Improving the quality of clinical trials in physical therapy

Bouter, L.M.

published in
Improving the quality of physical therapy: invited lectures
1995

document version
Publisher's PDF, also known as Version of record

5. IMPROVING THE QUALITY OF CLINICAL TRIALS IN PHYSICAL THERAPY

L.M. Bouter

In the Netherlands every year 12% of the population is treated at least once by a physical therapist, most frequently for back and neck pain. Each series of treatment consists on an average of 20 sessions, and the number of physical therapy sessions has increased by 16% over the last decade. About 60% of the sessions involve exercise therapy, while in 70% or more physical therapy modalities are used, such as ultrasound or laser therapy (1). In order to obtain an insight into what is known about the efficacy of the most prevalent forms of physical therapy for the most common indications, we summarized the literature in a number of meta-analyses (2). This presentation will start by explaining the strategy we used in reviewing the literature and will formulate some impressions regarding the state-of-the-art. Secondly, I will focus on the four most prevalent methodological flaws we identified (3), and discuss options to prevent them. Thirdly, some concluding remarks will be made on the desirable future course of efficacy research in physical therapy (1,2,3).

Eligibility for meta-analyses

From the literature we identified both explanatory and pragmatic randomized clinical trials dealing with indications and interventions relevant to physical therapy. Outcome measures should include pain, mobility, functional capacity or activities of daily living. Although we identified a large number of RCTs for a wide variety of indications, I will restrict myself in this presentation to the global results of 10 meta-analyses covering some 200 RCTs in total (4-10).

Search strategy

Starting from an explicit question regarding efficacy, we initially search bibliographical data-bases, such as Medline and Embase. In our experience, this typically identifies 40-50% of the eligible studies. Subsequently, this is supplemented by screening of non-indexed journals or proceedings, and by citation-tracking. Finally we ask experts and authors in the field whether they consider our list to be complete.

Review methods

Although randomized clinical trials offer the best chance for a valid study of efficacy, even RCTs can be seriously biased (11,12). Therefore, in reviewing we used a predefined set of methodological criteria and corresponding weights (13). These criteria were operationalized explicitly for every meta-analysis, and were applied independently by 2 or 3 reviewers. These were blinded for the authors, the journal and the outcomes of the study. Typically, there was 70 to 80% initial agreement on item level. Differences were resolved by discussion, and for every trial a methodological score was calculated on a scale of 0 tot 100. In the resulting article only the outcomes of the best studies were discussed. Sometimes for each study the difference in success rate between the groups and the corresponding confidence interval was calculated. Due to the fact that populations, interventions and outcomes always differed substantially over the studies, we never decided to pool the data statistically (14).

Table 1: Methodological criteria and weights used in the meta-analyses

Criteria	Weights
Study population e.g. < 10% loss to follow-up	35 (4)
Interventions e.g. co-interventions avoided	25 (5)
Measurement of effect e.g. blinded outcome assessment	30 (10)
Data presentation and analysis e.g. intention-to-treat	10 (5)

Table 1 shows the four categories of methodological criteria, each with an example, and the corresponding weights. Of course, these weights are arbitrary to some extent, but all our meta-analyses allow for recalculation using different weights. It usually turns out that the methodological ranking is not very sensitive to a change in weights (7).

Table 2: Methodological scores in the meta-analysis on traction for back and neck pain (6)

study	total score	indication	overall conclusion
1	68	chronic LBP	neg
2	52	acute LBP	neg
3	51	chronic cerival pain	neg
4	46	acute LBP	pos
5	45	prolapsed disc	neg
6	44	acute LBP	neg
7	41	cervical pain	neg
8	39	LBP	neg
9	36	subacute cervical pain	pos
10	36	prolapsed disc	neg
11	36	prolapsed disc	neg
12	34	acute LBP	neg
13	34	acute LBP	neg
14	28	prolapsed disc	neg
15	25	chronic LBP	pos
16	24	LBP	neg
17	23	chronic LBP	pos

The range of methodological scores is often large. For traction in back and neck pain, for instance, the total score ranges from 23 to 68 points (6). Four out of 17 trials conclude that traction is more effective than the reference treatment or a placebo. Among the 7 studies with the highest score, only one is positive. Of course, the cut-off at 40 points is completely arbitrary. Therefore, it is preferable to present the relation between the methodological quality and the outcome of the studies as a cumulative frequency distribution.

