Challenges in design and interpretation of chronic pain trials

R. A. Moore*, S. Derry and P. J. Wiffen

Pain Research and Nuffield Division of Anaesthetics, Nuffield Department of Clinical Neurosciences, University of Oxford, The Churchill, Oxford OX3 7LE, UK
* Corresponding author. E-mail: andrew.moore@ndcn.ox.ac.uk

Summary. The process of systematic review has shone a light on the methodology of randomized controlled trials. Notably, a range of potential biases hinders the interpretation of chronic pain trials. These include a consistent bias favouring active over placebo in trials that are small and of short duration. The use of the ‘last observation carried forward’ imputation method is known to inflate results, often generating statistically significance when adverse event withdrawals are high; in clinical practice terms, this is the wrong answer. Patients want outcomes of low pain scores, large reductions in pain and relief from associated symptoms, with improvements in ability to function and in quality of life. Some patients achieve this, but many do not. The distribution of benefit is skewed and the use of average pain scores, or change in pain, can be misleading compared with responder analysis in which withdrawal is regarded as non-response.

Historically, chronic pain trials have had a simple classic or a crossover design. They have been small and short, and used inappropriate imputation and outcomes unconnected to the experiences of most patients. While these designs are useful for answering some questions, they may be insensitive for many interventions. Newer designs, like enriched enrolment randomized withdrawal (EERW) trials or clinical effectiveness trials, are potentially more interesting and informative.

Keywords: bias; chronic pain; clinical trial

Editor’s key points

- Improving clinical trial design is important in the process of developing effective novel treatments.
- Factors introducing potential bias must be considered, particularly in small, short-duration trials.
- Using ‘average’ response to assess analgesic interventions may not equate to clinical effectiveness.
- Further high-quality trials with careful consideration of trials design and analysis are urgently required.

Chronic pain is prevalent, affecting an estimated one in five adults,¹ and has a large negative impact on quality of life² and ability to work.³ There is, however, a dearth of good evidence relating to how to treat chronic pain. This brief review aims to provide a realistic assessment of the situation now, and is a good place to consider how chronic pain trials of the future might be done, in the light of recent advances in understanding the deficiencies of chronic pain trials of the past. The point of view is principally that of drug trials, but the general principles apply to trials for any intervention in chronic pain.

A number of themes need to be considered together in any trial to achieve a result that makes sense. First is the question that you want a trial to answer. Second is the attention paid to issues of bias and quality; it cannot be overstressed that inattention to detail here can easily lead to getting a result that is quite wrong. Third is the trial design. There are not many different designs but they differ markedly. Some are explanatory, telling that an intervention works; others are pragmatic, telling us how to use the intervention to best advantage in our practice;⁴ other designs combine these properties.

What is the question?

More than one question can be posed for a chronic pain trial to answer. Examples include:

(i) Does this drug (intervention) work? This leads to some interesting secondary questions (e.g. what do you mean by ‘work’?). Indeed, whole chapters could be written on this subject, and are summarized below.
(ii) Is this drug (intervention) safe? Even more than for work, does the question of what is meant by ‘safe’ raise a range of secondary questions?
(iii) How well does this drug (intervention) work?
(iv) How well does this drug (intervention) work compared with the other drugs (interventions) available for this condition?
(v) For whom does this drug (intervention) work?
(vi) Is this drug cost-effective? This leads to another ‘what do you mean by...’ set of sub-questions.
(vii) What is the right place for this drug (intervention) in clinical practice, guidelines or formularies?

Evidence background

There is no point undertaking a trial, or even reading about a trial, if it is fundamentally flawed; it is worth remembering that there are numerous ways in which a trial can be wrong. The work of the Consolidated Standards of Reporting Trials (CONSORT) Group has done much to improve the conduct and reporting of trials. The CONSORT statement and checklists

© The Author [2013]. Published by Oxford University Press on behalf of the British Journal of Anaesthesia. All rights reserved. For Permissions, please email: journals.permissions@oup.com
are designed to be an aide-mémoire for all of us who look at trials (www.consort-statement.org/). But even CONSORT does not cover everything we need to have in mind when looking at chronic pain.

The most important types of bias are listed in Table 1. Some of these will be familiar, but several less familiar types of potential bias surround chronic pain trials.

