Low back pain and traction

Low back pain and traction

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Beurskens, Anna Johanna Helena Maria

Low back pain and traction / Anna Johanna Helana Maria Beurskens. - [S.l. : s.n.]
Thesis Rijksuniversiteit Limburg, Maastricht. - With summary in Dutch.
ISBN 90-74130-18-6 bound
NUGI 751
Subject headings: low back pain / physical therapy / traction.

Lay-out: Thum Aarts, Epidemiologie RL, Maastricht
Tekening omslag: Albère Köke, Munstergeleen
Omslag en produktie: Datawyse | Universitaire Pers Maastricht

Het onderzoek dat de grondslag vormt van dit proefschrift, werd verricht met financiële ondersteuning van het fonds Ontwikkelingsgeneeskunde van de Ziekenfondsraad. Dit proefschrift kwam tot stand binnen het instituut ExTra, dat deel uitmaakt van de door de erkenningscommissie van de KNAW geaccrediteerde Netherlands School of Primary Care Research (CaRe).

In de drukkosten van het proefschrift werd bijgedragen door Enraf-Nonius.

Low back pain and traction

Proefschrift

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof. mr. M.J. Cohen,
volgens het bestuur van het College van Decanen,
in het openbaar te verdedigen
op 28 maart 1996
om 14.00 uur

door

Anna Johanna Helena Maria Beurskens

geboren te Roermond op 22 april 1966

Promotor:	Prof. dr. P.G. Knipschild
Co-promotor:	Dr. ir. H.C.W. de Vet
Beoordelingscommissie:	Prof. dr. J.A. Knottnerus (voorzitter) Prof. dr. M.P.F. Berger Dr. M. Boers Prof. dr. L.M. Bouter (Vrije Universiteit, Amsterdam) Dr. G.H.I.M. Walenkamp

Contents

Hi, Ha, Ho. Let's go!

Chapter 1

Introduction

Low back pain is a common disorder. Epidemiologic studies indicate that about 80% of the population experience back pain during their lives.[1,2] Each year 5% of the population in industrialized countries experience an episode of low back pain.[2,3] Although low back pain occurs frequently, there is no consensus about the management. A large variety of therapeutic interventions exist, but none seem to be clearly superior to others. The efficacy of many (physio)therapeutic interventions for low back pain remains questionable.[3-5] One of the treatment options is traction.

Traction is a very old treatment modality. Since the days of Hippocrates (B.C.) it has been used to correct spinal deformalities, such as scoliosis. In 1934 Mixter and Ban[6] presented a paper which mentioned intervertebral disk pathology as a possible cause for low back pain problems with radiating pain. This led to a new interest in traction therapy. Traction was given on the assumption that low back pain could be relieved by means of widening of the intervertebral spaces.[7] Since then the techniques used to apply spinal traction varied considerably and have been used for various forms of low back pain. Traction can be exerted by gravitational forces through the body weight of the patient, manually by the therapist, by a motorized pulley or even by hanging the patient upside down.[8]

Literature review

We performed a systematic review on the available Randomized Clinical Trials (RCTs) about the efficacy of lumbar traction.[9] Relevant articles were found through Medline and Embase searches (until June 1992), additional screening of journals not covered by these computerized databases, and by citation-tracking of the selected articles. The articles to be reviewed were first blinded for author(s), journal, and outcome. Thereafter the quality of the studies' method was assessed by two blinded reviewers using a standardized set of methodological criteria. These criteria were based on generally accepted principles of intervention research.[10,11] Studies could earn points in four main categories: [1] study population, [2] interventions, [3] measurement of effect, and [4] data presentation and analysis. A study could earn a maximum methodological score of 100 points.

We traced 14 RCTs in the literature prior to June 1992.[9] Three studies showed favorable effects of traction. The methodological score of these 3 studies was low. The 2 RCTs with the best methodological score showed no favorable effects of traction.

Most studies lacked information about crucial features of the study design, the numbers of patients were often very small, and all studies suffered from severe methodological flaws. The most common flaws concerned poor prognostic comparability at baseline; lack of information about the randomization procedure and

poor operationalization of effect measures; incomparability of the co-interventions; and lack of blinding of the patients during treatment and of the assessors who performed the effect measurements.

The available studies on the efficacy of traction do not allow clear conclusions due to severe methodological flaws. Therefore, the evidence of the effectiveness of lumbar traction is not very sound. Three recent RCTs[12-14] did not change the results of the earlier review. We found that a better trial was needed with greater attention to proper design and conduct, as well as to a clear description of crucial methodological features and results and a large number of patients.

Pilot study

In a pilot study, we were able to avoid the most common flaws of published studies about the efficacy of lumbar traction.[15] After 5 weeks, 7 of the 11 patients (64%) in the group with high dose traction reported being totally recovered or much improved versus 4 of the 12 patients (34%) in the group with low dose traction. Because it was possible to overcome the methodological flaws of earlier studies and of the rather substantial effects that were found, we decided to conduct a larger trial with 150 patients.

Research question

The central question of the trial described in this thesis is whether traction is an effective treatment for patients with low back pain. The research question has been focused on the efficacy of continuous motorized traction for patients with at least 6 weeks non-specific low back pain. The research was funded by The Development Health Care Fund of The Sickfund Council.

Functional status

In our trial on the efficacy of lumbar traction we used a wide range of outcome measures. Several instruments are used for measuring functional status. Therefore, we are able to compare and evaluate the methodological quality of these functional status instruments in the same patient group.

In research it is very important to choose relevant outcome measures. Generally, there must be a clear relationship between the research question and the outcome measures chosen. The aim of treatment is often to improve the function of patients. For this reason, functional status is an important outcome in trials on low back pain. We define functional status as the ability of a person to perform common activities of daily living.

Organization of the thesis

Chapter 2 describes the design of the trial on the efficacy of lumbar traction in detail. Information about the design and justification of the choices made is valuable because it permits critical assessment of the methods independent of the outcome of the study.

Chapter 3 presents the 5-week results of the clinical trial. We carried out an intention-to-treat analysis as well as a per-protocol analysis. The latter analysis was restricted to a group of patients in which everything went the way we had planned it. Furthermore, we performed some subgroup analyses to explore the possibility that the effect of traction differed over special groups.

Chapter 4 shows the results of 12-week and 6-month follow-up in our trial. These follow-ups were included to detect longer term effects of lumbar traction.
Chapter 5 reviews the methodological quality of four instruments for measuring functional status of patients with low back pain. The questionnaires are discussed in terms of general description, scale structure, reliability, validity, responsiveness and clinical research applications.
Chapter 6 describes a comparison of the responsiveness of three instruments for evaluating functional status and an instrument for measuring pain. Two strategies for evaluating the responsiveness in terms of sensitivity to change and specificity to change are used: effect size statistics and receiver operating characteristics curves.
Chapter 7 presents the development and evaluation of a patient-specific approach for measuring functional status in low back pain. Activities of patients and their importance vary widely between patients. Therefore, it is sensible to focus in each patient on the complaints he or she presents as being most disturbing.
Chapter 8 consists of a general discussion, divided into two sections. In the first we pay attention to some methodological issues of the trial. In the second section we will look back at the efficacy of lumbar traction.

References

1. Haanen HCM. An epidemiologic survey on low back pain. Dissertation, Erasmus University Rotterdam, 1984.
2. Kelsey JL, White AA. Epidemiology and impact of low back pain. Spine 1980; 5: 133-142.
3. Frymoyer JW. Back pain and sciatica. New Engl J Med 1988; 318: 291-300.
4. Koes BW, Bouter LM, Heijden GJMG van der. Methodological quality of randomized clinical trials on treatment efficacy in low back pain. Spine 1995; 20: 228-235.
5. Spitzer WO, Leblanc FE, Dupuis M (eds). Scientific approach to the assessment and management of activity related spinal disorders. Spine 1987; 12 (Suppl): 1-59.
6. Mixter WJ, Ban JS. Rupture of the intervertebral disc with involvement of the spinal canal. New Eng J Med 1934; 211: 210-215.
7. Judovich BD. Lumbar traction therapy: elimination of physical factors that prevent stretch. JAMA 1955: 159: 549-610.
8. Geiringer SR, Kincaid CB, Rechtien JJ. Traction, Manipulation and Massage. In: DeLisa JA (eds). Rehabilitation medicine: principles and practice. Philadelphia: JB Lippincott, 1985: 276-294.
9. Heijden GJMG van der, Beurskens AJHM, Koes BW, Assendelft WJJ, Vet HCW de, Bouter LM. The efficacy of traction for back and neck pain. A blinded review of randomized clinical trial methods. Physical Therapy 1995; 75: 93-104.
10. Meinert CL. Clinical Trials: design, conduct and analysis. New York, NY: Oxford University Press, 1986.
11. Pocock SJ. Clinical trials. A practical approach. New York: John Wiley & Sons, 1983.
12. Konrad K, Tatrai T, Hunka A, Vereckei E, Korondi I. Controlled trial of balneotherapy in treatment of low back pain. Ann Rheum Dis 1992; 51: 820-822.
13. Letchuman R, Deusinger RH. Comparison of sacrospinalis myoelectric activity and pain level in patients undergoing static and intermittent lumbar traction. Spine 1993; 18: 1361-1365.
14. Ljunggren AE, Walker L, Weber H, Amundsen T. Manual traction versus isometric exercises in patients with herniated intervertebral lumbar discs. Physiotherapy Theory and Practice 1992; 8: 207-213.
15. Heijden GJMG van der, Beurskens AJHM, Dirx MJM, Bouter LM, Lindeman E. Efficacy of lumbar traction: a randomized clinical trial. Physiotherapy 1995; 81: 29-35.

Chapter 2

Efficacy of traction for low back pain

Design of a randomized clinical trial

AJHM Beurskens,[1] GJMG van der Heijden,[1] HCW de Vet,[1] AJA Köke,[2] E Lindeman,[3] W Regtop,[4] PG Knipschild.[1]

1 Department of Epidemiology, University of Limburg, Maastricht
2 Department of Physiotherapy, University Hospital, Maastricht
3 Department of Rehabilitation Medicine, University Hospital, Maastricht
4 'Hogeschool Heerlen', Department of Physiotherapy, Heerlen

J Manipulative Physiol Ther 1995; 18: 141-147.

Abstract

Objective - To present the design of a trial on the efficacy of lumbar traction.
Design - Randomized clinical trial.
Patients - Patients with a minimum of 6 weeks non-specific low back pain.
Intervention - High dose motorized continuous traction with a force between 35% and 50% of the total body weight will be compared with sham or low dose traction with a force between 0 and 20% of body weight. The sham traction is given with a specially developed brace which becomes tighter in the back during traction. This is experienced by patients as if traction were exerted.
Outcome measures - Primary measures are the patient's global impression of the effect and the severity of three main complaints. Secondary effect measures are functional status, pain, range of motion, work absence and recurrences. The effect measures will be rated before randomization and 4 weeks, 12 weeks and 6 months later.
Discussion - There have been a number of earlier trials on the efficacy of lumbar traction, but these suffered from severe methodological flaws. This trial aims to avoid these shortcomings.

Introduction

Back pain and traction

Each year 5% of the population in industrialized countries experience an episode of low back pain.[1,2] Of these incident cases 50% recover within 1 month, and 90% within 6 weeks. Only in a few cases (approximately 5%) does pain last longer than 3 months.[1] After recovery, 60% of all incident cases experience a relapse.[2,3] Low back pain is almost always non-specific, which means that no underlying disease can be established and the cause of the complaints remains unknown.

Although back pain is the most frequently presented disorder of the musculoskeletal system in general practice, there is no consensus about its management. There is a growing interest in the efficacy of physiotherapeutic interventions, which until now have remained questionable.[4-8] General practitioners in the Netherlands often refer patients with low back pain for physiotherapy. In these cases, traction is one of the treatment options. Patients in the Netherlands with back pain do not receive lumbar traction very often, and if they do, traction is often combined with other treatment modalities (e.g. massage, exercises, electrotherapy or heat).[8]

There are different traction modalities. Of the available modalities motorized and manual traction are the most frequently applied, but only motorized lumbar traction can be standardized satisfactorily.[9-11] The duration and magnitude of force exerted during motorized lumbar traction can be varied in a continuous or intermittent mode.[9-12]

Effect studies

We traced 14 Randomized Clinical Trials (RCTs) in the literature prior to June 1992 examining the efficacy of lumbar traction; 3 studies reported positive results of traction.[13] Most publications lack information about crucial features of the study design, the numbers of patients are often very small, and all studies suffer from severe methodological flaws. The most common flaws concern: poor prognostic comparability at baseline; a lack of information about the randomization procedure and poor operationalization of effect measures; incomparability of the co-interventions; and lack of blinding of the patients during treatment and of the assessors who perform the effect measurements. Therefore, the evidence of the effectiveness of lumbar traction is not very sound.

On the other hand, there are no indications that lumbar traction is an ineffective therapy for low back pain. In a pilot study (n=25) we were able to avoid the most common flaws of published studies about the efficacy of lumbar traction.[14] After 5 weeks, the recovery was 64% in the high dose traction group and 34% in the low dose traction group. With an alpha of 5% this difference would have reached statistical significance with 50 patients in each group. In view of the rather substantial effects that were found, the pilot study led to a larger trial with 150 patients. This article presents the design of this trial.

Preparing a trial in the field of physiotherapy or manual therapy is complex and is subject to many methodological problems. Articles reporting the results of randomized clinical trials rarely describe the design in detail. For readers information about the design and justification of the choices made can be very valuable because

it permits critical assessment of the methods independent of the outcome of the study.

Working mechanisms

The available literature is not very clear about the mechanisms by which lumbar traction could be effective. Rationales for the use of traction therapy are mainly based on the mechanical effects of traction: the elongation of the spine and the stretch on structures. These mechanisms cause the following actions: delordosis of the spine;[15,16] separation of the vertebrae;[9,15-22] widening of the intervertebral foramina;[15,16] a combination of distraction and gliding of the facet joints;[9,15,16,23] stretching of the spinal musculature[9,15,16,20] and of spinal ligaments.[9,12,15,16,19,22,23]

Various authors have reported that a certain amount of traction force is necessary to achieve separation of the vertebrae and widening of intervertebral foramina. If traction is of sufficient force to widen the intervertebral foramina, it would allow the nerve root more space, reduce a protrusion or prolaps or release an entrapped fold of capsule.

Judovich[24] reported that a pull equal to about one-half the weight of a body is needed to overcome friction of the body on the table-top; for the lower body this converts to 26% of total body weight. This amount of pull either needs to be achieved before "true" traction on the spine is accomplished, or a split table must be used.[24-26] In addition, muscle contraction and tone will account for a variable amount of the traction force, and some could be used in overcoming spinal curvatures, ligamentous resistance and friction of the machinery.[27]

If one believes that spinal traction is effective when bony separation occurs and joint surfaces are therefore distracted, more resistance has to be overcome than due merely to friction of the body on the table-top. Another 25% or more of body weight is the minimal force needed to cause vertebral separation.[16,24-27] For this reason, lumbar traction forces that are below 25% of the total body weight and use a split table can be regarded as sham (or low dose) traction. In our trial, the maximum sham traction force is defined as 20% of the total body weight.

Until now the proposed mechanisms by which traction could be effective have not been supported by sufficient research. We will briefly summarize the information available about the several theories on the possible effectiveness of traction.

To start with, the findings of Mathews[21] and Onel et al.[22] support the view that vertebral distraction does occur with static lumbar traction (with traction forces of 100-120 pounds) is applied to patients. Colachis and Strohm[18] demonstrated significant vertebral separation in normal subjects during static and intermittent traction with a traction force of 100 pounds. Bridger et al.[17] showed a greater increase in height after continuous traction compared with periods of lying supine with the knees flexed at 90°. There is no agreement as to the magnitude of this lengthening and no studies of whether the lengthening remains when the traction force is removed.

According to Cyriax,[19] the vertebrae in the spine are distracted during traction and a negative pressure develops in the disc that sucks back a protrusion. Pressures in the third lumbar discs were measured in vivo during traction in healthy volunteers.[28] During traction the pressure remained close to the resting pressure. Because of this, a suction effect of the disc is not a likely explanation of a therapeutic effect of

passive traction. On the other hand, it is unknown whether the pressure is increased, decreased or unchanged in patients with back symptoms.

Epidurograms and CT-scan showed that a lumbar disc protrusion can be reduced, at least temporarily, during intermittent traction[29] and continuous traction.[22,27] It is important to bear in mind that pain reduction is also possible without reduction of disc protrusions[30-32] and that disc protrusions can reduce without traction therapy.[33]

One reason for prescribing traction has been relaxation of spinal muscles.[9] Hood et al.[34] studied the amount of EMG activity in the sacrospinalis musculature of volunteers and Letchuman and Deusinger[35] did the same with low back patients, both during intermittent and continuous traction. There was an initial increase in sacrospinalis muscle activity at the onset of traction, regardless of the type of traction. The activity began to decrease 3-4 minutes after the onset of traction. After approximately 6-7 minutes (the termination of traction) muscle activity was only slightly above the initial resting readings.[34,35] Thus, the rationale for prescribing traction for the purpose of decreasing muscle activity[9,15,23,36] seems to be unfounded.

So far, the theory that continuous lumbar traction causes vertebral distraction and widening of the intervertebral foramen is better founded than the other proposed rationales. However, the argument is often raised that although traction can cause a separation and widening of the intervertebral foramen, the effect will only be temporary.[11,12] The anatomic corrections or effects produced by traction are unstable. If a patient with a degenerative, narrow disc space is given traction, it will not restore that disc space to its original size and structure.

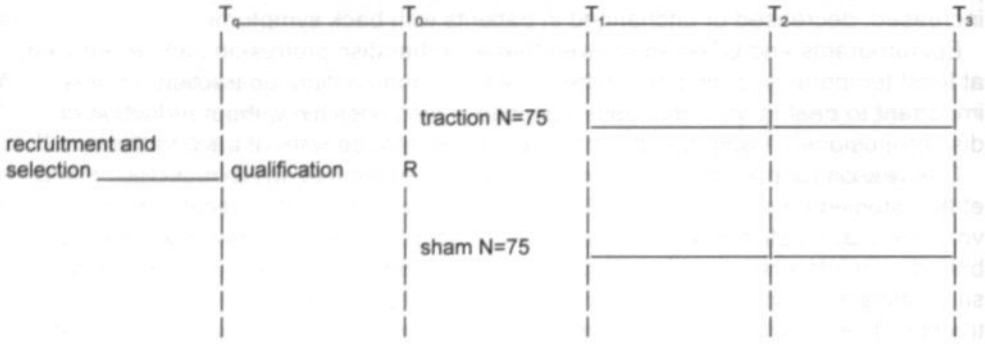
The rationales presented give some evidence for the explanation of short term effects of traction but they offer no explanations for long term effects of traction. Maybe we have to look for neuroreflectory mechanisms by which traction can cause an effect. For example, traction might exert a beneficial effect on some patients by its stretching influence on the mechanoreceptors present in discs, ligaments, and facet joints.[37] Until now, the effects of these neuroreflective mechanisms during traction have been unresearched.

Further scientific research is needed into working mechanisms of lumbar traction is needed. This research should precisely specify the body weights of the patients, the traction weights applied, the traction modalities used and the treatment duration.

Discussion of methods

Study design

The efficacy of lumbar traction for patients with subacute and chronic non-specific back complaints will be assessed with a randomized clinical trial. We will compare the effect of continuous motorized lumbar traction and of sham traction on the magnitude of recovery in patients with low back pain. In designing this trial special attention has been given to inclusion and exclusion criteria, (blind) outcome measurements and a credible sham treatment. The study protocol was approved by the Medical Ethics Committee of the University of Limburg and the University Hospital Maastricht. A brief outline of the design of the experiment is presented in figure 1.



T_q = Intake 1: followed by a 1-to-2-week qualification period
T_0 = Intake 2: measurements of baseline status, randomization (R)
T_1 = Outcome measurement 4 weeks after randomization
T_2 = Outcome measurement 12 weeks after randomization
T_3 = Outcome measurement 6 months after randomization

Figure 1. Study design.

Selection of patients and informed consent

Patients are being recruited by physiotherapists and general practitioners in the South of the Netherlands who are participating in the study. Patients showing interest are referred to an experienced physiotherapist, the research assistant. During the first intake he performs a physical examination and checks the inclusion and exclusion criteria. He explains the goal of the study and the chance of receiving the sham traction.

Patients (N=150) are selected if they have suffered for at least 6 weeks from non-specific low back pain, are at least 18 years old and if they have never before had any form of lumbar traction treatment. By non-specific we mean that no evidence is available for underlying disease or anatomical abnormalities (e.g., malignancy, osteoporosis) and that the exact cause of the complaints remains unknown after clinical examination. We will select no patients with complaints of less than 6 weeks duration because most of them recover irrespective of the type of treatment given.[1,38] Patients with complaints or symptoms related to the following four categories will not be selected. First, if there are indications for co-interventions: patients who receive a treatment for their back pain elsewhere or if another treatment is indicated within 6 months. For example, patients with S4 symptoms (incontinence or sensitivity disorders of the perineum). Second, if there are problems with fixing the canvas braces or if patients cannot lie down for 20 minutes. For example, patients who are obese or patients with severe respiratory complaints. Third, if the risks for complications are too high. For example, lesions suspect for bony metastasis. Fourthly, if traction is expected to have no effect. For example, patients with rheumatic diseases.

The first visit is followed by a qualification period of 1-2 weeks.[39] During this period, the patients will keep a pain diary. In this way we will be able to test the patients' compliance and evaluate whether there is a major improvement of the back pain.

When a patient meets the selection criteria and is willing to participate, the informed consent procedure will be completed at the second visit to the research assistant. After oral consent, the patient will sign a letter that explains all relevant information about the study, including the chance for receiving sham treatment. The patient will be allowed to withdraw from the study at any time. The measurement of baseline status and relevant prognostic variables will also be performed at this visit. This information enables us to assess later whether the randomization has been successful and to make subgroup analyses feasible. Important prognostic factors that will be measured are duration of complaints, baseline scores for outcome measures, age, recurrence status (number of relapses) at baseline, and radiating pain.

Randomization

Patients will be randomly allocated by computer to high dose and low dose traction. Sealed envelopes prepared by an independent person will contain the treatment code for either high or low dose traction. An envelope will be handed to the treating physiotherapist. He will open the envelope at the first treatment session, and will therefore not be blinded for the assigned treatment. Through randomization we hope to establish prognostic comparability among the contrasted groups (no difference in outcome if the intervention does not work). However, due to random errors unequal distributions may easily occur. In order to enhance prognostic comparability we stratify on duration of the complaints (less than or longer than 6 months. In addition, prestratification will be applied according to the physiotherapy practice (n=10) ensuring that each practice treats an equal number of patients with sham and traction. To ensure that more or less equal numbers of patients are assigned to high and low dose traction within each stratum, we use permuted blocks of 2 patients.

Interventions

All patients will be treated with the same traction apparatus (Eltrac, DIMEC Delft Instruments, the Netherlands). The traction apparatuses are calibrated before the start of the study and will be calibrated every half year.

The patient will be asked to lie down in the semi-fowler position (figure 2). The canvas braces will be attached as the patient lies down on the traction table, one around the iliacal crest and the other around the lower thoracic cage. The patients in the sham group will be explicitly told that they only have to feel a little pulling from the belts during the 20-minute session. After unlocking the sliding table top, the traction force will be slowly increased starting from zero, until the patient indicates that a little pulling is felt. In the sham traction group, force may not exceed 20% of the total body weight. Patients who receive the sham traction will get a special brace around the iliacal crest which becomes tighter in the back during the traction treatment (figures 3 and 4). This is experienced by patients as if traction were exerted. Experience has shown that the patients will feel the pulling before 20% is reached. Therefore, we hope that in practice, the traction force remains far less than 20% of body weight. Our brace allows for little longitudinal force, thereby increasing the contrast between the intervention groups with respect to spinal elongation as the postulated effective mechanism.

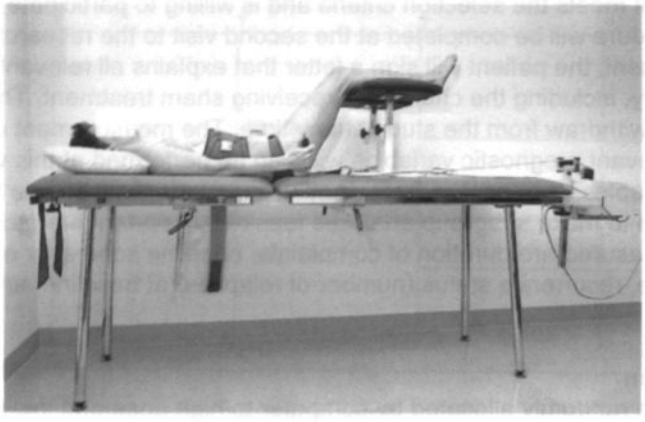


Figure 2. Patient lying in the semi-fowler position on the traction table.

The patients in the traction group will be explicitly told that they have to feel a distinct but tolerable pulling in the lumbar region during the 20-minute session. After unlocking the sliding table top, the traction force in the traction group is increased up to 35% of the total body weight; in the first two sessions, at least 30% of the total body weight is used. From this point, the traction force will be increased until the patient indicates that the tolerance for pulling has been reached. The maximum traction force is defined as 50% of the total body weight.

