Methodology

Considerations for the design and analysis of experimental studies in physical activity and exercise promotion: advantages of the randomised controlled trial

Sharon See Tai, Steve Iliffe

A review of all published articles in the *British Journal of Sports Medicine* from 1991 to 1995¹ showed that a wide range of subjects had been covered, but that randomised controlled trials (RCTs) accounted for only 3% of the total number of articles. Common study designs for evaluating (exercise) interventions are “before and after” or “uncontrolled” trials. In these, outcomes of interest in subjects are measured before and after the intervention or exposure—for example, saliva tests before and after an exercise test, or the trials may be unmatched or matched case-control studies, in which subjects under exposure or measurement are compared with controls—for example, by comparison of a questionnaire survey completed by athletes and controls.

These designs, though valid and important, are open to potential bias. The scientific outsider may be sceptical about the results and conclusions of such studies in the absence of the authors’ detailed description of—for example, recruitment strategies, inclusion criteria, loss to follow up, and rules for interpretation of results of the intervention. Perhaps more importantly, interventions like exercise programmes or advice on physical activity may seem to produce positive results in research projects, but fail to deliver them in real life because of bias hidden in the original study. This can leave practitioners sceptical about research claims and prone to reject a rigorous approach to the evaluation of new ideas and techniques. Minimising bias is important in making sure that we really know what works.

In this paper we try to indicate possible sources of bias in the design and analysis of experimental studies and remind sports scientists about the benefits of RCTs for reducing bias.

Potential sources of bias

Bias in statistical terms refers to the situation in which the statistical method does not estimate the quantity that is thought to be estimated, or does not test the hypothesis that is thought to be tested.²

A study is vulnerable to bias at all stages of its design, execution and analysis:

- (1) The selection stage—that is, how participants are selected for the study and how they are allocated to the study groups (in controlled designs). Are the subjects who ultimately receive the intervention selected (or self selected) because of the outcome they are likely to get?
- (2) The intervention stage—that is, if interventions are not standardised, crucial components (for example, the charisma of the exercise instructor) may affect the results. In controlled designs can we be sure that the study groups are treated in the same way?
- (3) The assessment of outcome stage—that is, what are the instruments for assessing outcome and is there a possibility of varied interpretation of the results derived from these instruments? Did more than one person assess the outcome? In controlled designs did the assessor know to which study group the participants were allocated?
- (4) The analysis stage—that is, were all enrolled participants accounted for at the end of the study? And in controlled designs were they analysed in the same groups as those to which they were originally allocated?

General considerations for the design and analysis of experimental studies

DEFINING THE RESEARCH QUESTION

What is the specific question to be answered and what instruments are you going to use to measure or answer this question? For example, “Does 18 weeks of weightbearing exercise increase bone mineral density (BMD) in adult men aged 35–64?”. This is a specific research question with a definable intervention, a measurable outcome, and a specifiable subject group.

Suppose “weightbearing exercise” (the intervention) has already been defined in physiological terms. We would need to be aware of possible “black box effects”. Is it the defined weightbearing exercise that alters BMD, or the motivating effect of the exercise instructor, who inadvertently encourages participants to do even more (unrecorded) exercise, or the social

Department of
Primary Care and
Population Sciences,
Royal Free and
University College
London Medical
Schools, Archway
Resource Centre,
Holborn Union
Building, Highgate
Hill, London N19 3UA,
UK
S See Tai
S Iliffe

Correspondence to:
Ms Sharon See Tai
Email: s.seetai@ucl.ac.uk

Accepted for publication
19 January 2000

effect of an exercise class that can change physical activity in a variety of ways? Describing the content of the intervention, and standardising it as much as possible—for example, by having one exercise instructor, or training a small group to the same standard, is important in all research into exercise promotion.³

We would also have to think more carefully about the inclusion criteria for our study group—that is, men. We could choose the more severe end of the population spectrum, which, in our example, would be sedentary subjects who have lower BMD than their more active counterparts⁴ and therefore have more to gain, and who may also respond more positively to an exercise programme.

