by the third surgeon) thirty five patients died.'[18]

Hamilton lived a scandalous life, and his biographer judges this account to be 'a fabrication, made up for the purpose of obtaining a degree and impressing his readers'.[19] Even if the account was fabricated, however, it is remarkable that Hamilton chose to describe the experiment in the terms that he did, particularly if he judged that his description of alternation and standard conditions would impress his examiners and other readers. Even today, it is often difficult to confirm that experiments have taken place in the ways described in reports. We may thus never know whether a controlled experiment showing the adverse effects of bloodletting was conducted during the Peninsular War.[20]

An account of an experiment conducted by another military surgeon, Thomas Graham Balfour, suggests that the notion that alternation could be used to create unbiased comparison groups might have been gaining ground in Britain during the first half of the early 19th century. Balfour tested the claim that homoeopathic belladonna could prevent scarlet fever in the orphan boys in his care at the Royal Military Asylum at Chelsea. His description of the experiment, conveyed in a communication to the author of a book of lectures on the diseases of infancy and childhood published in 1854, must rate as one of the most succinct and careful accounts of a clinical experiment ever written:

> 'There were 151 boys of whom I had tolerably satisfactory evidence that they had not had scarlatina; I divided them in two sections, taking them alternately from the list, to prevent the imputation of selection. To the first section (76) I gave belladonna; to the second (75) I gave none; the result was that two in each section were attacked by the disease. The numbers are too small to justify deductions as to the prophylactic power of belladonna, but the observation is good, because it shows how apt we are to be misled by imperfect observation. Had I given the remedy to all the boys, I should probably have attributed to it the cessation of the epidemic.'[21]

In these four sentences, Balfour addresses the application of eligibility criteria, control of selection bias, the problem of Type 2 statistical errors (that is, false negatives), and the dangers of reliance on uncontrolled case series as a basis for causal inferences about the effects of treatment. Balfour's caution in referring to the numbers of cases being 'too small to justify deductions as to the prophylactic power of belladonna' is especially noteworthy, particularly as William Guy, in a commentary on the experiment, deemed it 'amply sufficient' to demolish the hypothesis that homoeopathy is useful.[22]

The best-known 19th century example of a controlled therapeutic experiment is probably the evaluation of anti-diphtheria serum reported in 1898 by a Danish physician, Johannes Fibiger.[23] The key passage in the report relevant to the creation of unbiased comparison groups reads as follows:

> 'In many cases a trustworthy verdict can only be reached when a large number of randomly selected patients are treated with the new remedy and, at the same time, an equally large number of randomly selected patients are treated as usual … I suggested to Professor Sørensen to treat all patients admitted on the one day with serum, but none of those who were admitted the following day'.[23,24]

The observation of a positive effect of the serum in this trial was particularly important because it confirmed a then recent recognition that bacteria produced toxins causing lesions far removed from the site of infection.

Nine years later, William Fletcher, a District Surgeon in Kuala Lumpur, in the Malay States, described his comparison of cured (polished) and uncured rice among mental patients who were at nutritional risk:

> 'The lunatics are housed in two exactly similar buildings on opposite sides of a quadrangle surrounded by a high wall. On Dec. 5th all the lunatics at that time in the hospital were drawn up in the dining shed and numbered off from the left. The odd numbers were subsequently domiciled in the ward on the east side of the courtyard and no alteration was made in their diet, they were still supplied with the same uncured rice (Siamese) as in 1905. The even numbers were quartered in the ward on the west of the quadrangle and received the same rations as the occupants of the other ward, with the exception that they were supplied with cured (Indian) rice instead of the Siamese variety.'[25]

Fletcher's experiment showed how deaths from beri-beri could be reduced, in practice, but it was an experimentalist working with animals, Christiaan Eijkman, who received the Nobel prize nearly quarter of a century later for showing that the disease was due to thiamine deficiency.[26]

In 1918, Adolf Bingel reported a controlled trial to evaluate the effects of diphtheria antitoxin involving nearly a thousand patients treated in Braunschweig, Germany.[27] Bingel took steps to avoid biases by alternate allocation to an active serum or to an indistinguishable control serum, so his study seems likely to be one of the earliest therapeutic trials designed to control both selection bias and observer bias. Because the original report is in German, it seems worth quoting a translation of two extended passages from it.