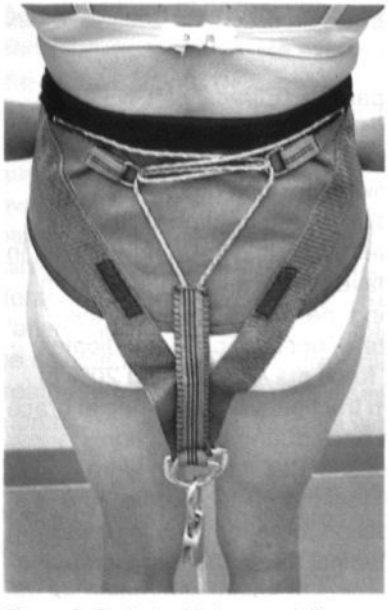


Figure 3. Patient with the specially developed sham traction brace around the iliacal crest.

Figure 4. The brace becomes tighter in the back during the sham traction treatment. This is experienced by patients as if traction were exerted.

No systematic research of the adverse effects of traction has been performed;[40] the few case reports available suggest that there is some danger in lumbar traction exceeding 50% of the total body weight.[40,41] Malignant diseases, pregnancy, spinal infections, osteoporosis, rheumatoid arthritis, and breathing disorders are mentioned as contraindications for traction;[10,40,41] consequently, we will exclude these patients in our trial. Therefore, although little is known about the harmful side-effects of motorized lumbar traction, a traction weight from 30-50% of the total body weight and exclusion of the above mentioned high risk groups seems to do no harm.[40,41]

The traction force, traction duration and date of each session will be noted by the physiotherapists. Both groups are treated 3 times a week for 4 weeks with least 10 sessions, and usually 12. The patients will be allowed to continue their regular pain medication. The compliance of pain medication is recorded by means of questionnaires that are completed by the patients on the 4-week, 12-week and 6-month follow-up measurements. Other co-interventions, e.g., injections, massage, exercises, physical modalities, will not allowed during the treatment period. To minimize variations between the participating physiotherapists in information and instructions on patient behavior and traction therapy, each patient will receive a booklet about low back pain and traction with instructions about activities of daily living. To minimize treatment variation among groups and to standardize additional care and attention, the physiotherapists are trained in giving the traction and sham traction therapy in a standardized way.

Blinding

The patients will be blinded; so too will be the research assistant who examines the patients, decides which patients are eligible for the trial and conducts outcome measurements, the researcher who analyzes the data will be blinded as well. The physiotherapists who perform the treatments cannot be blinded.

The success of blinding will be evaluated after the intervention at the 4 weeks follow-up measurement. The patients will be asked whether they think they have been treated with a sham traction or not and how sure they are about this. In a pilot study, only 6 of the 25 patients thought that they had been treated with sham traction, 3 patients in the sham traction group and 3 patients in the traction group.[14] In the pilot study, the sham traction was not performed with the special iliacal brace that becomes tighter in the back during traction treatment. Volunteers were used to test the credibility of the special brace. They found it very difficult to determine whether they had been treated with the sham, or not. We expect that in this study blinding will be at least as good as in the pilot study. Because of the special brace we expect that the contrast between the traction forces used in the sham traction and real traction group will be greater and that blinding will be maintained.

Outcome measurements

It is important to choose adequate outcome measures in an intervention study. First, the choice should be based on the main goal of the intervention. Second, all relevant fields have to be represented. Third, outcome measures should be able to detect small but important clinical changes; they should be responsive.[42,43] Fourth, if several outcomes are important, it is desirable to establish a hierarchy of importance. Table 1 shows the measures of effect in the hierarchical order of our preference.

Although there is disagreement about which outcome measures are the most important and responsive, in our opinion the most adequate outcome measures are those with the most relevance for the patient and clinician: the primary outcome measures. They are the patients' global impression of the perceived effect (recovery) and the severity of the three main complaints of the individual patient.

Global perceived effect is measured by self-assessment on a seven-point scale (1=completely recovered, 7=vastly worsened). From both patients' and clinicians' viewpoint, it is sensible to ask the patients to assess their perceived benefit which was used as a primary outcome measure in the pilot study.[14]

At the baseline, the patient will select the three main complaints in a standardized way. He will select three activities which he does frequently, which are important in his day-to-day life, and which are made difficult for him by low back pain.[42] The severity of the three main complaints will be rated on a visual analog scale (VAS). Although activities of patients and their importance varies widely among patients, it is sensible to focus on the complaints that they present as being the most disturbing.[42,44,45] Guyatt et al.[42] suggest to ask which activities are difficult to perform and to examine the effect of the intervention on performance of those activities. In this way outcome measures are tailored to the individual patient (each patient has his or her specific treatment goal). This approach is being used increasingly to measure disease-specific functional status and it appears to be responsive.[42,44]

Table 1. Overview of measurement and schedule of data collection

	Intake 1	Intake 2	4 weeks	12 weeks	6 months
Pain diary	X				
Demographic data	X				
Eligibility criteria	X	X			
Informed consent		X			
Medical history		X			
Randomization		X			
Effect measures					
I Global perceived effect			X	X	X
II Severity of three main complaints		X	X	X	X
III Functional status		X	X	X	X
IV Pain		X	X	X	X
V Severity of low back pain (assessed by the research assistant)		X	X	X	
VI Range of motion		X	X	X	
VII Work absence		X	X	X	X
VIII Disability		X	X	X	X
IX Medical consumption			X	X	X
X Recurrence				X	X
Blinding			X		
Satisfaction			X		

Other outcome measures are considered to be secondary. This does not mean that these measures are unimportant. Pain and functional status are also very important

outcome measures in our trial. The measures focus partly on the same concepts: patients' global impression of the perceived effect, severity of the main complaints measure and aspects of pain and functional status.

Functional status is measured with the Roland Disability Questionnaire, which has been designed for patients with low back pain.[46-48] Pain will be measured by the patient on a VAS. Relevance, validity and reliability of the VAS have been tested for the domain of back pain.[49-51] After a standardized anamnesis and physical examination, the severity of the low back pain will be evaluated by the research assistant on a 10-point scale (1=minimal severity, 10=maximal severity). Range of motion of the spine will be measured by using the inclinometer EDI 320-CYBEX. The reproducibility of spinal measurement with the EDI 320 seems to be satisfactory.[52] Length of absence from work, disability, and medical consumption as well as recurrences will be recorded by means of questionnaires that are completed by the patients. Almost all measurements will be carried out by the same research assistant to avoid inter-observer variation.

The effect on all outcome measures will be assessed separately. Primary and secondary measures will be similarly treated in the analysis. Readers may have other opinions as to the importance of the chosen outcome measures. It is up to them to interpret the results of the study according to their chosen relevant outcome measures.

Table 1 shows the schedule of the data collection. Evidence for short term effects of traction is discussed in the section about working mechanisms The 4-week and 12-week follow-up measurements are included to detect short term effects of treatment. Until now, the rationales offer no explanations for long term effects of traction. Nevertheless, a 6-month measurement is included to detect how the effects last and whether recurrences of the complaints occur. Measurements after a longer period are useless, since too many co-interventions are expected.

Analysis

The statistical analysis will be carried out according to the 'intention-to-treat' principle. This means that all patients remain in the group to which they were assigned by randomization. This includes drop-outs and patients with low compliance. We will also perform a 'per-protocol analysis' that is restricted to a group of patients in which everything went as planned. We will remain ignorant of treatment allocation until all decisions are made about cut-off points for protocol deviations and patients to be excluded from the 'per-protocol analysis'.

A one-way analysis of covariance (ANCOVA) will be used to estimate group differences, adjusted for differences at the baseline for important prognostic indicators. In addition, co-variables will be used for co-interventions, degree of compliance and strata of randomization.

For the outcome measures recorded at baseline and at follow-up, we will compute the difference between the follow-up score and the baseline score for each individual. In this way we are able to compare for differences in mean improvement or deterioration scores. With regard to the outcome measures for which baseline information is lacking (e.g. global perceived effect) only differences between the two groups in the scores can be calculated. Group differences and 95% confidence intervals will be calculated for all outcomes.

Discussion

There have been a number of trials on the efficacy of traction for low back pain, but these have been criticized on methodological grounds. This study aims to overcome earlier shortcomings. First, the outcome measures will be assessed blindly by the patients or the blinded research assistant. Second, no co-interventions will be allowed during the treatment period (except pain medication) and any interventions that may occur are accurately recorded. Third, the size of the study population seems to be sufficiently large to detect treatment differences which are clinically important. Fourth, we have chosen an appropriate sham traction which is believable for patients.

A major problem in randomized clinical trials is the recruitment of enough patients who fulfil the entry criteria. For our patient recruitment, we are dependent on the participating physiotherapists and general practitioners. So, it is very important to motivate and interest the participators.

We know that the theoretical design is always far better than the practical implementation. We hope to keep as close as possible to the design and to report the results within a short time.

References

1. Frymoyer JW. Back pain and sciatica. New Engl J Med 1988; 318: 291-300.
2. Kelsey JL, AA White. Epidemiology and impact of low back pain. Spine 1980; 5: 133-142.
3. Haanen HCM. An epidemiologic survey on low back pain (dissertation). Rotterdam: Erasmus University, 1984.
4. Deyo RA. Conservative therapy for low back pain. Distinguishing useful from useless therapy. Spine 1983; 11: 28-30.
5. Koes BW, Bouter LM, Beckerman H, Heijden GJMG van der, Knipschild PG. Physiotherapy exercises and back pain: a blinded review. BMJ 1991; 302: 1572-1576.
6. Koes BW, Assendelft WJJ, Heijden GJMG van der, Bouter LM, Knipschild PG. Spinal manipulation and mobilisation for back and neck pain: a blinded review. BMJ 1991; 303: 1298-1303.
7. Spitzer WO, Leblanc FE, Dupuis M (eds). Scientific approach to the assessment and management of activity related spinal disorders. Spine 1987; 12(Suppl): 1-59.
8. Beckerman H, Bouter LM, Heijden GJMG van der, Bie RA de, Koes BW. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? Br J Gen Pract 1993; 43: 73-77.
9. Harris R. Traction. In: Licht S (eds). Massage, manipulation and traction. New York: Krieger Publishing Company, 1976.
10. Geiringer SR, Kincaid CB, Rechtien JJ. Traction, manipulation, and massage. In: DeLisa JA eds. Rehabilitation Medicine. Principles and Practice. Philadelphia: JB. Lippincott, 1985.
11. Saunders HD. Use of spinal traction in the treatment of neck and back conditions. Clin Orthop 1983; 179: 31-38.
12. Swezey RL. The modern thrust of manipulation and traction therapy. Semin Arthritis Rheum 1983; 12: 322-331.
13. Heijden GJMG van der, Beurskens AJHM, Koes BW, Assendelft WJJ, Vet de HCW, Bouter LM. The efficacy of traction for back and neck pain. A blinded review of randomized clinical trial methods. Physical Therapy 1995; 75: 93-104.
14. Heijden GJMG van der, Bouter LM, Terpstra-Lindeman E, Essers AHH, Waltjé EMH, Köke AJA, Waelen AMW. De effectiviteit van tractie bij lage rugklachten. Nederlands Tijdschrift voor Fysiotherapie 1991; 101: 37-41.
15. Neuwirth E, Hilde W, Campbell R. Tables for vertebral elongation in the treatment of sciatica. Arch Phys Med 1952; August: 455-460.
16. Saunders HD. Lumbar traction. JOSPT 1979; 11: 36-45.

17. Bridger RS, Ossey S, Fourie G. Effect of lumbar traction on stature. Spine 1990; 156: 522-524.
18. Colachis SC, Strohm BR. Effects of intermittent traction on separation of lumbar vertebrae. Arch Phys Med Rehabil 1969; 44: 251-258.
19. Cyriax JH. Textbook of Orthopaedic Medicine, 9th ed, Vol 2. London: Cassell and Collier Macmillan Publishers Ltd, 1977.
20. Lehmann JF, Brunner GD. A device for the application of heavy lumbar traction: its mechanical effects. Arch Phys Med Rehabil 1958; November: 696-700.
21. Mathews JA. Dynamic discography: a study of lumbar traction. Ann Phys Med 1968; 97: 275-279.
22. Onel D, Tuzlaci M, Sari H, Demir K. Computed tomographic investigation of the effect of traction on lumbar disc herniations. Spine 1989; 141: 82-90.
23. Kekosz VN, Hilbert L, Tepperman PS. Cervical and lumbopelvic traction. To stretch or not to stretch. Postgrad Med 1986; 808: 187-194.
24. Judovich BD. Lumbar traction therapy - elimination of physical factors that prevent lumbar stretch. JAMA 1955; 159: 549-550.
25. Judovich B, Nobel GR. Traction therapy, a study of resistance forces. Am J Surg 1957; 93: 282-286.
26. Judovich BD. Lumbar traction therapy and dissipated force factors. Journal Lancet 1954; October; 411-414.
27. Mathews JA. The effects of spinal traction. Physiotherapy 1972; 58: 64-66.
28. Andersson GBJ, Schultz AB, Nachemson AL. Intervertebral disc pressure during traction. Scand J Rehab Med 1983; 9(Suppl): 88-91.
29. Gupta RC, Ramarao SV. Epidurography in reduction of lumbar disc prolapse by traction. Arch Phys Med Rehabil 1978; 59: 322-327.
30. Gillström P, Ericson K, Hindmarsh T. Autotraction in lumbar disc herniation. A myelographic study before and after treatment. Arch Orthop Trauma Surg 1985; 104: 207-210.
31. Ljunggren AE, Eldevik OP. Autotraction in lumbar disc herniation with CT examination before and after treatment, showing no change in appearance of the herniated tissue. J Oslo City Hosp 1986; 36: 87-91.
32. Natchev E, Valentino V. Low back pain and disc hernia observation during auto-traction treatment. Manual Medicine 1984; 1: 39-42.
33. Saal JA, Saal JS. Nonoperative treatment of herniated lumbar intervertebral disc with radiculopathy. An outcome study. Spine 1989; 144: 431-437.
34. Hood J, Hart DL, Smith HG, Davis H. Comparison of electromyographic activity in normal lumbar sacrospinalis musculature during continuous and intermittent pelvic traction. J Orthop Sports Phys Ther 1981; 2: 137-141.
35. Letchuman R, Deusinger Rh. Comparison of sacrospinalis myoelectric activity and pain levels in patients undergoing static and intermittent lumbar traction. Spine 1993; 18: 1361-1365.
36. Twomey LT. Sustained lumbar traction. An experimental study of long spine segments. Spine 1985; 10: 146-149.
37. Wyke BD. The neurology of low back pain. In: Jayson MIV. (eds). The lumbar spine and back pain. 2nd ed. London: Pitman, 1980.
38. Waddell G. A new clinical model for the treatment of low-back pain. Spine 1987; 12: 632-644.
39. Knipschild P, Leffers P, Feinstein AR. The qualification period. J Clin Epidemiol 1991; 44; 461-464.
40. Yates DAH. Indications and contra-indications for spinal traction. Physiotherapy 1972; 58: 55-57.
41. Eie N, Kristiansen K. Complications and hazard of traction in the treatment of ruptured lumbar intervertebral disks. J Oslo City Hosp 1962; 12: 4-12.
42. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40: 171-178.
43. Kirshner B. Guyatt GH. A methodological framework for assessing health indices. J Chron Dis 1985; 38: 27-36.
44. Koes BW. Efficacy of manual therapy and physiotherapy for back and neck complaints. (dissertation). Maastricht: University of Limburg, 1992.
45. Feinstein AR, Josephy BR, Wells CK. Scientific and clinical problems in indexes of functional disability. Ann Intern Med 1986; 1986; 105: 413-420.
46. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983; 8: 141-144.

47. Deyo RA. Comparative validity of the sickness impact profile and shorter scales for functional assessment for low-back pain. Spine 1986; 11: 951-954.
48. Deyo RA. Measuring the functional status of patients with low back pain. Arch Phys Med Rehabil 1988; 69: 1044-1053.
49. Carlsson AM. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analog scale. Pain 1983; 16: 87-101.
50. Revill SI, Robinson JO, Hogg MIJ. The reliability of a linear analog for evaluating pain. Anaesthesia 1976; 31: 1191-1198.
51. Sriwatanakul K, Kelvie W, Lasagna L, Calimlim JF, Weis OF, Mehta G. Studies with different types of visual analog scales for measurement of pain. Clin Pharmacol Ther 1983; 34: 234-239.
52. Koes BW, Mameren van H, Bouter LM, et al, Reproducibility of range of motion measurement of the spine with the CYBEX EDI 320. Proceedings of the 3rd International Physiotherapy Congress, Hong Kong, Link Printing, Sydney, 1990; 442-447.

Chapter 3

Efficacy of traction for non-specific low back pain

5-week results of a randomized clinical trial

AJHM Beurskens,[1] HCW de Vet,[1] AJA Köke,[2] E Lindeman,[3] W Regtop,[4] GJMG van der Heijden,[1] PG Knipschild.[1]

1 Department of Epidemiology, University of Limburg, Maastricht
2 Department of Physiotherapy, University Hospital, Maastricht
3 Department of Rehabilitation Medicine, University Hospital, Maastricht
4 'Hogeschool Heerlen', Department of Physiotherapy, Heerlen

Lancet 1995; 346: 1596-1600.

Abstract

Objective - To assess the efficacy of motorized physiotherapeutic traction for low back pain. Earlier trials on the efficacy of lumbar traction suffered from severe methodological flaws. This trial aimed to avoid these shortcomings.
Design - A randomized controlled trial. High dose traction was compared with sham traction. The sham traction was given with a specially developed brace that tightens in the back during traction. This was experienced as if traction were exerted. The patients and outcome assessor were blinded for the assigned treatment.
Subjects - Patients with at least 6 weeks non-specific low back pain.
Main outcome measures - Patients' global perceived effect, severity main complaints, functional status and pain.
Results - 151 patients were randomized. Intention-to-treat analysis showed no differences between the groups on all outcome measures; all 95% confidence intervals included the value zero. The number of withdrawals from treatment, loss to follow-up and protocol deviations was low. Consequently, the per-protocol analysis showed results similar to the intention-to-treat analysis. Subgroup analyses did not show any group for which traction might seem promising.
Conclusions - These data do not support the claim that traction is effective for patients with low back pain.

Introduction

Each year 5% of the population in industrialized countries experience an episode of low back pain (LBP).[1,2] There is no consensus about the management of LBP. The efficacy of many physiotherapeutic interventions remains questionable.[3-5] One of the treatment options is traction, often combined with other treatment modalities (e.g., massage, exercises, electrotherapy or heat).[3] There are different traction modalities of which motorized and manual traction are most frequently applied.

The available literature is not clear about the working mechanisms by which lumbar traction could be effective. Supposed mechanical effects of traction are: vertebral separation and widening of the intervertebral foramen.[6-9] These mechanisms suggest short-term rather than long-term effects.[10,11] Part of the applied traction force is needed to overcome opposing forces, namely friction of the body on the table-top, muscle contraction, spinal curvatures, ligamentous resistance and friction of the machinery.[12-14] Lumbar traction forces below 20% of the total body weight using a split table, to overcome friction of the body on the table-top, can be regarded as sham (or low dose) traction.

No systematic research has been performed into the adverse effects of traction;[15] the few case reports available suggest some danger in lumbar traction exceeding 50% of the total body weight.[15,16] To reach sufficient contrast with sham traction while avoiding adverse effects we chose to use a traction force between 35-50% of the total body weight for the active intervention. We preferred motorized lumbar traction over manual traction because the former can be standardized satisfactorily.

We traced 14 Randomized Clinical Trials (RCTs) in the literature prior to June 1992 about the efficacy of different lumbar traction modalities.[17] The available studies on the efficacy of lumbar traction do not allow clear conclusions due to methodological flaws, lack information about crucial features of the study design, and the small numbers of patients. Three recent RCTs[18-20] published after June 1992 did not change the conclusions of our review.

In a pilot experiment (n=25) we were able to avoid the most common flaws.[21] The recovery rate after 5 weeks in the high dose (intervention) traction group was 64% and in the low dose (sham) traction group it was 34%. With an alpha of 5% this difference would have reached statistical significance with 50 patients in each group. In view of the substantial effects that were found, we started a larger randomized trial with 150 patients. This article presents the results of this trial.

Methods

Study design

We compared the effect of continuous motorized lumbar traction and sham traction on the magnitude of recovery in patients with persistent non-specific LBP 5 and 12 weeks and 6 months after randomization. The 5-week follow-up was included to detect short-term effects. The 12-week and 6-month follow-ups were included to detect long-term effects and recurrences of LBP. In this article we present findings from 5-week of follow-up. The Medical Ethics Committee of the University of Limburg and the University Hospital Maastricht in the Netherlands approved the study protocol. An elaborate description of the study design can be found elsewhere.[22]

Selection of patients

Physiotherapists in 10 practises and 6 general practitioners in the Maastricht area recruited and selected the patients. Patients were selected if they had suffered for at least 6 weeks from non-specific LBP with or without radiation, if they were at least 18 years old and had never before had any form of lumbar traction treatment. Patients with evidence for underlying disease or anatomical abnormalities (such as rheumatic diseases, previous surgery or fracture, herniated disc) were excluded. Furthermore, patients were also excluded if they were receiving treatment (other than medication) for their LBP elsewhere at the moment of selection, if they improved much during the previous 2 weeks, if there were problems with fixing the canvas braces, or if patients were unable to lie down for 20 minutes. One of us (AJK), an experienced research physiotherapist, checked the inclusion and exclusion criteria. He gave information about the goal of the study, the chance of receiving the sham traction, and the right to withdraw from the study at any time. After receiving a letter that included this information, the patients signed a consent to participate.

Randomization

Patients were randomly allocated to high dose (intervention) and low dose (sham) traction with the help of random numbered list generated by computer. To enhance prognostic comparability we prestratified on duration of the complaints (less than, or longer than 6 months) and on physiotherapy practice (n=10). Use of permuted blocks of 2 patients ensured almost equal numbers of patients within each stratum. In order to assess the success of randomization, we measured several important prognostic factors: duration of complaints, baseline scores for outcome measures, age, recurrence status (number of relapses) at baseline, and radiation pain.

After the inclusion of a patient into the trial, the treating physiotherapist received a sealed envelope that contained the treatment code. He opened the envelope at the first treatment session, and was therefore not blinded for the assigned treatment.

Interventions

All patients were treated with similar traction apparatus (Eltrac, DIMEC Delft Instruments, the Netherlands), which were calibrated before the start of the study. Both intervention groups were treated 12 times in 5 weeks, 20 minutes per session. As the patient lay down on the traction table in the semi-fowler position the canvas braces were attached, one around the iliacal crest and the other around the lower thoracic cage. After unlocking the sliding table top, the physiotherapist increased the traction force.

Patients in the traction group were explicitly told that they had to feel a distinct but tolerable pulling in the lumbar region. The traction force was increased until the patient indicated that the tolerance for pulling was reached, with a minimum traction force of 35% and a maximum of 50% of the total body weight.

Patients in the sham group were explicitly told that they had only to feel a little pulling from the braces. The traction force was slowly increased until the patient indicated that he felt little pulling, with a maximum traction force of 20% of the total body weight. Patients received sham traction with a special brace around the iliacal crest, which became tighter in the back during the treatment. In this way all patients were blinded for the treatment allocation.

For each session the physiotherapists recorded date, applied duration and force of traction. To minimize treatment variation between both groups all patients received a booklet about LBP and traction, including instructions about activities of daily living. For the same reason, we trained the participating physiotherapists in giving the traction and sham traction therapy in a standardized way. The patients were allowed to continue the pain medication they used before entry into the study, i.e., non-narcotic analgesics or non-steroidal anti-inflammatory drugs. Other co-interventions were not allowed during the treatment period. We asked the patients to take no pain medication during the 24 hours before the effect measurement.

Outcome measurements and follow-up

Our primary outcome measures after 5 weeks of treatment were [1] global impression of the perceived effect (recovery) rated by the patient on a seven-point scale and [2] severity of the three main complaints. At baseline, the patients selected their three main complaints in a standardized way: each individual selected three activities he performed frequently, which he perceived as important in their daily life, and which LBP made difficult for him. The severity of these main complaints was rated on a visual analog scale (100 mms. VAS).

Secondary outcome measures were: [3] functional status measured with the Roland Disability Questionnaire (RDQ);[23] [4] pain and [5] Activities of Daily Living (ADL) disability marked by the patient on a 100 mms. VAS; [6] length of absence from work and [7] medical consumption recorded by questionnaires; [8] evaluation of the severity of LBP by the research physiotherapist indicated on an eleven-point scale after a standardized interview and physical examination; and [9] range of motion of the spine measured with an inclinometer (EDI 320-CYBEX).

Occurrence of side-effects and success of blinding of the patients were evaluated at end of the treatment. The research physiotherapist (AJK) who conducted outcome measurements, was blinded. For periods when the observer was not available the blinded outcome measurements were performed by AJB, also a physiotherapist. To minimize interobserver variation the two observers were trained in performing the measurements in a standardized way.

Data analysis

An independent person (HCdV) checked the treatment registration forms for protocol deviations. The researcher (AJB) who analyzed the data was blinded for treatment identity, until all decisions were taken about cut-off points and protocol deviations, and during the primary analyses.

The statistical analysis was carried out according to the 'intention-to-treat' principle: all patients, including withdrawals from treatment and patients with poor compliance, remained in the group to which they are assigned by randomization. Besides this, we present a 'per-protocol analysis' which is restricted to a group of patients in which study protocols were followed throughout.