Subjects who are known to have low BMD and who are already receiving some other form of treatment (for example, extra calcium and vitamin D) may also be included, but the effect of the exercise is confounded. For example, in a two arm study, if one group has exercise and other treatment and the other has no intervention, this study would be evaluating a research question that differs from the case where both groups have the other treatment but only one has the exercise intervention. Any intervention, including participation in a research project, may have an effect (the Hawthorne effect). It is only in the latter example that the precise effect of the exercise component can be isolated.

This research question can be investigated as an uncontrolled or controlled design.

The “before and after” or uncontrolled design

This design has the advantage of direct assessment of outcome on the feature and group of interest. In our example we could recruit a group of inactive volunteers and measure BMD before and after they had participated in the exercise programme. The obvious problem here is that we can determine the change in BMD in the group chosen, but in the absence of a comparison arm we cannot determine the “size of the effect”—the effect relative to that which might have occurred in the absence of any intervention. From the scientific outsider’s point of view one might wonder if some bias was involved in the (self) selection of these participants. Possibly, also, our dedication to the investigation and our enthusiasm for exercise promotion might affect our measurement of the outcomes? An uncontrolled design is not likely to be the best approach to answering our research question, given these potential biases.

The unmatched or matched case-control design

How do we choose a comparable control group? Unless we are specifically interested in, for example, comparing BMD between sedentary subjects recruited from the community and athletes, baseline BMD is likely to be different in the two groups and a comparison of the effect of weightbearing exercise on sedentary subjects (who may be more responsive to an exercise intervention) compared with athletes will be underestimated in this design.

Historical controls are sometimes used, and Pocock discusses in great detail the many sources of bias in this design.⁵ There may be

bias (a) in patient selection—for example, less clearly defined inclusion criteria for the historical controls than for the new subjects; (b) in environmental conditions—for example, interpreting the response to interventions which might have had different criteria for historical subjects. In this case the intervention effect may be overestimated.

Matching is often used to make the groups more “comparable”. How do we choose the criteria with which to match? It would seem sensible to match the factors that we think may influence the effect of the intervention—age, sex, and social class are popular choices. In our example, the current level of physical activity, height, weight, and BMD may also be factors. It is difficult to think of all the possible factors and also to find a sufficient number of controls when several variables have to be matched. These variables are “confounding” variables,⁶ and may be adjusted for in the subsequent statistical analysis, but it is advisable to consider and plan this analysis at the start of the study.

We might also be interested in answering secondary questions—for example, how does weightbearing exercise impact balance and muscle strength, quality of life, and the number of injurious falls? We may wish to compare subgroups—for example, changes in BMD in participants aged under and over 50. These questions should also be considered at the start of the study.

In these designs then, can we be sure that our intervention and control group participants are not selected (or excluded) in such a way that our subsequent evaluation of the effect of the intervention is compromised? Can our research question be answered more convincingly by an RCT? And, is it reasonable and practical to allocate participants randomly to intervention and control groups? If it is, and the biases inherent in research studies can be avoided, how should we design the trial? We need to consider recruitment, inclusion and exclusion criteria, sample size, confounding variables, outcome assessment, and data analysis.

RECRUITMENT AND INCLUSION/EXCLUSION CRITERIA

How and from what source are the subjects going to be recruited? This is not always explicitly explained in published studies, which gives the sceptic the notion that there may be some bias in the way that the subjects are selected.

The timescale for recruitment should be taken into consideration if subjects are invited to participate by letter or opportunistically—for example, through general practice. There is some evidence that low response rates to research questionnaires can be increased by using recorded delivery, if resources are available.⁷ Other strategies to increase the response rate to mailed questionnaires (for example, prepaid envelopes and number of contacts) have also been investigated.^{8–10}

One should be aware that the subject “source” will have implications for the research question and vice versa. That is, the (baseline) potential for change due to the intervention

may be different depending on the source of recruitment. In our example, subjects recruited from a general practice waiting room may have a different baseline BMD (maybe due to concurrent illness that brought them to the doctor) than a community sample recruited from the electoral register.

SAMPLE SIZE CALCULATION

How many subjects will be needed? We calculate the number needed taking into consideration the difference in outcome between the groups (or change in outcome if using an uncontrolled design) that we think it will be important to detect, which will be subsequently detected using a particular statistical technique! In our example we may be interested not only in the change in BMD but also in the rate of injurious falls and the change in perceived health status of participants using the SF36 (a measure of perceived health status), so we should calculate the number needed to measure important changes on all these outcomes. Several texts^{5 6 11–13} describe the methods used to calculate the number of subjects needed for quantitative outcomes—for example, BMD, or qualitative outcomes—for example, the proportion of subjects whose injurious falls were decreased by one during the study period.

