

Designing clinical trials in women's health

Jane Daniels^{*}, Robert K. Hills

University of Birmingham Clinical Trials Unit, Park Grange, 1 Somerset Road, Edgbaston, Birmingham B15 2RR, UK

Received 21 December 2005; accepted 8 February 2006

Abstract

Although outcomes in women's health are not as spectacular as in conditions like cancer, the large number of women who present each year means that the overall impact of these conditions is enormous. Similarly, although suboptimal therapies may not individually be much worse than best practice, the overall effect on a nation's health, wealth and happiness is substantial. There is therefore a real need to gather evidence as to which, if any, women, benefit from any particular therapy. Well-designed randomised controlled trials (RCTs) help provide reliable evidence on a treatment's effectiveness. In this article, we consider important aspects of designing a good clinical trial; and in particular their application to women's health issues. Designed as an overview of the subject, we consider how large trials need to be; the choice of endpoints; how they should be analysed; and also more practical considerations in running a successful trial. The considerations given here are of use not only to clinicians or researchers preparing to run their own trial, but also to anyone who reads reports of trials, and should help clinicians make informed judgements about evidence presented in published reports.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Evidence based medicine; Clinical trials; Meta-analysis

1. Introduction

All members of the medical profession have an obligation to their patients to provide the best possible care. It goes without saying that a surgeon would not nowadays operate under the conditions present 150 years ago before the advent of current hygiene and aseptic procedures. With this need to do one's best comes the requirement to know what best practice is, and hence the need for evidence based practice.

But precisely how does one determine best practice? We might have an idea that a particular treatment seems to relieve pain, or a plausible mechanism of action for a drug, but can we really be sure? Generally speaking, unless the action of a particular treatment is both immediate and breathtaking (such as insulin for diabetic coma), we cannot be absolutely certain which treatment is best for which women. As we shall see, historical comparisons, or other

database dependent methods, can prove misleading. What is required is a method that will provide reliable, convincing evidence to inform future practice.

Fortunately, there is such a tool: the randomised controlled trial (RCT). At its heart are two principles. First, through randomisation, any differences between the women receiving one treatment, and those receiving another are purely down to chance: so, if a sufficiently large difference is detected, then it is almost certainly due to the only thing that is systematically different between the two groups, namely the treatment. Second, with large numbers of patients, it becomes easier to detect smaller treatment effects and to say that any differences are not the result of chance. This, the statistical aspect of RCTs, is effectively a formalisation of common sense. If one tosses a coin 10 times and gets 6 heads and 4 tails, it is not out of the ordinary: but if one saw 6000 heads and 4000 tails from 10,000 tosses, then one would be concerned that the coin may be biased. The proportion of heads is the same: but larger numbers give stronger evidence of an unfair coin.

In this article we shall set out some of the factors that make a good trial. Much excellent literature has already been devoted to the theory of the RCT [1–3] (and much that is less

^{*} Corresponding author. Tel.: +44 121 687 2314.

E-mail addresses: j.p.daniels@bham.ac.uk (J. Daniels),
r.k.hills@bham.ac.uk (R.K. Hills).

excellent). Likewise, medical statistics has been well covered in a number of books and articles [4,5]. Here, instead, we shall concentrate on the underlying principles, with particular emphasis on their applications in women's health. Additionally, there will be some practical tips on how to run a successful trial, and examples to show how attractive, but incorrect methods can lead to misleading results. Although this chapter is aimed at researchers planning to conduct a randomised controlled trial, it also provides useful pointers for clinicians wanting to critically appraise a trial for reliability.

We therefore present our own “ten habits of highly effective trialists”.

1.1. Do not expect miracles: even moderate improvements are often clinically meaningful

Too often, early results from small case series lead to highly optimistic claims regarding the efficacy of an intervention. New “miracle breakthroughs” make headline news, and then little more is ever heard of them. The main reason for this is that case studies do not provide reliable evidence. There is a potentially serious risk of selection bias: it is impossible to tell whether only the most promising patients are selected for the new treatment, and without a comparator group, the true scale of any “breakthrough” cannot be gauged. The actual likelihood of a new intervention having a big treatment effect, or being vastly superior to an existing therapy, is fairly low. It is more realistic to expect a moderate difference between interventions, or a moderate effect compared with placebo.