All data analyses were done with SPSS statistical software.[24] For the outcome measures recorded at baseline, we computed the difference between the post-treatment and the baseline score for each individual and compared these between the two groups using Student *t* test for statistical significance. With regard to the outcome measures for which there was no baseline information (e.g., global perceived effect) only differences between the two groups in the post-treatment

scores could be calculated. Group differences and two-tailed 95 percent confidence intervals were calculated for all outcomes.

One-way analysis of covariance (ANCOVA) was used to estimate group differences adjusted for differences at baseline of important prognostic indicators. Co-variables for co-interventions, degree of compliance, use of regular pain medication, and strata of randomization were added to this model.

Results

Table 1. Comparability of treatment groups with respect to distribution of prognostic variables.

Characteristic	Traction	Sham traction
No. of patients	77	74
Mean age (yrs,sd)	39 ± 10	42 ± 11
Gender (% female)	34 (44%)	32 (43%)
Present episode		
25, 50, 75 percentile (wk)	8, 20, 52	8, 24, 52
chronic (> 6 months)	40 (52%)	40 (54%)
sub-acute (6 weeks <> 6 months)	37 (48%)	34 (46%)
radiation below the knee	28 (36%)	22 (30%)
previous treatment	47 (61%)	37 (50%)
previous physiotherapy	39 (51%)	30 (41%)
Previous low back pain	66 (86%)	57 (77%)
always the same pain	16 (21%)	10 (14%)
always back pain with periods of more pain	16 (21%)	14 (19%)
sometimes periods of back pain	34 (44%)	33 (45%)
Number of low back pain episodes		
ever 25, 50, 75 percentile	4, 6, 20	4, 10, 20
last year 25, 50, 75 percentile	1, 1, 2	0, 1, 2
Mean General Health Questionnaire*	8.3	8.6
Mean severity†		
first main complaint	75	73
second main complaint	74	70
Mean Roland Disability Questionnaire‡	12	12
Mean pain score†		
during measurement	61	55
last week	62	62
Mean severity of low back pain‖	5	5
Range of motion (degrees)	54	54
ADL-disability†	67	70

* Scored on a 36-item questionnaire (best score 0, worst score 36).
† Scored on a 100 mms. VAS (best score 0, worst score 100).
‡ Scored on a 24-item questionnaire (best score 0, worst score 24).
‖ Scored on an 11-point scale by the research assistant (best score 0, worst score 10).

Study sample

In the period June 1993 - December 1994 243 patients were invited for an appointment with the research physiotherapist to check their eligibility based on inclusion and exclusion criteria. Of these, 92 patients (38%) were excluded. The most common reasons for exclusion were: not enough motivation (i.e., no time,

chance of sham traction) (n=24) or less than 6 weeks LBP (n=16). Finally, 151 patients signed informed consent and were randomized: 77 to traction and 74 to sham traction.

Comparability of treatment groups

The two treatment groups were similar regarding most of the demographic and clinical baseline characteristics (table 1). The traction group contained a few more patients with pain radiating below the knee, previous treatment and previous LBP. On the other hand, the median number of LBP periods ever was higher in the sham traction group.

Compliance and blinding

Of the 151 randomized participants, 150 had complete follow-up; one patient in the sham traction group went to work abroad and was lost to follow-up. Table 2 provides information about compliance and success of blinding. The average traction force was calculated from the individual average traction force (percentage of the total body weight) over all treatment sessions: 42% in the traction group and 15% in the sham group.

Table 2. Compliance and success of blinding.

	Traction (n=77)	Sham traction (n=74)
Traction force (mean and SD)	42% (5)	15% (4)
Traction force inadequate*	8 (10%)	0
Treatment sessions (mean)†	12	12
Co-interventions‡	2 (3%)	2 (3%)
Withdrew from treatment		
increasing back pain	3 (4%)	2 (3%)
other reason‖	1 (1%)	2 (3%)
Major spacing in treatment days		1 (1%)
Blinding: patient's opinion¶		
real traction given	53 (74%)	51 (71%)
sham traction given	1 (1%)	4 (6%)
does not know	18 (25%)	17 (24%)

* Traction force less than 35% of body weight (two of these also withdrew from treatment).
† The number of treatment sessions was 11 or 12 in 93% of the patients.
‡ Traction: 2 patients received strong pain medication prescribed by their general practitioner (1 withdrew from treatment). Sham traction: 2 patients received physical therapy (these 2 patients also withdrew from treatment).
‖ Traction: fell off scaffolding. Sham traction: 1 patient got psychiatric problems and 1 patient went to work abroad (this patient was also lost to follow-up).
¶ Lacking for 4 patients in traction group.

The sham traction was not systematically unmasked by the patients who received it. Stratification on whether the patients had improved during treatment did not influence the results of blinding.

Intention-to-treat analysis

Table 3 shows the results of outcome measures in hierarchical order of our preference. Both groups improved, but all confidence intervals for the differences between groups included the value zero. This means that they were not statistically significant at 5% level (two-sided test). For all outcome measures the differences between the groups were very small.

Table 3. Intention-to-treat analysis at 5 weeks: improvement and difference between intervention groups with 95% confidence interval (CI).

Outcome measure	Traction (n=77)	Sham traction (n=73)	Difference (95% CI)* Traction minus Sham traction	P value
Global perceived effect** (no. and %)	34 (44%)	37 (51%)	-7% (-23%; 9%)	0.42
First main complaint†	28.5	28.4	0.1 (- 9.0; 9.2)	0.99
Second main complaint†	27.0	24.6	2.4 (- 6.8;11.5)	0.61
Roland Disability Questionnaire‡	3.5	4.8	-1.3 (- 2.9; 0.3)	0.11
Pain during measurement†	21.2	22.5	-1.3 (-10.4; 7.8)	0.78
Pain last week†	20.6	23.7	-3.0 (-11.8; 5.8)	0.50
Severity of low back pain‖	1.6	1.8	-0.3 (- 0.9; 0.3)	0.32
Range of motion ¶	- 2.1	0.1	-2.2 (- 6.3; 1.8)	0.28
ADL-disability†	26.7	33.8	-7.2 (-17.9; 3.6)	0.19
Work absence in days§	21.0	22.8	-1.8 (- 5.5; 1.9)	0.32

* Two-tailed 95% confidence intervals were calculated using the standard formulae for the differences of two percentages and Student *t* distribution.

** Ratings on a 7-point scale are dichotomized as improved (completely recovered and much improved) and not-improved (slightly improved, not changed, slightly worsened, much worsened and vastly worsened).

† Mean, scored on a 100 mms. VAS (best score 0, worst score 100).

‡ Mean, scored on a 24-item questionnaire (best score 0, worst score 24).

‖ Mean, scored on an 11-point scale by the research assistant (best score 0, worst score 10).

¶ Measured in degrees (lacking for 4 patients in the traction group).

§ Mean number of days divided by number of patients who were absent.

The ratings of the global perceived effect on a seven-point scale were dichotomized as improved (completely recovered and much improved); and not-improved (slightly improved, not changed, slightly worsened, much worsened and vastly worsened). Five patients (7%) in the traction group and 1 patient (1%) in the sham traction group rated themselves as much or vastly worsened. Almost half of all patients (47%) recovered completely or were much improved: 44% in the traction group and 51% in the sham group.

At baseline we asked the patients to select their 3 main complaints. Some patients identified only 2 complaints. For that reason only the scores of the first 2 main complaints were evaluated. Both groups showed a clear improvement on their 2 main complaints, but we could not discern a significant difference between groups.

Patients in both groups also improved on all secondary outcome measures, except on the range of motion. The number of side effects (e.g., headache n=6, 'strange' feeling in legs n=9) was evenly distributed between the groups.

Adjustments in the ANCOVA for baseline differences and the use of co-variables for co-interventions, degree of compliance, use of regular pain medication, and strata of randomization only slightly changed the estimates (and confidence intervals) of the group differences. Therefore, we present only the unadjusted data.

On basis of the mechanical rationale of traction a positive effect of high dose traction (higher than 35% of body weight) was expected, compared to low dose traction (lower than 20% of body weight). But, the figure shows that there was no relation between global perceived effect and the percentage traction force applied.

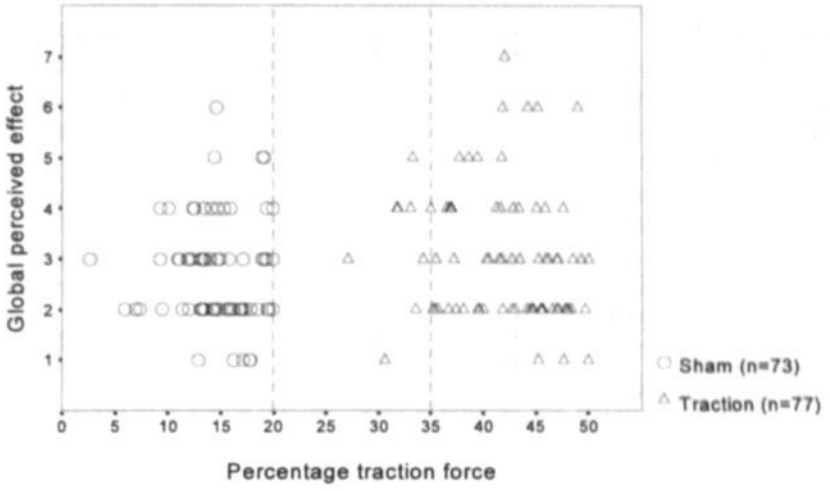


Figure. Percentage traction force of body weight (20%=maximum for sham traction; 35%=minimum for traction) versus global perceived effect (1=completely recovered; 7= vastly worsened). There is no relation between traction force applied and effect.

Per-protocol analysis

The per-protocol analysis was restricted to 135 patients: 66 patients in the traction group and 69 in the sham traction group. These patients were excluded if they withdrew from treatment, received co-interventions, received a traction force less that 35% of the total body weight (in the traction group), or had large intervals between treatment days (table 2). Some patients were excluded for more than one reason. The results of the per-protocol analysis were similar to the results of the intention-to-treat analysis: we could not show any difference in outcome measures between the traction and sham traction group.

Subgroup analysis

To explore whether particular subgroups of the population did benefit from traction we performed 8 subgroup analyses, using only our primary outcome measures. Subgroups were formed based on the following baseline characteristics, all dichotomized: age (cut-off: 40 years), gender, duration of the present episode (cut-off: 6 months), radiation below the knee (yes/no), baseline score on the General Health Questionnaire (GHQ; cut-off: 11 points), severity of the first main complaint (cut-off: 70 mms. on a VAS), appropriateness of traction therapy according to

treating physiotherapist (yes/no), and treatment in primary or hospital setting. In all subgroups we found no significant benefit of traction compared to sham traction (table 4).

Table 4. Subgroup analysis of global perceived effect (% improved).*

	Traction	Sham traction	Difference	95% CI**	P value
Age					
< 40 years (n=69)	49	53	- 4	-28; 20	0.70
≥ 40 years (n=81)	40	49	- 9	-31; 13	0.40
Gender					
male (n=84)	44	51	- 7	-28; 14	0.52
female (n=66)	44	50	- 6	-30; 18	0.63
Duration complaints					
< 6 months (n=71)	51	47	4	-19; 27	0.72
≥ 6 months (n=79)	38	54	-16	-38; 6	0.15
Radiation below the knee					
Yes (n=50)	50	46	4	-24; 32	0.75
No (n=100)	41	53	-12	-31; 7	0.22
General Health Questionnaire					
< 11 points (n=91)	43	48	- 5	-25; 16	0.65
≥ 11 points (n=59)	46	55	- 9	-34; 16	0.52
Score main compl baseline					
< 70 mm (n=105)	41	47	- 6	-25; 13	0.66
≥ 70 mm (n=45)	46	52	- 6	-35; 23	0.55
Appropriateness of traction					
appropriate (n=84)	46	55	- 9	-30; 12	0.38
not appropriate (n=64)	41	46	- 5	-29; 19	0.73
Treated					
in primary setting (n=94)	43	55	-12	-32; 8	0.22
in hospital setting (n=56)	47	42	5	-21; 31	0.74

* Ratings on a 7-point scale are dichotomized in improved (completely recovered and much improved) and not-improved (slightly improved, not changed, slightly worsened, much worsened and vastly worsened).

** Two-tailed 95% confidence intervals were calculated using the standard formulae for the differences of two percentages.

Discussion

This trial does not provide evidence for the efficacy of high dose traction in patients with persistent LBP. The results of the outcome measures were very consistent: there was no difference between the two groups on any of the outcome measures. Adjustments for baseline differences in the intention-to-treat analysis did not change the results. The number of withdrawals from treatment, loss to follow-up and protocol deviations was low. Consequently, the per-protocol analysis showed the same results as the intention-to-treat analysis. Furthermore, subgroup analysis did not show any group for which traction might seem promising.

In an earlier pilot experiment (n=25) we found substantial effects: the difference between the intervention groups was 30%.[21] This difference was not statistically

significant because of the small sample size. A trial with a small number of patients carries a considerable risk of failing to show a significant treatment effect although one is really present (Type II error).[25] Our larger trial, however, indicates that the results of the pilot were due to chance.

The large improvement of the patients is remarkable: one of every 2 patients was much improved or recovered completely. The total group of patients had LBP for a least 6 weeks, with a median duration of 20 weeks. They represent the type of patient for whom we expected little improvement without an efficacious treatment. We could attribute the improvements purely to placebo effect, but there are other reasons why patients improve after an ineffective treatment, such as spontaneous recovery or, in the statistical context, regression to the mean.[26]

In this study we could overcome most common flaws in earlier studies on traction therapy.[17] The prognostic comparability at baseline was good, the effect measures were clinically relevant and they were measured blindly. We were also able to blind the patients for treatment allocation. The sham traction was performed with a specially developed iliacal brace that simulates traction, but lacks the specific component (large traction force). We think that this study provides a valid estimate of the effect of lumbar traction for LBP. It may be wondered why we could not show any difference between low and high dose traction, as the trial was performed under rigid optimal protocols. Several explanations for the results can be given.

One explanation is that the rationales forming the basis of our contrast of traction therapy are wrong. Various authors have reported that a certain amount of traction force is necessary to achieve the mechanical effects. We expected 35-50% of the bodyweight to be efficacious and sham traction (<20%) as non efficacious. It can be argued that the contrast between the intervention groups was not large enough or that the sham traction was not a real placebo. However, we saw no relation between the applied traction force and effect of traction (figure). In other words, the effect of traction did not depend on the amount of traction force.

Another explanation could be that we selected only patients with non-specific LBP. Because the ambiguity in diagnosis of LBP is enormous[27], we made no distinction between patients with discus, facet joint or muscular problems. To explore whether particular subgroups of the population did benefit from traction we performed several subgroup analyses (for example, appropriateness for traction according to the physiotherapist and radiation below the knee). We restricted these analyses to broad patient characteristics, but not to possible diagnoses and found no effect of traction in any of our subgroups.

We only selected patients who had suffered for at least 6 weeks from non-specific LBP. Patients with acute complaints or herniated disc problems, for example, were excluded. We cannot infer that traction is ineffective for these patient groups, but the likelihood is small.

There might be positive effects if traction is given in combination with exercises. The exercises might strengthen the effect of traction. Because we found no effect at all in this trial, we think that an interaction effect with exercises is unlikely.

To sum up, this randomized trial provides a valid estimate of the effect of lumbar traction for patients with at least 6 weeks non-specific LBP. Almost half of the patients improved irrespective of the treatment given, this improvement could not be explained by the specific effect of traction. In our opinion, future research on traction therapy should not receive high priority.

Acknowledgements

The authors thank F. Phillipens, W. Simonis, W. van Baal, T. Belgers, R. Hoen, J. van Beurden, C. Gulikers, J. Coenjaerts, T. Dols, R. Maessen, W. Lahaye, W. Schmetz, R. Valkenburg, H. Creusen, C. v/d Velde, A. Hamelers for recruiting and treating the patients.

References

1. Frymoyer JW. Back pain and sciatica. New Engl J Med 1988; 318: 291-300.
2. Kelsey JL, AA White. Epidemiology and impact of low back pain. Spine 1980; 5: 133-142.
3. Beckerman H, Bouter LM, Heijden GJMG van der, Bie RA de, Koes BW. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? Br J Gen Pract 1993; 43: 73-77.
4. Deyo RA. Conservative therapy for low back pain. Distinguishing useful from useless therapy. Spine 1983; 11: 28-30.
5. Spitzer WO, Leblanc FE, Dupuis M (eds). Scientific approach to the assessment and management of activity related spinal disorders. Spine 1987; 12(Suppl): 1- 59.
6. Bridger RS, Ossey S, Fourie G. Effect of lumbar traction on stature. Spine 1990; 156: 522-524.
7. Colachis SC, Strohm BR. Effects of intermittent traction on separation of lumbar vertebrae. Arch Phys Med Rehabil 1969; 44: 251-258.
8. Mathews JA. Dynamic discography: a study of lumbar traction. Ann Phys Med 1968; 97: 275-279.
9. Onel D, Tuzlaci M, Sari H, Demir K. Computed tomographic investigation of the effect of traction on lumbar disc herniations. Spine 1989; 141: 82-90.
10. Saunders HD. Use of spinal traction in the treatment of neck and back conditions. Clin Orthop 1983; 179: 31-38.
11. Swezey RL. The modern thrust of manipulation and traction therapy. Semin Arthritis Rheum 1983; 12: 322-331.
12. Judovich BD. Lumbar traction therapy - elimination of physical factors that prevent lumbar stretch. JAMA 1955; 159; 549-550.
13. Judovich B, Nobel GR. Traction therapy, a study of resistance forces. Am J Surg 1957; 93: 282-286.
14. Mathews JA. The effects of spinal traction. Physiotherapy 1972; 58: 64-66.
15. Yates DAH. Indications and contra-indications for spinal traction. Physiotherapy 1972; 58: 55-57.
16. Eie N, Kristiansen K. Complications and hazard of traction in the treatment of ruptured lumbar intervertebral disks. J Oslo City Hosp 1962; 12: 4-12.
17. Heijden GJMG van der, Beurskens AJHM, Koes BW, Assendelft WJJ, Vet HCW de, Bouter LM. The efficacy of traction for back and neck pain. A blinded review of randomized clinical trial methods. Phys Ther 1995; 75: 93-104.
18. Ljunggren AE, Walker L, Weber H, Amundsen T. Manual traction versus isometric exercises in patients with herniated intervertebral lumbar discs. Physiotherapy Theory and Practice 1992; 8: 207-213
19. Konrad K, Tatrai T, Hunka A, Vereckei E, Korondi I. Controlled trial of balneotherapy in treatment of low back pain. Ann Rheum Dis 1992; 51: 820-822.
20. Letchuman R, Deusinger RH. Comparison of sacrospinalis myoelectric activity and pain level in patients undergoing static and intermittent lumbar traction. Spine 1993; 18: 1361-1365.
21. Heijden GJMG van der, Beurskens AJHM, Dirx MJM, Bouter LM, Lindeman E. Efficacy of lumbar traction: a randomised clinical trial. Physiotherapy 1995; 81: 29-35.
22. Beurskens AJHM, Heijden GJMG van der, Vet HCW de, Köke AJA, Lindeman E, Regtop W, Knipschild PG. The efficacy of traction for lumbar back pain. Design of a randomized clinical trial, J Manipulative Physiol Ther 1995; 18: 141-147.
23. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983; 8: 141-144.
24. Norusis MJ. SPSS for Windows. Base system user's Guide Release 5.0. Chicago: SPSS inc,1992.
25. Pocock SJ. Clinical trials. A practical approach. New York: John Wiley & Sons, 1983.

26. Kienle GS. Über das auftreten des Placeboeffekts. Eine Analyse des Materials von H.K. Beechers "The Powerful Placebo" (1955) und der Faktoren, die einen Placeboeffekt vortäuschen können. [dissertation]. Berlin: Freien Universität Berlin, 1995.
27. Deyo RA. Practice variations, treatment fads, rising disability. Spine 1993; 18: 2153-2162.

Chapter 4

Efficacy of traction for non-specific low back pain

12-week and 6-month results of a randomized clinical trial

AJHM Beurskens,[1] HCW de Vet,[1] AJA Köke,[2] W Regtop,[3] GJMG van der Heijden,[1] E Lindeman,[4] PG Knipschild.[1]

1 Department of Epidemiology, University of Limburg, Maastricht
2 Department of Physiotherapy, University Hospital, Maastricht
3 'Hogeschool Heerlen', Department of Physiotherapy, Heerlen
4 Department of Rehabilitation Medicine, University Hospital, Maastricht

Submitted for publication.

Abstract

Study Design - A randomized clinical trial.
Objective - To assess the efficacy of motorized continuous traction for low back pain.
Background - The available studies on the efficacy of lumbar traction do not allow clear conclusions due to severe methodological flaws. This trial aimed to overcome these shortcomings.
Methods - Patients with at least 6 weeks non-specific low back pain were selected. High dose traction was compared with sham (or low dose) traction. Sham traction was given with a specially developed brace that becomes tighter in the back during traction. This was experienced as if real traction were exerted. The patients and outcome assessor were blinded for treatment allocation. Outcome measures were: patient's global perceived effect, severity of main complaints, functional status, pain, range of motion, work absence and medical treatment. Results for the outcome measures 12 weeks and 6 months after randomization are presented.
Results - 151 patients were randomized. Intention-to-treat analysis of the 12-week and 6-month results showed no statistically significant differences between the groups on all outcome measures; all 95% confidence intervals included the value zero. The number of patients lost to follow-up was very low. Other analyses showed the same results.
Conclusions - Most common flaws of earlier studies on traction therapy could be overcome. This trial does not support the claim that traction is efficacious for patients with low back pain.

Introduction

Epidemiological studies have indicated that about 80% of the population experience low back pain (LBP) during their active lives.[1,2] There is no consensus about the management of LBP. The efficacy of many (physio)therapeutic interventions remains questionable.[3,4] One of the treatment options is traction, often combined with other treatment modalities (e.g., massage, exercises, electrotherapy or heat).[5]

We traced 14 Randomized Clinical Trials (RCTs) in the literature prior to June 1992 about the efficacy of different lumbar traction modalities.[6] These studies do not allow clear conclusions due to methodological flaws, such as poor prognostic comparability at baseline, lack of blinding of patients and outcome assessors. Three recent RCTs[7-9] published after June 1992 did not change the conclusions of our review.

In a pilot study (n=25) we were able to avoid the most common flaws.[10] The recovery rate after 5 weeks in the high dose traction was 64% and in the low dose (sham) traction it was 34%. In view of these substantial effects we decided to conduct a larger trial with 150 patients.

The available literature is not clear on the working mechanisms by which continuous lumbar traction could be effective. Supposed mechanical effects of traction are: vertebral distraction and widening of the intervertebral foramen.[11-14] These mechanisms suggest short-term rather than long-term effects.[15,16] A certain amount of traction force is necessary to achieve separation of the vertebrae and widening of intervertebral foramina, and a part of the force is needed to overcome muscle contraction, spinal curvatures, ligamentous resistance and friction of the body on the table-top and of the machinery.[17-19] Lumbar traction forces below 20% of the total body weight using a split table, to overcome friction of the body on the table-top, can be regarded as sham traction. To reach sufficient contrast with sham traction we chose to use a traction force between 35-50% of the total body weight for the real traction.

Methods

Study design

We compared the effect of continuous motorized lumbar traction and sham traction on the magnitude of recovery in patients with persistent non-specific LBP 5 and 12 weeks and 6 months after randomization. The 5-week follow-up was included to detect short-term effects. The 12-week and 6-month follow-up were included to detect longer term effects. In this article we present the results of the 12-week and 6-month follow-up. The results of the short-term effects are reported elsewhere.[20] The Medical Ethics Committee of the University of Limburg and the University Hospital Maastricht in the Netherlands approved the study protocol. An elaborate description of the study design has been published earlier.[21]

Patient selection and randomization

Physiotherapists in 10 practises and 6 general practitioners recruited and selected the patients. An experienced research physiotherapist (AJK) checked the inclusion and exclusion criteria. He gave information about the goal of the study and the chance of receiving the sham traction. After receiving a letter that included this

information, the patients signed informed consent. Patients were selected if they had suffered for at least 6 weeks from non-specific LBP, if they were at least 18 years old and had never had any form of lumbar traction treatment before. Patients with evidence for underlying diseases or anatomical abnormalities were excluded. Furthermore, patients were excluded if they had improved much during the previous 2 weeks. Subsequently, patients were blindly randomized to high dose or low dose traction based on a computer-generated list of random numbers. We prestratified on duration of complaints (less than, or longer than 6 months) and on physiotherapy practice (n=10).

Interventions

All patients were treated with similar traction apparatus (Eltrac, DIMEC Delft Instruments, the Netherlands). Both intervention groups were treated 12 times in 5 weeks, 20 minutes per session. After the inclusion of a patient into the trial, the treating physiotherapist received a sealed envelope that contained the treatment code. At the first treatment session he opened the envelope containing the treatment code, and was therefore not blinded for the assigned treatment.

As the patient lay down on the traction table in the semi-fowler position, braces were attached around the iliacal crest and the lower thoracic cage. After unlocking the sliding table top, the physiotherapist increased the traction force. In the traction group the force was increased until the patient indicated that the tolerance for pulling was reached, with a minimum traction force of 35% and a maximum of 50% of the total body weight. In the sham group the force was slowly increased until the patient indicated that he felt little pulling, with a maximum traction force of 20% of the total body weight. Patients received sham traction with a special brace around the iliacal crest that became tighter in the back during treatment. This was experienced as if traction were exerted. In this way we hoped that the patients would feel the pulling earlier and that the traction force in the sham group could be restricted to a minimum.