Epi Info software¹⁴ is in the public domain (free to copy and distribute) and can be downloaded from <http://www.cdc.gov/epo/epi/downepi6.htm>. It can be used to calculate sample size for both these methods. For complex trials with three arms (for example, intervention 1: weight bearing exercise, intervention 2: high impact aerobics, control: no exercise) or mixed model designs Cohen provides tables for sample size and power.¹⁵

So far we have assumed that the unit of analysis is the individual subject. In some cases it is necessary to randomise groups or centres—for example, leisure centres. Power is reduced under group randomisation and the minimum number of patients required under individual randomisation needs to be inflated to obtain the same power in this case,^{16–18} though there is probably a ceiling effect to the number of subjects in each centre.¹⁹

RANDOMISATION OF PARTICIPANTS AND ALLOCATION CONCEALMENT

We know that by randomising subjects into groups we eliminate potential selection bias and allow for the appropriate statistical analysis to be conducted on (hopefully) comparable independent groups.⁵ However, randomising, as it is often described in published studies, is not sufficiently explicit. The method of randomisation refers to whether we use, for example, simple, block, or stratified randomisation methods (described in several texts—for example, Altman¹²).

Briefly, for simple randomisation one list is used to assign participants to study groups. However, this method may result in unequal numbers in groups, in which case block randomisation can be used. To avoid bias due to researchers (who are in contact with participants) anticipating subsequent allocations, the

size of the blocks should not be disclosed. Stratified randomisation can be used to adjust for confounding variables. In our example we may want to stratify by men aged under and over 50 and create separate random lists for men in each strata. Minimisation is also an allocation procedure used to minimise the imbalance between numbers and certain variables of interest in study groups. It is a valid alternative to randomisation,¹² and is particularly useful in small trials where there are too many variables for stratified randomisation to be feasible.

Random number lists¹³ or computer generated lists can be used. STATA statistical software (<http://www.stata.com>) has an add-in facility for generating lists for any size and combination of blocks.

Another decision is whether to randomise before or after recruitment. If randomising before recruitment—for example, from patients lists in general practice, one is then in the situation where patients are targeted for a particular study group. They are then not as comparable as if they were recruited first (knowing of their 50% chance of being in the intervention arm) and one risks unequally sized groups if there is an imbalance in the number of people who refuse to take part.

The term “allocation concealment” refers to whether researchers in direct contact with participants know in advance to which group they have been allocated. Several articles have been published promoting the methodological quality of randomised trials,^{20 21} sometimes showing bias towards the treatment or intervention effect as a result of non-randomised controls²² and inadequate allocation concealment,²³ and criticising the inadequate reporting of the method of randomisation used.²⁴

The Cochrane Collaboration,²⁵ which promotes the use of systematic reviews to pool the evidence of RCTs, recommends certain methods of allocation as adequate: centralised allocation (assignment from an independent office that is unaware of the participants’ characteristics); precoded identical containers (which contain the study group assignment) that are administered serially to participants; sequentially numbered, sealed opaque envelopes. Other methods of allocation, such as alternate study group assignment, dates of birth, days of the week, are inadequate as clearly these assignments are all transparent before allocation.

BASELINE COMPARISONS AND CONFOUNDING VARIABLES

Why should we perform (and report) baseline comparisons? Basically, to see if the randomisation worked, and that similarity between the groups was achieved. The variables being compared will include confounding variables and should be decided at the start of the study. In our example the direct effect of weightbearing exercise on BMD may be confounded by age and current level of physical activity, and similarity between study groups for these variables should be ascertained. If we had controlled for age by stratified randomisation, however, an

imbalance on this variable would be unlucky!! Altman recommends that one should assess similarity by considering the prognostic strength of the variables and the magnitude of any imbalance rather than relying on hypothesis tests.²⁴ If there are large imbalances these can be adjusted for in the statistical analysis.