> 'After I had treated some adult diphtheria patients with ordinary horse serum in 1911, I began in 1912 to treat alternate adult patients with antitoxin serum and with ordinary serum, exactly in the temporal sequence in which they were admitted to the ward. The children all received antitoxin serum. In the second half of the year 1912 and in the first half of 1913, I gradually lowered the age of those to be treated with ordinary horse serum, and from 1 July 1913 every second case was treated with ordinary horse serum, whether child or adult, regardless of the severity of the illness or the presence of complications. I note that it is absolutely inadmissible to compare the results for different time periods, for example to give antitoxin serum during one year, and then to give only ordinary horse serum during a second year, and then to compare the results. That would lead to seriously wrong conclusions, for in no infectious disease is the nature of the epidemic so changeable as in diphtheria.'

> 'To make the trial as objective as possible, I have not relied on my own judgement alone, but have sought the views of the assistant physicians of the diphtheria ward, without informing them about the nature of the serum under test (namely the ordinary horse serum). Their judgement was thus completely without prejudice. I am keen to see my observations checked independently, and most warmly recommend this "blind" method for the purpose. Even the chief physician may try to draw conclusions about the nature of the serum (unknown to him) that has been used in a particular case: he will be astonished to see how little he is able to do this … Neither I nor my assistants Dr Reusz, Dr Schwab, Dr Weber, Dr Lube could detect a difference between the two sera. Dr Koennecke thought the old (antitoxin) serum had a certain advantage, while Dr Rehder declared that if he were to fall ill, he would wish to be treated with the new (horse) serum. The views of these two gentlemen thus neutralised each other.'[27]

After 1920, reports stating that alternation or 'random' allocation had been used to generate comparison groups become increasingly numerous.[28] Whereas explicit reference to alternation leaves little doubt about the basis of the allocation schedule, use of the word 'random' cannot be assumed to refer to a random process. A 1938 report of controlled trials of cold vaccines by Harold Diehl and his colleagues, for example, is often quoted as an example of the early use of randomization (as opposed to alternation) because the authors stated that participants 'were assigned at random and without selection to a control or to an experimental group'.[29] However, Lance Waller[30] has noted that Diehl had assigned participants to different treatments using alternation in earlier trials of treatments for the common cold published in 1933 and 1935, and that, in an unpublished manuscript summarizing an address he gave in 1941, he had stated that 'At the beginning of the (1938) study, students who volunteered to take these treatments were assigned alternately and without selection to control groups and experimental groups'.

A clinical trial reported by Amberson, McMahon and Pinner in 1931 described how a coin was flipped to decide which of two matched groups of patients would receive an anti-tuberculous drug,[31] in other words, randomizing the clusters, as Van Helmont may have been proposing nearly three centuries earlier. A report by Theobold published in 1937, however, gives a clear description of a random process to generate an allocation schedule for assigning individual participants in a comparative study:

> 'Apparently healthy women, not more than twenty-four weeks' pregnant, were divided by the sister into two groups when they first attended at the clinic, no attention being paid to their previous obstetric histories. They were divided at random in the following manner:
> An equal number of blue and white beads were placed in a box. Each woman accepted for the experiment was asked to draw a bead from the box. Those who drew blue beads were placed in Group A while those who drew white beads were placed in Group B. The beads drawn out were placed in a separate container.
> The patients in Group A were requested to take daily, for the remainder of their pregnancies, calcium lactate 20 grains, vitamin A (11 000 international units) and naturally occurring vitamin D (450 units); while those in Group B served as controls.'[32]

These reports in the 1930s of a single coin toss and selection of different coloured beads from a box are descriptions of random allocation processes. Nevertheless, it is conceivable that the allocation may have been manipulated, and bias thus introduced. Theobald's account comes nearest to providing the kind of detailed description of the process that we require: selection bias would indeed have been abolished successfully if the coloured beads in the box had remained concealed from the women until each was drawn out of the box, and each woman had then joined 'irrevocably' whichever of the two comparison groups the bead colour indicated.