The patients were allowed to continue the pain medication they used before entry of the study (regular pain medication). Other co-interventions were not allowed during the treatment period. We asked the patients to take no pain medication during the 24 hours before the effect measurement. After the treatment period the treating physiotherapists and research physiotherapist asked the patients to restrict further treatment as much as possible.

Outcome measurements and follow-up

The 12-week measurement included a physical examination. At 6-month follow-up the patients were asked to complete a postal questionnaire. The primary outcome measures were [1] global impression of the perceived effect (recovery) rated by the patient on a seven-point scale and [2] the severity of the three main complaints. At baseline the patients selected their three main complaints in a standardized way: each patient selected three activities he performed frequently, which he perceived as important in his daily life, and which LBP made difficult for him. The severities of these main complaints were rated on a 100 mm. visual analog scale (VAS).

Secondary outcome measures were: [3] functional status measured with the Roland Disability Questionnaire (RDQ)[22]; [4] pain (100 mm. VAS) and [5] severity of LBP as evaluated by the research physiotherapist indicated on an eleven-point scale

after a standardized interview and physical examination; [6] range of motion of the spine measured with the inclinometer EDI 320-CYBEX (degrees); [7] ADL disability (100 mm. VAS); [8] work absence in days and [9] medical consumption sought by the patient recorded by questionnaires. Medical consumption is an important outcome measure for long term follow-up. Therefore, we will pay special attention to describing the results of this outcome measure. The research physiotherapist (AJK) who conducted almost all outcome measurements was blinded for treatment allocation. In periods when the research physiotherapist was on holiday, the researcher (AJB), who is also a physiotherapist, performed the blinded outcome measurements. To minimize interobserver variation, the two observers were trained in performing the measurements in a standardized way.

Data analysis

The researcher (AJB) who analyzed the data was blinded for treatment allocation, until all decisions were taken about cut-off points and protocol deviations. She was also blinded during the primary analyses.

The primary statistical analysis was carried out according to the 'intention-to-treat' principle: all patients, including withdrawals from treatment and patients with poor compliance, remained in the group to which they had been assigned by randomization. Besides this, we also performed a 'per-protocol' analysis which is restricted to a group of patients in which everything went the way it was planned. Patients were excluded from the per-protocol analysis if they: [1] withdrew during the treatment period; [2] were treated with inadequate traction forces (too low or too high); [3] received co-interventions during the 5-week treatment period; or [4] received medical treatment after the 5-week treatment period.

All data analyses were done with SPSS statistical software.[23] For the outcome measures that were also recorded at baseline, we computed the differences between the post-treatment and the baseline score for each individual and compared these between the two groups. With regard to the outcome measures without baseline information (e.g., global perceived effect and medical treatment), only the post-treatment scores of the two groups could be compared. Group differences and two-sided 95 percent confidence intervals were calculated for all outcomes. One-way analysis of covariance (ANCOVA) was used to estimate group differences adjusted for differences at baseline of important prognostic indicators, co-interventions, degree of compliance, use of regular pain medication, and strata of randomization.

Results

Study sample

In the period of June 1993-December 1994, 243 patients were invited for an appointment with the research physiotherapist to check their eligibility based on inclusion and exclusion criteria. Of these, 92 patients (38%) were excluded. The most common reasons for exclusion were: not enough motivation (n=24) (i.e., no time, chance of sham traction) or less than 6 weeks LBP (n=16). Finally, 151 patients signed informed consent, of these 130 patients were selected by physiotherapists and 21 by general practitioners. The patients were randomized: 77 patients to traction and 74 to sham traction. Of the 151 patients, 150 patients

completed the 12-week follow-up and 148 completed the 6-month follow-up. After 12 weeks one patient in the sham group went to work abroad and was lost to follow-up. In addition, 4 patients failed to attend the 12 weeks visit for physical measurement by the research assistant. However, they returned the complete questionnaires by post and could therefore be included in those analyses for outcome measures for which their data were not missing. After 6 months 3 patients did not return the questionnaires: the patient who went to work abroad and 2 patients who moved without leaving their addresses.

The two treatment groups were similar regarding most of the demographic and clinical baseline characteristics (table 1). The traction group contained a few more patients with pain radiating below the knee, previous treatment and previous LBP. On the other hand, the median number of previous LBP episodes was higher in the sham traction group.

Table 1. Comparability of treatment groups with respect to distribution of prognostic variables.

Characteristic	Traction	Sham traction
No. of patients	77	74
Mean age (yrs,sd)	39 ± 10	42 ± 11
Gender (% female)	34 (44%)	32 (43%)
Present episode		
25, 50, 75 percentile (wk)	8, 20, 52	8, 24, 52
chronic (> 6 months)	40 (52%)	40 (54%)
sub-acute (6 weeks <> 6 months)	37 (48%)	34 (46%)
radiation below the knee	28 (36%)	22 (30%)
previous treatment	47 (61%)	37 (50%)
Previous low back pain	66 (86%)	57 (77%)
Number of low back pain episodes ever		
25, 50, 75 percentile	4, 6, 20	4, 10, 20
Mean General Health Questionnaire (0-36)	8.3	8.6
Mean severity (100 mms. VAS)		
first main complaint	75	73
second main complaint	74	70
Mean Roland Disability Questionnaire (0-24)	12	12
Mean pain score (100 mms. VAS)		
during measurement	61	55
last week	62	62
Mean severity of low back pain (0-10)	5	5
Range of motion (degrees)	54	54
ADL disability (100 mms. VAS)	67	70

Compliance and blinding

The average traction force was 42% of the total body weight in the traction group and 15% in the sham group. The average traction force was calculated from the individual average traction force (percentage of the total body weight) over all treatment sessions. During the treatment period 8 patients (4 in every group) withdrew from treatment. Directly after the 5-week traction treatment we asked the patients to guess their treatment allocation. The traction was not systematically

unmasked; only 6% of the patients in the sham group and 1% in the traction group thought that they had received sham traction.[20]

Outcomes

Tables 2 and 3 show the 12-week and 6-month results of the intention-to-treat analyses of outcome measures, respectively. Both the traction and sham traction groups improved, but at both measuring points the differences between the groups were very small for all outcome measures. All confidence intervals for the differences between groups included the value zero. This means that they were not statistically significant at 5% level (two-sided test).

The ratings of the global perceived effect on a 7-point scale were dichotomized in improved (completely recovered and much improved) and not-improved (slightly improved, not changed, slightly worsened, much worsened and vastly worsened). After 12 weeks, 50% of the patients in the traction group and 48% in the sham group had recovered completely or were much improved. After 6 months these percentages were 47% and 44% respectively.

Table 2. Intention-to-treat analysis at 12 weeks: improvement and difference between intervention groups with 95% confidence interval (CI).

Outcome measure*	Traction (n=77)	Sham traction (n=73)	Difference (95% CI) Traction minus Sham traction	P value
Global perceived effect** (no. and %)	38 (50%)	35 (48%)	2% (-14%;18%)	0.80
First main complaint §	33.7	31.5	2.2 (- 8.5;13.0)	0.68
Second main complaint §	35.4	30.7	4.7 (- 5.2;14.7)	0.35
Roland disability questionnaire ¶	4.4	4.3	0.1 (- 1.8; 1.9)	0.97
Pain at the moment §	28.5	22.8	5.7 (- 4.6;15.9)	0.28
Pain during last week §	24.2	23.9	0.3 (- 9.9;10.5)	0.95
Severity of low back pain †	2.3	2.2	0.1 (- 0.6; 0.9)	0.69
Range of motion ǂ	- 1.1	1.2	-2.4 (- 6.9; 2.2)	0.31
ADL disability §	27.1	29.4	-2.4 (-13.6; 8.9)	0.68
Work absence in days‡	23.5	27.8	-4.3 (-14.7; 6.1)	0.41
Medical consumption (no. and %)¥	26 (34%)	18 (25%)	9% (-6%; 24%)	0.22

* Lost to follow-up: 1 in sham group. Resulting numbers: 77 patients in traction group and 73 patients in sham group. Missing values: Global perceived effect: 1 in traction; Severity of low back pain: 3 in traction and 3 in sham; Range of motion: 6 in traction and 3 in sham.

** Ratings on a 7-point scale are dichotomized in improved (completely recovered and much improved) and not-improved (slightly improved, no changed, slightly worsened, much worsened and vastly worsened).

§ Mean, scored on a 100 mm. VAS (best score 0, worst score 100).

¶ Mean, scored on a 24-item questionnaire (best score 0, worst score 24).

† Mean, scored on a 11-point scale by the research assistant (best score 0, worst score 10).

ǂ Measured in degrees.

‡ Mean number of days between baseline and 12-week follow-up divided by number of patients with payed work.

¥ Medical consumption between baseline and 12-week follow-up.

At baseline we asked the patients to select their 3 main complaints, but some patients (5%) could identify only 2 complaints. For that reason only the scores of the first 2 main complaints were evaluated. Both groups showed a clear improvement on their 2 main complaints, but we could not find any difference between groups.

Adjustments in the ANCOVA for baseline differences, use of regular pain medication, and the use of co-variables for strata of randomization only slightly changed the estimates (and confidence intervals) of the group differences. Therefore, we only present the unadjusted data. The results of the per-protocol analysis were similar to the results of the intention-to-treat analysis: we could not show any difference in outcome measures between the traction and sham traction group.

Medical consumption

The number of additional treatments sought by the patient between baseline and 12-week follow-up in the traction group was higher than in the sham group; 34% versus 25% (table 2). But, after 6 months the number of treatments was similar for both groups; 45% of the patients in the traction and 42% in the sham group had at least one additional treatment (table 3).

Table 3. Intention-to-treat analysis at 6 months: improvement and difference between intervention groups with 95% confidence interval (CI).

Outcome measure*	Traction (n=76)	Sham traction (n=73)	Difference (95% CI) Traction minus Sham traction	P value
Global perceived effect** (no. and %)	35 (47%)	32 (44%)	3% (-13%;19%)	0.79
First main complaint §	36.7	36.0	0.7 (- 9.8;11.3)	0.89
Second main complaint §	35.8	32.8	3.0 (- 7.3;13.3)	0.56
Roland disability questionnaire ¶	4.7	4.0	0.7 (- 1.1; 2.6)	0.44
Pain at the moment §	23.8	20.1	3.7 (- 8.4;15.8)	0.55
Pain during last week §	25.0	25.5	-0.5 (-11.5;10.6)	0.93
ADL disability §	25.7	25.8	0.1 (-11.5;11.2)	0.98
Work absence in days‡	35.7	43.7	-8.0 (-27.0;11.0)	0.40
Medical consumption (no. and %)¥	34 (45%)	30 (42%)	3% (-13%;19%)	0.71

* Lost to follow-up: 1 in traction and 2 in sham. Resulting numbers: 76 patients in traction group and 72 patients in sham group. Missing values: Global perceived effect: 1 in traction; Pain at the moment: 2 in sham; Pain during last week: 1 in traction and 1 in sham; First and second main complaint: 1 in sham; ADL-disability: 8 in traction and 10 in sham.

** Ratings on a 7-point scale are dichotomized in improved (completely recovered and much improved) and not-improved (slightly improved, no changed, slightly worsened, much worsened and vastly worsened).

§ Mean, scored on a 100 mm. VAS (best score 0, worst score 100).

¶ Mean, scored on a 24-item questionnaire (best score 0, worst score 24).

‡ Mean, number of days between baseline and 6-month follow-up divided by number of patients with payed work.

¥ Medical consumption between baseline and 6-month follow-up.

The number of additional treatment gives no information about the content of the treatments. It is, therefore, important to evaluate the types of treatments patients received. Table 4 presents the cumulative number of medical consumption at follow-up. The types of treatments for LBP diverged during the 6-month follow-up, but were not remarkably different for the two groups. Some patients had more than one type of treatment (e.g., medication and physiotherapy).

When many patients have additional treatments, it is not possible any more to separate effects of traction and other treatments. Only 4 patients had an additional treatment besides traction during the 5-week intervention period. The additional treatments sought by the patients occurred mainly between 5 weeks and 6 months after randomization. Therefore, at 5-week follow-up the pure effect of traction was assessed. To evaluate which patients had treatments, we studied the relation between global perceived effect assessed after 5-week traction treatment and medical consumption at 6-month follow-up (figure). Because high dose traction and sham-traction did not differ, we only show the relation with global perceived effect for the whole study population. None of the 9 patients who rated themselves as recovered after the traction treatment had an additional treatment during the 6-month follow-up (figure). But 20 of the 61 patients (33%) who were much recovered after 5 weeks had one or more treatments in the follow-up period. The majority of the patients who rated themselves as slightly improved, not changed or deteriorated received additional treatment for LBP.

Table 4. Cumulative number of medical consumption at follow-up.

	5 weeks	12 weeks	6 months
Traction	2 medication	4 medication 5 corset 14 physiotherapy 2 manual therapy 2 cesar therapy 1 HNP operation	5 medication 6 corset 20 physiotherapy 3 manual therapy 3 cesar therapy 2 HNP operation 1 chiropractor 2 nerve block 1 health resort 1 behavioral program
Sham traction	2 physiotherapy	3 medication 1 corset 11 physiotherapy 4 manual therapy 1 HNP operation	4 medication 3 corset 22 physiotherapy 8 manual therapy 1 HNP operation 1 nerve block 1 cesar therapy 1 homeopathy 1 rest cure

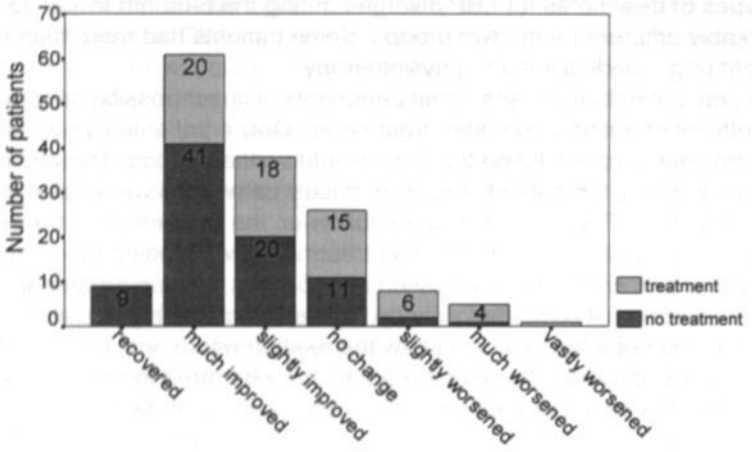


Figure Additional treatment at 6-month follow-up across global perceived effect assessed after 5-week traction treatment.

Discussion

Supposed rationales of traction are based on mechanical effects. These mechanisms suggest that short-term rather than long-term effects can be expected.[15,16] Until now, the rationales offer no explanations for long-term effects of traction. The 5-week results showed no statistically significant differences between the intervention groups.[20] After 12-week and 6-month follow-up there were also no statistically significant differences between traction and sham traction treatment on all outcome measures.

We compared the number of patients who had some kind of additional treatment for their LBP during the follow-up period. We regarded additional treatment after the 5-week intervention period as an outcome measure and not as a protocol deviation. The underlying assumption is that patients will turn to other therapies when the allocated therapy is not effective (any more) or not effective enough. In this trial the types of additional treatments during the follow-up period diverged and the numbers were relatively high but not remarkably different for the two groups. When the number of patients with additional treatments is high, the results of the other outcome measures are difficult to interpret: it is no longer clear to what extent the effect can be attributed to traction therapy or to the other treatment. We saw that one third of the patients who rated themselves as much improved after 5-week traction treatment had additional treatment for LBP within 6 months. This is an indication that the improvement was only temporary, or not satisfactory enough for the patient.

The number of recurrences is considered to be an important outcome measure in long term follow-up of intervention trials for LBP. In the design of this trial, the number of recurrences was intended to serve as an outcome measure for long-term

follow-up.[21] Recurrences were recorded by means of patient questionnaires at 12-week and 6-month follow-up. However, during the follow-up it appeared to be very difficult for patients to identify and recall accurately the beginning and end of periods with LBP, and the number of these periods. Thus, due to recall bias, the answers of the patients were almost certainly imprecise and probably not valid. For this reason, we decided not to use the data on recurrences in the analysis.

To ascertain the number and duration of recurrences precisely, more frequent measurements are necessary,[24] e.g., with short telephone interviews monthly. It seems quite likely that there is a relation between recurrences and medical treatments. The fact that totally recovered patients had no additional treatment indicates that these patients probably had no recurrences.

A serious form of bias may occur when patients are lost to follow-up. Patients who do not show up for their follow-up measurement might be a selected group of patients in which therapy was very successful or unsuccessful. Frequently, as patients become more ill and symptomatic, they will be incapable of completing a questionnaire or unwilling to do so.[25] Large numbers of patients lost to follow-up are often overlooked as a serious source of bias for evaluating the effect of a therapy.[26] At each contact, the research physiotherapist explained to the patients that it was very important for them to come to the effect measurements and to fill out the questionnaires, also in case their complaint did not improve. We put in great effort to get in touch (by phone or post) with patients who did not respond. As a result we were able to limit the number of patients lost to follow-up. We had a very complete data set which made the analyses simple and clear.

In conclusion, in this trial we were able to keep close to the original design of this study. Thereby we could overcome most common flaws in earlier studies on traction therapy. This trial does not support the claim that traction is efficacious for patients with low back pain.

Acknowledgements

The authors thank F. Phillipens, W. Simonis, W. van Baal, T. Belgers, R. Hoen, J. van Beurden, C. Gulikers, J. Coenjaerts, T. Dols, R. Maessen, W. Lahaye, W. Schmetz, R. Valkenburg, H. Creusen, C. v/d Velde, A. Hamelers for recruiting and treating the patients.

References

1. Haanen HCM. An epidemiologic survey on low back pain. Dissertation, Erasmus University, Rotterdam, 1984.
2. Kelsey JL. Epidemiology of musculoskeletal disorders. New York: Oxford University Press, 1982: 145-167.
3. Koes BW, Bouter LM, Heijden GJMG van der. Methodological quality of randomized clinical trials on treatment efficacy in low back pain. Spine 1995; 20: 228-235.
4. Spitzer WO, Leblanc FE, Dupuis M (eds). Scientific approach to the assessment and management of activity related spinal disorders. Spine 1987; (Suppl) 12: 1-59.
5. Beckerman H, Bouter LM, Heijden GJMG van der, Bie RA de, Koes BW. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? Br J Gen Pract 1993; 43: 73-77.
6. Heijden GJMG van der, Beurskens AJHM, Koes BW, Assendelft WJJ, Vet HCW de, Bouter LM. The efficacy of traction for back and neck pain. A blinded review of randomized clinical trial methods. Phys Ther 1995; 75: 93-104.

7. Konrad K, Tatrai T, Hunka A, Vereckei E, Korondi I. Controlled trial of balneotherapy in treatment of low back pain. Ann Rheum Dis 1992; 51: 820-822.
8. Letchuman R, Deusinger RH. Comparison of sacrospinalis myoelectric activity and pain level in patients undergoing static and intermittent lumbar traction. Spine 1993; 18: 1361-1365.
9. Ljunggren AE, Walker L, Weber H, Amundsen T. Manual traction versus isometric exercises in patients with herniated intervertebral lumbar discs. Physiotherapy Theory and Practice 1992; 8: 207-213.
10. Heijden GJMG van der, Beurskens AJHM, Dirx MJM, Bouter LM, Lindeman E. Efficacy of lumbar traction: a randomized clinical trial. Physiotherapy 1995; 81: 29-35.
11. Bridger RS, Ossey S, Fourie G. Effect of lumbar traction on stature. Spine 1990; 156: 522-524.
12. Colachis SC, Strohm BR. Effects of intermittent traction on separation of lumbar vertebrae. Arch Phys Med Rehabil 1969; 44: 251-258.
13. Mathews JA. Dynamic discography: a study of lumbar traction. Ann Phys Med 1968; 97: 275-279.
14. Onel D, Tuzlaci M, Sari H, Demir K. Computed tomographic investigation of the effect of traction on lumbar disc herniations. Spine 1989; 141: 82-90.
15. Saunders HD. Use of spinal traction in the treatment of neck and back conditions. Clin Orthop 1983; 179: 31-38.
16. Swezey RL. The modern thrust of manipulation and traction therapy. Semin Arthritis Rheum 1983; 12: 322-331.
17. Judovich BD. Lumbar traction therapy - elimination of physical factors that prevent lumbar stretch. JAMA 1955; 159; 549-550.
18. Judovich B, Nobel GR. Traction therapy, a study of resistance forces. Am J Surg 1957; 93: 282-286.
19. Mathews JA. The effects of spinal traction. Physiotherapy 1972; 58: 64-66.
20. Beurskens AJHM, Vet HCW de, Köke AJA, Lindeman E, Regtop W, Heijden GJMG van der, Knipschild PG. The efficacy of traction for low back pain. A randomized clinical trial. Lancet 1995; 346: 1596-1600.
21. Beurskens AJHM, Heijden GJMG van der, Vet HCW de, Köke AJA, Lindeman E, Regtop W, Knipschild PG. The efficacy of traction for lumbar back pain. Design of a randomized clinical trial. J Manipulative Physiol Ther 1995; 18: 141-147.
22. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low back pain. Spine 1983; 8: 141-144.
23. Norusis MJ. SPSS for Windows. Base system user's Guide Release 5.0. Chicago: SPSS inc, 1992.
24. Hoogen HJM van der, Koes BW, Eijk JThM van, Bouter LM, Devillé W. On the course of low-back pain in general practice: a one-year follow-up study. (Submitted).
25. Deyo RA. Practice variations, treatment fads, rising disability. Do we need a new clinical research paradigm? Spine 1993; 18: 2153-2162.
26. Bloch R. Methodology in clinical back pain trials. Spine 1987; 12: 430-432.

Chapter 5

Review article

Measuring the functional status of patients with low back pain

Assessment of the quality of four disease-specific questionnaires

AJHM Beurskens,[1] HCW de Vet,[1] AJA Köke,[2] GJMG van der Heijden,[1] PG Knipschild.[1]

1 Department of Epidemiology, University of Limburg, Maastricht
2 Department of Physiotherapy, University Hospital, Maastricht

Spine 1995; 20: 1017-1028.

Abstract

Study design - This study was a literature review of the quality of four disease-specific functional status questionnaires for patients with low back pain: Oswestry; Million; Roland; and Waddell disability questionnaire.
Objectives - The questionnaires were evaluated in terms of general description, scale structure, reliability, validity, responsiveness, and clinical research applications.
Summary of Background Data - Functional status is an outcome of great interest for clinical trials of low back pain.
Methods - A computer aided search was conducted of articles published between 1981-1993 and references given in selected relevant publications. Articles were selected if at least one of the four functional status questionnaires was used or if the authors gave relevant information about the methodology of these questionnaires.
Results - There was not enough information available about the criteria of item selection used for the development of the questionnaires. The test-retest reproducibility of the questionnaires seemed satisfactory. The Oswestry and Roland disability questionnaires have been used and evaluated more frequently than the Million and Waddell. Therefore, we can be more certain about the validity and responsiveness of the former pair of questionnaires.
Conclusion - In the absence of a gold standard, direct comparisons of evaluative functional status questionnaires in a single patient group are needed. Through direct comparisons, comparative validity and responsiveness can be assessed. Functional status measures are not currently used in many settings in which they would be valuable. It is important to encourage their wider use in clinical trials. Additional research is needed to compare and improve the existing questionnaires.

Introduction

Epidemiologic studies have indicated that about 80% of the population experience back pain during their active lives.[1-3] The impact of low back pain is related to the function of patients. The goal of the treatment is to optimize patients' quality of life in terms of symptoms and function.[4,5] In research, it is important to choose relevant outcome parameters. Outcome parameters should fit in with central research questions. For trials of low back pain, symptoms and functional status are of great interest.[6-8]

Patients want to be free of pain and perform their daily activities. Therefore, "hard" measures, such as laboratory measures, muscle strength and spinal mobility, have no direct clinical importance to patients.[7] These physical measures are useful depending on the degree that they correlate with symptoms and functional status, but the correlation between physical measures on the one hand and symptoms and functional status on the other hand is often very poor.[9] Rather than to infer them, functional status questionnaires seek to quantify symptoms and function directly.[10] The more they do, the better they reflect the concerns of the patients.

Functional status is a patient-referenced concept and different for each individual. Some patients make higher demands on functions than others. However, certain categories of activities are common to everyone, such as sitting and sleeping. Disease specific questionnaires are used to assess functional status in patients with low back pain in clinical trials. In line with Deyo[7] we will use the term "functional status questionnaires" to denote questionnaires that assess a patient's limitations in performing usual human tasks of living.

Measurement instruments can be used for three purposes: 1) discriminating among subjects, 2) predicting prognosis, and 3) evaluating change over time.[11,12] Questionnaires with different purposes require different measurement properties.[12] This paper was restricted to evaluative questionnaires. These questionnaires are used to measure the magnitude of longitudinal change in an individual or a group.