But moderate effects are still important. In common conditions, even small differences can add up to a large number of lives saved, or pain free days, when aggregated over the population. There are 800,000 live births in the UK every year, and about 350,000 women suffer perineal trauma requiring suturing. If, as the MoMS trial showed [6] a new, easy to learn suturing technique can reduce the duration of perineal pain by 2 days on average, then this amounts to almost 2000 pain-free woman-years every year in the UK alone. These sorts of differences, when aggregated, affect not only women's quality of life, but also a country's economy, by reducing absenteeism and burden on the health system. Trials therefore need to be able to distinguish between this kind of moderate, but still worthwhile, effect and a difference that is too small to be of importance in practice.

1.2. Get reliable results: avoid biases

In order to detect moderate, but meaningful, differences between treatments, it is important to ensure that the results are not distorted by any systematic differences between the treatments being compared. For example, infant mortality rates have generally improved over time. So, any comparison that looked at today's figures as opposed to

those in the past would show a benefit, even for a placebo. Similarly, and more light-heartedly, a comparison of A level results today, and those of 1980 show an improved pass rate. Students sitting A levels in 1980 received free school milk while at primary school: those sitting them today did not. So are we to conclude that milk makes you less intelligent? If surgeons are more likely to give a surgical procedure to women with more severe symptoms, then those who actually receive surgery are of worse prognosis than those who do not. A straight comparison would show that surgery appeared worse. (This is one reason why specialist hospitals may do less well in raw league tables than non-specialist hospitals—the case mix is different.)

The only way to make sure we reduce this kind of selection bias is to randomise patients: that is, allocate treatments in a fashion that produces equivalent groups, and in such a way that there is no possibility of predicting the next outcome. Allocation by date of birth, for example, may not produce equivalent groups, because it is easy to predict what a person's treatment is going to be, and then make a decision on whether to enter the trial based on this knowledge.

There are several different methods of allocating patients that can produce equivalent groups [7,8]. Some methods ensure that the groups are balanced for important prognostic factors, although it is important to prevent any discernable patterns in the randomisation that could allow prediction of the next allocation. It is a fallacy that strict random allocation is required: far more important is concealment of the treatment allocation until the patient is irreversibly committed to the trial. Furthermore, there cannot be any opportunity to re-randomise a participant who was not allocated the preferred treatment. Clinicians can and do attempt to subvert some randomisation approaches, and care needs to be taken to ensure that this is made as difficult as possible [9]. **Box 1** gives guidance on good practice in randomisation.

1.3. Get reliable results: make sure your trial is large enough

If the different treatment groups are equivalent, then it follows that any differences seen are the result either of chance, or of a genuine difference between the treatments. The probability that the effect seen is just the result of chance differences between groups is given by the p -value: it tells you how often a trial of an ineffective treatment would be expected to produce this sort of result. If the p -value is sufficiently small, then we conclude that the data are inconsistent with the treatment being ineffective, rather in the same way that a jury starts with the presumption of innocence and asks if there is enough evidence to overturn it beyond reasonable doubt. Alternatively, one can use the accumulating trial data to produce an estimate of the size of any treatment effect, together with a measure of a likely range, based on the natural variability between patients.

Box 1. Good practice in randomisation

- Use third party randomisation
An independent randomisation service with a degree of separation from clinician and patient provides greater security. Envelopes may be attractive, but need to be policed to stop clinicians opening several envelopes until they find the preferred treatment.
- Collect all important prognostic factors prior to randomisation
The recorded value of a prognostic factor may change or be influenced by treatment allocation.
- Balance prognostic groups by stratification or minimisation
This ensures similar numbers of different types of patient in each group. At its simplest, this could be achieved by a randomisation list for each category, but for more than a few variables, this becomes cumbersome and a computerised system is preferable and allows minimisation across many variables simultaneously.
- Do not stratify by surgeon or center
Whilst seeming desirable to ensure equal numbers per centre or clinician, if a trial cannot be effectively blinded, balancing at this level makes the randomisation schedule much more predictable [10]. Post-hoc analysis by centre or surgeon is not precluded.

Clearly, in order to detect moderate differences, one must discriminate between the signal and ambient noise: although one cannot reduce differences between people, by using larger numbers one can improve the signal to noise ratio. For example, suppose the number of women in a particular group with dysmenorrhoea who eventually opt for a hysterectomy is 50%. Then in order to detect reliably whether a new treatment reduces this proportion to 25% (i.e. halves the hysterectomy rate) needs about 150 women. But, to detect an improvement from 50 to 45%, which would still be worthwhile, needs about 4000 women. It also follows that trials which do not give significant results do not necessarily mean that the treatment is no good: the trial may just have been too small to detect a moderate, but relevant difference [11]. There is clearly a need to balance the desire to detect moderate differences with the feasibility of recruiting a large number of women: hence the need to foster large collaborative groups in women's health.