Duration bias describes the propensity of effect sizes to decrease with trial duration. Longitudinal individual patient data analyses show decreasing effectiveness over time, particularly for analgesics with lesser effectiveness as with ibuprofen in osteoarthritis and pregabalin in fibromyalgia. Imputation bias relates to what happens when patients withdraw from chronic pain trials. It has become the norm to ‘impute’ (ascribe) data for the time points after withdrawal by taking the last pain measurement before withdrawal and carrying it forward to the end of the trial which can be several months later. This last observation carried forward (LOCF) method has been shown to grossly overestimate treatment effects when adverse event withdrawals are larger with active than placebo treatment. This is particularly the case with opioids (except tapentadol), where any efficacy in chronic non-cancer pain is probably due to this artefact.

Small studies historically show greater treatment effects than larger studies. This is probably attributable to small studies having other biases but there may also be a contribution from fraudulent data. In any event, much larger amounts of data are needed to be sure about both the direction of a result (does it work?) and the magnitude (how well does it work?).

**Starting point**

In order to measure analgesia, you must start with patients with pain. If there is no pain, or very little pain, analgesic effects of drugs cannot be measured. For this reason, chronic pain trials usually recruit patients with moderate or severe initial pain before treatment, or pain of at least 40/100 mm on a visual analogue pain intensity scale (VAS).

It is usual to take patients with established painful conditions off existing medication at the point where they have been screened to ensure that they meet other inclusion criteria. Pain scores are expected to increase at this point, and those who meet the pain intensity criteria are then randomized. The increase in pain on stopping established medicines is called a flare and pain scores throughout a typical flare trial in musculoskeletal pain are shown in Figure 1. Studies using a flare design can show larger effect sizes of treatment than those that do not use a flare.

**Outcome**

The single most important issue in chronic pain trials is that of starting point. There has been much discussion over the years about minimally important or clinically important differences in pain

<table>
<thead>
<tr>
<th>Table 1 Types of bias found in trials in chronic pain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias</strong></td>
</tr>
<tr>
<td><strong>Detail</strong></td>
</tr>
<tr>
<td>Selection bias</td>
</tr>
<tr>
<td>Randomization. Sequence generation needs to be truly random, using random number tables or computer-generated randomization. Reduces selection bias. Quasi randomization using date of birth or hospital number is not considered reliable</td>
</tr>
<tr>
<td>Allocation concealment. Ensures that those conducting the study do not know to which group participants are assigned. Not the same as blinding as it happens before treatment commences. It should be explicitly reported, typically to state that none of the investigators were aware of the allocation</td>
</tr>
<tr>
<td>Performance bias</td>
</tr>
<tr>
<td>Blinding of participants, personnel and outcome assessors. Measures taken to ensure that participants and personnel do not know which intervention a participant received. The methods used need to be described to show that blinding was effective. For double blind trials treatments may be described as being identical in appearance, size and taste. The use of double dummy or more complicated techniques may be needed</td>
</tr>
<tr>
<td>Attrition bias</td>
</tr>
<tr>
<td>Incomplete outcome data. There should be a description of the completeness of the outcome data for each main outcome for all intervention groups. Withdrawals from the study should be described with reasons together with any assumptions made in the analyses</td>
</tr>
<tr>
<td>Duration bias</td>
</tr>
<tr>
<td>In chronic pain, studies should ideally be 12 weeks long. The minimum should be 8 weeks</td>
</tr>
<tr>
<td>Imputation bias</td>
</tr>
<tr>
<td>Dealing with data when patients have withdrawn from treatment. This is typically dealt with by carrying results from the last observation forward to the end of the trial (LOCF). As many as 30–60% of patients withdraw in chronic pain trials</td>
</tr>
<tr>
<td>Size</td>
</tr>
<tr>
<td>Only a minority of patients benefit from chronic therapy, so larger trials are needed to overcome the random play of chance. This usually means having 100–200 patients per group as a minimum, and sometimes more. Even then, this only tells about direction (whether something is different from placebo for instance). Uncertainty about the magnitude of the effect size remains</td>
</tr>
</tbody>
</table>

| **Reasons**                                         |
| Non-randomized, or improperly randomized trials overestimate the effects of treatment |
| Where allocation is not effectively concealed, treatment effects tend to be higher |
| Unblinded, or open, trials overestimate the effects of treatment |
| There is a tendency for studies with high withdrawals to overestimate effects, or with incomplete reporting to report only those outcomes with significant benefits |
| Shorter durations, especially <4 weeks, overestimate effectiveness compared with longer trials |
| LOCF produces major overestimation of treatment effect when adverse event withdrawals are high |
| Historically, small trials in chronic pain have been shown to overestimate treatment effects. This is probably because small trials have other quality problems |
trials. This is interesting from an academic perspective but the key should be what the patient wants. Studies of what patients seek from treatment of chronic pain are few but report a consensus that patients want:

