Grouping in individually randomised trials

Talk presented at the 4th Annual Conference on Randomised Controlled Trials in the Social Sciences, September 2009, York, UK:

Martin Bland
Prof. of Health Statistics
Dept. of Health Sciences
University of York, York, England

Abstract

Methods for the design and analysis of trials where participants are allocated to treatment in clusters are now well established. Clustering also happens in trials where participants are allocated individually, when the intervention is provided by individual operators, such as surgeons or therapists. These operators form a hidden sample, whose effect is usually ignored. Recently, trial designers have been considering how they should allow for this clustering effect and funders have been asking applicants the same question. In this talk I examine some of these issues and suggest one simple method of analysis.

Hidden samples

In 1998, I gave my Chairman's Address to the Medical Section of the Royal Statistical Society on the subject of hidden samples in medical research. Today I am going to return to some of the ideas in that talk, because their time seems to have come. I shall start with some examples of research studies.

The Wandsworth Terminal Care Study

In my final example, there were operators for one group but not for the other. The Wandsworth Terminal Care Study was a trial of the effectiveness of Terminal Care Coordinating Nurses (Addington-Hall *et al.*, 1992). The proposal was that nurses would be employed to improve the care of patients dying of cancer in their homes. The theory was that the NHS, local authority, and voluntary sector already provided resources to look after these patients, but that the patients might not get all the services to which they were entitled because the providers did not know that they were needed. The coordinator would contact terminally ill patients in hospital, then after their discharge would visit the patients in their homes and assess their needs. They would then inform the appropriate service providers.

The trial was randomized at the level of the general practice, so all of one practices patients would be allocated to the same arm of the trial. As there were expected to be 200 patients per group and about 200 practices were randomised, I decided to ignore the practice in the analysis. Much more important, and also ignored, were the coordinators. There were only two of them. There were coordinators for one arm of the trial, and nobody at all for the other arm.

The trial was broadly negative, with few differences between the groups. The conclusion was that 'This coordinating service made little difference to patient or family outcomes, perhaps because the service did not have a budget with which it could obtain services or because the professional skills of the nurse-coordinators may have conflicted with the requirements of the coordinating role.' Or as the research fellow commented to me when I said 'So it doesn't work then.', 'Not with these coordinators.'

The coordinators are a sample from a large population of potential care coordinators. They are also a very small sample. I think that we should include the uncertainty this produces in our analysis (which we did not). It is almost always ignored.

Sometimes, the way operators relate their patients is different in the two treatment groups. The Know Your Midwife Study (KYM) was a randomised trial of continuity of care (Flint and Poulengeris, 1986, Flint et al, 1989). Five midwives ran the KYM clinic, and each mother would have the same midwife for all anti-natal care, delivery, and post-natal care. The originator of the scheme, Caroline Flint, was very enthusiastic and was convinced that no woman who knew about KYM would be willing to accept anything else. She was therefore worried that if she asked women if they were willing to be randomised, those allocated to standard care would be very disappointed. We therefore did what is now called a randomised consent design, allocating women to be offered KYM and then asking if they would have this modality, or to receive standard care without the option. Perhaps to her surprise, some women refused KYM, so they had standard care too. For the analysis, the KYM refusers were combined with the KYM acceptors in an intention to treat analysis, which required many long discussions between myself and the researchers as I explained it again.

What was not taken into account in the analysis was the variability between midwives. For each mother receiving KYM, there was a named midwife who did all the care. The success of the care must depend to some extent on the skill of the midwife and her relationship with the patients. (Now that this type of scheme is widespread, women have remarked to me that it can be unsatisfactory if the mother does not like the midwife to whom they are allocated!) This is clearly a sample of midwives, and one problem might be that the midwives who opt for KYM are different in some way from midwives who provide the standard care. The latter will also vary in their skill, etc., but do not have named patients to whom they can be linked. How then can we take this variation into account?

Surgery

In a trial of surgery, we might have a comparison of two operations. Surgeons often have strong preferences for one type of operation. Different surgeons might give the treatments in the two arms of the trial. Should we include surgeon in the analysis?