A milestone in preventing foreknowledge of allocation schedules

Recognition of the importance of concealing the allocation schedule from those (including patients) involved in assessing eligibility to participate in a trial and to which group eligible people should be allocated represents a crucially important development in the evolution of methods to create unbiased comparison groups in therapeutic experiments.

A British Medical Research Council trial of serum treatment for lobar pneumonia reported in 1934 had used an (unconcealed) allocation schedule based on alternation,[33] and important imbalances in the characteristics of patients in the treatment and control groups had occurred. In an unpublished critique of the study for the Council, the medical statistician Austin Bradford Hill suggested that greater effort should be taken 'that the division of cases really did ensure a random selection'.[34] This experience appears to have been an important stimulus (Joan Austoker, personal communication) for thinking about how to conceal allocation schedules and thus prevent foreknowledge among those involved in deciding eligibility and assigning treatments in controlled trials.

Reports published in the 1930s and early 1940s may have referred to controlled trials in which steps had been taken to conceal allocation schedules successfully, particularly if placebo controls had been used, for example, in the British Medical Research Council's trial of patulin for the common cold.[35] The earliest clear description of concealment of the allocation schedule of which I am aware, however, is contained in the celebrated 1948 report of the British Medical Research Council's trial of streptomycin for pulmonary tuberculosis:

> 'Determination of whether a patient would be treated by streptomycin and bed-rest (S case) or by bed-rest alone (C case) was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each centre by Professor Bradford Hill; the details of the series were unknown to any of the investigators or to the co-ordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel, and before admission to the streptomycin centre, the appropriate numbered envelope was opened at the central office; the card inside told if the patient was to be an S or a C case, and this information was then given to the medical officer of the centre'.[9]

The only surviving member of the team that designed the streptomycin trial, Philip D'Arcy Hart (who celebrated his 101st birthday on 25 June 2001), did history a great service by reporting in 1999 that Bradford Hill's motivation for replacing alternation with randomization was 'to better conceal the allocation schedule'.[36] This is what Bradford Hill had told William Silverman and me when we visited him on 3 April 1982, and what Guy Scadding (another member of the team that designed the trial) told Mike Clarke and me when we visited him on 10 June 1999.[7]

The reason that the Medical Research Council's controlled trial of streptomycin for pulmonary tuberculosis[9] should be regarded as a landmark is thus not, as is often suggested, because random number tables were used to generate the allocation schedule (as shown above, random allocation had been used at least a decade earlier). Rather it is because of the clearly described precautions that were taken to conceal the allocation schedule from those involved in entering patients. The results of the streptomycin trial would have been no less valid if the trial had used a system of alternation as a basis for the allocation schedule and—against the odds—had succeeded in concealing this from those taking decisions about eligibility and allocation of patients. In view of the methodological importance of allocation concealment, it is surprising that an unambiguous term to distinguish the process from other design features of controlled trials was only introduced very recently.[2,37–41]

Discussion

It is surprising that histories of controlled trials only rarely identify as a distinct theme the development of efforts to control biases. An exception is Kaptchuk's recent account of the history of blinding and placebos for reducing observer biases.[3] In this complementary paper I have introduced and discussed some milestones between 1662 and 1948 in the development of methods to control selection biases when assembling therapeutic comparison groups, to ensure, as far as possible, that 'like is compared with like'. I have noted (i) that treatment allocation based on strict alternation abolishes selection bias as effectively as treatment allocation based on strict random allocation; (ii) that use of schedules based on random numbers is more likely to prevent foreknowledge of allocation schedules, and thus the risk of introducing selection bias at the point of recruitment to trials; and (iii) that a concern to conceal allocation schedules was the rationale for using schedules based on random numbers in the Medical Research Council trials of vaccination for whooping cough and streptomycin for pulmonary tuberculosis.