The purpose of this paper was to evaluate four disease-specific functional status questionnaires for patients with low back pain. We selected the most widely applied and evaluated questionnaires.13,[7,13-17] We restricted ourselves to: the Oswestry low back disability questionnaire,[18] the Million visual analog scale,[9] the Roland disability questionnaire,[19] and the Waddell disability index.[20]

Method

We evaluated the four questionnaires in terms of general description, scale structure, reliability, validity, responsiveness, and clinical research applications using criteria which have been suggested by several authors.[4,7,5,11,12,21,22] For easy reference we will shortly describe the theoretical backgrounds of the sections regarding reliability, validity, responsiveness and research applications before the results of each section.

We conducted a MEDLINE search of articles published between 1981-1993 (keywords: activities of daily living, backache, disability evaluation, low back pain, pain measurement, questionnaires, psychometrics). In addition, references given in selected relevant publications were further examined. An article was selected if at

least one of the four functional status questionnaires was used or if the authors gave relevant information about the methodology of these questionnaires.

General description

A general description of the four functional status questionnaires and information about the scale structure are given in table 1.

Table 1. General description and scale structure of four functional status questionnaires for low back pain.

	Oswestry Disability Questionnaire	Million Visual Analog Scale	Roland Disability Questionnaire	Waddell Disability Index
Author, year of first publication	Fairbank, 1980	Million, 1982	Roland, 1983	Waddell, 1984
Number of items	10	15	24	9
Response possibilities	ordinal scale; each item contains six statements	interval scale; visual analog scales	nominal; yes/no	nominal; yes/no
Subscales	pain analgetic, personal care, lifting, walking, sitting, standing, sleeping, sex life, social life, traveling	back pain, pain night, stressful circumstances, pain analgetic, stiffness, freedom and discomfort walking, standing, turning and twisting, sitting soft/ hard chair; lying, working, overall handicap, life-style	sentences that describe the patient's situation today; items drawn from 8 different Sickness Impact Profile categories	lifting, sitting, standing, travelling, walking, sleeping, social activity, sexual life, putting on footwear
Time specific items	not restricted to a time period	not restricted to a time period	today	not restricted to a time period
Time for administration	3½ to 5 minutes	not reported, about 10 minutes	5 minutes	not reported, about 5 minutes

The Oswestry questionnaire indicates the extent to which a person's functional level is restricted by pain.[18] The scale consists of 10 sections that refer to activities of daily living that might be disrupted by low back pain. The sections have been selected from a series of experimental questionnaires designed to assess limitations of various activities of daily living. The chosen sections are those found to be most relevant to the problems experienced by people with low back pain receiving physical therapy. Each section contains six statements. Each statement describes a greater degree of difficulty in that activity than the preceding statement.[18] In each section, the six statements are scored from 0 to 5. The scores of all sections are added up, giving a possible score of 50. The total score is doubled and expressed as a percentage.

A slight modification of the original questionnaire was made by the Medical Research Council (MRC) in a feasibility study of a randomized controlled trial of chiropractic and hospital out-patient management for low back pain.[23] Pain was originally assessed in terms of analgesic intake. Because this presented difficulties for patients who were not using analgesics, they replaced this question with one about 'pain severity'. The levels of disability indicated by the Oswestry (initial scores around 30%) were rather low in the feasibility study.[23] If functional status scores at baseline are low, there is not much room for improvement in functional status. It

becomes increasingly difficult to detect small but clinical important improvements. For this reason Hudson-Cook et al[8] designed the revised Oswestry disability questionnaire. The wording of all sections was altered, and one section was omitted entirely due to poor compliance (section 8: sex life disability). A new section was created dealing with the patient's interpretation of their own changing pain pattern. The revised Oswestry produced higher mean disability scores and a wider representation of scores in each section.[8] This is useful when monitoring changes in patients suffering from less disabling forms of low back pain, because the baseline scores will be higher, leaving more room for improvement. The Oswestry and the revised Oswestry were found to be equally reliable and valid.[8] Moreover, it is often unclear which version of the Oswestry has been used in the studies. For these purposes, we do not make a distinction between the two versions of the Oswestry in this present report.

Million and colleagues[9] developed a 15-item checklist about disability and pain intensity for measuring progress among patients with back pain. They selected a series of questions relating to various aspects of pain, such as the influence of the circumstances and activities on the pain, and the effects the pain has upon the patient's capacity to lead a normal life. There is no information available about the process of item selection. The score is integrated by adding up the equally weighted scores.

The Roland disability questionnaire (also named the RDQ or the St. Thomas questionnaire) has been derived from the Sickness Impact Profile (SIP). The SIP was developed as a generic health status questionnaire for use in a variety of chronic diseases.[24] The Roland questionnaire was constructed by choosing 24 yes/no items relevant for back pain from the SIP that cover a range of aspects of daily living.[19] The phrase "because of my back pain" was added to each statement in order to distinguish disability resulting from back pain and disability resulting from other causes. The scoring of the Roland is achieved simply by adding the equally weighted number of positive responses. An individual patient's score can vary from zero (no disability) to 24 (severe disability).[19] Patrick et al[25] have developed a slightly modified version of the Roland questionnaire that excludes 5 items that were relatively non-responsive in two different studies of acute and chronic low back pain. It adds 4 new items from the SIP. There are no published data regarding the use of this version of the Roland questionnaire.

The Waddell disability index is a short yes/no checklist. Waddell and Main[20] defined disability as the resulting loss of function based on general activities of daily living. Loss of function is assessed on nine basic physical activities of daily living commonly restricted by low back pain. These activities were partly identified from a previous questionnaire[26] and from pilot interviews. The score is integrated by adding the equally weighted positive items and can vary from 0 to 9.[20]

There is a lack of clarity about the populations of patients and the criteria of item selection and item reduction used for the development of the four questionnaires. The population on which a questionnaire is based determines the application possibilities of that questionnaire. For example, patients with chronic low back pain may have difficulties with activities of daily living that are different from those of patients with acute pain. Important criteria for item selection and item reduction include the number of patients who listed the item as an important problem and the potential responsiveness of the items.

Scale structure

Questions in an evaluative instrument must be specific about time.[27] The time frame should be limited. If the question time frame is too long or is left undefined, patients may be confused as to which period to report.[21] The time frame of the Roland is defined as today. The time frames of the other three questionnaires are not defined by the authors.

It is important to evaluate whether the questions are capacity or performance based. If patients actually performed the movements or actions, the question is in the form: "Did you do ....?" Or the other hand, the question may be about the patient's estimation of whether he is able to perform the movements or actions: "Are you able to ....?" The problem with capacity based questions is that patients can overestimate or underestimate their abilities. The purpose of functional status questionnaires is to assess limitations in performing movements and actions. In this context, it is more relevant to ask what a patient is really doing than what the patient thinks he can do. The Roland, Million, and Waddell contain only performance based questions. The statements in some sections of the Oswestry are mixed up capacity and performance based. In this way it is not clear what is being asked - whether the patient can perform the movement or action described in a statement or whether the patient thinks he can perform the movement or action.

To ensure questionnaire responsiveness, small changes on each item must be detected if they occur. Response options refer to the categories or range that patients have in responding to questionnaire items.[27] The response options should have sufficient graduations for registering change. If yes/no response options are used, the effect of an intervention that reduced, but did not eliminate difficulty with a movement or action could not be detected (for example climbing stairs).[28] The answers on the Million scale are scored on visual analog scales. Each question of the Oswestry is scored on a hierarchical 6-point scale. The Roland and Waddell questionnaires use yes/no questions. If a patient improves only a little, it may be difficult to choose between yes and no. As a result, if there are only small changes, the Million and Oswestry questionnaires will be more responsive than the Waddell or Roland.

The items of the four questionnaires are easy to understand, and the time required to administer the questionnaires is very short - about 5 to 10 minutes (table 1). The four questionnaires are designed to be self-administered. The Roland and Waddell can also be administered by phone, because the questions are very short and simple. The Oswestry and Million, in contrast, are difficult to administer over the phone because the response options are difficult to fill in.

Reliability

Theoretical background

Reliability is a generic term that is used by psychometricians to indicate both reproducibility (or precision) of scores and internal consistency of a scale.[10,12,29] The former typically includes test-retest reproducibility, especially for self-administered questionnaires, and inter-observer reproducibility for questionnaires by interviewers.

Test-retest and inter-observer reproducibility estimate the extent to which the same results are obtained on repeated administrations of the same instrument when

no change in functional status is expected; the absence of random error.[29] For an evaluative instrument, the only requirement is that replicate measurements on each individual remain stable over time, that is, the magnitude of the within-person variance is small.[12,30]

For continuous data, the traditional measure of reproducibility is the product-moment correlation (r). The disadvantage of this statistic is that repeated measures may be systematically different, and yet highly or (perfectly) correlated. The test-retest correlation could be combined with a t-test for the difference in means. Another measure for reproducibility is the intra-class correlation coefficient (ICC). This statistic for repeated measures assesses not only the strength of correlation, but also whether the slope and intercept vary from those expected.[31] Like a product-moment correlation, values of the ICC vary from -1 to +1, with 0 indicating no association at all.[10] According to Guyatt and Jaeschke,[30] a high reproducibility coefficient is not always a necessary condition for the responsiveness of an evaluative instrument.

Reproducibility of functional status questionnaires intended for evaluative use may be best measured at intervals of 1-2 weeks, if one can assure or estimate that clinical status is reasonably stable. As opposed to same day or next day testing, this strategy minimizes subjects' recall for the previous answers, and also provides a more realistic view of the degree of score change that may occur for non-specific reasons (random error) in an evaluative study.[32] However, in situations wherein the clinical status is not stable (e.g., in (sub)acute patients), a time interval of 1-2 weeks might be too long.

Internal consistency is an estimate of the homogeneity of an instrument or the extent to which the instrument measures a single trait or characteristic.[29] Internal consistency is more important for questionnaires with discriminative purposes than for questionnaires used for evaluation of longitudinal change.[12] To discriminate between (groups of) patients, the questions asked in the questionnaire should be relevant for all subjects. The purpose of evaluative questionnaires is measurement of the magnitude of longitudinal change in patients.[7,12,33] It is not important whether all questions are relevant for all patients. However, items not relevant for all patients cannot be left out, because they can be of great importance to some patients.[12] To be responsive to change, the components of a multiple-item index are not expected to measure a single homogeneous attribute.[8,12,27,34]

Therefore, statistical tests for measuring internal consistency, such as Cronbach's alpha, are not helpful for evaluative measures.[12,34] Some authors have calculated Cronbach's alpha for the Waddell disability index,[35] the Oswestry,[16] and the Roland[16,36]. Because of the irrelevance of internal consistency for evaluative measurements, it will not be discussed further.

Results reliability

The test-retest and inter-observer correlation coefficients are given in table 2.[40,41] The inter-observer reproducibility has only been assessed for the Million questionnaire. The correlation was 0.92, which is satisfactory.[9] The Pearson's correlation coefficient is more often used to assess the test-retest correlations than the intra-class correlation coefficient (table 2). Pearson's correlations vary between 0.72 and 0.99. The intra-class correlation coefficient has only been calculated for the

Oswestry, which was 0.83.[15] The authors used different time-intervals for measuring the test-retest correlations; they ranged from 2 hours to 6 months.

Table 2. Reliability results of four functional status questionnaires for low back pain.

	Oswestry Disability Questionnaire	Million Visual Analog Scale	Roland Disability Questionnaire	Waddell Disability Index
Test-retest correlation	r*=0.99 (1 day)[18]	r=0.97 (same day)[9,38]	r=0.91 (same day)[19]	0.73<r<0.90 individual items (time between measurements unknown)[35,41]
	ICC♦: 0.83 (1 week)[15]		r=0.83 (3 weeks)[39]	
	r=0.94 (2 hours)[37]		r=0.72 (within 6 months; range: 2 days-6 months)[40]	
Inter-observer correlation			r=0.92 (2 observers)[9]	

* r = Pearson's correlation coefficient
♦ ICC = Intra-class correlation coefficient

Notwithstanding the limitations of using Pearson's correlation coefficients and the wide range of time-intervals, the test-retest reproducibilities of the Oswestry, Million, Roland and Waddell questionnaires seem satisfactory. This means that, apart from possible systematic error, these functional status measures show only a small amount of random error in relatively stable subjects. We are more certain about the test-retest quality of the Oswestry than for the other questionnaires because both the intra-class correlation coefficient and the Pearson's correlation coefficients have been calculated.

Validity

Theoretical background

A gold standard for measuring functional status in low back pain patients is not available. The main methods for establishing whether evaluative questionnaires are measuring what they are supposed to measure are called construct and content validity.

Construct validity means that the instrument relates to other tests or measures in the way one would expect if it is really measuring what it is supposed to measure.[7,12,42] Content validity means that all relevant aspects of the domain of low back pain are represented.[42,43]

Results validity

Table 3[40,41,44,46-48] summarizes the results about construct validity. The correlations between the four questionnaires on the one hand and physical tests and signs on the other hand are low; they vary from 0.003 to 0.55, with a outlier of 0.74. Correlations between pain and the functional status questionnaires are higher but moderate; they vary from 0.27 to 0.62. In addition, the correlations with psychosocial questionnaires are low.

Table 3. Construct validity results of four functional status questionnaires for low back pain.

	Oswestry Disability Questionnaire	Million Visual Analog Scale	Roland Disability Questionnaire	Waddell Disability Index
Physical test/signs	Related to presence or absence of relaxation in back muscles during flexion: r=0.74[17] Trunk strength ratios inversely related: r=-0.47[17] Trunk mobility was inversely related[17] Distribution paraspinal muscle atrophy with CT-scan findings L5-S1 males r=0.55; females r=0.12; overall r=0.33 females r=0.12; overall r=0.33[44]		Little agreement[15] Chance in spinal mobility only significantly associated in women[47] Combination of physical tests r=0.51; combination explained 24.4% of the variance[48] Spinal flexion r=low[49] Change in spinal flexion r=0.29; improvement in SLR r=0.003[50] Spine flexion r=0.42; SLR r=-0.30[36]	Physical impairment index explained 46% of the variance[20] Physical impairment explained 40.3% of the variance[51] Spinal flexion r=low[49] Combination of physical tests r=0.52; combination explained 26.3% of the variance[48]
Pain	VAS r=0.62[15] VAS r=0.47[14]	VAS r=0.44[45]	VAS r=0.38[14] Six-point pain rating scale: good agreement[15] Self rated pain improvement r=0.41; professional's rating of improvement r= 0.30[50] Self-rated pain severity r=0.42[36] McGill pain questionnaire r=0.27[46] Pain drawing r=0.28[46]	In range of 0.3-0.6[51]
Psychosocial findings	MMPI scales: on average 0.33[46]	Posttreatment Back Depression Inventory (BDI) r=0.46 [45]	Back Depression Inventory (BDI) r=0.47[46] Psychiatric problems r=-0.09[36]	Low back outcome scale r=0.74[28]
Disability assessments	Low back outcome scale r=0.87[28] Pain Disability Index r=0.83[15]	Posttreatment measure: PAIRS (pain and impairment scale) r=0.79[45]	SIP r=0.78[46]; r=0.85[39] SIP physical dimension r=0.60[46]; r=0.59[39] SIP psychosocial dimension r=0.60[46]; r=0.59[39] Pain Disability Index r=0.63[46] Functional assessment screening questionnaire r=0.50[46] Functional Rating scale r=0.54[46] Resumption of full activity r=0.38[50]	
Oswestry **Million** **RDQ** **Waddell**	- - r=0.77[14] r=0.70[20]	- -	r=0.77[14] - -	r=0.70[20] - - - -

VAS = visual analog scale, SIP = Sickness Impact Profile, MMPI = Minnesota Multiphasic Personality Inventory.

We found higher correlations among the four questionnaires and other disability assessments (table 3). For example, correlations between the Low Back Outcome Scale and the Oswestry and Waddell are 0.87 and 0.74 respectively.[28] The correlation between the Pain Disability Index and the Oswestry is 0.83.[28] Correlations among the three functional status questionnaires are high: between Oswestry and Roland 0.77[14]; and between Oswestry and Waddell 0.70.[20] All disability assessments contain questions about aspects of daily living. The response

categories, the degree of limitations of functional status, and the accents of the measured aspects differ from questionnaire to questionnaire, but the purpose is in general the same. Because of this, it is logical that the correlations among disability assessments are moderate to high. As table 3 shows, the construct validity of the Oswestry and Roland questionnaires have been more frequently studied than that of the Waddell and Million.

It is reasonable to presume that functional status questionnaires cover a broad field of aspects, including pain, physiological, and psychosocial aspects. However, they also measure other relevant dimensions belonging to the field of functional status. If correlations between pain, physiological and psychosocial measures and functional status were extremely high, either functional status assessment or one of the other measures would be redundant.[50] It seems that the functional status questionnaires do not measure or only partly measure pain and physiological and psychosocial aspects. In cases where these factors are considered relevant, other questionnaires such as the McGill pain questionnaire,[52] physical impairment scale,[20] or the Zung psychological questionnaire[53] can be used.

Only a few authors have reported on the content validity of the four questionnaires. In their study, Baker et al[13] demonstrated that the Roland tended to score higher in lower ranges of disability than the Oswestry, and thus reaching maximum before the Oswestry. This means that the Roland seems more sensitive than the Oswestry in detecting changes when patients have minor disabilities, but seems less sensitive when there are severe disabilities.[13] Co et al[14] found a moderate linear relationship between the two scores obtained from the Roland and Oswestry questionnaires. The standardized mean values of the Roland scores were significantly higher than the Oswestry scores. For the assessment of disability in low back pain patients with different degrees of disability, it may be useful to employ both the Roland and Oswestry questionnaires because they complement each other.[14]

The emphasis of the questions in the Roland, Oswestry, and Waddell questionnaires is mainly about the loss of function due to low back pain: the extent to which a person's functional level is restricted by the back pain. In contrast, the emphasis of the questions of the Million is mainly about the influence of activities on the level of pain. This difference has consequences for the interpretation of the results of the questionnaires.

Responsiveness

Theoretical background

Responsiveness is the ability of an instrument to detect small but clinically important changes.[42,50] It concerns the power of the index to detect a difference when one is present.[12] The issue of responsiveness depends on changeability and degree of variability.[30,54] Changeability refers to the extent to which subjects' scores increase or decrease with change in the underlying state: the "signal". The degree of variability that cannot be attributed to true change is called the "noise". The signal is the difference one wishes to detect (if it is truly present). The noise is the measurement error over which the differences must be detected.[54]

A number of strategies have been suggested for quantifying responsiveness of evaluative questionnaires. First, a common strategy for demonstrating responsiveness is to administer a scale before and after some intervention that is

expected to cause substantial clinical improvement. Ideally, an intervention of known efficacy is studied in a randomized clinical trial with a placebo control group. If scale scores improve, the inference is made that the scale is responsive.[7,10,12] The resulting t-statistic could be compared with simultaneous measures of competing instruments or other outcome indicators for the target condition.

When a relatively new outcome measure is used in a situation where the effectiveness of the treatment is unknown, it is difficult to determine whether the findings, either positive or negative, are due primarily to the properties of the treatment or the properties of the measurement approach.[55] If an outcome measure is used in a trial with known efficacy, then this will provide information about the true properties of the new outcome measure, because the approximate effect of an intervention is already known. Unfortunately, an intervention of known efficacy is not always available. As a compromise it is possible to study the natural course of acute low back pain. Patients with acute back pain tend to improve spontaneously with time.[56] Unfortunately, this method does not account well for the score variability that may occur in apparently stable subjects.[10] Systematic score changes could occur because of learning effects, for example, which may vary from instrument to instrument.[10]

A second strategy is studying the correlations between scale change with changes in other measures. If functional scores change over time, and those changes correlate with changes in other measures, this suggests that the functional status questionnaires have some degree of responsiveness.[7,12,57] It may be more appropriate to consider such correlations as a form of longitudinal construct validity, rather than as direct measures of responsiveness.

A third strategy is described by Deyo and Centor.[50] They have drawn an analogy between health status assessment and diagnostic tests. In some ways, determining scale responsiveness is analogous to assessing a diagnostic test. In this case, the condition to be "diagnosed" is whether or not some clinically important change has occurred. From this perspective, a functional score change could be either a true positive outcome (sensitivity), reflecting real patient change, or a false positive outcome (1-specificity), reflecting random or other non-specific score variability.[50] Like a diagnostic test, a functional scale could be described in terms of its sensitivity and specificity for detecting change as established by some external criterion. The sensitivity and specificity can be calculated for each of several cut-off points in change score.[50] This strategy permits quantification of responsiveness and formal statistical comparisons between scales. Unfortunately, receiver operating characteristic (ROC) analysis requires that the external outcome criterion is dichotomous (e.g. improved versus unimproved) rather than preserving information about degree of improvement or deterioration.[50]

Guyatt et al[42] have proposed a fourth strategy. They presented a statistic to describe responsiveness if the degree of score change which is clinically important is known. The ratio of the clinically important difference to the score variability observed in stable subjects would represent responsiveness.[42] However, for most functional scales, we do not know how much change in score constitutes a clinically important difference. Guyatt et al argued that an instrument's responsiveness could be initially estimated by the alternative of comparing within person standard deviation to the change in score observed after an intervention of known efficacy.[42] The best estimate of clinically important difference may require empirical evidence

from a variety of sources, including clinical trials.[32] In a clinical trial, the placebo group could provide the denominator for Guyatt's responsiveness statistic.[32]

The four strategies depend on some external criterion for judging improvement. Because no gold standard for functional status exists, no single observation can prove critical or definitive responsiveness. One may compare scales against several external criteria in one study population. If results are consistent on the basis of several external criteria, confidence increases about the correct ranking of the several scales.[50]

Determination of differences between treatment groups not only depends on the effect measures used but also on the power and design of the study. For example, in small populations it is very difficult to determine a small but important clinical change.

Results responsiveness

Table 4[37,40,47,59,61,63,65] lists the studies which have used one of the four functional status questionnaires as an effect measure and have given information about responsiveness. We divided the studies in four categories which correspond with the strategies used for assessing responsiveness.

The Oswestry, Waddell, and Roland have been used in trials with known efficacy.[18,20,50,56] It turned out that these three questionnaires could discriminate between clinical success and failure. Fairbank et al,[18] for example, studied the expected improvement in a group of 25 patients with first episodes of low back pain.

The second strategy, correlation of scale changes with changes in other measures, is more often used than the other strategies. The improvement in functional status scores does not always correlate with changes in other measures. Million et al[9,38] and Ongley et al[49] have found changes in functional status scores but no changes in objective measures. Two other studies have shown no differences between the treatment groups on all effect measures.[60,64] However, Deyo and Centor,[50] Hazard et al,[62] and Rainville et al,[19] showed an improvement on functional status questionnaires and objective measures using the Roland, Oswestry, and Million, respectively.

The receiver operating characteristic curves have only been used for the Roland.[50] The Guyatt statistic has not yet been used. For two studies it was not possible to fit them in with one of the four strategies. These studies only used the questionnaire as an effect measure and the intervention investigated was of unknown efficacy.[66,67]

Table 4 shows that the Roland and Oswestry questionnaires are most frequently used in studies. If an instrument is responsive on several strategies, the chance that an instrument is able to detect important clinical changes becomes higher. Three strategies for assessing responsiveness have been used for the Roland questionnaire and two have been used for the Oswestry questionnaire. It may be concluded that there is more evidence for the responsiveness of the Roland and Oswestry questionnaires than for the Waddell and Million. The Million questionnaire has been used in only four studies, and only one strategy was used to assess responsiveness.

Table 4. Responsiveness results of four functional status questionnaires for low back pain.