1. Major reductions in pain of 50–70%.
2. Ideally no worse than mild pain.
3. Major improvements in the considerable concomitant problems with chronic pain, including poor sleep, fatigue, depression, function, and impaired work and quality of life.

Historically, chronic pain trials have not reported these outcomes, but tended to report average pain scores or average changes in pain scores. This is problematical for several reasons. One is that pain responses in chronic pain are probably not normally distributed but have a bimodal distribution with the majority of patients having either very good pain relief or very little. Consequently, the ‘average’ pain score or change represents a result that few patients experience. Another is that small ‘average’ pain differences between active and placebo hide the fact that a substantial minority achieve extremely good levels of pain relief.

Moreover, good pain relief is now being shown to be consistent with large improvements in sleep, fatigue, and depression, and also function and work, so that quality of life gains upon successful treatment (at least 50% pain intensity reduction, or pain no worse than mild) are significant, and essentially reverse the quality of life loss associated with having chronic pain.

The bottom line is that the preferred outcome in chronic pain trials is now generally one of:

- Pain intensity reduction of at least 50% or no worse than mild pain.
- Tolerable adverse events.
- Able to continue with medication without withdrawal for (ideally) 12 weeks.

The outcomes of chronic pain trials and patient expectations or desires from treatment now match very well. Unsurprisingly, given this background, any trial of any design that reports only average results will be of little value and its results almost impossible to interpret except in the most general terms.

**Control**

The majority of chronic pain trials will use a placebo control. For pain, but virtually no other clinical condition, placebo can be better than a no-treatment control. Patients given placebo or active therapy will have recourse to additional analgesic (also called rescue or escape analgesia) but this is often only paracetamol up to 2 g daily. Lack of efficacy withdrawal is frequent with placebo.

In neuropathic pain or fibromyalgia, almost all new therapeutic drug trials use only a placebo control. In musculoskeletal pain, almost all new therapeutic trials use both a placebo and an active control.

**Classic designs**

A classic design is shown in Figure 2. These trials are done primarily to demonstrate that an intervention works—better than placebo.

In this design, patients properly diagnosed with a chronic pain condition are initially screened for suitability. This will include initial pain intensity, perhaps recorded daily over one or two weeks to ensure that it is both consistent and of a sufficient magnitude. It will also involve checking against pre-set criteria for inclusion and exclusion.

The whole population is then randomized to treatment and a range of assessments made over the period of the trial,
ideally weekly over about 12 weeks. The assessments can include pain intensity, pain relief, global impression of change and other measures like sleep, depression, fatigue, symptoms, adverse events, and quality of life measures. Disease-specific measures, of which there are many, are also used (e.g. Fibromyalgia Impact Questionnaire, Western Ontario and McMaster Universities Arthritis Index).

Figure 3 shows the percentage of patients with at least 50% pain intensity reduction at 12 weeks (or equivalent) from systematic reviews and meta-analyses seeking to use the highest quality data. There are two key points: responses with placebo can be substantial, and gains over placebo with active therapy are at best moderate. NNTs calculated from these data are between 5 and 10.

The reason for the high response rates with placebo is uncertain. Part will be regression to the mean with pain reduction after a flare, part a consequence of using newspaper and other advertisements to recruit participants rather than clinic patients, and part will be the acknowledged benefits of being in a trial and receiving additional attention. The fact of high placebo response rates and modest benefit with active treatment is that classic trials will lack sensitivity to determine whether or not a drug works. Attempts to minimize placebo responses do not yet seem to have any great success.
Given the inherent insensitivity of the classic design, it will continue to be a common occurrence that two identically designed trials will differ in the level of significance achieved. Such a result is not unexpected. Higher rates of response with placebo and active-specific benefit of 10% or less will result in no significant difference from placebo unless trial size is very large.