In my view, Bradford Hill's explicit recognition of the importance of allocation concealment represents the most recent substantive milestone in a history of efforts to create unbiased comparison groups in therapeutic experiments which goes back over three centuries. Although Fisher is properly regarded as a key figure in developing the principles of experimental design,[42] including randomization, his influence on the history of efforts to create unbiased comparison groups in therapeutic experiments is minimal. As he was concerned mainly with agricultural field trials and animal experiments, he had little reason to be concerned about allocation concealment. Indeed, the report of one of the few trials in humans of which Fisher is a co-author does not provide any evidence that allocation concealment was considered or achieved.[43]

Peter Armitage, Bradford Hill's successor at the London School of Hygiene and Tropical Medicine, has noted[44] that Hill (who knew Fisher well) would have been unimpressed by one of Fisher's two reasons for promoting randomization—to guarantee the validity of tests of statistical significance; but he would have wholeheartedly endorsed the other—to abolish bias from many unmeasured (and often unidentified) factors of prognostic importance.[45]

As Stephen Lock has suggested,[46] Bradford Hill deserved to receive a Nobel Prize for this immensely important methodological contribution to the process of assessing the beneficial and harmful effects of medical care. In Hill's wonderfully readable expository papers on the clinical trial published in the early 1950s,[47,48] he notes the circumstances in which carefully controlled trials are unnecessary; he discusses the ethics of doing and not doing trials; and he covers virtually all the methodological aspects of the subject matter that are judged important today. He even remarks, for those who perceive some antithesis between controlled trials and the collection of qualitative data, that as long as the studies have been appropriately designed to control biases, subjective impressions can be given full weight in analyses of controlled trials.

Since Bradford Hill, there have been no substantive milestones related to the control of selection biases in assembling comparison groups in therapeutic experiments (although there has been important progress in recognizing the need to reduce statistical imprecision, an advance for which Richard Peto[49] deserves special credit). The next substantive milestone in the history of efforts to create unbiased comparison groups may be erected when someone solves the interesting methodological conundrum presented by biases resulting from patient preferences.[45,50] In some circumstances, randomization may lead some participants to be pleased with their allocated treatment and others disappointed, thus generating baseline imbalances in psychological states, which could have implications for interpreting subsequent comparisons between the randomized groups.

Although this paper has documented some milestones in the development of methods to create unbiased comparison groups in therapeutic experiments, controversy about the conduct and interpretation of these studies has existed for at least two centuries,[15] and seems set to continue.[51] Indeed, unless the public begins to demand unbiased, reliable estimates of the effects of treatments from researchers and health professionals,[52–56] there is a real possibility that the next 'milestone' in the history of controlled trials may be a gravestone.

Acknowledgements

I am indebted to Doug Altman, Peter Armitage, Joan Austoker, Mike Clarke, Philip D'Arcy Hart, George Davey Smith, Shah Ebrahim, Richard Doll, Ted Kaptchuk, Stephen Lock, Irving Loudon, Guy Scadding, Ken Schulz, William Silverman, Ben Toth and Ulrich Tröhler for help in preparing this paper; to Andrew Herxheimer for translating passages from Bingel's paper; and to Iain Milne for help in many ways, and particularly for spotting Alexander Hamilton's 1816 report of the use of alternation in Lisa Rosner's biography. I thank the Bodleian Library, University of Oxford, for permission to publish the Frontispiece and title page from Oriatrike of Physick Refined … by JB Van Helmont, 1662. Reference (shelfmark) Lister D.46.

References

[1] Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *Br Med J* 1998;**317:**1185–90.

[2] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273:**408–12.

[3] Kaptchuk TJ. Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine* 1998;**72:**389–433.

[4] Lavoisier A, Franklin B. *Rapport des Commissaires Chargés par le Roi du Magnétisme Animal*. Paris: Imprimerie Royale, 1784.

[5] Franklin B, Majault, Le Roy *et al*. *Report of Dr Benjamin Franklin, and Other Commissioners, Charged by the King of France, with the Examination of Animal Magnetism, as Now Practiced in Paris*. Trans. William Godwin. London: J Johnson, 1785.

[6] Haygarth J. *Of the Imagination, as a Cause and as a Cure of Disorders of the Body: Exemplified by Fictitious Tractors, and Epidemical Convulsions*. Bath: R Crutwell, 1800.