The CAVATAS trial

CAVATAS (CAVATAS Investigators, 2001) was a multi-centre, multinational trial comparing balloon angioplasty (the new treatment) with surgical graft for stenosis of the carotid artery. In each centre, there were one or more radiologists doing the angioplasty and one or more surgeons doing the surgery. Randomisation was at the level of the patient. The main outcome was survival to death or disabling stroke.

Here we have several samples, and mostly they are not random. Centres were recruited by asking any neurologist who thought they could run the trial in their centre to volunteer. Thus we do not have a representative sample of countries, or of hospitals within countries. This is normal in such studies.

We can deal with the effect of centre by fitting it as a fixed effect. We could also fit it as a random effect, since it is essentially a nuisance parameter. If we wanted to we could have two levels of random effect: country and centre. However, the source of variation represented by the operators is more difficult to deal with. Within a centre, there may be one, two or several surgeons, and one, two or several radiologists. Some may do many procedures, some only one or two. There is no reason to suppose that the variability in skill between surgeons is the same as the variability between radiologists. In a centre with two surgeons one may take all the difficult cases. Allowing for surgeon effect under these circumstances will remove variation which is really between patients.

In each centre we have a varying number of surgeons and radiologists. Here operator and treatment are totally confounded, as surgeons do one arm of the trial and radiologists the other. We can analyse by operator as a cluster. We can do a three level multi-level model. Both treat operator as a random effect but neither can distinguish between neurologist and surgeon. One problem here is that there is no reason to suppose that the variability in skill between surgeons is the same as the variability between radiologists. Another is that, for example, in a centre with two surgeons one may take all the difficult cases. Allowing for surgeon effect under these circumstances will remove variation which is really between patients.

We should also ask whether our samples of surgeons and neurologists are representative or comparable. This is recognized, in a way, by the clinicians themselves. The CAVATAS trial has been criticised by rival surgeons for the perceived high mortality. They claim that mortality in this trial is ‘much higher than in our hands’ (M Brown, personal communication).

Summing up

In many trials there is a hidden sample of operators, which is almost always ignored.

It should not be ignored.

Hidden samples return

I did not get round to preparing this paper for publication and did not see many other references to this issue. Then, recently, it surfaced.

In a grant committee (the Health Technology Assessment Programme Clinical Trials Board) we had a proposal for a clinical trial which involved therapists of some kind in the new treatment arm, but not in the control arm. They proposed to allow for this in their sample size calculation by using an intra-cluster correlation coefficient to compute a design effect, as if it were a cluster randomised trial. I asked Keith Abrams, Prof. of Medical Statistics at the University of Leicester, who was sitting next me, how he would analyse that. He said that he did not know. Neither did I. Over the next two or three meetings we had other similar proposals. There was never any detail as to the proposed analysis. I often raised the question as to what the analysis would be, we never got any answer. As the sample sizes were being increased, which I think is almost always desirable, this gap did not impede funding.

Acupuncture for depression

My colleagues and I proposed, to a different funding body, a trial of acupuncture for the treatment of depression. We were asked by the funders how we would deal with differences between acupuncturists? I had to come up with an answer.

VenUS IV: high compression hosiery in the treatment of venous leg ulcers

Around the same time, colleagues and I proposed to the Health Technology Assessment Programme Clinical Trials Board a trial of high compression hosiery (elastic stockings) for the treatment of venous leg ulcers. This was the fourth in a series of trials in this area. (Of course, I had to leave the room for the discussion and so didn’t know what was decided.)

Venous leg ulcers are wounds which do not heal because of poor blood flow in the leg. They can persist for years and may cover the entire lower limb. The only effective treatment is to compress the limb with a series of bandages to narrow the veins and make it easier for blood to pumped back to the heart. An earlier trial, VenUS I, had compared a four layer bandage system with the usual short stretch bandage system. (ref) The four layer bandage system shortened healing times and is now accepted as the standard treatment. One problem with these bandages is the patient cannot remove and replace the bandage. If the bandage is removed, they must do without compression until they see the nurse again. The VenUS IV trial was to compare newly-introduced high-compression elastic stockings, which could be removed, washed, and replaced by the patient, with the four layer bandage.