Used in a trial with known efficacy

Oswestry Disability Questionnaire

- Non-RCT: expected improvement observed in a group of 25 patients with first episode of low back pain. The score after 3 weeks was significantly better than on admission (t-test, $p < 0.05$). Disability declined on average about 28% during the recovery period.[18]

Roland Disability Questionnaire

- Non-RCT: 40 patients were examined to see whether the scores on the RDQ and pain rating were significantly related to poor outcome after 4 weeks. RDQ seemed to be a more discriminating indicator of outcome than a six-point pain rating scale.[56]
- RCT: retrospectively assessed in a trial of bed rest for LBP. Mean score changes for RDQ for three samples: the entire patient sample (t=6.35); those for whom patient and clinician agreed that pain had improved (t=6.70); and patients who stated that they had resumed their full activities (t=6.96). In each subgroup the RDQ appeared less responsive than the SIP physical dimension, and in some cases less responsive than the overall SIP.[50]

Waddell Disability Index

- Non-RCT: 30 patients (acute attack of low back pain), before and 1 month after conservative or surgical treatment. The assessments of Waddell disability and physical impairment discriminated very well between clinical success and failure.[20]

Correlation of scale change with changes in other measures

Oswestry Disability Index

- RCT: chiropractic versus shortwave diathermy. Outcome measures: Oswestry, SLR, lumbar flexion. Manipulation not better after 6 weeks but significantly better Oswestry scores after 6 months and 2 years. The change in SLR and flexion was greater in those treated by chiropractor.[58]
- Case-study: chiropractic distraction and manipulation and exercises. Results: Oswestry and RDQ improvement after 1 week, a 7% reduction in the size of the disk protrusion, complete relief of sciatica.[59]
- Non-RCT: population of patients, treatment by chiropractor. Six weeks after intake: means for initial Oswestry and pain VAS scores and 6 weeks later statistically significantly different.[37]
- Non-RCT: multi disciplinary rehabilitation program. No significant differences between the groups on the Oswestry and McGill pain questionnaire.[60]
- RCT: autotraction versus passive traction. Non-responders were crossed-over to the other modality. The favorable response to AT was 75% versus 22% to PT. After 3 months 63% of the responders to AT reported continued improvement. In these patients, pain ratings remained stable and the disability (Oswestry) scores decreased to 0 to 23% of the pretreatment value.[61]
- Non-RCT: functional restoration with behavioral support. Significant improvement in Oswestry and Million score, pain, depression (after 3 weeks and 1 year) and physical capacities (3 weeks).[62]

Million Visual Analog Scale

- RCT: trial of corset with or without rigid insert. Corset with the spinal support provided a significantly decrease of Million score than corset without the spinal support. No differences in the objective changes of spinal movement and straight leg raising between the two groups.[9,36]
- Non-RCT: functional orientated treatment program compared drop-out subjects and those who completed program. The PAIRS, Million scale and physical performance tests improved significantly during treatment.[45]
- Non-RCT: functional restoration with behavioral support. Significant improvement in Oswestry and Million score, pain, depression (after 3 weeks and 1 year) and physical capacities (3 weeks).[62]
- Non-RCT: work tolerance program. Large improvement on return to work and spinal range of motion, small improvement on Million score.[63]

Roland Disability Questionnaire

- RCT: lumbar traction versus sham traction. Outcome measures: global perceived recovery, RDQ, severity of pain, physical impairment. Except for physical impairment and functional status at 5 weeks and pain during effect measurement at 9 weeks, there was always a small difference (not significant) between the groups.[64]
- Non-RCT: 52 subjects completed SIP questionnaires 3 months following inpatient treatment for chronic pain. A series of paired t-tests was performed. The sensitivity of the RDQ and other SIP scales were very similar.[40]
- Non-RCT: comparison of two multi modal back treatment programs. Long term follow-up (12 months). The results of the group with a higher amount and intensity of physical exercise were better in the Back pain index, the RDQ, and the subjective state of health.[36]
- RCT: tests efficacy of low-energy laser biostimulation combined with exercises. Subjective data: VAS-pain, RDQ. Results: objective and subjective data: significant improvement in both groups, but no relative advantage was found for either group.[65]
- Case-study: chiropractic distraction and manipulation and exercises. Results: Oswestry and RDQ improvement after 1 week, a 7% reduction in the size of the disk protrusion, complete relief of sciatica.[59]
- RCT: retrospectively assessed in a trial of bed rest for LBP. Correlations between RDQ score changes and changes in self-rated improvement (r=0.41), clinician rated improvement (r=0.30), spine flexion (r=0.29), and resumption of full activities (r=0.38) are significant.[50]

Waddell Disability Index

- RCT: Bourdillon manipulation versus non-forceful manipulation. Mean pain on VAS and RDQ/Waddell all differences significant. Spinal flexion no significant difference.[49]

Receiver operating characteristic curves

Roland Disability Questionnaire

- RCT: retrospectively assessed in a trial of bed rest for LBP. Using "return to full activities" or consensus patient and clinician judgement as the external criterions: the RDQ appeared somewhat better than the SIP and SIP physical dimension and SIP psychosocial dimension.[50]

Guyatt statistic: comparison of scales in a clinical trial

– – –

RCT = Randomized clinical trial, AT = autotraction, PT = passive traction

Research applications

We determined the extent and nature of functional status measurements in randomized clinical trials about low back pain. For this purpose, RCTs used in three blinded reviews about the efficacy of physiotherapy exercises,[68] mobilization,[69] and traction[70] were examined. In these reviews the methodological quality of the trials was assessed. The criteria used were based on generally accepted principles of intervention research.[71,72]

We used these three reviews because an extensive literature research had been conducted by the authors. They put much effort into obtaining all available randomized clinical trials; a total of 53 studies were included in the three reviews.

We evaluated whether functional status was used as an effect parameter in these 53 studies. The extent to which functional status was actually measured was evaluated according to the following criteria. First, no attempt was made to measure functional status. Second, an ad hoc or untested questionnaire was used to measure functional status. Thirdly, at least one questionnaire of demonstrated reproducibility or validity or responsiveness was used.

In 33 (62 %) of the 53 RCTs, functional status was used as an effect measure. In 18 (34%) RCTs, the authors used a questionnaire to measure functional status. In the selected RCTs, many unknown questionnaires have been used. In a number of articles, the effect measures used were inadequately described. It was often unclear what was meant by the effect measure functional status. It was also unclear what the questionnaires used looked like.

Only two well known questionnaires were used. The Roland disability questionnaire[19] was used in 3 studies and the Oswestry pain disability questionnaire[18] was used in one study.

It is remarkable that the four RCTs[49,58,64,67] in which a well known questionnaire was used to measure functional status had high methodological scores in the reviews compared with the other studies.[68-70] Most studies using functional status questionnaires have been published recently. It seems that there is a tendency of using functional status questionnaires more frequently in RCTs.

Discussion and conclusion

The purpose of the four evaluated functional status questionnaires is the same: to measure functions of daily living. Only the degree of limitations of functional status investigated differs among the questionnaires. Furthermore, the number of questions and the response possibilities of the questions are different. The emphasis of the Oswestry, Roland and Waddell questionnaires is on the loss of function due to low back pain, whereas the focus of the Million questionnaire is mainly the influence of activities on the level of pain.

Table 5 presents a summary of our results. All four questionnaires have negative and positive aspects. To start, there is not enough information available about the population of low back pain patients and the criteria of item selection used for the development of the questionnaires. These procedures determine the use of a questionnaire. Second, the time frame of the Oswestry, Million, and Waddell is left undefined. Patients may be confused regarding which period to report. In addition, the response possibilities of the Waddell and Roland questionnaires are only 'yes' or

'no'. This may have consequences for the responsiveness of these questionnaires. The test-retest reproducibility of the questionnaires seems satisfactory.

Table 5. Summary of the results of four evaluative functional status questionnaires for low back pain.

	Oswestry Disability Quesionnaire	Million Visual Analog Scale	Roland Disability Quesionnaire	Waddell Disability Index
Item selection	+/-	-	+/-	+/-
Number of items	+	+	+	+
Time specific	-	-	+	-
Response possibilities	+	+	-	-
Reliability	+	+	+	+
Validity	+	+/-	+	+/-
Responsiveness	+	+/-	+	+/-
Research applications	+	-	+	-

++ very good; + good; +/- doubtful; - negative.

The associations between functional status on the one hand and physical, psychosocial, and pain measurements on the other hand are often weak, suggesting that functional status should be assessed in its own right. It is reasonable to suppose that functional status questionnaires cover a broad field of aspects, including pain, physiological, and psychosocial aspects. However, they also measure other relevant dimensions belonging to the field of functional status.

As discussed in the section on validity, Baker et al[13] demonstrated that the Roland tends to score higher in the lower ranges of disability than the Oswestry, and reached a maximum before the Oswestry. As a consequence, it is likely that the Roland is more sensitive than the Oswestry in detecting changes when patients have minor disabilities, but is less sensitive when they have severe self-rated disability. A broader range of response categories probably would increase the responsiveness of the Roland further.

Tables 3 and 4 indicate that the Oswestry and Roland disability questionnaires have been used and evaluated more frequently than the Million and Waddell. As a consequence, we can be more certain about the validity and responsiveness of the Oswestry and Roland questionnaires. It is difficult to determine the validity and responsiveness of questionnaires which are not frequently used, such as the Million and Waddell questionnaires.

When a relatively new outcome measure is used in randomized clinical trials or other longitudinal studies it makes sense to use a number of outcome measures. This provides a direct assessment of the questionnaire's performance and also a comparison with other measures. There is more evidence for the effect of an intervention if the effect has been shown on several outcome measures and if there is a high correlation among the outcome measures.

In designing a trial, researchers should decide whether functional status is a relevant outcome to assess. If improvement of the function of the patients is a major purpose of the treatment, functional status should be used as an outcome measure. Ideally, the choices among the available questionnaires are made upon information about the instruments' methodological quality.[11,12,22,73] Unfortunately, there are still important gaps in the evaluation of functional status questionnaires. The main reason for this is that no gold standard exists for functional status. Therefore, we will

never be able to prove the validity and responsiveness of functional status questionnaires. In the absence of a gold standard, direct comparisons of functional status questionnaires are needed. They must be judged against the same validity and responsiveness criteria in the same patient group. Through direct comparisons we are able to assess comparative validity and responsiveness. On the other hand, a valid and responsive questionnaire may be sound for one population but not for another. As a consequence, it is also important to evaluate the questionnaires in different patient groups, for example, in patients with chronic and acute low back pain.

If no appropriate questionnaire is available for a study, researchers can decide to develop a new questionnaire. The use of untested, ad hoc measures may generate uncertainty about the methodological quality and therefore about the relevance of the outcome.[33] It is very expensive and time-consuming to develop a new questionnaire. New questionnaires should not be generated unless a careful literature review points out that a new one is needed.[10,11,]

Back pain tends to improve spontaneously with time and, therefore, the difference in improvement between a treatment group and a control group may be relatively small. If an insensitive outcome measure is used, the true effect of a treatment may not be detected.[56] The functional status questionnaires discussed probably offer good responsiveness to clinical changes over time. Functional status measures are not currently used in many settings in which they would be valuable. Therefore, it is important to encourage their wider use in clinical trials. Additional research is needed to compare and improve the existing questionnaires.

References

1. Haanen HCM. An epidemiologic survey on low back pain. Dissertation, Erasumus University, Rotterdam, 1984.
2. Kelsey JL. Epidemiology of musculoskeletal disorders. New York: Oxford University Press, 1982: 145-167.
3. Kelsey JL, White AA. Epidemiology and impact of low back pain. Spine 1980; 5: 133-142.
4. Deyo RA. Measuring functional outcomes in therapeutic trials for chronic disease. Controlled Clin Trials 1984; 5: 223-240.
5. Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: Design, analysis, and interpretation. BMJ 1992; 305: 1145-1148.
6. Bouter LM, Linden Sj van der, Koes B. Effectmeting in de fysiotherapie. Nederlands Tijdschrift voor Fysiotherapie 1991; 101: 46-8.
7. Deyo RA. Measuring the functional status of patients with low back pain. Arch Phys Med Rehabil 1988; 69: 1044-1053.
8. Hudson-Cook N, Tomes-Nicholson K, Breen A. A revised Oswestry disability questionnaire. In: Roland OM, Jenner JR (eds). Backpain new approaches to rehabilitation and education. Manchester: University Press, 1989: 187-204.
9. Million R, Hall W, Nilsen KH, Baker RD, Jayson MI. Assessment of the progress of the back pain patient. Spine 1982; 7: 204-212.
10. Deyo RA. The quality of life, research, and care. Ann Intern Med 1991; 114: 695-697.
11. Bombardier C, Tugwell P. Methodological considerations in functional assessment. J Rheumatol 1987; 15(Suppl): 6-10.
12. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. J Chron Dis 1985; 38: 27-36.
13. Baker JD, Pynsent PB, Fairbank JCT. The Oswestry disability index revisited: its reliability, repeatability and validity, and a comparison with the St Thomas's disability index. In: Roland MO, Jenner JR (eds). Backpain new approaches to rehabilitation and education. Manchester: University Press, 1989: 174-186.

14. Co YY, Eaton S, Maxwell MW. The relationship between the St. Thomas and Oswestry disability scores and the severity of low back pain. J Manipulative Physiol Ther 1993; 16: 14-18.
15. Grönblad M, Hupli M, Wennerstrand P, Järvinen E, Lukinmaa A, Kouri JP, Karaharju EO. Intercorrelation and test-retest reliability of the pain disability index (PDI) and the Oswestry disability questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. Clin J Pain 1993; 9: 189-195.
16. Hsieh CJ, Phillips RB, Adams AH, Pope MH. Functional outcomes of low back pain: Comparison of four treatment groups in a randomized controlled trial. J Manipulative Physiol Ther 1992; 15: 4-9.
17. Triano JJ, Schultz AB. Correlation of objective measure of trunk motion and muscle function with low-back disability ratings. Spine 1987; 12: 561-565.
18. Fairbank JCT, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66: 271-273.
19. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983; 8: 141-144.
20. Waddell G, Main CJ. Assessment of severity in low-back disorders. Spine 1984; 9: 204-208.
21. Aaronson NK. Quality of life assessment in clinical trials: methodologic issues. Controlled Clin Trials 1989; 10(Suppl): 195-208.
22. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. New York, Oxford: Oxford University Press, 1987.
23. Meade TW, et al. (Report of a working group) Comparison of chiropractic and hospital outpatient management of low back pain: a feasibility study. J Epidemiol Community Health 1986; 40: 12-17.
24. Bergner M, Bobbitt RA, Carter WB, Gilson BS. Sickness Impact Profile: development and final revision of health status measure. Med Care 1981; 19: 787-805.
25. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995; 20: 1899-1909.
26. Wing PC, Wilfling FJ, Kokan PJ. Psychological demographic and orthopaedic factors associated with prediction of outcome of spinal fusion. Clin Orthop 1973; 90: 153-160.
27. Guyatt GH, Bombardier C, Tugwell P. Measuring disease-specific quality of life in clinical trials CMAJ 1986; 134: 889-895.
28. Greenough CG, Fraser RD. Assessment of outcome in patients with low-back pain. Spine 1992; 17: 36-41.
29. Bergner M, Rothman ML. Health status measures: an overview and guide for selection. Ann Rev Public Health 1987; 8: 191-210.
30. Guyatt GH, Jaeschke R. Measurement in clinical trials: choosing the appropriate approach. In: Spilker B (eds). Quality of life assessment in clinical trials. New York: Raven Press Ltd, 1990: 37-46.
31. Kramer MS, Feinstein AR. Clinical biostatistics LIV. The biostatistics of concordance. Clinical Pharmacol Ther 1981; 29: 111-123.
32. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clin Trials 1991; 12(Suppl): 142-158.
33. Guyatt GH, Veldhuyzen Van Zanten SJO, Feeny DH, Patrick DL. Measuring quality of life in clinical trials: a taxonomy and review. CMAJ 1989; 140: 1441-1448.
34. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J Clin Epidemiol 1992; 45: 1201-1218.
35. Waddell G, Main CJ, Morris EW, Venner RM, Rae PS, Sharmy SH, Galloway H: Normality and reliability in clinical assessment of backache. BMJ 1982; 284: 1519-1523.
36. Järvikoski A, Mellin G, Estlander A, Härkäpää, Vanharanta H, Hupli Markku, Heinonen R. Outcome of two multi modal back treatment programs with and without intensive physical training. J Spinal Dis 1993; 6: 93-98.
37. Triano JJ, McGregor M, Cramer GD, Emde D. A comparison of outcome measures for use with back pain patients: results of a feasibility study. J Manipulative Physiol Ther 1993; 16: 67-73.
38. Million R, Nilsen KH, Jayson MIV, Baker RD. Evaluation of low back pain and assessment of lumbar corsets with and without back supports. Ann Rheum Dis 1981; 40: 449-454.
39. Deyo RA. Comparative validity of the sickness impact profile and shorter scales for functional assessment in low-back pain. Spine 1986; 11: 951-954.
40. Jensen MP, Strom SE, Turner J, Romano JM. Validity of the sickness impact profile Roland scale as a measure of dysfunction in chronic pain patients. Pain 1992; 50: 157-162.

41. Waddell G, Main CJ, Morris EW, Paola MD, Gray ICM. Chronic low-back pain, psychologic distress, and illness behaviour. Spine 1984; 9: 209-213.
42. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40: 171-178.
43. Bombardier C, Tugwell P. A methodological framework to develop and select indices for clinical trials: statistical and judgement approaches. J Rheumatol 1982; 9: 753-757.
44. Alaranta H, Tallroth K, Soukka A, Heliövaara M. Fat content of lumbar extensor muscles and low back disability: A radiographic and clinical comparison. J Spinal Dis 1993; 6: 137-140.
45. Rainville J, Ahern DK, Phalen L. Altering beliefs about pain and impairment in a functionally oriented treatment program for chronic low back pain. Clin J Pain 1993; 9: 196-201.
46. Millard RW, Jones RH. Construct validity of practical questionnaires for assessing disability of low-back pain. Spine 1991; 16: 835-838.
47. Mellin G, Härkäpää K, Vanharanta H, Hupli M, Heinonen R, Järvikoski A. Outcome of a multi modal treatment including intensive physical training of patients with chronic low back pain. Spine 1993; 18: 825-829.
48. Waddell G, Somerville D, Henderson I, Newton M. Objective clinical evaluation of physical impairment in chronic low back pain. Spine 1992; 17: 617-628.
49. Ongley MJ, Klein RG, Dorman ThA, Eek BC, Hubert LJ. A new approach to the treatment of chronic low back pain. Lancet 1987; 18: 143-146.
50. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. J Chron Dis 1986; 39: 897-906.
51. Waddell G. Clinical assessment of lumbar impairment. Clin Orthop 1987; 221: 110-120.
52. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain 1975; 1: 277-299.
53. Zung WWK. A self-rating depression scale in an outpatient clinic. Arch Gen Psychiatry 1965; 13: 508-515.
54. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992; 45: 1341-1345.
55. Meenan RF, Anderson JJ, Kazis LE, Egger MJ, Altz-Smith M, Samuelson CO, Willkens RF, Solsky MA, Hayes SP, Blocka KL, Weinstein A, Guttadauria M, Kaplan B, Klippel J. Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. Arthritis Rheum 1984; 27: 1344-1352.
56. Roland M, Morris R. A study of the natural history of back pain. Part II: Development of guidelines for trials of treatment in primary care. Spine 1983; 8: 145-150.
57. Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. Health Serv Res 1984; 19: 275-289.
58. Meade TW, Dyer S, Browne W, Townsend J, Frank AO. Low back pain of mechanical origin: randomised comparison of chiropractic and hospital outpatient treatment. BMJ 1990; 300: 1431-1436.
59. Cox JM, Hazen LJ, Mungovan M. Distraction manipulation reduction of an L5-S1 disk herniation. J Manipulative Physiol Ther 1993; 16: 342-346.
60. Cassisi JE, Sypert GW, Salamon A, Kapel L. Independent evaluation of a multi disciplinary rehabilitation program for chronic low back pain. Neurosurgery 1989; 25: 877-83.
61. Tesio L, Merlo A. Autotraction versus passive traction: An open controlled study in lumbar disc herniation. Arch Phys Med Rehabil 1993; 74: 871-876.
62. Hazard RG, Fenwick JW, Kalisch SM, Reddmond J, Reeves V, Reid S, Frymoyer JW. Functional restoration with behavioral support. A one-year prospective study of patients with chronic low-back pain. Spine 1989; 14: 157-161.
63. Sachs BL, David JAF, Olimpio D, Scala AD, Lacroix M. Spine rehabilitation by work tolerance based on objective physical capacity assessment of dysfunction. A prospective study with control subjects and twelve-month review. Spine 1990; 15: 1325-1332.
64. Heijden GJMG van der, Beurskens AJHM, Dirx MJM, Terpstra-Lindeman E. The efficacy of lumbar traction: A randomized clinical trial. Design and results of a pilot study. Physiotherapy 1995; 81: 29-35.
65. Klein RG, Eek BC. Low-energy laser treatment and exercise for chronic low back pain: Double-blind controlled trial. Arch Phys Med Rehabil 1990; 71: 34-37.

66. Delitto A, Cibulka MT, Erhard RE, Bowling RW, Thehuia JA. Evidence for use of an extension-mobilization category in acute low back syndrome: A prescriptive validation pilot study. Phys Ther 1993; 73: 216-222.
67. Hadler NM, Curtis P, Gillings DB, Stinnett S. A benefit of spinal manipulation as adjunctive therapy for acute low-back pain: A stratified controlled trial. Spine 1987; 12: 703-705.
68. Koes BW, Bouter LM, Beckerman H, Heijden van der GJMG, Knipschild PG. Physiotherapy exercises and back pain. A blinded review. BMJ 1991; 302: 1572-1576.
69. Koes BW, Assendelft WJJ, Heijden van der GJMG, Bouter LM, Knipschild PG. Spinal manipulation and mobilization for back and neck pain. A blinded review. BMJ 1991; 303: 1298-1303.
70. Heijden GJMG van der, Beurskens AJHM, Koes BW, Assendelft WJJ, Vet HCW, Bouter LM. Traction for back and neck pain: a blinded review. Physical Therapy 1995; 75: 93-104.
71. Feinstein AR. Clinical Epidemiology: the architecture of clinical research. Philadelphia: WB Saunders, 1985.
72. Meinert CL. Clinical trials: design, conduct and analysis. New York: Oxford University Press, 1986.
73. Feinstein AR. Clinical biostatistics XLI. Hard science, soft data, and the challenge of choosing clinical variables in research. Clin Pharmacol Ther 1977; 22: 485-498.

86. Deyo A, Diehl AK, Ehrhart PE, Bowles BW, Thelen
mobilization category in acute low back syndrome. A
1991; 73: 814-822.

87. Hadler NM, Curtis P, Gillings DB, Stinnett S. A benefit of spinal manipulation as adjunctive therapy for acute low back pain: A stochastic controlled trial. Spine 1987; 12: 703-708.

88. Koes BW, Bouter LM, Beckerman H, Headley van der
exercises and back pain. A blinded review. BMJ 1991;

89. Koes BW, Assendelft WW, Heijden van der G, Bouter
manipulation and mobilization for back and neck pain.

90. Heijden GJMG van der, Beurskens AJHM, Koes BW, Assendelft WJJ, Hondeghem LM, Bouter LM. The efficacy of traction for back and neck pain: a blinded review. Physiotherapy 1995; 81: 67-71.

91. Fordyce WE. Behavioral methods for chronic pain and illness. St Louis: Mosby, 1976.

92. Meenan CL. Clinical trials: design, conduct and analysis. New York: Oxford University Press, 1986.

93. Feinstein AR. Clinical biostatistics XLI. Hard science, soft data, and the challenge of choosing multivariable models. Clin Pharmacol Ther 1981; 30: 555-574.

Chapter 6

Responsiveness of functional status in low back pain

A comparison of different instruments

AJHM Beurskens,[1] HCW de Vet,[1] AJA Köke.[2]

1 Department of Epidemiology, University of Limburg, Maastricht
2 Department of Physiotherapy, University Hospital, Maastricht

Pain (in press).

Abstract

This study compares the responsiveness of three instruments of functional status: two disease-specific questionnaires (Oswestry and Roland disability questionnaires), and a patient-specific method (severity of the main complaint). We compared changes over time of functional status instruments with pain rated on a visual analog scale. Two strategies for evaluating the responsiveness in terms of sensitivity to change and specificity to change were used: effect size statistics and receiver operating characteristic method. We choose global perceived effect as external criterion.

A cohort of 81 patients with non-specific low back pain for at least 6 weeks assessed these measures before and after 5 weeks of treatment. According to the external criterion 38 patients improved. The results of both strategies were the same. All instruments were able to discriminate between improvement and non-improvement. The effect size statistics of the instruments were higher in the improved group than in the non-improved group. For each instrument the receiver operating characteristic curves showed some discriminative ability. The curves for the Roland questionnaire and pain were closer to the upper left than the curves for the other instruments. The sensitivity to change of the rating of Oswestry questionnaire was lower than that of the other instruments. The main complaint was not very specific to change. The two strategies for evaluating the responsiveness were very useful and appeared to complement each other.

Introduction

Low back pain has an impact on patients' functioning. The aim of treatment is often improving the function of patients.[1-3] Therefore, functional status is an important outcome in trials on low back pain.

Restriction of functional status is a patient-referenced concept and different for each individual. Some patients make higher demands on functions than others. However, there are certain categories of activities that are performed by everyone, such as sitting and sleeping. Questionnaires have been developed for measuring functional status specifically in low back pain. The Oswestry[4] and Roland[5] low back disability questionnaires have been widely used for assessing functional status in low back pain.[1,2,6-8] The number of items and response categories differ, but the purpose of both questionnaires is in general the same.

Although many activities of daily living are important to everyone, the degree of importance varies widely between patients. So, it seems sensible, and it may well reveal a greater responsiveness, when we only focus on activities that an individual patient experiences as most disturbing.[9-11] These activities differ from patient to patient. Guyatt et al[10] suggest asking each patient which activities are difficult to perform and examining the effect of the intervention on performance of those activities. In this way, we tailor outcome measures to the individual patient: each patient has his or her specific treatment goal. This approach is being used increasingly to measure functional status and it is responsive.[10,11]

This paper deals with the ability of instruments for functional status to detect change over time, often called responsiveness. In this article we define responsiveness as the ability of an instrument to detect clinically important changes.[10,12,13] Responsiveness means that an instrument can discriminate between clinically important and clinically unimportant changes. This is analogous to assessing the discriminating properties of a diagnostic test.[12] The condition to be diagnosed is whether a clinically important change has occurred. One may describe the responsiveness of an instrument in terms of sensitivity to change (true positive) and specificity to change (true negative) in detecting improvement or non-improvement as established by external criteria.

The purpose of this article is to compare the responsiveness of three instruments for evaluating the functional status and the severity of pain. The functional status instruments are the Oswestry[4] and Roland[5] disability questionnaires (two disease-specific approaches) and the severity of the main complaint of the patient (patient-specific approach). Reduction of pain is often a treatment goal and pain is a frequently used effect measure in research. In addition, patients often report pain as a major complaint. Therefore, change over time of the pain rated on a visual analog scale (VAS) will be compared with the changes on the functional status instruments.