Figure 1: Spinal manipulation for back and neck pain: Relation between methods score of trials and their results (4)

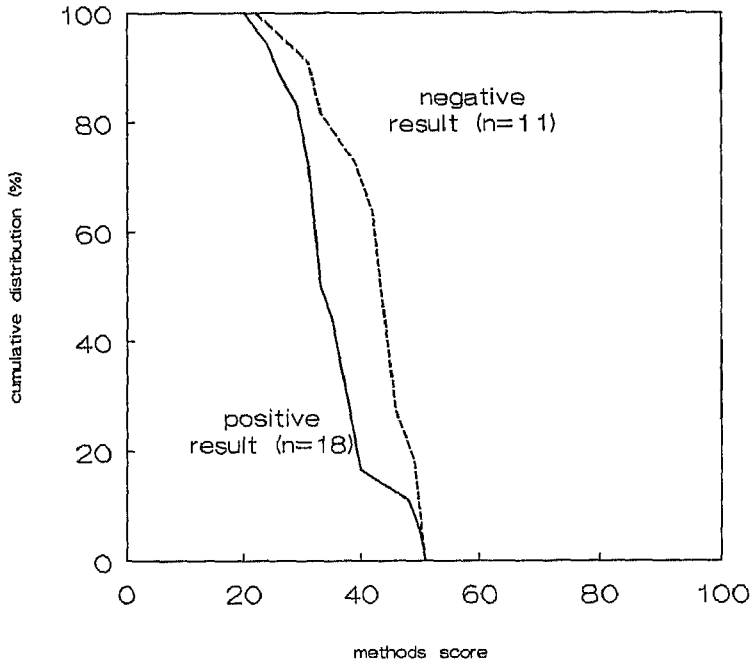

Figure 1 shows such a distribution for 18 positive and 11 negative trials dealing with the efficacy of spinal manipulation for back and neck pain (4). If the boundary between acceptable and unacceptable methodological quality would again be drawn at 40 points, it would be concluded that less than 20% of the positive trials were of acceptable quality, while this was the case for about 70% of the negative trials. From the graph you can see that for the whole range of possible cut-off points the negative studies tend to be better.

This is by no means necessarily so. As you can see, when studying the efficacy of physical therapy exercises for back pain (5), the positive studies tend to be of higher methodological quality (see figure 2).

Figure 2: Physiotherapy exercises and back pain: Relation between methods score of trials and their results. (Positive result shows exercise is better than reference treatment, negative result shows exercise is no better or worse than reference treatment) (5)

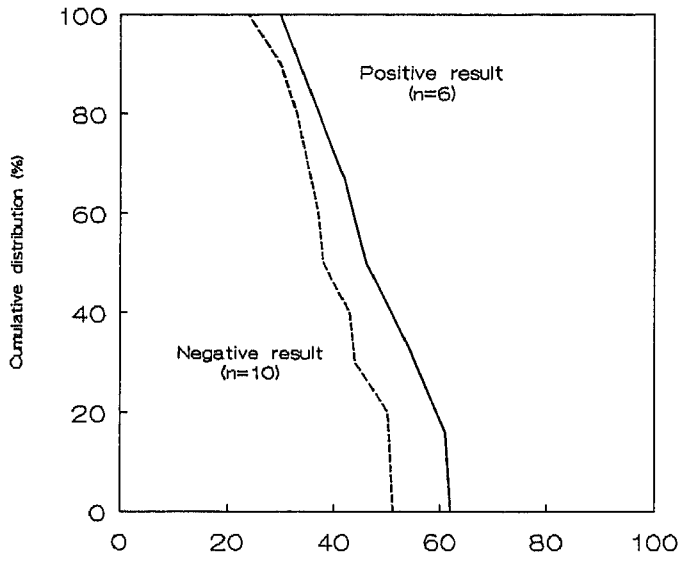

Table 3: Overview of meta-analyses

Indication	Intervention (reference)	Number of RCTs	Median score (range)
Back pain	Spinal manipulation (4)	30	35 (20-56)
	Exercise therapy (5)	16	40 (24-61)
	Traction (6)	17	36 (23-68)
	Back schools (7)	16	36 (16-70)
Neck pain	Spinal manipulation (4)	5	39 (26-50)
	Traction (6)	3	39 (36-51)
Shoulder complaints	Physiotherapy (8)	18	49 (22-76)
Knee disorders	Physiotherapy (1)	63	29 (6-52)
Musculoskeletal disorders	Ultrasound therapy (9)	16	41 (17-70)
	Laser therapy (10)	33	40 (4-72)

In table 3 an overview of the 10 meta-analyses is presented (4-10). Although we were surprised by the large number of RCTs available in the literature, we were generally disappointed by the methodological quality of the studies. While the best studies included in a meta-analysis usually scored something between 60 and 70 points, the median scores were typically low and in the range between 30 and 40 points. Consequently, we hesitate to draw strong conclusions regarding efficacy in the resulting meta-analyses, and we tend to concentrate on the most prevalent methodological flaws.