We received a letter saying that we would be funded, but that any impact of clustering due to nurses having different skills at bandaging needs to be taken into account in the sample size and analysis.

There is indeed evidence to suggest that bandager skill varies considerably between nurses (Feben 2003). We had to find an answer. I did wonder whether the committee had decided to call my bluff.

Summing up

There is a new awareness of potential operator effects and applicants for research grants are including proposals to do this, grant bodies are asking for it.

Some research grant applicants now include a sample size calculation using an ICC as if for a cluster randomised trial. But how are they going to do the analysis?

Dealing with operator effects

VenUS IV: high compression hosiery in the treatment of venous leg ulcers

The solution I proposed for this trial was to use robust standard errors. We would treat the trial centre as a cluster. How much would this increase our standard error?

The sample size calculations were based on VenUS I. This was trial in a similar patient population, where the new intervention was the four layer high compression bandage system which would be the control treatment in VenUS IV. This trial recruited 386 participants over 20 months from 9 UK sites.

The primary outcome of the trial was time to healing and the adjusted hazard ratio = 1.33 (95% CI 1.05 to 1.67). This analysis treated centre as a fixed effect. We repeated this analysis using robust standard errors to allow for centre as a cluster, hence as a random effect. Using robust standard errors inflated the variance compared to a fixed effects model.

Fixed effect: standard error of log hazard ratio = 0.119.

Random effect: standard error of log hazard ratio = 0.129.

Variance was increased by factor $0.129^2/0.119^2 = 1.19$.

This is the ratio in which we think we should increase the sample size to give the same power. We should increase the sample size by factor 1.19.

We then argued that this would be too large an increase, for the following reasons. In VenUS I there were good reasons to suspect that there would be variation in bandaging skill. Some centres had prior experience using the short stretch bandage control treatment and some did not, some had prior experience in using the four layer bandage intervention and some did not. In VenUS IV, the point of the intervention, stockings, is that skill is not required. Any variation in skill will be in the application of the four layer bandage which is the control treatment. This is now the standard treatment and all centres should be experienced in its use. We therefore expect that there will be much less variation between centres than in VenUS I. We therefore proposed to inflate the sample size by 10% to 489 patients. This was accepted by the funding board.

Note that this problem is essentially the same as CAVATAS, a small number of operators within each centre, different operators for the two treatments (nurses vs. patients themselves), several centres.

Acupuncture for depression

I came up with a simple summary statistics solution for the acupuncture trial. I think that my brain had been working on this problem on and off, but perhaps I just know more now. Anyway, I came up with a solution immediately.

We have clusters around acupuncturists in the intervention group, not in the control group. I suggested that when a participant is recruited to this trial, we will allocate not just to acupuncture or control, but to an acupuncturist. Each participant gets an acupuncturist, whether they get acupuncture or not. If they are controls, they do not know about this, nor does the acupuncturist.

Each acupuncturist has their own cluster of controls and acupuncture patients. For each acupuncturist, we will then have a mean outcome depression score for acupuncture and a mean for controls. We could do a paired t test on these summary statistics. We could also adjust for baseline mean depression score using analysis of covariance with acupuncturist as a random effect. More simply (but less powerfully) we could use the means of change in depression score from baseline for acupuncture patients and controls in a paired t analysis.

We proposed this as a sensitivity analysis rather than as a primary analysis and did not adjust the sample size. The plan was accepted by funders.

I decided to call this the artificial cluster method.

Acupuncture for irritable bowel syndrome

We are now implementing the artificial cluster method in a trial which was already funded and about to begin. This is a trial of acupuncture for the treatment of irritable bowel syndrome. This trial has now closed its recruitment and we expect to have the first outcome data fairly soon. We shall then apply the artificial cluster technique to the trial.

Like the Terminal Care Coordinators Study and the Know Your Midwife, these acupuncture trials have defined operators in one treatment arm only. The other arm has either no operator or a mixture of many operators. This applies to almost any complex intervention vs. usual care.

