# Bias in treatment trials

*Huw TO Davies*

***Rigorous treatment trials exhibit a number of key features to guard against bias. Randomization, blinding (of patients, care staff and assessors), full follow-up on all patients, and 'intention-to-treat' analyses all contribute to removing bias so that any true treatment effect can be revealed. This article outlines the rationale behind these features and advises how bias in treatment trials can be assessed.***

Clinicians, health-care managers and even policy makers are increasingly questioning the effectiveness of health-care interventions. At the same time, a growing reservoir of research evidence provides some guidance. However, not all research evidence on the effectiveness of treatments is a reliable guide: treatment trials can be flawed in design, execution or analysis. This article outlines the key areas where bias may arise in treatment trials. It is in these areas that discerning readers will search for problems when appraising the quality of research evidence.

## WHY TRIAL?

Personal experience provides a poor guide for physicians when it comes to assessing the effectiveness of treatment interventions (Crombie and Davies, 1996; Davies and Nutley, 1999). Patients get better by themselves, often despite not because of interventions. The placebo effect (benefits arising which are non-specific to the therapy) and the clouding effects of chance variability also conspire to obscure any true estimate of treatment effect.

History has shown that treatment studies without appropriate controls are apt to mislead (Gilbert et al, 1977; Pocock, 1983). Historical controls (comparing current treatment success with previous patient outcomes), and concurrent non-randomized controls (comparing new treatments with existing practice but with no experimental allocation of patients between these two groups), both prove susceptible to bias. Control groups need to be concurrent and randomly assigned.

**Dr Huw TO Davies** is Reader in Health Care Policy and Management, Department of Management, University of St Andrews, St Andrews, Fife KY16 9AL

## CONCEALED RANDOMIZATION

Random allocation of patients to either the new treatment under test or some comparative therapy is the cornerstone of a good quality trial. Such random allocation provides no guarantees. However, given sufficient patient numbers, randomization makes it likely that the two groups will be evenly matched — both for factors known to be important in determining outcome (e.g. age, severity) and also for those factors whose prognostic value may be unknown but nonetheless crucial (e.g. genetic makeup). It is randomization that leads (on average) to balanced groups and thus fair comparisons.

Randomization can be achieved in a number of different ways: envelopes containing group allocation may be provided in advance, or clinicians recruiting new patients may telephone a central coordinator to learn the group allocation of the next patient. Thus some methods of randomization are more secure than others; that is, they are less prone to manipulation by clinicians who favour one treatment avenue over another. There is some evidence that randomization that is not tamper proof may not be immune from bias (Schulz et al, 1995). Thus, in appraising the quality of a trial, it may be worth asking not only was the allocation to new treatment and control carried out randomly, but also was the random allocation secure from interference or prediction.

## EQUIVALENCE AT BASELINE

Random allocation of patients to 'new treatment' and 'control' groups will on average lead to fair and balanced groups. This tendency increases as the number of patients increases. However, it is possible that, even when the randomization was well conducted, just by chance, the two groups may not be as evenly balanced as hoped.

Thus, in analysing trial data, it is usual to first of all compare and present the composition of the two groups 'at baseline' (i.e. just after randomization and before treatments have been applied). This comparison seeks reassurance that the two groups are indeed similar on all known patient variables. Any significant differences found between the two groups may need to be taken into consideration when assessing the treatment outcomes.

## BLINDING

The key features of rigorous trials after randomization try to ensure that both groups are handled identically thereafter to prevent any imbalances creeping in. A key tool here is blinding.

'Single blinding' is achieved when patients are unaware of whether they are receiving the new treatment under test or the control treatment. This protects against the placebo effect and patient misreport arising from differential expectations about the likely impact of treatment. Of course, such concealment should also extend to those assessing the treatment outcomes, otherwise knowledge of which treatment group patients are in may bias the assessor's judgment (Noseworthy et al, 1994).

'Double blinding' is achieved when not only are the patients unaware of their treatment group, but so too are their physicians and other care staff. This approach guards against patients being treated differently because of knowledge about their group allocation. Good studies often try to conceal group allocation even during the final analysis (sometimes called 'triple blinding') to prevent any bias creeping in during data analysis and interpretation.

Although blinding in all its forms is an essential bulwark against bias, it may be difficult or impossible to achieve (for example, in trials of physical therapies or surgery; Deyo et al, 1990). Even when possible, blinding may be incomplete, for example, patients may be able to tell which group they are in because of the distinctive characteristics of new and old treatments. Thus presentation of some data on the success or otherwise of blinding (among both patients and staff) may help to clarify the potential for bias.

