

## **Invited commentary: Methodological issues in the design and analysis of randomised trials**

Mohammad Ali Mansournia<sup>1</sup>, Douglas G Altman<sup>2</sup>

<sup>1</sup> Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

<sup>2</sup> Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK.

Randomised trials are widely considered the “gold standard” for causal inference, because *on average* randomisation balances covariates between treatment groups, even if those covariates are unobserved. However, trials are not immune to random confounding as well as selection bias and measurement bias. Therefore, special care is needed in the design and analysis stages of randomised trials. Here, we review some important methodological aspects of randomised controlled trials in the context of recently published paper in the BJSM, which assessed the effect of McKenzie method of mechanical diagnosis and therapy on pain and disability in patients with chronic non-specific low back pain using a randomized placebo-controlled trial.<sup>1</sup>

i) *Method of randomisation:* Garcia et al state that they randomly assigned 148 participants to two groups of similar sizes i.e., 74 patients per group using simple randomisation. However, simple (unrestricted) randomisation, equivalent to repeated fair coin-tossing, can lead to treatment groups of markedly different sizes in small trials and thus imprecise effect estimates. In fact, the authors were very fortunate, as the probability of complete balance in their study is just 6.5%, and the probability of imbalances equal or greater than 10 (i.e., 79 vs. 69) is non-negligible (46.0%). Balanced block randomisation with, say, 37 blocks of size 4, would have insured balance in number of patients being allocated to intervention or placebo in this study, though its sequence is more predictable than simple randomisation (to fix the latter problem, one can use larger block sizes, and randomly varying the block size). More importantly, balanced blocking prevents from substantial periodic imbalance and thus is often recommended for assignment in randomised trials.<sup>2</sup> Successful randomisation also depends on allocation concealment which authors achieved by use of sealed,

opaque, numbered envelopes. Failure to conceal the assignments at the point of enrolment is a well-known cause of bias.<sup>3</sup>

ii) *Adjustment for the baseline values of the outcome*: Randomised trials are subject to random (chance) confounding as randomisation balances all baseline covariates only in expectation and a particular allocation could be imbalanced with respect to baseline risk factors.<sup>4</sup> An important potential confounder is the outcome at baseline, thus *analysis of covariance (ANCOVA)* with baseline values of the outcome as covariate has been recommended for the analysis of randomised trials with one follow-up visit.<sup>5,6</sup> Unknown to many researchers, *the analysis of change scores* i.e., the difference between follow-up and baseline values does not adjust for the baseline values as the differences are clearly correlated with baseline values, sometimes known as *regression to the mean*. Note that the effect estimates from *ANCOVA* and *the analysis of change scores* are the same if the outcome baseline values are exactly balanced, though *ANCOVA* may still be preferred in terms of the precision of the effect estimate. Extending *ANCOVA* to the analysis of randomised trials with more than one follow-up visit will be discussed in the next point.

The baseline values of the outcome of pain intensity are not substantially imbalanced (Table 1 of the paper) so we are not very concerned about confounding bias, but adjustment for baseline values can still be helpful for increasing efficiency as the model used was linear. The imbalance in baseline values of disability does not seem to be negligible. It is good that the authors did not report the significance tests of baseline difference, a still common misuse in the literature; adjustment for variables which differ significantly at baseline is likely to bias the treatment effect estimate.<sup>3</sup>

iii) *Analysis of longitudinal data from randomised trials with more than one follow-up visit*: According to the design of Garcia et al. paper, patients had to attend 4 follow-up visits: at the end of treatment (5 weeks), and 3, 6 and 12 months after randomisation. In this trial, they pre-specified primary outcome was 5 weeks which doesn't preclude an analysis using all time points. An important point is that the analysis method should account for within-subject correlation in repeated outcome measurements. One possibility is using *random-effect models* of which *repeated measures analysis of variance (ANOVA)* is a special case; another alternative is *generalized estimating*

*equation (GEE) method.*<sup>7</sup> There is also a simpler approach, called *summary measures or response feature analysis*, where, in the first step, one get ride of within-subject correlation by combining the repeated measures on individuals into a suitable summary measure (e.g., average outcome over time) which is then, in a second step, analysed.<sup>8</sup>