ASSESSMENT OF OUTCOME

Lack of blind assessment of outcome has also been shown consistently to overestimate the effect of interventions.^{23–26} Strictly speaking, double blind designs are those in which neither the person assessing the participant nor the participant can identify the intervention being assessed.²⁶ This scenario is obviously easier to achieve in drug-placebo trials than in studies such as our example, in which the intervention is an exercise programme. However, the assessor can and should be blind to the study group allocations especially in uncontrolled and case-control designs. A single assessor can reduce bias by controlling for variable assessment by different assessors. When, and how many, assessments should be made will be decided by the research team.

ANALYSIS OF DATA AND INTERPRETATION OF RESULTS

In an uncontrolled design paired statistical techniques (for example, paired *t* test, Wilcoxon test) are appropriate for assessing the change before and after the intervention. In matched case-control studies a paired analysis is also appropriate for determining the difference between the groups. This type of design is popular in cancer research, and analysis of such data can be seen in Breslow and Day.²⁷ Strictly speaking, standard statistical techniques should be used on variables which are statistically independent—that is, standard techniques apply to two groups that differ only in their responses to the intervention. This would have been achieved by randomisation but not necessarily in unmatched case-control designs.

Independent statistical techniques (for example, *t* tests, Mann-Whitney U tests) are appropriate for determining differences in outcomes between groups. In our example, to determine the size of the effect of the intervention we calculate the change in BMD in each group and compare this change between groups. Confounders can be adjusted for using regression analysis.^{12–24} The precision of the effect²⁸ can be determined by calculating confidence intervals, which is recommended.²⁹

In group randomisation designs it is not appropriate to analyse by individual subject, and each centre provides one datum. A number of articles consider the analysis of quantitative and qualitative data in group randomised trials.^{30–34}

At this stage it is worth mentioning why we constantly advise preplanning of statistical analysis—for example, analysis of confounders and subgroup analysis at the start of the study. It is always tempting to perform many hypothesis tests on all variables to see if these are statistically significant between the study groups, which will lead to misleading results as we can

obtain significant results by chance. Also in designs where there are more than two groups, special techniques, such as the Bonferroni method,¹² can be used to adjust the *p* value when performing multiple pairwise comparisons. In our example (in the “Sample size calculation” section), we might want to test the difference in BMD between the two intervention groups and between each of the two intervention groups and the control group.

Another potential area of bias is loss to follow up or exclusions after randomisation, which can influence the intervention effect. One can be sceptical about studies that do not account for participants who were not available at the end of the study. There may have been valid reasons for missing data (for example, death) but in the absence of an explanation the reader may wonder if there was some bias in their exclusion. A rule of thumb is to analyse as you randomise, that is, an intention to treat analysis is appropriate.³⁵

For intention to treat analysis one can take a conservative (pessimistic) or optimistic approach. For a qualitative response (for example, success (reduction in the number of injurious falls by one) or failure (no reduction)) the former assumes that those missing at the end of the study in the intervention arm would have had the poor outcome (failure) and missing participants in the control arm would have had the good outcome (success). One may then assume that there is good evidence for a positive intervention effect if one is detected under these circumstances. For the optimistic approach the converse would be assumed. For a quantitative response (for example, BMD) one might assume that for missing participants there was no change from baseline or that the change was the average of the outcomes of the remaining participants.

It is also worthwhile checking to see if there are any differences between responders and non-responders as is sometimes the case in questionnaire surveys.³⁶

In an unbiased study, assessment and statistical analysis of the predefined outcomes of a prespecified research question should seem straightforward. However, we must not forget that we sample to make an inference on the population and assume that it is representative. In our example it would not be advisable to conclude that the results found for changes in BMD in adult men were the same for young men or women.

Final suggestion

Becoming proficient at designing and executing experimental studies is somewhat a case of trial and error. The CONSORT statement³⁷ guides authors through adequate reporting of RCTs, but can also serve as a useful check on your design.

1 Thompson B. A review of the British Journal of Sports Medicine 1991–1995. *Br J Sports Med* 1996;30:354–5.

2 Fisher LD, Van Belle G. Biostatistics. A methodology for the health sciences. New York: Wiley, 1993.

3 See Tai S, Gould M, Iliffe S. Promoting healthy exercise among older people in general practice: issues in designing and evaluating therapeutic interventions. *Br J Gen Pract* 1997;47:119–22.