For the comparison of responsiveness of instruments, an external criterion is necessary to distinguish between improved and non-improved patients. In this study, an estimate of a clinically important change was obtained by global perceived effect assessed by the patient on a seven-point scale (1=completely recovered, 7=vastly worsened). We defined completely recovered and much improved as a clinically important change. From both patients' and clinicians' viewpoint it is relevant and sensible to ask the patient to assess his perceived benefit.[14,15]

There is no consensus on the most appropriate strategy for quantifying responsiveness. In this study responsiveness has been operationalized using two strategies: effect size statistics and Receiver Operating Characteristic (ROC) method. Apart from the main question regarding which instrument is most responsive, we are interested in the level of agreement between these two strategies for the four instruments.

Effect size statistics relate the magnitude of the change to the variability in score.[16] The statistic is the relation between "signal" and "noise". The effect size statistic is calculated by taking the mean change found in a variable (the signal) and dividing it by the standard deviation of that variable (the noise).[17] However, controversy exists on which standard deviation to take. Kazis et al[17] took the standard deviation of baseline scores. Guyatt et al[10] used the standard deviation of the mean score change in stable patients. They relate clinically meaningful score changes on a functional status instrument to the variability of scores in stable subjects. A potential disadvantage of this method is that the numerator and denominator were based on different samples.[18] We think that a measure of change is not a function of the standard deviation of stable patients or of baseline scores but a function of the standard deviation of the score changes. Thus, the standard deviation of the mean score change of the same group should be used as suggested by Cohen.[16]

Deyo and Centor[12] described the second strategy. They have drawn an analogy between health status assessment and diagnostic tests. In some ways, determining instrument responsiveness is analogous to evaluating a diagnostic test, in which the functional status instrument is the diagnostic test and the gold standard is global perceived effect. Like a diagnostic test, then, an instrument could be described in terms of its sensitivity and specificity for detecting change as established by some gold standard. The ROC curve is a graph of 'true positive' (sensitivity) versus 'false positive' (1-specificity) for each of several cut-off points in score change.[12] The area under the ROC curve can be interpreted as the probability of correctly discriminating between improved and non-improved patients. This area theoretically ranges from 0.5 (no accuracy in discriminating improved from non-improved) to 1.0 (perfect accuracy). The ROC curve can provide an indication of what score change represents the best cut-off points for making the distinction of improved or non-improved patients.[19]

Methods

Patients

Patients were participants in a randomized controlled trial on the efficacy of continuous motorized traction for low back pain. We selected patients if they had suffered for at least 6 weeks from non-specific low back pain. The details of study design of the randomized controlled trial have been published elsewhere.[20]

The Roland disability questionnaire, main complaint, pain and global perceived effect were assessed for all 151 patients. We asked a selection of the patients (the last 81) to fill out the Oswestry questionnaire at baseline and 5 weeks later, after the treatment period. This report is limited to this subgroup.

Outcome measures

The Oswestry questionnaire has 10 sections that refer to activities of daily living that might be disrupted by low back pain.[4] The Roland disability questionnaire (also named the RDQ or the St. Thomas questionnaire) has been derived by choosing 24 yes/no items relevant for low back pain from the Sickness Impact Profile (SIP).[5] At baseline patients selected their three main complaints in a standardized way: they selected three activities they performed frequently, which they perceived as important in their day-to-day life, and which low back pain made difficult for them. Patients rated the severity of the three main complaints on a 100 mm. VAS. In this study only the first main complaint was used. Patients evaluated the average severity of pain during the last week on a 100 mm. VAS.

Global perceived effect was measured by self-assessment on a seven-point scale (1=completely recovered, 7=vastly worsened). We asked the patients in what way their low back pain changed during the last 5 weeks. If a patient indicated completely recovery or much improvement, we coded the patient as improved; we coded slightly improved, no change, slightly worsened as non-improved. Patients who assessed themselves as much worsened and vastly worsened were interpreted as deteriorated. We excluded these patients from the analysis.

Analysis

The data were analyzed using the Statistical Package for Social Sciences (SPSS/PC+). The scores of the Oswestry, pain, and severity of the main complaints range from 0-100, the sum score of the Roland ranges from 0-24. For a better comparison of the scores we also present a standardized Roland score of 0-100, by multiplying each score by 100/24.

To compare the absolute score levels, means and standard deviations were calculated the baseline and 5-week scores. We calculated differences between the measurements by subtracting the score after 5 weeks from the baseline score. A positive score difference thus indicates an improvement, a negative difference a deterioration. Effect size statistics, based on within patients changes, are presented for evaluating the average change. For different score changes of each instrument a ROC curve is presented by plotting the true-positive rate (sensitivity) against false-positive rate (1-specificity).

Results

Scores on the instruments

Of all 81 participants we had complete data on the pre-treatment and post-treatment measurements. Table 1 provides information on characteristics of the participants. By self-assessment of global perceived effect on a seven-point scale 6 patients (7%) rated themselves completely recovered and 32 patients (40%) as much improved (table 2). There were only 5 patients who indicated deterioration (much worsened or vastly worsened). Therefore, we did not calculate the effect size of patients who deteriorated and made no ROC curve for detecting deterioration. The 5 patients were excluded from further analysis.

Table 1. Characteristics of patients (n=81)

Mean age (yrs,sd)	41 ± 10
Gender (% female)	37 (46%)
Present episode	
median duration (wks)	24
mean duration (wks, sd)	70 ± 119
chronic (> 6 months)	47 (58%)
sub-acute (6 weeks <> 6 months)	34 (42%)
Previous low back pain	68 (84%)
Number of low back pain periods ever	
median	10
mean (sd)	15 ± 19
Radiating into lower leg	28 (35%)

Table 2. Patients' assessment of global perceived effect (n=81).

Answer	Number	(%)
completely recovered	6	(7)
much improved	32	(40)
slightly improved	19	(24)
no change	15	(19)
slightly worsened	4	(5)
much worsened	4	(5)
vastly worsened	1	(1)

Table 3 shows the mean scores and the standard deviations of the improved and non-improved patients at baseline and after 5 weeks. The functional status instruments register different mean scores at baseline; they range from 26.2 on the Oswestry to 71.4 on the main complaint. The baseline scores of the improved and non-improved group are comparable.

Table 3. Mean scores (SD) of the improved and non-improved patients and the total group.

Instruments	Baseline score		5-week score	
	improved n=38	non-improved n=38	improved n=38	non-improved n=38
Oswestry	26.2 (13.5)	29.1 (15.2)	14.3 (15.1)	29.5 (17.4)
Roland (0-24)	12.1 (4.7)	11.8 (5.1)	4.3 (4.2)	10.6 (5.5)
Roland std.*	50.4 (19.4)	49.3 (21.3)	17.9 (17.6)	44.2 (22.8)
Main complaint	71.4 (17.2)	69.6 (16.5)	30.1 (24.3)	56.0 (25.1)
Pain last week	55.7 (20.2)	57.6 (20.0)	19.1 (17.9)	55.7 (21.5)

* Roland score has been standardized to a scale of 0-100.

Effect size statistics

The mean score changes of the instruments differed between the patients in the improved and non-improved group (table 4). As was expected, the effect size statistics in the improved group are much higher than those in the non-improved group. The effect size statistic of the Roland questionnaire was the highest (2.02) and the effect size of the Oswestry questionnaire the lowest (0.80). The effect size statistics in the non-improved group were mostly small or even negative. The only exception was that of the main complaint, which was 0.73. This means that some patients who had not improved according to their global perceived effect changed on their main complaint.

Table 4. Mean changes, standard deviations (SD) and effect size statistics in the improved (n=38) and non-improved (n=38) patients.

Instruments	Mean changes	SD	Effect size*
Oswestry			
improved	11.9	14.9	0.80
non-improved	- 0.4	9.2	-0.04
Roland (0-24)			
improved	7.8	3.9	2.02
non-improved	1.2	3.0	0.41
Roland std.**			
improved	32.6	16.1	2.02
non-improved	5.1	12.5	0.41
Main complaint			
improved	41.3	25.4	1.63
non-improved	13.6	18.5	0.73
Pain last week			
improved	36.6	23.1	1.58
non-improved	1.9	14.5	0.13

* Effect size is calculated as mean change score divided by the standard deviation of the mean change score.

** Roland score has been standardized to a scale of 0-100.

The differences between the effect size statistics in the improved and non-improved group allow the conclusion that all instruments can discriminate between improved and non-improved patients. If patients indicated that they improved according to the external criterion, the amount of improvement on the Oswestry was low compared with improvement on the other instruments: the sensitivity to change was low. But in the non-improved group the effect size statistic was the lowest for the Oswestry. Consequently we can say that the Oswestry was the most specific to change. For the main complaint the gain in functional status in the non-improved group was still moderate: the specificity to change was lower than for ratings of the other instruments.

ROC method

Figure 1 shows the ROC plot for the four instruments, using global perceived effect as the gold standard. The best discrimination occurs with a curve that reaches most

to the upper left: a instrument whose curve has a sharp initial rise (high true positive rate and low false positive rate), then flattens while the false positive rate is still small.[12] For each instrument the curve was to the left above the diagonal, showing some discriminative ability. The curves for the Roland and pain were closer to the upper left than the curves for the other instruments. The areas under the ROC curves of the Roland and pain were very high; 0.93 and 0.91 respectively (table 5). They showed the best discrimination between improved and non-improved patients.

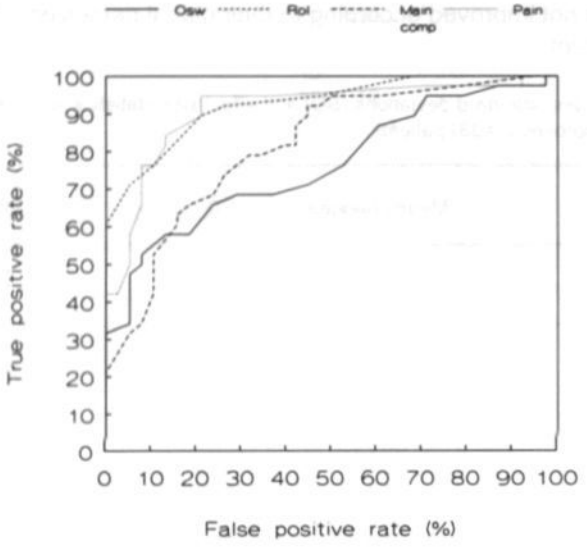


Figure 1. ROC curves of the score changes for the Oswestry, Roland, main complaint and pain.

Table 5. Areas under the ROC curves for each instrument (n=76). Patients who were completely recovered or much improved were coded as improved.

Instruments	ROC area
Oswestry	0.76
Roland	0.93
Main complaint	0.82
Pain last week	0.91

If we use the instruments in clinical practice for individuals, it may be useful to choose cut-off points that best discriminate between improved and non-improved. These are the points on the curve closest to the upper left corner of the plot. The importance of false positive and false negative errors may vary depending on the situation. If we assume that they are equally important, score changes with the best cut-off points in this study were: 4 to 6 points of the 100 points for the Oswestry, 2.5

to 5 points of the 24 points for the Roland, 18 to 24 mm. on a VAS for the main complaint, 10 to 18 mm. on a VAS for the pain.

Less stringent gold standard

We saw that in the non-improved group the effect size statistic was still moderate for the main complaint. Therefore, we performed a second analysis in which we changed our gold standard defining slightly improved also as a clinically relevant improvement. The effect size statistic of the main complaint in the improved group decreased from 1.63 to 1.39 and in the non-improved group decreased from 0.73 to 0.38. The sensitivity to change of the rating of main complaint decreased slightly and the specificity for change increased. For the other instruments the ability to discriminate between improved and non-improved patients was unchanged or decreased (data not presented).

In addition, Figure 2 shows the ROC curves of the main complaint with the two different gold standards. When we used the less stringent gold standard it turned out that for a level of true positive rate between 90% and 70%, the false positive rate decreased substantially: the curve of the main compliant with the changed gold standard was to the left above the other curve.

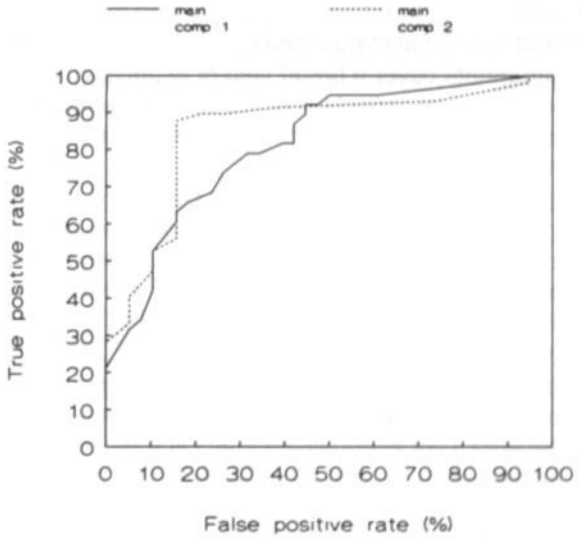


Figure 2. ROC curves of the score changes for the main complaint with two gold standards: Main complaint 1 (completely recovered, much improved), Main complaint 2 (completely recovered, much and slightly improved).

Discussion

We defined responsiveness as the ability of an instrument to detect clinically important changes. An estimate of clinically important change was obtained by global perceived effect. We used effect size statistics and the ROC method to

evaluate responsiveness in terms of sensitivity to change and specificity to change. The results of both strategies were the same. All instruments were able to discriminate between improvement and non-improvement. The Oswestry appeared to be less sensitive to change and the main complaint less specific to change.

Showing that instrument scores improve in the improved group addresses the sensitivity to change of a scale, but not its specificity to change. The concept specificity to change is also important, since changes without clinical relevance may occur in functional scale scores.[12] In this study, the score change for the main complaint was high (mean=41.3 mm.) in the improved group but it was also moderate (mean=13.6 mm.) in the non-improved group. This means that some patients who were not improved according to the gold standard changed significantly on their main complaint. In the second analysis in which we changed our external criterion (patients who improved slightly were also indicated as improved), the discriminative ability of the main complaint increased. It seems that the patient-specific approach is so sensitive to change that patients who only slightly improved on the external criterion registered modest improvement on the main complaint. One can wonder if only slight improvement is clinically relevant.

The results of this study suggest that the responsiveness of the rating of pain during last week is as high as the responsiveness of the rating of Roland questionnaire and better than the rating of Oswestry and main complaint. This does not mean that pain and functional status measure the same aspects of recovery. Functional status instruments cover a broad field of aspects including pain, physiological, and psychosocial aspects. Yet, it seems that patients who evaluated themselves as improved on the global perceived effect measure, evaluated both their pain and function as improved.

In this study we used the original versions of the Oswestry[4] and Roland[5] questionnaires. Patrick et al[21] have developed a slightly modified version of the Roland questionnaire. Hudson-Cook et al[22] designed a revised Oswestry that produced higher mean disability scores. It is possible that the responsiveness of the revised questionnaires is different from the responsiveness of the original version. This has to be assessed in studies comparable to this one.

The strategies used to assess responsiveness depend on some external criterion for judging improvement. We have used the patients' judgement to select the clinically improved patients. Global perceived effect is a measure for improvement that includes pain, functional status and other aspects that patients classify as important. Although it is not a "gold standard" for change, most people would be reluctant to label a patient as improved or worse contrary to this personal assessment. For any method of measuring responsiveness, the choice of external criteria for change is problematic. Since our external criterion cannot be regarded as a "gold standard", more comparisons of functional status measures against several external criteria are needed.

Two strategies, effect size statistics and ROC method, were used to evaluate the responsiveness of four instruments. They provided the same information in different ways. This improved the understanding of the data. Both strategies permit quantification of responsiveness; the effect size statistic gives a ratio of the mean change divided by the standard deviation of that change (signal to noise ratio) and the ROC method gives the areas under the ROC curves. The effect size statistic is easier to calculate and to understand than the ROC method. ROC curves visualize

the relation between the true positive and false positive rates at different cut-off points of change scores. With the ROC curves one can identify cut-off points in score changes, which provide the best combination of sensitivity and specificity. Unfortunately, both strategies require that the external criterion is dichotomous instead of degree of improvement or deterioration. To sum up, both strategies appeared to complement each other and were very useful for evaluating the responsiveness.

Acknowledgements

The authors thank P Nelemans, GJMG van der Heijden and AGH Kessels for their helpful comments.

References

1. Beurskens AJHM, Vet HCW de, Köke AJA, Heijden GJMG van der, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease specific questionnaires. Spine 1995; 20: 1017-1028.
2. Deyo RA. Measuring the functional status of patients with low back pain. Arch Phys Med Rehabil 1988; 69: 1044-1053.
3. Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: Design, analysis and interpretation. BMJ 1992; 305: 1145-1148.
4. Fairbank JCT, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66: 271-273.
5. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain Spine 1983; 8: 141-144.
6. Baker JD, Pynsent PB, Fairbank JCT. The Oswestry disability index revisited: its reliability, repeatability and validity, and a comparison with the St Thomas's disability index. In: Roland MO, Jenner JR (eds). Backpain: new approaches to rehabilitation and education. Manchester: University Press, 1989: 174-186.
7. Co YY, Eaton S, Maxwell MW. The relationship between the St. Thomas and Oswestry disability scores and the severity of low back pain. J Manipulative Physiol Ther 1993; 16: 14-18.
8. Grönblad M, Hupli M, Wennerstrand P, Järvinen E, Lukinmaa A, Kouri JP, Karaharju EO. Intercorrelation and test-retest reliability of the pain disability index (PDI) and the Oswestry disability questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. Clin J Pain 1983; 9: 189-195.
9. Feinstein AR, Josephy BR, Wells CK. Scientific and clinical problems in indexes of functional disability. Ann Intern Med 1986; 105: 413-420.
10. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40: 171-178.
11. Koes BW, Bouter LM, Mameren H van, Essers AHM, Verstegen GMJR, Hofhuizen DM, Houben JP, Knipschild PG. The effectiveness of manual therapy, physiotherapy and treatment of the general practitioner for non-specific back and neck complaints. A randomized clinical trial. Spine 1992; 17: 28-35.
12. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. J Chron Dis 1986; 39: 897-906.
13. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. J Chron Dis 1985; 38: 27-36.
14. Bombardier C, Tugwell P, Sinclair A, Dok C, Anderson G, Buchanan WW. Preference for endpoint measures in clinical trials: results of structured workshops. J Rheumatol 1982; 9: 798-801.
15. Fries JF. Toward an understanding of patient outcome measurement. Arthritis Rheum 1983; 26: 697-704.
16. Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press, 1977: 1-27.

17. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Medical Care 1989; 27: 178-189.
18. Norman GR. Issues in the use of change scores in randomized trials. J Clin Epidemiol 1989; 42: 1097-1105.
19. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clin Trials 1991; 12: 142-158.
20. Beurskens AJHM, Heijden GJMG van der, Vet HCW de, Köke AJA, Lindeman E, Regtop W, Knipschild PG. The efficacy of traction for lumbar back pain. Design of a randomized clinical trial, J Manipulative Physiol Ther 1995; 18: 141-147.
21. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995; 20: 1899-1909.
22. Hudson-Cook N, Tomes-Nicholson K, Breen A. A revised Oswestry disability questionnaire. In: Roland MO, Jenner JR (eds). Backpain: new approaches to rehabilitation and education. Manchester: University Press, 1989: 187-204.

Chapter 7

A patient-specific approach for measuring functional status in low back pain

AJHM Beurskens,[1] HCW de Vet,[1] AJA Köke,[2] E Lindeman,[3] GJMG van der Heijden,[1] W Regtop,[4] PG Knipschild.[1]

1 Department of Epidemiology, University of Limburg, Maastricht
2 Department of Physiotherapy, University Hospital, Maastricht
3 Department of Rehabilitation Medicine, University Hospital, Maastricht
4 'Hogeschool Heerlen', Department of Physiotherapy, Heerlen

Submitted for publication.

Abstract

Background and purpose - The purpose of this study was to develop and evaluate a patient-specific approach for measuring functional status in low back pain.
Subjects - A cohort of 150 patients with non-specific low back pain were measured at baseline and 12 weeks later.
Methods - The feasibility of the patient-specific approach was evaluated. The influence of baseline levels was studied by comparing absolute and relative change scores of the instruments. Effect size statistics and correlations between change scores were used to evaluate the responsiveness. The patient-specific approach was compared with more established instruments such as the Roland disability questionnaire and pain evaluated on a visual analog scale.
Results - The patient-specific approach was feasible. It appeared that, although the baseline score of patient-specific instrument was higher than of the other instruments, the relative change scores were similar. The patient-specific approach was able to discriminate between improved and non-improved patients: the effect size statistics were higher in the improved group than in the non-improved group. The correlations between change scores of the main complaints on the one hand and global perceived effect, Roland disability questionnaire and pain on the other were high: they varied from 0.69 to 0.81.
Conclusion and Discussion - The responsiveness of the patient-specific approach was comparable with more established outcome measures. What is more, the approach was able to detect changes on complaints that were highly relevant for the patients.

Introduction

Measuring functional status has received considerable attention over the last few years. Patients want to perform the usual activities of daily living. Since activities and their perceived importance varies widely between patients it is sensible to focus on the complaints a patient presents as being the most disturbing and important.[1,2] In this way outcome measurement is tailored to the individual patient (each patient has his or her specific main treatment goal). This patient-specific approach has been evaluated in the fields of rheumatoid arthritis,[3,4] the elderly,[5] lung disease,[6] heart failure,[7] general health,[8] mental health,[9] and hip replacement.[10] The approach has proved to be at least as responsive than existing questionnaires with which it was compared.

In our opinion the patient-specific approach is also very useful in the field of low back pain. This article presents the development and evaluation of a patient-specific approach for assessing functional status in patients with low back pain. First, we will evaluate the feasibility of the selection procedure of main complaints. Second, the influence of baseline levels will be studied by comparing absolute and relative change scores of the instruments. Third, we will evaluate the responsiveness of the instruments using effect size statistics and correlations between change scores. In this article we define responsiveness as the ability of an instrument to discriminate between clinically important and clinically unimportant changes on group level. The results of the patient-specific approach will be compared with more established instruments such as the Roland Disability Questionnaire (RDQ)[11] and pain during the last week evaluated on a visual analog scale (VAS).[12,13]

Method

Patients

Patients were participants in a randomized clinical trial on the efficacy of continuous motorized traction for low back pain. The Medical Ethics Committee of the University of Limburg and the University Hospital Maastricht in the Netherlands approved the study protocol. All patients signed informed consent. In total 150 patients were selected with at least 6 weeks non-specific low back pain. The mean age was 41 years and 44% were women. The median duration of low back pain was 20 weeks, 82% had a previous episode of low back pain, and 33% had radiating pain into the lower leg. The details of the study design have been published elsewhere.[14]

Patient-specific approach

The main complaints of the patients were selected using a standardized procedure (figure 1). Before randomization 2 intake visits were planned, separated by a 1-2 week qualification period.[15]

During the first intake visit a patient selected his or her main complaints in the following way. The research physiotherapist asked the patient: "Please tell me which activities you perceive as important, and which were hampered by low back pain during last week?" A list with 36 activities (appendix 1) was offered as suggestions to support recall. Patients were also allowed to select activities that were not on the list. The patient was asked to select 5 activities which he or she perceived as problematic, scoring for each complaint the severity, importance, and frequency of

performance on a VAS (figure 2). On the basis of these scores the patient was asked to choose 3 out of the 5 complaints and rank them in order of importance.

Procedure

Intake visit 1
- ➜ List of 36 activities
- ➜ Select 5 complaints
- ➜ Score for these complaints on a VAS:
  - ✱ difficulty
  - ✱ importance
  - ✱ frequency of performance
- ➜ Select 3 complaints on the basis of these criteria

Qualification period of 1-2 week
- ➜ Pay attention to activities that are hampered by low back pain

Intake visit 2
- ➜ List of 36 activities
- ➜ Select 5 complaints
- ➜ Score for these complaints on a VAS:
  - ✱ difficulty
  - ✱ importance
  - ✱ frequency of performance
- ➜ Select 3 complaints on the basis of these criteria

Definitive selection
- ➜ Compare the complaints selected during both intake visits
- ➜ Select of 3 definite main complaints

Figure 1. Procedure: selection of the 3 main complaints.

Complaint 1: *Standing for a long time (30 minutes)*

How difficult was it to perform this movement or activity during the last week?

no problems — impossible

How important was it for you to perform this movement or activity during the last week?

not important — very important

How often did you perform this movement or activity during the last week?

never — very often

Figure 2. An example of difficulty, importance and frequency of performance of a complaint. Filled in by every patient for 5 complaints during both intake visits.

Patients were not allowed to select complaints that could be easily avoided, because during the follow-up it would be impossible to evaluate the severity of complaints that were not performed. At the end of the first intake visit the research physiotherapist asked the patient to pay attention to activities that are hampered by low back pain in the qualification period.

During the second intake visit the whole procedure was repeated; the patients selected again 3 complaints. During the selection procedure no information was revealed about the 3 complaints selected during the first intake visit. Finally, the selected main complaints during the first and second intake visit were compared. Out of these complaints the patient selected 3 definitive complaints and ranked them in order of importance. These 3 definitive complaints constituted the main complaints for the duration of the study. The patient was asked to be as specific as possible in describing the complaints. In this way the wording of the complaints could specified. For each outcome measurement the difficulty of performance of the main complaints was scored by self-assessment on a VAS (0=no problems; 100=impossible). No interviewer is needed for the scoring of the complaints during outcome measurements. In appendix 2 an example of a selection procedure is given.