The Knee Arthroplasty Trial (KAT)

I tried to find studies which had allowed for operator effects, but couldn't. It is a difficult search to do. I tried the Medstats email list and got only one reply, from Graeme Maclennan, to who I am very grateful. He had done a surgical trial of knee replacements (KAT Trial Group, 2009). It was a complex trial with three treatment comparisons:

- with or without a metal backing of the tibial component,
- with or without patellar resurfacing,
- with or without a mobile bearing.

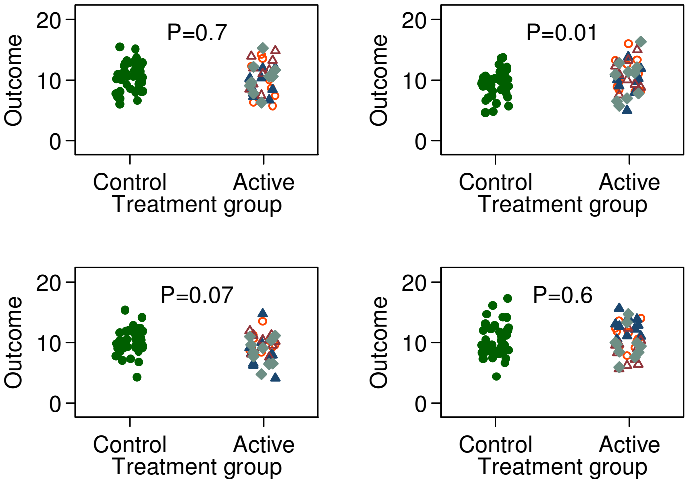
Randomization to more than one comparison was allowed. One hundred and sixteen surgeons in thirty-four centres participated. The randomization was stratified by surgeon, with minimization according to the patient's age, sex, and site of disease.

Effect sizes were presented with the associated 95% confidence intervals estimated with robust standard errors to account for potential surgeon effects. So this trial used the approach which I plan for VenUS IV.

Simulation studies

I decided to concentrate on the problem of defined operators in one group only and carry out some simulation studies of possible analyses.

Figure 1. Four simulations of a four therapist, 40 participants per group trial, ignoring the effect of therapists



Ignoring the operators

First I asked can we ignore this? Ignoring the operator has been the usual approach for the past fifty or sixty years. My first simulation was of a trial with four therapists, each with 10 participants, and a similar sized group of 40 controls. I set the individual patients variance = 4 and added an additional therapist variance = 1, in the therapist group only. All random variables were Normal. The null hypothesis is true, so the average effect of therapy is zero. Some therapists give benefit, some give harm.

Figure 1 shows the result of four runs of the simulation. The first shows no significant difference, P = 0.7, the second a significant difference with active greater than control, the third shows a difference which is almost significant, P = 0.07, with control greater than active, the fourth no significance, P= = 0.6. So all does not look completely well. I ran 1000 runs and obtained figure 2 as the distribution of P values.

If the test were valid, we should get P less than 0.05 is 5% of trials and the distribution of P values should be uniform. Clearly we have spurious significant differences. We should not ignore the therapists.

P values are a bit old fashioned and we should have interval estimates, but P values are easy and the corresponding confidence intervals would be too narrow.

The artificial cluster method

Does the artificial cluster approach work? I ran the same simulation, 4 therapists, 10 participants each, 40 controls, participants variance = 4, therapist variance = 1, null hypothesis true. Figure 3 shows the P values from 1000 runs.

As Figure 3 shows, the distribution is uniform and nothing suggests that we will get spurious significant differences. The artificial cluster method works. (I knew it would, of course, but it was a great relief when this graph appeared!)