**EERW designs**

Enriched enrolment randomized withdrawal (EERW) designs (Fig. 4) are helpful in determining whether an intervention works. Classic trials have used a form of partial enrichment to try and improve sensitivity. This has consisted of trying to exclude patients who may have previously had adverse events with a particular drug or class of drugs, and to include patients who have responded favourably to a drug class. This has not been successful.

More useful has been to use complete enrichment by performing an open titration with the drug under test after initial screening. Some patients will have intolerable adverse events and for others the drug will lack efficacy. These will be withdrawn and only those for whom the drug has proven efficacy with tolerable adverse events enter the trial properly (Fig. 4). Randomization is then between continuing on the proven dose of active or placebo in a double blind test of persisting efficacy.

These EERW trials are few in number. Probably the best to date is the FREEDOM trial of pregabalin in fibromyalgia. It was large, long (6 months) and had stringent entry criteria; to be randomized, patients must have had ≥50% decrease in pain and a self-rating of ‘much’ or ‘very much’ improved on Patient Global Impression of Change. The outcome was loss of therapeutic response, defined as <30% reduction in pain from baseline or worsening of fibromyalgia. For pregabalin in fibromyalgia, the FREEDOM study rejected about 50% of patients before randomization but 68% of those continuing on pregabalin continued to have effective analgesia over 6 months compared with 39% on placebo. The long-term drug specific benefit was 29% more than 6 months in the randomized patients, equivalent to about 13% in the whole fibromyalgia population. Overall, the drug specific benefit was similar to that in classic trials, but the sensitivity of the trial was increased.

Using current best methods neither classic nor EERW trial designs have shown benefit for opioids in chronic non-cancer pain. EERW trials have been suggested as a useful design in the early phases of drug testing in humans.

**Effectiveness designs**

An outline of a clinical effectiveness design is shown in Figure 5. This design can answer questions about how well an intervention works in clinical practice, how well it works compared with the other interventions available and may also answer questions about what order of interventions may be best in a care pathway. However, broader issues can also be involved. The general principle is that patients with a particular clinical diagnosis are randomized to one of several active therapies where titration of dose to pain relief with tolerable adverse events takes place. With treatment success, patients continue on treatment. With failure, they are re-randomized to another option and so on until treatment success is achieved.

While the goal of treatment is good pain relief in the individual, the trial identifies the care pathway providing the highest proportion of successes in the shortest time and at the lowest cost (cost-effectiveness measures are easily included in such a design). These are critically important questions that classic and EERW designs do not answer. Unfortunately,
there is not a single example of a trial using this design in chronic pain.

In depression, clinical effectiveness trials of various designs have been used to tease out best initial treatment choices and the effects of formulary limitation. For example, a randomized trial of several initial therapies showed broad similarity between different drug classes in effectiveness and cost. Another agreed that single agents might be equally effective but that the availability of several drugs led to higher success in treating a population with depression than any one drug could do.

Crossover designs

In crossover trials, patients receive a sequence of alternative therapies (Fig. 6); patients act as their own controls to eliminate confounding factors. The sequence can include placebo or a sequence of active drugs, including combinations. They are used to answer questions about whether drugs work or whether one drug may be better than another.

A major problem with crossover trials is that they tend to be short (usually 2–4 weeks) which limits their applicability. There are issues about the time needed (if any) for washout between treatment periods. Withdrawal rates tend to be high with multiple crossovers, which can be tedious for patients, and that limits applicability or use of paired data between treatments for the same patients. And poor reporting, inter alia, limits their use in meta-analysis, possibly with some biases.

However, even small crossover trials can have rewarding results. They can show that patients with chronic pain who react poorly with one drug may do well with another closely related drug, as with amitriptyline and nortriptyline.

Adaptive designs

Clinical trials can be designed that adapt with changes in the trial after interim examination of the accumulating data at interim times. It is considered to be useful by making trials more efficient by needing fewer patients for a shorter duration. It is thought (without much evidence yet) that adaptive designs are more likely to demonstrate an effect of the drug if one exists, or be more informative.

