Methodological and Statistical Aspects


What Are the Minimal Methodological and Statistical Requirements for a Good Trial?

I. The Clinician's View

Sjef van der Linden², Lex Bouter³, Peter Tugwell⁴

²Department of Internal Medicine, Division of Rheumatology, and ³Department of Epidemiology and Health Care Research, University of Limburg, Maastricht, The Netherlands; ⁴Department of Medicine and Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont., Canada

Physiotherapy (PT) is one of the most frequently prescribed therapeutic regimens. For most of the patients involved, PT is presumed to reduce pain or to restore (or improve) function. In addition to this, the aim of PT is often to prevent further deterioration of function, new episodes of pain, or episodes of dysfunction. For the clinician considering a PT treatment for a patient the crucial question is whether this intervention will be effective in terms of these goals. There are usually several options for intervention (physiotherapeutic or non-physiotherapeutic), between which a choice has to be made. The question has to be answered whether the PT intervention at issue is better than doing nothing (no intervention) or doing something else. The scientific question behind this process of medical decision-making is whether the PT intervention has any specific effect on pain or function.

In other words, the main issue is whether the difference between the pre-treatment and the post-treatment conditions can really be attributed to the PT, i.e., does a better clinical condition after a certain PT treatment really imply that this effect is at least partially due to the PT provided? The answers to these questions are important in providing effective and efficient health care. Which type of PT (e.g. massage, exercises, electrotherapy or ultrasound) should be prescribed to patients with a certain disorder? Which regimen (e.g. daily or twice a week) gives the best results, in terms of both effects and costs?
Answers to questions like these can only be found by performing properly designed studies. It is only such studies which are able to distinguish between specific PT effects and effects due to other causes. If they are to provide valid answers, these studies should be precise (lack of random error) and valid (lack of systematic error, i.e., unbiased).

In this chapter, we will discuss why randomized controlled trials (RCT's) are the most powerful instruments for addressing the important questions about the effects of PT. We will also present a set of guidelines which may be useful in the critical appraisal of articles dealing with the effects of PT interventions. These methodological criteria can also be used in designing new clinical trials of PT applications.

**Efficacy and Effectiveness**

It is useful to distinguish between efficacy studies and effectiveness studies. Efficacy asks the question 'Can it work?' It is defined as the extent to which a health intervention does more good than harm to patients who are diagnosed correctly, and appropriately cared for, and who comply fully with recommendations for treatment. That is, efficacy assumes optimal diagnostic accuracy and compliance of both health providers and patients. Therefore, it yields the maximum benefit that can be achieved under 'ideal circumstances' [1]. Effectiveness relates to the benefit that can be obtained under the circumstances in 'real life'. In the community, diagnostic accuracy and patient and provider compliance are frequently suboptimal.

**Randomized Controlled Trials**

Why do we need (randomized controlled) clinical trials in PT? Evaluation of the specific effects of PT is often difficult because the difference between the post-treatment state and the initial state might be a consequence of the PT, but also of many other factors. That is, the effect observed must be regarded as being composed of several underlying phenomena. In addition to the specific effect of the treatment at issue, the observed effect can also reflect effects due to other causes. For example, spontaneous changes due to the natural history of the disease will influence the observed effect as well. This means that observed improvements (or deteriorations) in clinical status may occur in the absence of any specific
effect of the therapeutic intervention. Examples of prognostic factors that can be responsible for differences in outcome between patients are: age; sex; aspects of lifestyle; duration and stage of the disorder, and treatment in the (recent) past. Imbalances in several factors may already be present at the start of a study (baseline differences). But also after the start of a trial, other factors may contribute to the observed effect of a treatment. Examples of these are compliance during the trial and co-interventions such as analgetic (self)medication. Therefore, it will be clear that the observed effect will usually differ quite considerably from the specific effect of a PT intervention. Consequently, the most fundamental method for assessing whether a given PT intervention has relevant specific effects on the outcome, is to perform an experimental study with two groups: an experimental group and a control group. The control group should have the same distribution of prognostic factors at baseline as the experimental group, and should be given the same treatment as the experimental group, except for the PT intervention in which we are interested. Any differences in effect between the two groups is then only due to the difference between the PT intervention at issue and the comparative maneuver. The latter can be another intervention (PT or non-PT), a placebo treatment or no intervention.