1.4. Get generalisable results: avoid restrictive entry criteria

All trials are expected to define the included population, with clear diagnostic criteria for the condition under investigation. Patients who could potentially be at risk from one of the trial interventions need to be excluded. However, when considering the appropriate type of patient for a potential trial, there is scope for too rigid and exclusive selection. A conventional definition of menorrhagia is menstrual blood loss (MBL) of >80 ml per cycle. Apart from the practical difficulties of determining MBL

objectively, what distinguishes heavy periods with 75 ml MBL from menorrhagia with 80 ml MBL? Can results from trial with this stringent criterion be extrapolated to women with lower MBL? Or is the emphasis on such criteria detracting from the more important diagnostic criterion—whether the women perceives herself to have heavy periods?

In similar ways, age can be applied as an exclusion criterion on arbitrary grounds. Whilst one may argue that an upper age limit maybe be important to exclude perimenopausal women, it fails to acknowledge the age range of onset of menopausal symptoms. It is far preferable to avoid prescriptive inclusion and exclusion criteria and aim for a representative sample from which generalisable results can be obtained.

The practice of evidence based medicine – identifying, assessing and implementing research evidence – requires the evidence to be present and for a clear answer to a clinical dilemma, conclusive. Unfortunately, this is seldom the case and the clinician is faced with the prospect of uncertainty regarding the relative merits of different courses of action. There may be differences of opinion between clinicians presented with the same circumstances. The ethical imperative then would be to try to reduce uncertainty by contributing to the evidence, and the most appropriate way of achieving this would be to recruit the patient into a well designed clinical trial. It would be unethical for a patient to have their treatment chosen at random if either they or their doctor are substantially certain what treatment they prefer. However, randomisation can, and should, be considered when both doctor and patient are uncertain as to which treatment is preferable. No further restrictions, other than diagnostic and safety-related criteria, need be applied, allowing a wide range of patients to be recruited. This sort of wide, pragmatic entry condition has an added benefit. If a wide range of women is randomised, it is possible to examine, albeit cautiously, whether or not different types of women respond differently to the treatment. Only by running large-scale randomised trials with wide entry criteria can questions of which women benefit be answered with any confidence. A heterogeneous population in a trial is therefore a strength: but care needs to be taken to perform appropriately stratified analyses, and to ensure that entry criteria are not so vague that the type of woman recruited cannot be defined.

1.5. Maintain comparability—blinding of treatments

Whilst strict randomisation is essential for creating comparable groups, systematic differences between the groups can emerge if supportive care or additional treatments are provided preferentially to one group or another. This performance bias is particularly difficult to control for if frequency of contact with clinical services is necessarily higher in one group: for example, if surgery required additional follow-up to inspect the wound during which further advice on symptom management was provided.

The recognised method of ensuring comparability of treatment, care and assessment is to blind clinician, patient and outcome assessor to the treatment allocation. This way, any beliefs regarding the intervention, negative or positive, should be expected not to influence the outcome. This is relatively straightforward for a drug versus placebo comparison and can be achieved with various other interventions with imagination. It is possible to blind some surgical procedures, from the woman at least, but questions regarding ethics are raised. With placebo controlled trials, there is a shared ignorance between the parties involved, whereas performing sham surgery requires deception on the part of the clinician. Sometimes it will also require an acceptance of an additional, clinically unnecessary procedure by the participant, for example, in a trial of laparoscopic uterosacral nerve ablation, where the control group have a superficial skin incision to mimic the lateral port [12]. It can be successfully argued that sham procedures are methodologically necessary to produce valid results and therefore as long as the participant has been adequately informed that she will receive either a real or a sham intervention and that the sham procedure will be indistinguishable from the real treatment under investigation, there is no deception involved [13].

1.6. Choose appropriate outcomes—meaningful, and not too burdensome

Perhaps the most difficult task in women's health trials is the choice of a suitable outcome variable. In trials in cancer, choice of outcome is easy: typically, the important thing is to reduce mortality. For chronic conditions, a number of different dimensions need to be considered, and important choices made. It is important to capture outcomes that are of relevance to the woman herself, as well as to clinicians, and to bodies such as NICE. Generally speaking the simplest outcomes are often the best: less pain, better mobility, fewer hospitalisations or days off work. One possible way of determining outcome measures was performed as part of the MoMS trial [6], where a focus group of mothers who had experienced perineal trauma were asked about their experiences and questionnaires adapted to include outcomes relevant to the mothers themselves.