[7] Chalmers I. Why transition from alternation to randomisation in clinical trials was made. *Br Med J* 1999;**319:**1372.

[8] Medical Research Council. The prevention of whooping-cough by vaccination: a Medical Research Council investigation. *Br Med J* 1951;**i:**1463–71.

[9] Medical Research Council. Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *Br Med J* 1948; **ii:**769–82.

[10] Doll R. *Presentation at 'Clinical Trials: Into the New Millennium'*. St Anne's College, Oxford, 25 September 2000.

[11] Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *Br Med J* 1999;**318:**1209.

[12] Altman DG, Schulz KF. Concealing treatment allocation in randomised trials. *Br Med J* (in press).

[13] Van Helmont JA. *Oriatrike, or Physick Refined: The Common Errors Therein Refuted and the Whole Art Reformed and Rectified*. London: Lodowick-Loyd, 1662, p.526.

[14] Silverman WA, Chalmers I. Casting and drawing lots. In: Chalmers I, Milne I, Tröhler U (eds). *Controlled Trials from History*. www.rcpe.ac.uk/controlled_trials. Accessed 24 July 2001.

[15] Tröhler U. *'To Improve the Evidence of Medicine': The 18th Century British Origins of a Critical Approach*. Edinburgh: Royal College of Physicians, 2000.

[16] Lind J. *A Treatise of the Scurvy. In Three Parts. Containing an Inquiry into the Nature, Causes and Cure, of that Disease. Together with a Critical and Chronological View of what has been Published on the Subject*. Edinburgh: Printed by Sands, Murray and Cochran for A Kincaid and A Donaldson, 1753, pp.145–46.

[17] Blanco RL. *Wellington's Surgeon General: Sir James McGrigor*. Durham, NC: Duke University Press, 1974, p.127.

[18] Hamilton AL. *Dissertatio Medica Inauguralis De Synocho Castrensi*. Edinburgh: J Ballantyne, 1816.

[19] Rosner L. *The Most Beautiful Man in Existence*. Philadelphia: University of Pennsylvania Press, 1999, p.133.

[20] Milne I, Chalmers I. Tackling bias in assessing the effects of health care interventions: early contributions from T Graham Balfour and Alexander Lesassier Hamilton. *Proceedings of the Royal College of Physicians of Edinburgh*. In press.

[21] Balfour TG. Quoted in West C. *Lectures on the Diseases of Infancy and Childhood.* London: Longman, Brown, Green and Longmans, 1854, p.600.

[22] Guy WA. The numerical method and its application to the science and art of medicine. *Br Med J* 1860;**No. CLXXXVI:**553–55 (21 July).

[23] Fibiger J. Om serumbehandling af difteri [On treatment of diptheria with serum]. *Hospitalstidende* 1898;**6:**309–25.

[24] Hróbjartsson A, Gøtzsche P, Gluud C. The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria. *Br Med J* 1998;**317:**1243–45.

[25] Fletcher W. Rice and beri-beri: preliminary report on an experiment conducted in the Kuala Lumpur Insane Asylum. *Lancet* 1907;**ii:**776–79.

[26] Fraser DW. Vitamins and vitriol: W.L. Braddon's epidemiology of beriberi. *Am J Epidemiol* 1998;**148:**519–27.

[27] Bingel A. Über Behandlung der Diphtherie mit gewöhnlichem Pferdeserum. *Deutsch Arch Klin Med* 1918;**125:**284–332.

[28] Chalmers I, Milne I, Tröhler U (eds). *Controlled Trials from History*. www.rcpe.ac.uk/controlled_trials. Accessed 24 July 2002.

[29] Diehl HS, Baker AB, Cowan DW. Cold vaccines: an evaluation based on a controlled study. *JAMA* 1938;**111:**1168–73.

[30] Waller LA. A note on Harold S. Diehl, randomization, and clinical trials. *Cont Clin Trials* 1997;**18:**180–83.

[31] Amberson JB, McMahon BT, Pinner M. A clinical trial of sanocrysin in pulmonary tuberculosis. *Am Rev Tuberc* 1931;**24:**401–35.

[32] Theobald GW. Effect of calcium and vitamin A and D on incidence of pregnancy toxaemia. *Lancet* 1937;**ii:**1397–99.