The wide range of variation regarding methodological quality suggests that there is certainly room for improvement. The current unclarity about the efficacy of widely used physical therapy interventions is, in my view, a very undesirable situation. I can see no good reason why new drugs should prove their efficacy and even superiority over agents which were already available before they entered the market, while physical therapy interventions can be introduced (and may even become very popular) without any substantial evidence regarding efficacy. I'm strongly in favour of the use of similar standards for all types of intervention, and will argue later that physical therapy trials can, indeed, meet most of the criteria currently applied to drug research.

Flaw 1: Heterogeneous population

A common problem is inclusion in the trial of a group of patients which is too heterogeneous, often on the basis of a vague diagnosis like lumbago or frozen shoulder. In such a population the susceptibility for the intervention at issue may vary substantially. Dilution of efficacy will be the consequence of this. More attention should be paid to diagnosis, and especially prognosis, to enable inclusion of homogeneous groups in a trial. In my view, this is one of the major challenges for physical therapy research today. The fact that most musculoskeletal disorders are still defined as syndromes consisting of a mix of signs and symptoms is clearly unsatisfactory.

Looking up-stream at etiology and pathogenesis, to enable the identification of causally defined diagnostic entities, has not been very successful, I might say. Alternatively, I would suggest considering the diagnostic problem in terms of predicting the prognosis. More specifically, identification of groups of patients who will most likely benefit from specific physical therapeutic interventions, seems to have a high priority. These questions regarding the crucial prognostic indicators cannot be answered by RCTs, but demand creative exploratory studies using, for instance, well-documented case-series or a single case design (15). Of course, once a prognostically homogeneous category of patients seems to be identified, the 'proof of the pudding' ought to be provided by another randomized clinical trial. In such future trials the similarity of relevant characteristics at baseline should be carefully evaluated and all drop-outs should be identified, including their reasons for leaving the study.

Flaw 2: Incompletely described intervention

A related problem is posed by the often incomplete description of the interventions used. This makes implementation of positive results unfeasible. Furthermore, we got the impression that the interventions were often clearly sub-optimal. For instance, it turned out that the dosages used in many laser trials were so low that a specific effect became very improbable (10). Of course, the solution to this problem would be the use of more explicit and optimal treatment protocols.

Often physical therapists participating in intervention research strongly object to the use of a strict treatment protocol, because they rightly consider it sub-optimal to treat all patients allocated to the intervention in exactly the same way. In my view, such strict uniformity is only necessary when an explanatory trial is at issue, which uses a placebo contrast to study the specific effect of the central component of the intervention. However, in many physical therapy trials the main question is of a pragmatic nature, involving a contrast between realistic treatment strategies. This excludes the use of a placebo intervention by definition, and positively asks for treatment protocols which allow treatment to be guided by relevant characteristics on entry into the trial, and

subsequently by early treatment responses. So, high quality trials do not depend on rigid treatment protocols, but will often involve flexible treatment strategies. Of course, the options available within a treatment strategy should be restricted to a workable optimum, and all decisions regarding treatment modification ought to be made according to explicit criteria.

Flaw 3: Wrong outcome parameters

The clinical relevance of the outcome parameters is often doubtful, for example sophisticated measures concerning range of motion or muscle strength in low back pain. Also the validity and precision of outcome measurement is typically unknown and probably not very impressive. We fear that a lot of the outcome parameters may be insensitive to a clinically meaningful change over time. Consequently, real treatment effects may be missed. It seems urgent to give more attention to the design of outcome measures in physical therapy trials (16,17).

Quite a lot of physical therapy trials still focus exclusively on the impairment level. In my opinion, the popularity of outcome measures dealing with impairments has two reasons. The first is a general reluctance to rely on 'soft' subjective data, which often leads to a 'hard' objective and precise measurement of the wrong phenomenon (16). The second reason seems to be confusion between the questions whether an intervention is effective and how it works. Once clinical efficacy is established on the disability level, data on the corresponding changes in impairments, of course, can offer an insight in the mechanism involved, but can never be a substitute for it. In designing outcome measures dealing with impairments, it is important not to rely on the standard psychometric criteria only, but to also pay special attention to sensitivity for clinically meaningful change over time (17).
Successful recent examples of the development of suitable outcome measures are the Quebec Back Pain Disability Scale and the Shoulder Pain Disability Questionnaire.