It is important, however, not to get too carried away, and introduce too many outcome measures. That way, a trial will lack clarity of focus, and one may be in a position where, by chance, one outcome indicates that one treatment is better, and for another outcome, the reverse is the case. Given enough outcomes, it is usually possible to find a significant result on one of them even in trials of the most unpromising treatments. For this reason, it is important to identify a small number of primary and secondary outcomes, considered the most important. These may be clinical outcomes obtained from notes, or patient-centred outcomes collected through questionnaires. If a questionnaire is being used, then it should be validated, to make sure that results are meaningful

and reproducible, and if possible to identify the minimally relevant treatment effect.

Sometimes the real outcome of interest occurs a long time in the future (e.g. hysterectomy, or even, in the case of perinatal studies long-term outcomes affecting the baby). Because it is impractical to wait many years for an answer, it is sometimes possible to use instead a surrogate outcome, which predicts this future outcome. However, this can be fraught with problems: can one really say that a change in a laboratory marker is really indicative of a better outcome in 10 years time? In general, if at all possible, it is better to measure the outcome directly, and not rely on surrogates: after all, it is less easy to argue with hard facts than the results of some modelling exercise.

If a surrogate outcome is the only alternative, then a good surrogate marker will be validated and:

- correlate with clinically relevant measure
- capture the whole of the clinically relevant effect.

After deciding the primary and secondary outcome measures for a trial, and translating them into a format that accurately captures the information required, it is often tempting to include extra “nice to know” data items. Maybe these additional data could form a second paper after the trial has been completed. Before committing yourself, and all other investigators, to collection of large amounts of data, though, the following questions should be asked:

- Is the information readily available for all participants?
- How likely are the data to influence the outcome of the trial?
- How will they be analysed?

Data that can only be collected on a minority of patients are unlikely to be influential on the outcome and if it is unclear how, or indeed, whether data are to be analysed it should not be collected. Collecting data that would not be analysed is putting an extra burden on participants and clinicians, without any ultimate purpose. Keeping data collection to a minimum also increases the likelihood of it being collected.

To obtain good follow-up:

- Keep it short—do not ask for unnecessary data
- Keep it simple—do not add unnecessary extra tests
- Keep it seldom—do not repeat assessments too many times.

1.7. Include every patient randomised—intention-to-treat analysis

Once someone is randomised into a trial, then they need to remain in the analyses, even if they stop taking the treatment, or never start in the first place. The whole point of randomisation is to create equivalent groups: once one starts excluding patients, then the groups cease to be equivalent.

Intention-to-treat analysis analyses every patient according to the treatment they were allocated, not what they received. In the case of protocol deviations, then ITT analyses tend to be conservative: they underestimate the treatment effect. So, if one sees a difference, then one can be reasonably sure that it really is there. Other sorts of analysis may look attractive, because they compare what actually happened, not what was meant to happen. But, as Box 2 shows, they can end up giving misleading results, and even making an ineffective therapy appear worthwhile. What ITT really measures is the effect of introducing the policy of giving the treatment: what is the real-world bottom line on introducing a new treatment?

1.8. Follow patients through to the end

Inevitably, there will be circumstances where women who have committed to the trial do not complete treatment

and follow-up. Sometimes this is treatment related: for example, if side effects are distressing or if there is no perceived benefit from the treatment. These drop-outs are not random and may appear more frequently in one arm than another, so it is crucial to note the reason for non-compliance or withdrawal, especially if no further follow-up information is obtainable. Failure to do this will introduce a systematic difference between the groups (known as attrition bias). Patients who withdraw from treatment should still be encouraged to contribute to data collection—it is a common misconception that a deviation from the protocol necessitates withdrawal.

In women’s health where participants tend to be young and relatively healthy, the population tends to be mobile. Collecting multiple identifiers and contact details at the start will assist in tracing participants over time, although this should be made clear at consent. A meta-analysis of methods of improving the response rate to postal questionnaires identified a number of successful strategies [14], such as a pre-contact before the questionnaire is sent, provision of a prepaid reply envelope and using a short, interesting questionnaire, in addition to obvious monetary incentives.