However, this is true only if the distribution of the prognostic factors at baseline can indeed be regarded as comparable. This is why randomization is such a powerful tool in controlled trials. Randomization means that every patient will have an equal (usually 50%) chance of being assigned to the experimental group or the control group. Consequently, prognostic factors will be randomly distributed over the groups. Randomization thus also helps in eliminating bias by removing patient or physician preferences that might systematically assign the PT intervention to patients with good (or bad) prognostic factors [2].

However, randomization cannot guarantee similarity between the groups. Especially if the numbers of patients included are rather small, some bias due to dissimilarity for one or more factors might easily arise by chance. Therefore, even in a randomized study the prognostic comparability at baseline should always be carefully evaluated.

Every step of a RCT should be properly designed. In the context of this paper, it is not possible to discuss the design and execution of RCTs in depth. Here, the reader is referred to outstanding textbooks on this subject [3, 4].

After being checked for eligibility, the patients who enter the trial will be randomly assigned to receive the PT intervention at issue or an alterna-
tive intervention (PT or non-PT), placebo treatment, or no intervention. After the PT regimen has been completed, all clinically relevant outcomes will have to be assessed. If a difference between the groups can be established (in favor of the experimental group) and if there is no indication that this difference is due to bias, then the effect can be attributed to the PT intervention given to the experimental group.

In developing RCTs for PT it should always be kept in mind that the skill of the performer (the physiotherapist) is an important aspect of the intervention, because this will usually determine the specific effect to a large extent.

RCTs constitute a powerful scientific tool. However, contradictory results from different RCTs on the same question may be encountered. Two frequent types of causes of contradictory results are: (1) methodological causes, when trials incorporate design errors that lead to bias and consequently to invalid conclusions, and (2) statistical causes, when too few patients are enrolled in a trial. The latter leads to imprecision in the estimation of the specific effect, so that true therapeutic differences of a relevant magnitude might easily be missed.

**Guidelines**

We will now present a set of guidelines for the critical assessment of articles on RCTs about the effects of PT. These methodological criteria can also be useful in designing new RCTs. These guidelines have been adapted from general guidelines for intervention studies [5]:

1. Was the assignment of patients to treatment really randomized?
2. Were all clinically relevant outcomes reported?
3. Were the patients studied reasonably similar to your own?
4. Were both statistical and clinical significance considered?
5. Is the physiotherapeutic intervention feasible in your practice?
6. Were all patients who entered the study accounted for the publication?

(1) Was the assignment of patients to treatment really randomized?

Every patient who entered the trial should have had the same chance (usually 50%) of receiving either the PT intervention of interest or the comparative maneuver. Thus, assignment to the (new) PT modality or to
the other treatment should have been carried out by random allocation, which eliminates many of the biases that usually lead to over-optimistic conclusions in non-randomized trials [6]. Furthermore, it is important to check whether the similarity of the groups at baseline for important prognostic factors has been documented. If one does not want to risk unequal distribution of an important prognostic factor by chance, it might be advisable to stratify for that particular prognostic factor first and to randomize within these strata.

(2) Were all clinically relevant outcomes reported?

The question the authors should have addressed is: which is the relevant spectrum of outcomes for the particular question this study is trying to answer? Generally, outcomes can be assessed according to the 5 Ds: discomfort (including pain); disability; death; dollars; drug (or adverse) effects [7].

Of these, discomfort, disability, costs and adverse effects are probably the most relevant with respect to PT interventions. The disability relates to physical, emotional (psychological) and social functioning. These aspects are frequently considered as an index of 'quality of life'.

Instruments for assessing outcomes should always be designed carefully, with the intention of answering the central questions of the study. Usually, these outcomes will reflect changes in pain or function. This means that outcome evaluation will often consist of checklists or questionnaires. Proper attention should be given to the validity and precision of these outcome measures [8, 9].

In PT trials it is not appropriate to consider only effects on pain or function. In addition, it is also important to quantify, among other things, the number of hours lost in travelling to the PT unit, the time spent there, any adverse reactions such as unpleasant sensations or pain during or after the intervention, the costs for the patients, but also the costs for the insurance company and society in general.

It is especially important to ensure that the outcome assessment has been performed 'blindly', i.e. without knowledge of which intervention the patient was actually receiving.

The reader should also be convinced, that explicit and objective criteria for the clinical outcomes of interest have been used. If possible, the patients should also be blinded with respect to the intervention they receive, in order to avoid bias due to possible preferences for one of the interventions applied in the study. This is especially important when patients have to rate the outcome scales themselves.
(3) Were the patients studied reasonably similar to the reader's own patients?