[33] Medical Research Council Therapeutic Trials Committee. The serum treatment of lobar pneumonia. *Br Med J* 1934;**i:**231–49.

[34] Medical Research Council 1487, VI: A. Bradford Hill. Serum treatment of pneumonia. 22 December 1933. Cited in: Austoker J, Bryder L. The National Institute for Medical Research and related activities of the MRC. In: Austoker J, Bryder L (eds). *Historical Perspectives on the Role of the MRC*. Oxford: Oxford University Press, 1989, pp.35–57.

[35] Medical Research Council. Clinical trial of patulin in the common cold: report of the Patulin Clinical Trials Committee, Medical Research Council. *Lancet* 1944;**ii:**373–75.

[36] D'Arcy Hart P. A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. *Br Med J* 1999;**319:**572–73.

[37] Schulz KF, Chalmers I, Grimes DA, Altman DG Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994;**272:**125–28.

[38] CONSORT Group. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;**276:**637–39.

[39] Moher D, Pham B, Jones A *et al.* Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;**352:**609–13.

[40] Haynes RB. Incorporating allocation concealment and blinding in randomised controlled trials. *Evidence-Based Medicine* 2000;**5:**38.

[41] Last J. *Dictionary of Epidemiology. 4th Edn*. Oxford: Oxford University Press, 2000.

[42] Fisher RA. *The Design of Experiments*. London: Oliver and Boyd, 1935.

[43] Atkins WRG, Fisher RA. The therapeutic use of Vitamin C. *J Roy Army Med Corps* 1944;**83:**251–52.

[44] Armitage P. Randomisation and alternation: a note on Diehl *et al*. (1938). In: Chalmers I, Milne I, Tröhler U (eds). *Controlled Trials from History*. www.rcpe.ac.uk/controlled_trials. Accessed 24 July 2001.

[45] Kleijnen J, Gøtzsche P, Kunz RH, Oxman AD, Chalmers I. So what's so special about randomisation? In: Maynard A, Chalmers I (eds). *Non-random Reflections on Health Services Research: On the 25th Anniversary of Archie Cochrane's Effectiveness and Efficiency*. London: BMJ Books, 1997, pp.93–106.

[46] Lock S. The randomised controlled trial—a British invention. In: Lawrence G (ed.). *Technologies of Modern Medicine*. London: Science Museum, 1994, pp.81–87.

[47] Bradford Hill, A. The clinical trial. *Br Med Bull* 1951;**7:**278–82.

[48] Bradford Hill, A. The clinical trial. *N Engl J Med* 1952;**247:**113–19.

[49] Peto R. Clinical trial methodology. *Biomedicine Special Issue* 1978;**28:** 24–36.

[50] McPherson K, Chalmers I. Incorporating patient preferences into clinical trials. *Br Med J* 1998;**317:**78.

[51] Swales J. The troublesome search for evidence: three cultures in need of integration. *J R Soc Med* 2000;**93:**402–07.

[52] Chalmers I. What do I want from health research and researchers when I am a patient? *Br Med J* 1995;**310:**1315–18.

[53] Hart JT. Response rates in south Wales 1950–96: changing requirements for mass participation in human research. In: Maynard A, Chalmers I (eds). *Non-random Reflections on Health Services Research: On the 25th Anniversary of Archie Cochrane's Effectiveness and Efficiency*. London: BMJ Books, 1997, pp.31–57.

[54] Oliver S. Exploring lay perspectives on questions of effectiveness. In: Maynard A, Chalmers I (eds). *Non-random Reflections on Health Services Research: On the 25th Anniversary of Archie Cochrane's Effectiveness and Efficiency*. London: BMJ Books, 1997, pp.272–91.

[55] Chalmers I. Assembling comparison groups to assess the effects of health care. *J R Soc Med* 1997;**90:**379–86.

[56] Chalmers I, Lindley R. Double standards on informed consent to treatment. In: Doyal L, Tobias JS (eds). *Informed Consent: Respecting Patients' Rights in Research, Teaching and Practice*. London: BMJ Publications, 2000, pp.266–75.