Similar to their attitude with respect to treatment protocols, physical therapists participating in a trial often object to using the same outcome phenomenon for all patients included in the trial. They argue that the main complaint may differ substantially among prognostically similar patients. In my view, this problem can be avoided by identifying the main complaint for each patient at randomization, and by making change in the severity of that complaint the primary outcome parameter. In our experience, these individualized outcome measures are often very sensitive to change over time. Because in pragmatic trials blinding by the use of a placebo intervention is not only often unfeasible, but also undesirable, special attention sould be paid to unbiased outcome assessment. For this purpose trials should exclude patients who strongly favour one of the interventions under comparison, and the treating physical therapist should have no role in effect measurement. Whenever possible, blinded research assistants ought to assess the main outcome phenomena. Also, more attention should be paid to the duration of follow-up: many trials provide data on short-term results only. Recurrences of the original complaint, and treatment outside the framework of the trial, deserve special attention towards the end of the follow-up.

Flaw 4: Insufficient sample sizes

Before concluding from a negative finding that there is no difference in effect, one should always have a look at the statistical power of the trial. The data in table 4 show

clearly that sample sizes in physical therapy trials are often very small (2). Consequently, the chance of making a Type II error is substantial. In other words: a fairly large proportion of the negative findings may be 'false negatives'. The solution to this is obvious: enlarging sample sizes to, say, more than 50 patients per group. I would like to add to this that, in my view, it is not a matter of sample size determination, but of recruitment. The formulas can be found in any textbook on biostatistics, and the main decision you have to make is about the magnitude of the difference in efficacy you would like to be able to detect. Once you have decided upon the minimal difference in effect you consider to be clinically relevant, there is no point in aiming at making smaller differences statistically significant by increasing the sample size further. The real problem is getting the number of patients you need, which will often make a multi-centre design obligatory.

Table 4: Sample sizes

Type of study	No. of studies	Median no. of patients in smallest group	(25% and 75% percentiles)
Explanatory trials	67	15	(12 and 33)
Pragmatic trials	107	20	(10 and 31)

Avoidance of Type II errors is not the only reason to advocate large sample sizes. In small studies randomization might result in unbalanced groups with respect to prognosis. For known and measured prognostic factors, adjustment for dysbalance is possible in the data-analysis. But for unknown, and therefore unmeasured factors this is, of course, not possible. Another argument in favour of large sample sizes is the expectation that large trials are likely to be published, irrespectieve of the study results. Small trials with a negative result might have little chance of appearing in print. This introduces publication bias, because small positive studies would be over-represented in the literature.

Future developments

This brings me to the third and last part of my presentation, in which I will formulate my views concerning the future course of intervention research in physical therapy. These views concern methodological developments, use in patient care, influence on policy decisions and the setting of research priorities.

While for the majority of the methodological problems identified the standard general solutions can be applied, in my view two important challenges seem to be fairly specific for the physical therapy domain. The first challenge consists of a meaningful demarcation of categories of patients which are prognostically homogeneous with respect to the intervention at issue. The second challenge deals with the development of outcome parameters that combine all desirable test characteristics, and offer a fair chance to physical therapy interventions to prove efficacy. For both challenges the RCT-design is inappropriate, although exploratory sub-group analyses may be of some use. In this field I can see an opportunity for a number of different study designs, of which single-case studies and longitudinal follow-up of case-series are illustrative examples (15).

Individual patient care will always involve decision-making in uncertainty, no matter what evidence from intervention studies is available. RCTs generate knowledge about average effects only, while the effects among the individual patients usually differ substantially, and some may benefit most from the on the average less effective intervention. For patients who do not meet the eligibility criteria of the trial at issue, even extrapolation of the average efficacy is uncertain. The validity of extrapolation in these instances will depend on the existence of differences compared with the trial population that will modify the efficacy. Of course, this is a matter of judgement, adding up to the amount of uncertainty. Although the state-of-the-art is currently rather sad, I can see no alternative to the accumulation of knowledge on average efficacy by RCTs. In my view, this knowledge should provide a firm basis for standardized physical tharpy care. Explicit standards, like those formulated by the Dutch general practitioners for a wide range of indications, seem to me to be clearly superior to the alarming variety of interventions given nowadays to some categories of patients.