Figure 2. 1000 simulations of a four therapist, 40 participants per group trial, ignoring the effect of therapists

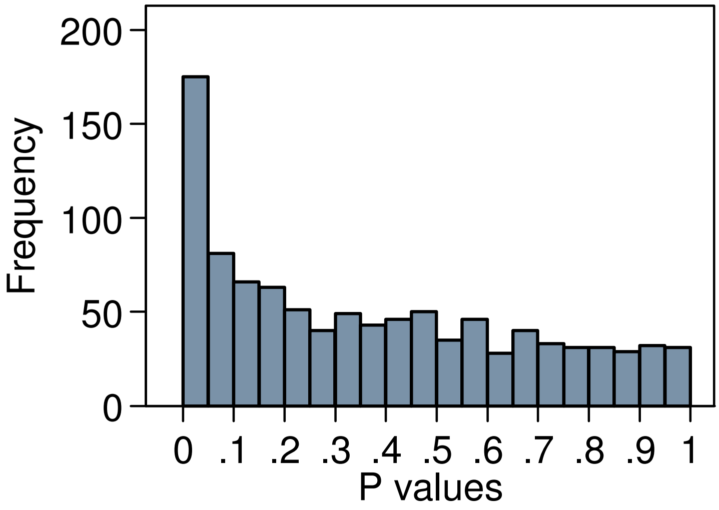
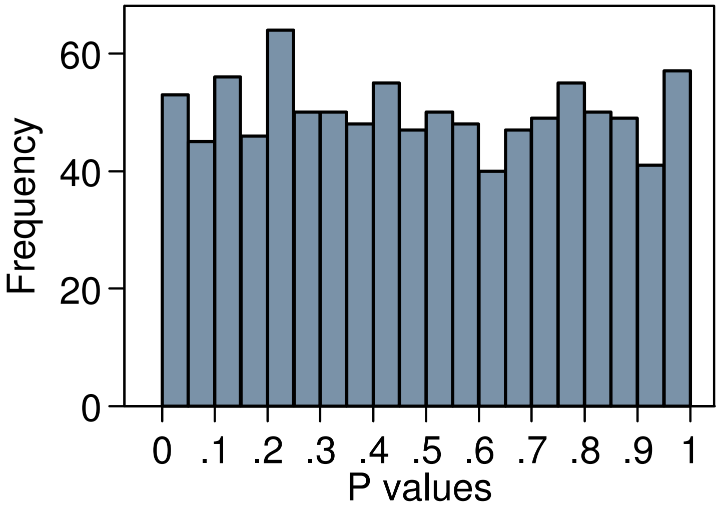


Figure 3. 1000 simulations of a four therapist, 40 participants per group trial, allowing for the effect of therapists by the artificial cluster method



Other approaches to clusters

There are several other approaches we could consider, as used in cluster randomised trials, including:

- robust standard errors
- multilevel modelling
- general estimating equations
- Bayesian hierarchical models

Figure 4. 1000 simulations of a four therapist, 40 participants per group trial, allowing for the effect of therapists by the robust standard errors method

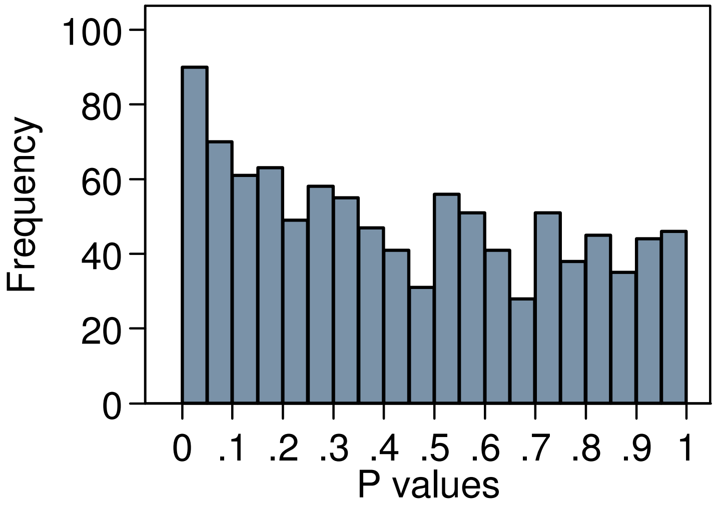


Figure 5. 1000 simulations of a four therapist, 40 participants per group trial, allowing for the effect of therapists by the robust standard errors method