An example of an adaptive design might be during the early stages of testing a new drug in humans, where the determination of dose is important. In an adaptive design, a very wide range of doses could be tested initially, and those doses with no early effect, or with effect but too severe adverse effects, are dropped while more patients are recruited into groups.
with doses closer to the optimal. This would be better than providing equal numbers to all doses irrespective of outcome. There is a little evidence that this approach can work, but only in identifying a drug without efficacy.34 The most useful discussion of the difficulties is in Food and Drug Administration guidance.35

There are few examples of adaptive designs in pain, especially chronic pain. Those claiming to be adaptive are actually enrichment designs.36

**Discussion**

To our knowledge, no design can directly address the question of for whom any particular intervention will work. The best hope would be individual patient data analysis in any of these designs where some specific patient characteristic could be reliably linked to good pain relief, to the extent that it was useful in clinical practice. Very large inter-individual genetic differences, together with complexities in handling pain in the central nervous system, make this unlikely.

Similarly, none of these designs can, alone or together, answer questions about safety relating to rare but serious adverse events. The largest body of evidence in chronic pain comes with pregabalin, with around 8000 patients exposed. If there are no rare but serious cases of harm, we can be 95% sure that something very bad will not occur more often than one patient in about 2000 after 3 months of exposure.37 Clearly, observational studies are the only designs that can elucidate this question.

Deciding what is cost effective probably does not need a specific trial design. Certainly, the capture of data for cost effectiveness analysis in pain trials can be highly advantageous and enlightening, and there are some examples of trials in acute pain having cost effectiveness as their main outcome.38 The use of individual patient data analyses will prove decisive in accurately determining cost effectiveness. Patients with very good pain relief, or whose pain is reduced to no worse than mild, record impressive improvements in sleep, fatigue, depression, functioning, and quality of life. Quality adjusted life year (QALY) improvements are 0.1–0.2 out of a maximum of 1. Because drug costs are usually modest (4p to about £2 a day), costs per QALY are below £6000, well below the maximum £20 000 per QALY. The constraints of health economics have to be addressed.39

Finally, it needs to be remembered that while all clinical trials have pros and cons (Table 2), individual trials can differ enormously in terms of the patients included and excluded, and interventions used. Even then, chronic pain treatment is often multimodal and the evidence about a single intervention needs to be tempered with wisdom and experience to be used sensibly in clinical practice.

**Declaration of interest**

R.A.M. has received payments and travel funding for the UK and overseas lectures from MSD, Pfizer, Eli-Lilly, Reckitt Benckiser, and Menarini, and has participated in contracted research (investigator initiated) with MSD, Pfizer, and Reckitt Benckiser.

**Funding**

No specific funding was received for this work. The Oxford Pain Relief Trust provided institution support.

**Authors’ contribution**

All authors contributed equally.

**References**


---

**Table 2 Pros and cons of different types of trial design**

<table>
<thead>
<tr>
<th>Trial design</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic</td>
<td>Determines fact of analgesic effect when compared with placebo</td>
<td>Unhelpful with small size, inadequate outcome, short duration, or results reported as average</td>
</tr>
<tr>
<td></td>
<td>Can determine magnitude of analgesic effect if very large, or sufficiently large and can be combined with other, similar, trials</td>
<td>Lacks sensitivity when intervention specific effect (active minus placebo) approaches 10%, unless of exceptional size</td>
</tr>
<tr>
<td>EERW</td>
<td>Increases sensitivity by removing definite non-responders</td>
<td>Not generally accepted, nor much understood</td>
</tr>
<tr>
<td></td>
<td>Can give useful indication of overall proportion of patients who benefit</td>
<td>Few good examples as yet</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>Should provide evidence for clinical practice</td>
<td>No examples of this design in chronic pain</td>
</tr>
<tr>
<td></td>
<td>Could be done using large numbers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ideal for primary care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Should provide evidence about responders to particular drugs</td>
<td></td>
</tr>
<tr>
<td>Crossover</td>
<td>Should provide evidence about responders to particular drugs</td>
<td>Studies tend to be small, of short duration, with selected patients, and suffer high withdrawal rates, making them less reliable</td>
</tr>
<tr>
<td></td>
<td>Can compare different active drugs</td>
<td>Need for washouts and effect of duration of washout controversial</td>
</tr>
</tbody>
</table>

---

44
11 Moore RA, Derry S, McQuay HJ. Fraud or flawed: adverse impact of fabricated or poor quality research. Anaesthesia 2010; 65: 327–30
12 Moore RA, Gavaghan D, Tramer MR, Collins SL, McQuay HJ. Size is everything–large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. Pain 1998; 78: 209–16
21 O’Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. Control Clin Trials 1997; 18: 550–6
28 Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003; 290: 1624–32