Even if the results of the study are valid internally, i.e. for the study population, this does not imply that the results are relevant for your practice. This guideline has two aspects. First, the patients studied must be recognizable; that is, their clinical status and the distribution of prognostic factors must be described in sufficient detail for you to be able to recognize the similarity between them and your own patients. Second, the patients in the study must be reasonably similar to those in your practice. If both recognizability and similarity are satisfactorily established, clinical readers will be able to apply the study results to their own patients.

This guideline, consequently, requires that reproducible inclusion and exclusion criteria have been applied and are properly described in the study together with appropriate references. Patients treated at baseline must also have had a suitable clinical condition warranting treatment. If pain or physical disability are clinically relevant outcomes, then the patient should have 'enough' pain or impairment before the start of the study. In other words, there should be enough potential for improvement.

(4) Were both statistical and clinical significance considered?

If statistically significant, was the difference important enough to be clinically significant? If not statistically significant, was the study big enough to show a clinically important difference if there was any?

Clinical significance here refers to the importance of a difference in clinical outcomes between the study groups (experimental intervention, control intervention, placebo, no intervention). It is often described in terms of magnitude of result and usually evaluated as the difference or the ratio between the scores on the effect parameter in the study groups. The magnitude of the result can be considered clinically relevant if it leads to a clear preference in clinical practice for one of the interventions.

By contrast, statistical significance merely tells us something about the precision of the study result, not whether it is important. More precisely, the statistical significance is nothing more than a statement about the likelihood that the study result is due to chance alone. Conversely, when the population in particular study is too small, even large differences in potential clinical significance may not be statistically significant. Consequently, when the study result is not statistically significant, the reader should always wonder whether the study was large enough to detect a true difference of clinical importance. The mathematics of this issue is called power or sample-size calculations, for which formulas are provided in most text-
books on RCTs [3, 4]. One should always bear in mind that statistical
significance, frequently expressed as a p value, is to a large extent a func-
tion of the study sample size and the variability in measuring the effect
parameter. Furthermore, p values refer to precision (lack of random error)
and have nothing to do with the validity (lack of systemic error or bias) of
the RCT [10].

(5) Is the physiotherapeutic intervention feasible in your practice?
(a) Available, affordable, sensible? The PT intervention should be
available, acceptable, affordable and executable in your setting. Which
equipment has been used? Which amount of energy has been applied? Was
it clearly stated which anatomic sites should be treated? Was it made clear
how long the application should last? Which instructions to the patients
were provided? What exercises were the patients asked to do at home?
Were exact, written instructions given to the patients? Does the interven-
tion make clinical and biological sense? What is known about the biologi-
cal (side) effects?

(b) Were contamination and co-interventions avoided? Do the authors
show that the effects are not due to contamination? This would be the case
if some control patients accidentally received the experimental treatment.
It would result in a spurious reduction in the difference in clinical out-
comes between the experimental and control groups. Is it proved that the
effects are not due to co-intervention, meaning that additional therapeutic
maneuvers are performed on experimental but not on control patients?
This would result in a spurious increase in the difference in clinical out-
comes observed between the experimental and control groups.

(c) Were the interventions in the study administered blindly? To what
degree was it possible to administer the PT intervention blindly to the
patient and was the assessor of the outcomes blinded as well? Could the
blindness be sustained throughout the study? Did the authors evaluate the
blinding properly, or was the placebo treatment frequently unmasked?

(d) Was compliance measured? How frequently did the patients miss
visits for the intervention? How satisfied were the patients (and the phy-
siotherapists) with the intervention? What were the reasons for non-com-
pliance?

(6) Were all patients who entered the study accounted for in the pub-
lication? Were drop-outs, withdrawals, and non-compliers identified and
appropriately dealt with in the analysis of the data?

The authors should carefully report results on outcome parameters for
all patients who entered the study and were randomized. Such a data anal-
ysis is called intention-to-treat and forms the best guarantee against bias due to self-selective processes during the trial. In addition to the intention-to-treat results, the results for subgroups of study participants (e.g. full compliers) can be given to get an impression of the effect under optimal conditions.

References


Sjef van der Linden, MD, Department of Internal Medicine, Division of Rheumatology, University of Limburg, PO Box 1918, NL–6201 BX Maastricht (The Netherlands)