## CLEAR STUDY POPULATION

The reader of any treatment trial needs to know about the patients studied so that they can decide whether or not the findings are applicable to other patient populations. Thus reports of treatment studies should provide a clear description not only of the patients included in the study but also how they were drawn. That is, they should describe the planned study population, the sites where individuals were recruited into the study, the inclusion or exclusion criteria that were used to signify suitability for the trial, and the number of rejections or refusals.

This kind of information is best presented as a flow chart, as illustrated in the *Journal of the American Medical Association*'s instructions to authors (JAMA, 1998). Close inspection of these numbers allows the reader to determine whether or not the patients studied are likely to be 'typical' and the implications this has for interpretation of any treatment effect found. This form of presentation helps account for any lost patients or transfers between the treatment groups — both of which have implications for the assessment of bias.

## FULL PATIENT FOLLOW-UP INTENTION-TO-TREAT ANALYSIS

Randomization helps lead to balanced groups at baseline. What happens to group composition after this can compromise any balance and thus lead to unfair comparisons.

First, patients lost to follow-up after they have been allocated to groups give cause for concern. Even if similar numbers of patients are lost from each treatment group, the worry is that different sorts of patients may be being lost. Suppose, for example, that patients with the more severe disease drop out from one group and those with milder disease drop out from the other. This would lead to imbalances between the groups and would bias the findings.

Similarly, concerns arise over patients who swap between treatment groups, i.e. those patients who, for whatever reason, subsequently receive care different from that originally allocated to them.

Because of these concerns about the balance between groups being upset, the key analysis should compare the groups as originally allocated. This is called an 'intention-to-treat analysis'. Such an analysis is the simplest way to prevent bias creeping in from differential loss of patients or differential movement between treatment groups. If no data on treatment outcomes are available for some patients, then an analysis that assumes that these outcomes were poor can help clarify whether or not the overall findings are susceptible to bias from these losses.

Intention-to-treat analyses answer a very precise question. They are not trying to answer the question 'Does treatment A produce better outcomes than treatment B?'. Instead, they address the much more pragmatic issue of 'Does a decision to use treatment A produce better outcomes on aggregate than a decision to use treatment B (even if that decision cannot always be seen through to completion)?'. Thus intention-to-treat analyses answer pragmatic questions of real interest to clinicians faced with making treatment choices.

## THE PLAY OF CHANCE

Preventing bias does not guarantee that trials will always reveal any real underlying treatment effects. Chance variability can conspire to mislead in even the best-planned clinical trials. *Table 1* shows how, just through chance, the findings from any single trial may mismatch the true situation. Trials may mislead in two ways: they may show an apparent effect which is not real, or they may appear to show no difference when in fact one treatment is truly superior in effect.

Detailed guidance on assessing the impact of chance in treatment trials has been given elsewhere in this series (Davies, 1998a). In brief, it is the statistical significance level that tells of the likelihood of making the first error, and it is the power calculation that provides information on the likelihood of the second. There is always a trade-off between these two errors: as we try to avoid one we become more likely to commit the other. The only way out of this bind is an increase in sample size.

A further complication lies in the question 'when is a difference a worth-

**TABLE 1.**
**Possible matches between 'truth' and study findings brought about by chance**

| | | Real situation | |
|---|---|---|---|
| | | No worthwhile difference between new and old treatment | Worthwhile difference between new and old treatment |
| What the study finds | No statistical difference between new and old treatment | Study correctly matches true situation | **Misleading result:** Can happen by chance. Power calculation tells how often |
| | Statistically significant difference between new and old treatment | **Misleading result:** Can happen by chance. Significance level tells how often | Study correctly matches true situation |

while difference?'. This is a question about clinical, as compared to statistical, significance. We cannot easily assess how likely we are to be led astray by chance until we have answered that question. Searching for small effects is more demanding than looking for large differences. What is more, chance variability can only be assessed once bias has been dismissed as a major factor in the findings (Brennan and Croft, 1994). A number of useful sources provide further elaboration of the tricky issues surrounding sample size, power and the assessment of the play of chance (Altman and Bland, 1995; Eypasch et al, 1995; Florey, 1993; Gardner and Altman, 1986).

## RELEVANT OUTCOMES AND SUBGROUP ANALYSES

Statistical considerations also impact on the range of variables that can be explored for differences in clinical trails. Good clinical trials specify in advance the key outcome of interest. This outcome is then the primary focus of the analysis (usually termed 'the primary endpoint'). This pre-specification is necessary because otherwise the basis for the statistical tests used to distinguish potentially real effects from chance differences is undermined.

For the same reason sub-group analyses should always be specified in advance — and be limited to few comparisons with some underlying biological plausibility. Comparing many different subgroups in an ad hoc manner is likely to lead to spurious findings.