Assuming that resources for health care are essentially limited, decision-making on the policy level cannot be avoided. Although there is clearly no place for ineffective treatments, setting priorities among budgets for interventions that do have some effect is certainly no easy task. Unfortunately, empirical data concerning efficacy, cost-effectiveness and safety usually play only a minor role in these policy decisions. However, this role is likely to increase in the near future. Several Dutch agencies have already stressed the importance of studying critically the benefits and the costs of currently very widely used forms of physical therapy for their most prevalent indications.

This brings me to my last point: to me this demand of evidence, similar to the evidence concerning drug efficacy, only makes sense when a fair chance is given to the physical therapy profession. The fairness of this chance will largely depend on the funding available for (1) the training of qualified investigators, (2) the necessary infrastructure, and (3) the execution of the intervention studies at issue. Although the funding of physical therapy trials is certainly not easy, some progress has definitely been made during the last few years. Currently, in the Netherlands a number of RCTs are being executed, aimed at avoiding most of the flaws identified in this presentation.

Let me end by expressing the hope that this challenging field of intervention research in physical therapy will, in future years, continue to blossom.

References

1. Beckerman H., Bouter L. (eds.). Effectiviteit van fysiotherapie: een literatuuronderzoek. Maastricht: Rijksuniversiteit Limburg. 1991; 1-194.
2. Beckerman H., Bouter L.M., Heijden G.J.M.G. van der, Bie R.A. de, Koes B.W. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? Br J Gen Pract 1993; 43: 73-7.
3. Koes B.W., Bouter L.M., Heijden G.J.M.G. Methodological quality of randomized clinical trails on treatment efficacy in low back pain. Spine 1995; 20: 228-35.
4. Koes B.W., Assendelft W.J.J., Heijden G.J.M.G. van der, Bouter L.M., Knipschild P.G. Spinal manipulation and mobilisation for back and neck pain: a blinded review. Br Med J 1991; 303: 1298-303.
5. Koes B.W., Bouter L.M., Beckerman H., Heijden G.J.M.G. van der, Knipschild P.G. Physiotherapy exercises and back pain: a blinded review. Br Med J 1991; 302: 1572-6.

6. Heijden G.J.M.G. van der, Beurskens A.J.H.M., Koes B.W., Assendelft W.J.J., Vet H.C.W. de, Bouter L.M. The efficacy of traction for back and neck pain: a systematic, blinded review of randomized clinical trial methods. Phys Ther 1995; 75: 93-104.
7. Koes B.W., Tulder M.W., Windt D.A.W.M. van der, Bouter L.M. The efficacy of back schools: a review of randomized clinical trials. J Clin Epidemiol 1994; 47: 851-62.
8. Heijden G.J.M.G. van der, Bouter L.M., Beckerman H., Bie R.A. de, Oostendorp R.A.B. De effectiviteit van fysiotherapie bij schouderklachten: een geblindeerd literatuuronderzoek. Ned Tijdschr Fysioth 1992; 102: 38-46.
9. Heijden G.J.M.G. van der, Bouter L.M., Beckmerman H., Bie R.A. de, Oostendorp R.A.B. De effectiviteit van ultrageluid bij aandoeningen van het bewegingsapparaat : een op methodologische criteria gebaseerde geblindeerde review van gerandomiseerd patiëntgebonden onderzoek. Ned Tijdschr Fysiother 1991; 101: 169-77.
10. Beckerman H., Bie R.A. de, Bouter L.M., Cuyper H.J. de, Oostendorp r.A.B. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. Phys Ther 1992; 72: 483-91.
11. Pocock S.J. Clinical trials. A practical approach. Chichester; John Wiley 1989.
12. Meinert C.L. Clinical trials. Monographs in epidemiology and biostatistics. Volume 8. New York: Oxford University Press, 1986.
13. Chalmers T.C., Smith H. Blackburn B. et al. A method for assessing the quality of a randomized control trial. Controlled Clin Trials 1981; 2: 31-49.
14. Bouter L.M., Riet G. ter, Koes B.W., Knipschild P.G. Meta-analysis for physiotherapists. In: Proceedings of the third international physiotherapy congress. Hong Kong 1990. Sydney. Australia: Link Printing, 1990.
15. Bouter L.M. Prevalence of methodologic errors in rehabilitation research. Journal of Rehabilitation Sciences. 1994; 7: 60-2.
16. Feinstein A.R. Clinimetrics. New Haven, CT: Yale Unviersity Press, 1987.
17. Guyatt G.H., Deyo R.A., Charlson M. et al. Responsivenss and validity in health status measurement: a clarification. J Clin Epidemiol. 1989; 42: 403-408.