There are two general approaches for assessing treatment effects. One approach is testing the *interaction* between treatment group and time which is a generalization of *the analysis of change scores* mentioned in the previous point. Like *the analysis of change scores*, a disadvantage of this approach is that it does not account for the baseline values of outcome. In the statistical methods section of the paper, the authors state that they used interaction terms between treatment groups and time in a linear mixed model. Unfortunately, they did not report the P-value for the global interaction test. A better approach starts the analysis from 5 weeks after treatment i.e., time is coded as month since the end of treatment (at week 5) and adjusts for the baseline values of the outcome. This approach can estimate the effect of treatment at week 5 and the mean change in the effect of treatment per month after the end of treatment.<sup>7</sup>

iv) *Clinically important effect size*: The effect size estimate for pain intensity was -1.00 (95% CI: -2.09 to -0.01) on a ten point scale. On the other hand, the authors used 1-point change in pain intensity in the sample size calculation at the design stage. The authors concluded that "We found a small and likely not clinically relevant difference in pain intensity..." as based on some references, a 2-point change was considered as the minimal detectable change for the numerical pain rating scale. There are two points worth mentioning: i) if 1-point change in pain intensity is not a clinically important effect size, it should not be used in the sample size calculation section of the paper. Of course, the required sample size with 2-point change is smaller than the actual sample size calculated for 1-point change, so there is no concern about the power of the study, and (ii) *clinically important effect size* should ideally be determined based on clinical considerations e.g., the important consequences of the outcome.<sup>9</sup> In the absence of such information, one can gauge the effect size in relation to its standard deviation in the studied population. As an example, the estimate of 1-point difference in pain intensity is more than one half of its baseline SD reported in Table 1, which would be at least a medium effect size based on Cohen's rule.<sup>10</sup>

v) *Post-randomisation exclusion and intention-to-treat analysis*: The paper states that the analysis was *intention-to-treat*. However, one participant was excluded after randomisation, because he had a diagnosis of cancer during the period of treatment. Any exclusion after randomisation violates the intention-to-treat principle and could introduce selection bias.<sup>11</sup> However, when ineligible participants are mistakenly included, investigators could safely remove them without violating the intention-to-treat principle, if the decision only relies on information that reflects the patient's status before randomisation. We note that here only one patient was excluded after randomisation, so impact would be negligible irrespective of the reason for exclusion.

It should be noted that the *intention-to-treat* effect can only be estimated in the absence of censoring and other forms of missing outcome as is the case for the Garcia et al study. In the presence of non-negligible amount of missing outcomes, the analysis of randomised trials requires appropriate adjustment for selection bias using *multiple imputation* or *inverse probability weighting* to estimate the *intention-to-treat* effects.<sup>12,13</sup>

vi) *Blinding of outcome assessors*: One important point about the Garcia et al paper is that assessment of outcome can be done blind even when the therapy cannot be delivered blinded. They assessed the success of blinding by asking the assessor after the trial, which is not a good idea: if the active intervention is indeed beneficial, his/her guesses are expected to be better than those produced by chance.<sup>3</sup>

We have addressed only a few main issues in this commentary. There are many other ways in which researchers need to take care in how they design, analyse and interpret trials.<sup>3</sup> Overall the Garcia et al randomised trial is good in terms of methodology, though the validity of their results would have been strengthened, if they addressed the critical issues mentioned in this commentary.

Acknowledgements: We thank Rasmus Østergaard Nielsen for helpful comments on an earlier draft of this commentary.

## References

- 1) Garcia AN, Costa LDCM, Hancock MJ, et al. McKenzie Method Mechanical Diagnosis and Therapy was slightly more effective than placebo for pain, but not for disability, in patients with chronic nonspecific low back pain: a randomized placebo controlled trial with short and longer-term follow-up. *Br J Sports Med*. In press.
- 2) Matthews JNS. *An Introduction to Randomised Controlled Clinical Trials*. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2006.
- 3) Altman DG, Schulz KF, Moher D, et al; CONSORT GROUP (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001 Apr 17;134(8):663-94.
- 4) Greenland S, Mansournia MA. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *Eur J Epidemiol*. 2015;30:1101-10.
- 5) Senn S. *Statistical issues in drug development*. 2nd ed. Chichester, England: John Wiley; 2008.
- 6) Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow-up measurements. *BMJ* 2001;323:1123-1124.
- 7) Kirkwood BR, Sterne JAC. *Essential Medical Statistics*. 2nd ed. Oxford: Blackwell Science Ltd; 2003.
- 8) Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *BMJ*. 1990;300:230-5.
- 9) Nielsen RO, Bertelsen ML, Verhagen E, et al. When is a study result important for athletes, clinicians and team coaches/staff? *Br J Sports Med*. 2017 May 16. pii: bjsports-2017-097759. doi: 10.1136/bjsports-2017-097759.
- 10) Cohen, J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
- 11) Mansournia MA, Higgins JPT, Sterne JAC, Hernán MA. Biases in randomized trials: a conversation between trialists and epidemiologists. *Epidemiology*. 2017;28:54-59.
- 12) Vickers AJ, Altman DG. Statistics notes: missing outcomes in randomised trials. *BMJ*. 2013;346:f3438.
- 13) Mansournia MA, Altman DG. Inverse probability weighting. *BMJ*. 2016;352:i189.