## CONCLUSIONS

This article has highlighted those areas to examine when going in search of bias in treatment trials. Many other equally important issues also arise when examining reports of clinical trials.

Are the benefits, such as they might be, worth the costs (Neilson and Davies, 1998)? Were all the possible benefits enumerated and assessed (Neilson and Davies, 1999)? What about issues of safety (Eypasch et al, 1995)? Were patient preferences incorporated (Tavakoli et al, 1999)? Were the benefits portrayed in both absolute and relative terms (Davies, 1998b)?

Assessment of bias must take place before any statistical judgments on the effects can be made (Brennan and Croft, 1994). A number of places in the design, execution and analysis of trials need careful inspection: randomization and its tamperproof nature, blinding and the success of concealment achieved in practice, the completeness of follow-up and the use of an intention-to-treat analysis, and the appropriate handling of chance variability with a clear declaration of the power of the study to uncover real and worthwhile effects. Further guidance on these issues is available from a number of published checklists (Crombie, 1996; Guyatt et al, 1993, 1994; Sackett et al, 1997).

Finally, individual treatment trials should not be viewed in isolation. As more and better quality systematic reviews and meta-analyses are becoming available, single trials need to be interpreted in the context of this wider body of information (Davies and Crombie, 1999). HM


Altman DG, Bland JM (1995) Absence of evidence is not evidence of absence. *Br Med J* **311:** 485

Brennan P, Croft P (1994) Interpreting the results of observational research: chance is not such a fine thing. *Br Med J* **309:** 727–30

Crombie IK (1996) *The Pocket Guide to Critical Appraisal.* BMJ Publishing, London

Crombie IK, Davies HTO (1996) *Research in Health Care: Design Conduct and Interpretation of Health Services Research.* John Wiley & Sons, Chichester

Davies HTO (1998a) Assessing chance variability in treatment trials. *Hosp Med* **59:** 650–2

Davies HTO (1998b) Interpreting measures of treatment effect. *Hosp Med* **59:** 499–501

Davies HTO, Crombie IK (1999) Getting to grips with systematic reviews and meta-analysis. *Hosp Med* **59:** 955–8

Davies HTO, Nutley SM (1999) The rise and rise of evidence in health care. *Public Money & Management* **19:** 9–16

Deyo RA, Walsh NE, Schoenfeld LS, Ramamurthy S (1990) Can trials of physical treatments be blinded? The example of transcutaneous electrical nerve stimulation for chronic pain. *Am J Phys Med Rehabil* **69:** 6–10

Eypasch E, Lefering R, Kum CK, Troidl H (1995) Probability of adverse events that have not yet occurred: a statistical reminder. *Br Med J* **311:** 619–20

Florey CdV (1993) Sample size for beginners. *Br Med J* **306:** 1181–4

Gardner MJ, Altman DG (1986) Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *Br Med J* **292:** 746–50

Gilbert JP, McPeek B, Mosteller F (1977) Statistics and ethics in surgery and anesthesia. *Science* **198:** 684–9

Guyatt GH, Sackett DL, Cook DJ (1993) Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* **270:** 2598–601

Guyatt GH, Sackett DL, Cook DJ (1994) Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA* **271:** 59–63

JAMA (1998) Instructions for authors. *JAMA* **279:** 71

Neilson AR, Davies HTO (1998) Interpreting reported health care costs. *Hosp Med* **59:** 803–6

Neilson AR, Davies HTO (1999) Interpreting reported health care benefits. *Hosp Med* **60:** 134–7

Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R (1994) The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology* **44:** 16–20

Pocock SJ (1983) *Clinical Trials: A Practical Approach.* John Wiley & Sons, Chichester

Sackett DL, Richardson WS, Rosenberg W, Haynes RB (1997) *Evidence Based Medicine: How to Practice and Teach EBM.* Churchill Livingstone, London

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

Tavakoli M, Davies HTO, Thomson R (1999) Aiding clinical decisions with decision analysis. *Hosp Med* **60:** 444–7


### KEY POINTS

- The key objective in any treatment trial is for fair comparisons.
- Initial balance between treatment groups is best achieved by random allocation between new and old treatments. Randomization should be concealed to avoid manipulation.
- Treatment groups should be assessed for their equivalence at baseline as randomization does not guarantee balance.
- Ensuring fair comparison thereafter depends on blinding all participants (patients, health-care staff, assessors and even analysts) to the individuals' group allocation.
- The success of the blinding should be assessed to investigate possible sources of bias.
- As far as possible, all patients should be followed up, and an 'intention-to-treat analysis' on the primary endpoint should be performed. Subgroup analyses need to be treated with circumspection.
- Chance variability can mislead by creating spurious effects or hiding real ones. Before chance effects can be estimated, bias must be ruled out.