1.9. Do not go looking for significant subgroups—no data-dredging

A clinical trial sets out to answer whether a particular treatment is effective on average, for a wide range of patients. In this sense, while a trial can give a best guess on what treatment to give an individual patient, it does so by considering the population as a whole, and not each individual separately. It is tempting to try and identify subgroups of patients who benefit from an intervention. But, remember that a significant result ($p < 0.05$) will be seen by chance 5% of the time. So in a trial of an ineffective treatment, if one looks at 20 different subgroups of patients there will be, on average, a significant result in 1 subgroup. It is laughable to say that for women born on one particular day of the month putting brown sugar in one’s coffee is any better (or worse!) for pelvic pain than white sugar. But chances are that if you run a large RCT you will see this kind of spurious subgroup effect purely by chance. If you look hard enough you will find something significant—but it may be a chance finding. There are statistical techniques to test whether treatment effects differ between subgroups [15], but trials which want to look at subgroups typically need to be much bigger. To reliably detect a difference within subgroups of the same size as the pre-specified treatment effect, four times as many participants are required. Subgroup analyses therefore tend to identify qualitative differences (treatment works in one group, not in another), not smaller, quantitative, differences in the size of effect. If there are subgroups where there are legitimate reasons for anticipating a different response to treatment, then these should be specified in the trial protocol in advance, together

Box 2. Intention to treat

Consider a trial of a surgical procedure, which has no effect whatever.

There are two types of patient

Good risk—with a successful outcome 70% of the time

Poor risk—with a successful outcome 30% of the time
The patients in the trial are equally split between good and poor risk.

Not all surgeons abide by the treatment allocation. In 20% of the poor risk patients allocated no surgery, the situation gets so bad that surgery is considered essential. Because there are only a finite number of surgical slots, this means that 20% of the good risk patients allocated surgery do not actually get it.

Then, we have two tables of proportions of patients in each group: the intention-to-treat analysis

	Allocated surgery (%)	Allocated no surgery (%)
Good risk	50	50
Poor risk	50	50
Successful outcome rate	50	50

And the as-treated analysis

	Receiving surgery (%)	Not receiving surgery (%)
Good risk	40	60
Poor risk	60	40
Successful outcome rate	46	54

The ITT analysis correctly shows that the surgery is ineffective, but the as-treated analysis gives an 8% effect size against surgery. In the as-treated analysis, the groups are no longer comparable. The difference arises because more poor risk patients get surgery. As-treated analyses will tend to show an artificial benefit for the treatment received by the better risk patients.

with a justification of the underlying mechanism and the influence it may have on the outcome. For example, the large international trial of magnesium for pre-eclampsia specified severity of pre-eclampsia, imminent eclampsia, gestational age, whether they had an anticonvulsant in the previous 48 h, and whether they had already given birth in advance and presented results graphically as a forest plot [16]. The trial showed a statistically significant effect within one subgroup (prior anticonvulsant drug use) for one outcome measure (neonatal mortality), but noted that this was irrespective of pre-trial magnesium use, and therefore unlikely to be anything other than a chance finding.

The authors in this study quite rightly emphasised the overall results of the study. Where significant subgroup differences exist, they should always be viewed in the context of the overall findings. It is highly unlikely that if the overall result is positive, that a subgroup finding of significant harm is to be relied upon. Such qualitative effects are to be viewed with great scepticism, unless again there is some plausible, pre-specified reason why some subgroups should behave differently.

1.10. Put your results in context—meta-analysis

With the ongoing controversy surrounding objective reporting of all clinical trial results – both negative and positive – study results should be viewed in context. For too long, new trials have failed to be designed with reference to the findings of a systematic review of previous research. Not only can this lead to repetition of previous mistakes, it could lead to unnecessary trials being conducted as the superiority of one treatment tested over another was already known from previous work. Examples where superfluous trials were conducted and introduction of good practice was delayed by decades are now being to surface. The evidence about positioning babies to sleep on their back rather than their front to avoid sudden infant death syndrome has recently been brought together in a meta-analysis [17]. Whilst the “Back to Sleep” campaign was successfully promoted in the early 1990s by both government and charities, the benefits of this strategy would have been apparent if systematic reviews of known risk factors for SIDS had been done as long ago as the 1970s. Such a review could have prevented tens of thousands of infant deaths worldwide.