- 4 Snow-Harter C, Whalen R, Myburgh K, *et al.* Bone mineral density, muscle strength, and recreational exercise in men. *J Bone Miner Res* 1992;7:1291-6.
- 5 Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.
- 6 Kirkwood B. *Essentials of medical statistics*. Oxford: Blackwell Scientific, 1988.
- 7 See Tai S, Nazareth I, Haines A, *et al.* A randomized trial of the impact of telephone and recorded delivery reminders on the response rate to research questionnaires. *J Publ Health Med* 1997;19:219-21.
- 8 Harvey L. Factors affecting response rates to mailed questionnaires: a comprehensive literature review. *Journal of the Market Research Society* 1987;29:341-53.
- 9 Heberlein TA, Baumgartner R. Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. *American Sociology Review* 1978;43:447-62.
- 10 Goyder JC. Further evidence on factors affecting response rates to mailed questionnaires. *American Sociology Review* 1982;47:550-3.
- 11 Armitage P, Berry G. *Statistical methods in medical research*. 3rd ed. Oxford: Blackwell Scientific, 1994.
- 12 Altman DG. *Practical statistics for medical research*. London: Chapman & Hall, 1991.
- 13 Machin D, Campbell MJ. *Statistical tables for the design of clinical trials* (table 13.1). Oxford: Blackwell Scientific, 1987.
- 14 Epi Info 6. *A word processing, database and statistics program for public health on IBM compatible microcomputers*. Atlanta, Georgia: The Division of Surveillance and Epidemiology, Epidemiology Program Office, Centers for Disease Control and Prevention (CDC), July 1996.
- 15 Cohen J. *Statistical power analysis for the behavioural sciences*. 2nd ed. Mahwah, New Jersey: Hillsdale, Lawrence Erlbaum Associates, 1988.
- 16 Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;108:100-2.
- 17 Kerry SM, Bland JM. Trials which randomize practices II: sample size. *Fam Pract* 1998;15:84-6.
- 18 Kerry SM, Bland JM. Statistics notes. Sample size in cluster randomization. *BMJ* 1998;316:549.
- 19 Underwood M, Barnett A, Hajioff S. Cluster randomization: a trap for the unwary. *Br J Gen Pract* 1998;48:1089-90.
- 20 Moher D, Jadad AR, Nichol G, *et al.* Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;16:62-73.
- 21 Emerson JD, Burdick E, Hoaglin DC, *et al.* An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trials* 1990;11:339-52.
- 22 Chalmers TC, Celano P, Sacks HS, *et al.* Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358-61.
- 23 Schulz KF, Chalmers I, Hayes RJ, *et al.* Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.
- 24 Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990;335:149-53.
- 25 Murrow CD, Oxman AD, eds. *The Cochrane collaboration handbook version 4.0*. In: *The Cochrane Library* [database on disk and CD-ROM]. Oxford: The Cochrane Collaboration. (Updated software 1996.)
- 26 Jadad AR, Moore R, Carroll D, *et al.* Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1-12.
- 27 Breslow NE, Day NE. *Statistical methods in cancer research*. Vol 1. *The analysis of case-control studies*. Lyon: International Agency for Research on Cancer, 1990. (IARC Scientific Publications, No 32.)
- 28 Guyatt GH, Sackett DL, Cook DJ. User's guide 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* 1994;271:59-63.
- 29 Gardner MJ, Altman DG. *Statistics with confidence—confidence intervals and statistical guidelines*. London: BMJ Publishing Group, 1989.
- 30 Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol* 1981;114:906-14.
- 31 Bland JM, Kerry SM. Statistics notes. Trials randomised in clusters. *BMJ* 1998;315:600.
- 32 Kerry SM, Bland JM. Statistics notes. Analysis of a trial randomised in clusters. *BMJ* 1998;316:54.
- 33 Bland JM, Kerry SM. Statistics notes. Weighted comparison of means. *BMJ* 1998;316:129.
- 34 Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Stat Med* 1997;16:2063-79.
- 35 Guyatt GH, Sackett DL, Cook DJ. User's guide to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 1993;270:2598-601.
- 36 Hill A, Roberts J, Ewings P, *et al.* Non-response bias in a lifestyle survey. *J Publ Health Med* 1997;19:203-7.
- 37 Begg C, Cho M, Eastwood S, *et al.* Special communication. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;276:637-9.