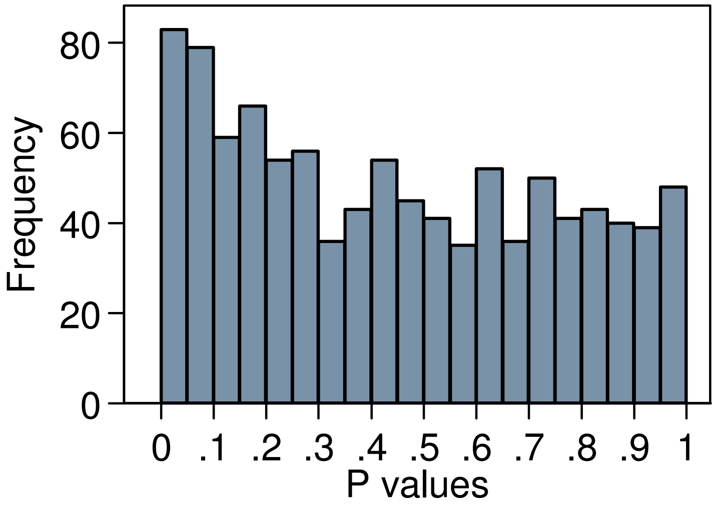


Figure 5 shows the P values for 1000 runs. The method does not work. The distribution of P is not uniform. It may be that the number of therapists is still too small, but 20 is quite large for a practical trial.

We could treat the control group as a single cluster and see what happens. Of course, the structure is wrong for the standard models. The therapist level is present in one arm and not the other. Would the control arm work as a single large cluster? I decided not to run the Bayesian analysis, as I think a deliberately wrong Bayesian model might be taken the wrong way by some of my friends!

Robust standard errors

Figure 4 shows the results of analysis using robust standard errors for the trials shown in Figures 2 and 3.

Clearly this does not work. The distribution of P is not uniform. However, the number of therapists is too small for a robust standard errors model to be reliable. I therefore increased the size of the simulation to 20 therapists with 5 participants each, and 100 controls. As before, the participant variance = 4, the therapist variance = 1, and the null hypothesis is true.

Multilevel modelling

Does the multilevel modelling approach work? I ran the same simulation, four therapists, 10 participants each, 40 controls. participants variance = 4, therapist variance = 1, null hypothesis true, treating the control group as a single cluster. It crashed on the second run, having failed to converge. The number of therapists is far too small, so I increased it to 20 therapists, 5 participants each, 100 controls. It crashed at run 357, having failed to converge.

Generalised estimating equations

Does the generalised estimating equations approach work? I ran the simulation with four therapists, 10 participants each, and 40 controls. It, too, crashed on the second run, reporting an exchangeable working correlation matrix which was not positive definite.

For a simulation with 20 therapists, 5 participants each, and 100 controls, the program crashed on run 73, again with exchangeable working correlation matrix not positive definite. This method does not work.

Conclusions

For the problem of defined operators in one group only, we should taken the operator into account. The artificial cluster method works. Other approaches tried, robust standard errors, multilevel modelling, and general estimating equations, all using the control group as a single cluster, all failed, even with a fairly large trial.

New models need to be developed for these applications. I am not aware of these at the moment.

References

Addington-Hall JM, Macdonald LD, Anderson HR, Chamberlain J, Freeling P, Bland JM, Raftery J. (1992) A randomised controlled trial of the effects of coordinating care for terminally ill cancer patients. *British Medical Journal* **305**, 1317-322.

CAVATAS investigators. (2001) Endovascular versus surgical treatment in patients with carotid stenosis in the Carotid and Vertebral Artery Transluminal Angioplasty study (CAVATAS): a randomised trial. *Lancet* **357**, 1729-37.

Feben K. (2003) How effective is training in compression bandaging techniques? *British Journal of Community Nursing* **8**, 80-4.

Flint C. and Poulengeris P. (1986) *The 'Know Your Midwife' Report.* Caroline Flint, London.

Flint C, Poulengeris P, Grant A. (1989) The 'Know Your Midwife' scheme -- a randomised trial of continuity of care by a team of midwives. *Midwifery* **5**, 11-16.

KAT Trial Group. (2009) The Knee Arthroplasty Trial (KAT): design features, baseline characteristics, and two-year functional outcomes after alternative approaches to knee replacement. *Journal of Bone and Joint Surgery of America* (2009) 91: 134-41.