“Unnecessary and badly presented clinical research injures volunteers and patients as surely as any other form of bad medicine, as well as wasting resources and abusing the trust placed in investigators by their trial participants”. That is the conclusion of the editors of the *Lancet* [18]. This prestigious journal now requires authors to include a clear summary of previous research findings. This means that researchers should always attempt to identify recent high quality systematic reviews prior to commencement of the study and where none exists, undertake one themselves. This may seem like an onerous burden upon the would-be

Box 3. Trials in context

- When developing the research question, conduct a systematic review or identify a relevant review done by someone else. Check the Cochrane Library first.
- Learn from the achievements and mistakes of past trials but read their results sceptically.
- Discuss the findings of your study in the context of an updated systematic review of relevant research.
- When writing up a study, update the meta-analysis and include it in the report.

investigator, and may require skills, time or resources they do not possess, but as the above example shows, a review could demonstrate that the trial is unnecessary and therefore unethical.

The corollary is that at the conclusion of the trial, new data should be considered in relation to previous trials and the overall impact of the results discussed. This is best achieved by adding the trial into the existing meta-analysis and interpreting any effect that has to the overall result. It is, after all, the totality of trial data that provides the best evidence and overemphasis on any particular trial, even the one you conducted, can give misleading findings. **Box 3** gives guidelines for putting results in context.

2. Summary

Without proper evidence, evidence based practice is impossible. In assessing the effects of new treatments, it is impossible to overstate the importance of randomised controlled trials, and meta-analyses of trials. But randomisation alone is not enough. Trials need to be well designed, use appropriate endpoints, and be properly analysed before their results can be fed into clinical practice. These “ten habits of highly effective trialists” only scratch the surface of the subject; but it is hoped that applying these guidelines, a healthy scepticism, and one’s innate common sense cannot only help produce better clinical trials, but make the interpretation of clinical trial results easier. It is often said that learning from one’s mistakes is fruitful: it is also true that learning from others’ mistakes is less painful, so critical reading of existing research is a very good way of starting to design one’s own trial.

References

- [1] Altman DG. Statistical methods for medical research, 2nd ed., London: Chapman and Hall; 1991.
- [2] Duley L, Farrell B. Clinical trials London: BMJ Books; 2002.
- [3] Pocock SJ. Clinical trials: a practical approach. New edition London: John Wiley and Sons; 1996.
- [4] Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence London: BMJ Books; 2000.

- [5] Swinscow TDV, Campbell MJ. *Statistics at square one* London: BMJ Books; 2002.
- [6] Kettle C, Hills RK, Jones P, Darby L, Gray R, Johanson R. Continuous versus interrupted perineal repair with standard or rapidly absorbed sutures after spontaneous vaginal birth: a randomised controlled trial. *Lancet* 2002;359(9325):2217–23.
- [7] Altman D, Bland JM. Treatment allocation by minimisation. *BMJ* 2005;330:843.
- [8] Altman D, Bland JM. How to randomise. *BMJ* 1999;319:703–4.
- [9] Schulz KF. Subverting randomisation in controlled trials. *JAMA* 1995;274(18):1456–8.
- [10] Hills R, Gray R, Wheatley K. High probability of guessing next treatment allocation with minimisation by clinician. *Control Clin Trials* 2003;24(Suppl. 3S):70S.
- [11] Altman D, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311(7003):485.
- [12] LUNA Trial Collaboration. A randomised controlled trial to assess the efficacy of Laparoscopic Uterosacral Nerve Ablation (LUNA) in the treatment of chronic pelvic pain: The trial protocol [ISRCTN41196151]. *BMC Women's Health* 2003;3(6). 1472-6874-3-6.
- [13] Miller FG, Kaptchuk TJ. Sham procedures and the ethics of clinical trials. *J R Soc Med* 2004;97(12):576–8.
- [14] Edwards P, Roberts I, Clarke M, DiGiuseppi C, Pratap S, Wentz R, et al. Increasing response rates to postal questionnaires: systematic review. *BMJ* 2002;324(7347):1183.
- [15] Assmann SF, Pocock S, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355(3209):1064–9.
- [16] The Magpie Trial Collaborative Group. Do women with pre-eclampsia, and their babies, benefit from magnesium sulphate? *Lancet* 2002;359(9321):1877–90.
- [17] Gilbert R, Salanti G, Haredn M, See S. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *Int J Epidemiol* 2005;34(4):874–87.
- [18] Young C, Horton R. Putting clinical trials in context. *Lancet* 2005;366(9480):107–8.