Other outcome measures

The patient-specific approach was compared with the following outcome measures:

[a] Global perceived effect was measured by self-assessment on a seven-point scale (1=completely recovered; 7=vastly worsened). For comparisons between patients who improved and who did not improved and those who deteriorated we recoded the answers in the following way: completely recovered or much improved were coded as improved; slightly improved, no change and slightly worsened as non-improved; and much and vastly worsened as deteriorated.

[b] The RDQ has been developed to measure the functional status in a disease specific way.[10] This questionnaire was derived by choosing 24 yes/no items relevant for low back pain from the Sickness Impact Profile. The scoring of the RDQ is achieved simply by counting the number of positive responses; a patient individual score can vary from zero (no disability) to 24 (severe disability). The RDQ has been used frequently and seems to be responsive to clinical changes.[16,17]

[c] Pain during the last week was evaluated on a 100 mm VAS (0=no pain, 100=unbearable pain). Reliability, validity and responsiveness of the VAS have been tested.[12,13]

All outcome measures were assessed at baseline, and after 5 and 12 weeks, and 6 months. Since the results at the different moments of measurements were the same, we report only the data from the 12-week measurement.

Data analysis

All data analyses were done with SPSS statistical software.[18] The height of the baseline scores of an instrument influences the possibility to improve (if we assume that high scores indicate more complaints or pain). It becomes more difficult to assess improvement if baseline scores are already low. For that reason, we compared the influence of baseline levels on change scores through comparing absolute and relative change scores. The absolute score change was the difference obtained by subtracting the follow-up score from the baseline score. A positive score difference indicates an improvement, a negative difference a deterioration. The relative score difference was calculated by dividing the absolute change scores by the baseline score. The scores of the main complaints and pain ranged from 0-100, the sumscore of the RDQ ranged from 0-24. For better comparison we present a standardized RDQ score of 0-100, by multiplying each score by 100/24.

One method for assessing responsiveness on group level is the use of effect size statistics. These statistics relate the magnitude of the change of the patients to the variability in change score.[19] Effect size is calculated by taking the mean change of a variable and dividing it by the standard deviation of that variable. The effect size statistic is the relation between "signal" and "noise" and is also named the efficiency (E).[20] For the comparison of responsiveness using effect size statistics an external criterion is needed to distinguish between groups of patients who improved, those who did not improve and those who deteriorated. An estimate of clinically important change was obtained by recoding the global perceived effect assessed by the patient.

Another frequently used method for assessing responsiveness is studying the correlations between change scores of measures[21] also called longitudinal construct validity.[22] If the severity of the main complaints decreases over time, it is expected to be associated with improvements in other measures. We assessed Spearman rank correlation coefficients between change scores of main complaints, RDQ, pain on the one hand and global perceived effect on the other. Pearsons correlation coefficients were also calculated between main complaints, RDQ and pain.

Results

Practical performance of the patient-specific approach

Patients highly appreciated it that we paid attention to their specific situation and that they could select the complaints of importance to them. The procedure for identification of the complaints during the intake visits took about 10 minutes. During the first intake visit a large number of patients indicated that it was difficult to select activities, while the selection during the second intake visit was much easier. Only 7 patients were not able to identify a third definitive complaint. None of the patients selected an activity that was not on the 36-item list. Of the 3 selected definitive main complaints the degree of importance evaluated on a VAS was 91, 89, and 87. The frequency of performance was 87, 82 and 78 respectively.

In 48 of the patients (32%) the 3 main complaints selected during both intake visits were the same; these complaints constituted the definitive complaints. Of the remaining 103 patients only 6% selected the 3 complaints of the first intake visit as definitive complaints, 69% selected the 3 complaints during the second intake visit, and in 25% the definitive complaints were a combination of the complaints selected during both intake visits.

We distinguished the selected main complaints into categories because the originally differentiation was too detailed. Table 1 shows the categories of complaints that patients selected most frequently and their ranking of importance. There is a large variation in the complaints selected. Complaints frequently selected are standing, sleeping, lying flat, sitting, housework and standing in a slightly bended position. What is particularly noticeable was that sleeping or lying flat was most often mentioned as first complaint. Other selected complaints were turning, running, turning in bed, getting in and out of the car.

Table 1. Ranking of importance of the selected definitive main complaints.

	Complaint	Ranked first	Ranked second	Ranked third	Total
1	Standing	13	23	18	54
2	Sleeping/lying flat	31	16	4	51
3	Sitting	10	15	20	45
4	Housework	17	8	15	40
5	Standing slightly bended	15	10	14	39
6	Lifting/carrying	7	16	15	38
7	Walking	10	13	10	33
8	Bending	13	12	5	30
9	Rising from bed	9	8	5	22
10	Rising from a chair	3	7	10	20
11	Driving a car	4	4	6	14
12	Working	7	6	1	14

If a patient reported having problems with housework or work, he or she was asked to select that activity or posture that caused the most complaints. For 40 patients regarding housework and 14 patients regarding work it was not possible to distinguish between problematic activities. These patients had problems with the housework and work in totality and not only with one activity.

Complaints between sub-acute and chronic patients did not differ. Only in the case of the patients who selected sleeping or lying flat did many have chronic, i.e. more than 6-months', low back pain. Complaints during static activities (i.e. standing, lying, sitting) were selected more frequently than during dynamic activities (i.e. lifting, rising from chair or bed). Especially the maintenance of a particular position for a long time was frequently mentioned as a problem.

Baseline and change scores

Table 2 shows the response to the question about global perceived effect: 71 patients (47%) were recoded as improved and 73 patients (49%) were recoded as non-improved. Only 6 patients (4%) indicated deterioration (much worsened or vastly worsened).

Table 2. Patients' assessment of global perceived effect after 12 weeks (n=150).*

Answer	Number	(%)
completely recovered	15	(10)
much improved	58	(38)
slightly improved	31	(21)
no change	27	(18)
slightly worsened	12	(8)
much worsened	5	(3)
vastly worsened	1	(1)

* missing for 1 patient.

Because of this small number these 6 patients were excluded from the analyses in which we used the recoded scores.
The baseline scores in the improved and non-improved group were similar for all outcome measures. In the improved group the baseline scores and absolute change scores of the main complaints were higher than of the other outcome measures (table 3). Yet, the relative change scores were the same for all instruments, ranging from 73-76%.

Table 3. Baseline scores, absolute and relative change scores for the improved (n=73) and non-improved (n=70) patients.

Instrument	Baseline score		Absolute change score		Relative change score (%)	
	improved	non-improved	improved	non-improved	improved	non-improved
Main complaint 1	76	72	57	11	75	15
Main complaint 2	75	70	57	12	76	16
Main complaint 3	75	73	56	8	75	11
RDQ*	48	51	34	4	73	8
Pain last week	61	63	44	6	73	10

* standardized RDQ (0-100)

Responsiveness

Table 4 shows mean changes and effect size statistics of the groups of patients who improved and those who did not improve.

Table 4. Mean changes, standard deviations (SD) and effect size statistics in the improved (N=73) and non-improved (n=70) patients.

Instruments	Mean changes	SD	Effect size*
Main complaint 1			
improved	57.0	24.8	2.30
non-improved	10.5	22.0	0.48
Main complaint 2			
improved	56.5	22.3	2.53
non-improved	11.6	19.9	0.58
Main complaint 3**			
improved	55.8	24.7	2.26
non-improved	8.0	19.5	0.41
RDQ (0-24)			
improved	8.3	4.8	1.73
non-improved	1.0	3.7	0.27
Pain last week			
improved	44.2	26.6	1.66
non-improved	6.1	23.2	0.26

* Effect size is calculated as mean change score divided by the standard deviation of the mean change score.
** 7 patients could not select a third main complaint.

For all outcome measures the effect size statistics in the improved group are higher than those in the non-improved group. This means that all measures can discriminate between groups of improved and non-improved patients. The effect size statistic for the main complaint in the improved and non-improved group was slightly higher than those for the RDQ and pain. In improved patients this means that some patients showed more improvement on the main complaints than on the other outcome measures. In non-improved patients this means that some patients who did not improve according to their global perceived effect changed more on their main complaints than on the RDQ and pain.

Table 5 shows that the correlation coefficients between the mean change scores of the main complaint on the one hand and RDQ, pain and global perceived effect on the other are high. They vary from r=0.69 to r=0.81.

Table 5. Correlations between the change scores of the instruments (n=150).

	Global perceived effect*	Main complaint 1	Main complaint 2	Main complaint 3**
Global perceived effect*	-	0.80	0.77	0.79
Roland	0.75	0.75	0.69	0.74
Pain	0.73	0.81	0.70	0.73

* Missing for 1 patient. Spearman rank correlation coefficients.
** 7 patients could not select a third main complaint.

Discussion

In this study we described the development and evaluation of a patient-specific approach for measuring functional status in patients with low back pain. The patient-specific approach was feasible, it was easy to understand and the time required to complete it was short. Patients highly appreciated it that we paid attention to their specific situation and that they could select the complaints of importance to them. The high scores for the degree of importance and frequency of performance indicated that the selected main complaints were highly relevant.

We evaluated the influence of the level of baseline scores. It appeared that, although the baseline score of patient-specific instrument was higher than of RDQ and pain, the relative change scores were similar. In addition, all instruments were able to discriminate between improved and non-improved patients and correlations between change scores of the patient-specific approach and the other instruments were high. Hence, the ability of the patients specific approach to discriminate between groups of patients who improved and those who did not improve was comparable to that of frequently used instruments such as the RDQ and pain evaluated on a VAS.

For the selection procedure described an interviewer is needed. The interviewer assists the patient in selecting the main complaints and allows selection of only activities that are difficult to avoid. For example, if a patient selects bending, it is necessary that he has to bend regularly during his daily life. We think it is worth the expense of an interviewer to document the disabilities that are most important and

difficult to avoid. However, no interviewer is needed for the scoring of the complaints during effect measurements.

One can wonder why it is necessary to have 2 intake visits. We saw that only 32% of the patients selected exactly the same 3 main complaints during the first and second intake visit. During the first intake visit patients were often surprised and needed time to think about the selection. Patients could use the qualification period to pay attention to activities which were hampered by their low back pain. In this way, the selection of the complaints during the second intake visit was better considered. We think that time for reflection between the 2 intake visits is advisable.

We asked the patients to rank three main complaints in order of importance. The responsiveness of the three complaints were similar. This implies that it would be possible to substitute 3 complaints by 1 or 2. However, 3 main complaints provide a more comprehensive profile of functional status. In this study a few patients were not able to select a third complaint. But it is possible that in another population all patients can do this.

Tugwell et al[3] suggest that it is important to take account of new disabilities that develop during the study period. It is possible to ask at each visit if there are new important complaints. However, the new complaints cannot be analyzed in the same way as the other complaints because baseline values are missing for newly arising complaints.

In our study the correlations between main complaints, RDQ, pain and global perceived effect were rather high. Deyo[23] found a much lower correlation (r=0.16) between change scores of the RDQ with a three-point clinical rating scale. This low correlation coefficient could be the result of the insensitivity of the three-point scale to register small changes. Stratford et al[24] have reported a correlation of 0.60 when RDQ change scores were correlated with a 15-point global rating of change scale.

The strategy used to assess responsiveness depends on some external criterion or gold standard that defines the smallest clinically relevant improvement or deterioration. Because a perfect external criterium for clinically relevant change of functional status does not exist, no single external criterion can determine absolute responsiveness. Using a sub-optimal external criterion the responsiveness of several instruments can be compared. In this way, the relative responsiveness of each instrument is assessed. It is important to compare instruments against several external criteria in one study population. If results are consistent on the basis of several external criteria, confidence increases about the correct ranking of the responsiveness of the instruments.[25]

For evaluating responsiveness in the field of low back pain, the choice of external criteria for change is difficult. We have chosen global perceived effect because it is an all-encompassing measure for improvement that includes pain, functional status and other aspects that patients classify as important. Although it defines not the smallest clinically relevant change, most people would be reluctant to label a patient as improved or worse contrary to this personal assessment. Global perceived effect can also be assessed by the observer or doctor. Using a sub-optimal external criterion, it has to be noted that an instrument is not able to discriminate better than the gold standard between improved and non-improved patients.

In this study the effect size statistics of the main complaints in the non-improved group ranged from 0.41 to 0.58. In another study in which we evaluated the 5 week follow-up of the first main complaint in a subgroup (n=81) of the patients in this

study, the effect size in the non-improved group was slightly higher (0.73).[26] In both studies the values of the effect size statistics in the non-improved were higher than the effect size statistics for the RDQ and pain, indicating that some non-improved patients changed more on their main complaints than on the RDQ and pain. In interpreting the results of the patient-specific approach we should be aware of these non-specific improvements.

The RDQ and the patient-specific approach are instruments for measuring functional status. The RDQ contains questions about a range of activities in which possible activities important for some patients are omitted and activities not relevant for some patients are represented. The advantage of the patient-specific approach is that every patient selects his or her main complaints. In this way only a few, but for patients very important aspects of functional status are evaluated.

In conclusion, the application of the patient-specific approach was feasible. The responsiveness was comparable to more established outcome measures. What is more, this approach was able to detect changes on complaints that were highly relevant for the patients. On the basis of these findings it would be valuable to apply the patient-specific approach in future effectivity studies to try to replicate our results and provide further evaluation.

References

1. Feinstein AR, Josephy BR, Wells CK. Scientific and clinical problems in indexes of functional disability. Ann Intern Med 1986; 105: 413-420.
2. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40: 171-178.
3. Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B. The MACTAR patient preference disability questionnaire. An individualized function priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987; 14: 446-451.
4. Scott PJ, Huskisson EC. Measurement of functional capacity with visual analogue scales. Rheumatology and Rehabilitation 1977; 16: 257-259.
5. Guyatt GH, Eagle DJ, Sackett B, William A, Griffith L, McIlroy W, Patterson CJ, Turpie I. Measuring quality of life in the frail elderly. J Clin Epidemiol 1993; 46: 1433-1444.
6. Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. Thorax 1987; 42: 773-778.
7. Guyatt GH, Nogradi S, Halcrow S, Singer J, Sullivan MJJ, Fallen EL. Development and testing of a new measure of health status for clinical trials in heart failure. J Gen Intern Med 1989; 4: 101-107.
8. MacKenzie CR, Charlson ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Arch Intern Med 1986; 146: 1325-1329.
9. Kiresuk TJ, Sherman RE. Goal attainment scaling: a general method for evaluating comprehensive community mental health programs. Community Ment Health J 1968; 4: 443-453.
10. O'Boyle CA, McGee H, Hickey A, O'Malley K, Joyce CRB. Individual quality of life in patients undergoing hip replacement. Lancet 1992; 339: 1088-91
11. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983; 8: 141-144.
12. Carlsson AM. Assessment of chronic pain: I. Aspects of the reliability and validity of the visual analog scale. Pain 1983; 16: 87-101.
13. Sriwatanakul K, Kelvie W, Lasagna L, Calimlim JF, Weis OF, Metha G. Studies with different types of visual analog scales for measurement of pain. Clin Pharmacol Ther 1983; 34: 234-239.
14. Beurskens AJHM, Heijden GJMG van der, Vet HCW de, Köke AJA, Lindeman E, Regtop W, Knipschild PG. The efficacy of traction for lumbar back pain. Design of a randomized clinical trial. J Manipulative Physiol Ther 1995; 18; 141-147.
15. Knipschild P, Leffers P, Feinstein A. The qualification period. J Clin Epidemiol 1991; 44: 461-464.

16. Beurskens AJ, Vet HC de, Köke AJ, Heijden GJ van der, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease specific questionnaires. Spine 1995; 20: 1017-1028.
17. Deyo RA. Measuring the functional status of patients with low back pain. Arch Phys Med Rehabil 1988; 69: 1044-1053.
18. Norusis MJ. SPSS for Windows. Base system user's Guide Release 5.0. Chicago: SPSS inc, 1992
19. Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1977, pp. 1-27.
20. Anderson JJ, Chernoff C. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. J Rheumatol 1993; 20: 535-537.
21. Meenan RF, Anderson JJ, Kazis LE, Egger MJ, Altz-Smith M, Samuelson CO, Willkens RF, Solsky MA, Hayes SP, Blocka KL, Weinstein A, Guttadauria M, Kaplan B, Klippel J. Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. Arthritis Rheum 1984; 27: 1344-52.
22. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. J Chron Dis 1985; 38: 27-36.
23. Deyo RA. Comparative validity of the Sickness Impact Profile and shorter scales for functional assessment in low back pain. Spine 1986; 11: 951-954.
24. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. Phys Ther 1994; 74: 528-534.
25. Deyo RA, Center RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test-performance. J Chron Dis 1986; 39: 897-906.
26. Beurskens AJHM, Vet HCW de, Köke AJA. Responsiveness of functional status in low back pain. A comparison of different instruments. Pain (in press).

Appendix 1

List of activities

Lying in bed
Turning in bed
Rising from bed
Sleeping

Rising from a chair
Get seated
Sitting for a long time
Get in or out a car
Driving a car of bus
Biking

Standing
Standing for a long time
Light work in and around the house
Heavy work in an around the house

Walking inside the house
Walking
Running
Going up stairs
Going down stairs

Bending for a long time
Standing slightly bended
Turning your back
Bending with a turned back

Carrying about 5 kilogram, e.g., a shopping bag
Carrying about 10 kilogram, e.g., a one year old child
Picking up something light from the floor, e.g., a handkerchief
Picking up something heavy from the floor, e.g. a rubbish bag
Repeated lifting

Visiting family or friends
Going out
Sexual activities
Working
Housework
Doing hobby's
Sporting
Travelling

Other activities or movements:

.....

.....

Appendix 2

Case study

Ms X is housewife, aged 37 and mother of two children (2 and 5 years old). For several years she has regularly had periods with low back pain. At present her back complaints are worse again. She selected her main complaints according to the standardized procedure. During the first intake visit, she selected the following 5 activities out of the list with 36 activities: rising from bed, climbing stairs, bearing her youngest child, bending forwards and sitting for a long time. She scored the severity, importance and frequency of performance of each activity on a VAS. After a short discussion with the research assistant about her scores she chose as her 3 main complaints in order of importance: bearing her youngest child, bending forwards and climbing stairs. The other two were not chosen because the severity of rising from bed was low and she could easily avoid sitting for a long time.

Ms X went home with the instruction to pay attention to activities which were hampered by her back pain during the next week. At the second intake visit she selected again 3 activities in the standardized way. She was not allowed to see her choices of the first visit. This time she selected: bearing her youngest child, driving a car and walking.

Then the complaints selected during the first and second intake visits were compared. Ms X explained that the differences between the selections were probably due to the greater attention she paid to activities that caused her back pain during the qualification period. Ms X was asked to decide which of the activities were her 3 definitive main complaints. She chose the following activities. First, bearing her youngest child. She performed this activity frequently and it was very important for her. Second, bending forwards. She also did this very frequently and bending was difficult to avoid. Finally, she selected driving a car, because she used her car frequently to bring her child to school and for shopping. Walking and climbing stairs were not selected as main complaints because they were less frequently done and could be avoided.

Chapter 8

General discussion

This final chapter discusses the findings of our study. The discussion is divided into two sections. In the first we pay attention to some methodological issues. In the second section we look back at the efficacy of lumbar traction. We restrict this discussion to general aspects. Specific problems have already been dealt with in the discussion sections of the chapters.

Section 1: Methodological issues

Participation of patients and physiotherapists

According to the time schedule we planned to randomize 8 patients every month. In this way 19 months were needed to randomize 150 patients. When we started recruiting patients colleagues warned us for Lasagna's law (named after the statistician Louis Lasagna). This law states that suitable patients seem to disappear when the recruitment of patients for a trial starts. The availability of suitable patients commonly is overestimated, even by a factor of 10.[1] However, we were able to finish the recruitment and selection of patients within the planned time period.

The number of patients selected each month varied between 2 and 17 (figure 1). In the summer months July and August the number of randomized patients was very low. Autumn, particularly October and November, was the best time period to select patients.

We tried to keep the interest and involvement of the participating physiotherapists high by informing them regularly about the progress of the study through newsletters and meetings. In addition, we gave surprises at St Nicholas (5 December) and small gifts during the duration of the study (e.g. ice-creams in the Summer; cake and home-made pie when we had selected 75 patients and 150 patients, respectively). In figure 1 we have indicated the time-periods when we took actions and the kind of action. We do not know whether these actions can explain our successful patient recruitment. But we think that involvement of the 'suppliers' of patients throughout the whole study is essential and that it is important to show them that their work is sincerely appreciated. Thanks to their cooperation, we were able to keep close to the treatment protocol and recruit enough patients within the planned time.

For our patient recruitment and compliance of traction therapy we were dependent on the participating physiotherapists. They were practitioners with busy jobs. When we asked them to participate in the study, we explained the goal of the study and informed them about the effort it would take of them. We asked them to consider participation in the study. If they had doubts about the cooperation, we asked them to take no part in the research.

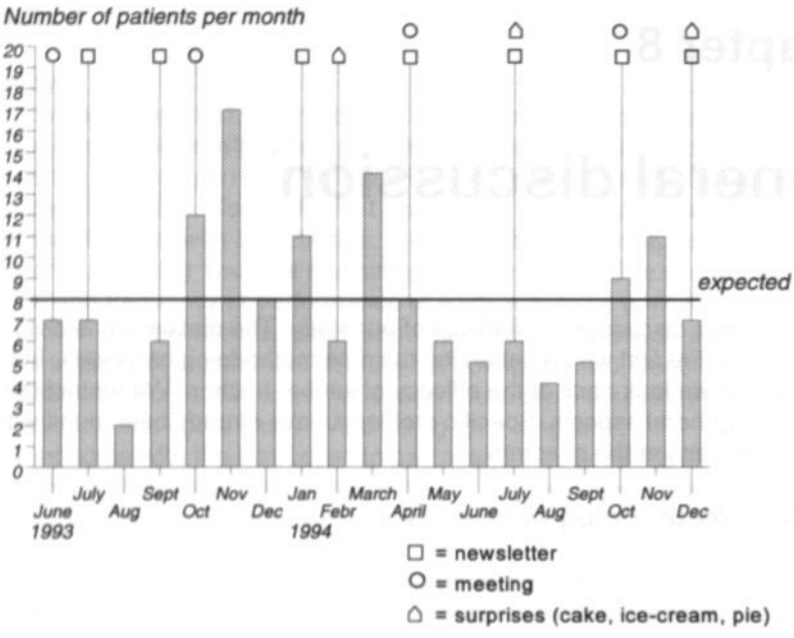


Figure 1. Number of patients selected per month and kind and timing of actions.

Before randomization we planned two intake visits. Between the visits was a qualification period[2] of one to two weeks. During this time period patients could think about their participation in the study and discuss it with their family. Patients who indicated that they did not have enough time or did not accept the risk of a placebo treatment were excluded. The research physiotherapist excluded patients who showed poor compliance, e.g., being not on time for the intake visits without a good reason. He also told the patients that it was very important to come to all effect measurements and to fill out the questionnaires, also in case their complaints did not improve.

In this trial the number of withdrawals during the treatment period and patients lost to follow-up were low. This is probably due to the careful selection of the participating physiotherapists and patients.

Placebo treatment

An optimal placebo treatment should mimic the "real" treatment as exactly as possible but lacking the specific effect. The sham traction was performed with a specially developed iliacal brace that tightens in the back during traction treatment. This was experienced as if traction were exerted. In this way we hoped that the patients in the sham group would feel the pulling earlier and that the traction force could be restricted to a minimum. Thereby increasing the contrast between the intervention groups with respect to vertebral distraction.

We developed and tested several prototypes. In order to discover whether the blinding was satisfactory the braces were tested on members of the research team and patients. We showed that with some creativity it is possible to design a sham

traction that appears real, but lacks a large traction force. Directly after the 5-week traction treatment we asked the patients to guess the treatment allocation. The traction was not systematically unmasked, only 4 patients in the sham group and 1 patient in the traction group thought that they had received sham traction.

Large improvement in both groups

Almost half of the patients in this trial improved irrespective of the kind of traction treatment given. As a consequence, the improvement could not be explained by the specific effect of traction. We can attribute the improvements to placebo effects, but there are other possible explanations.[3]

Firstly, the natural course of back pain is typical to improve regardless of therapy.[4] Furthermore, there is the problem of regression to the mean (also called statistical regression). Statistical regression describes a tendency of extreme measures to move closer to the mean when they are repeated.[5] Patients have symptoms that fluctuate from time to time. Patients visit a physiotherapist when their symptoms are extreme, but the evaluation of the effect of the therapy is at a later point in time. The chance that the patients have improved is then substantial.[6]

In addition, attention and concern of the physiotherapist, and the enthusiasm and conviction of the researchers must not be underestimated. Improvements may be a result of the care for patients and not of specific effects of the intervention. If patients know that they are taken care of, this alone may already lead to changes in the way they perceive their complaints.

Moreover, patients often evaluate the physiotherapist instead of the perceived effect, or may give a positive answer to please the research assistant. To anticipate socially desired answers, we classified for the outcome measure global perceived effect, totally recovered and much improved as improvement and classified slightly recovered as not-improved. But still many patients improved.

Finally, innovation of new technologies often causes an effect (novelty effect); the effect decreases when the special circumstance becomes usual. Our patients had no previous experience with traction therapy.

We gave some possible explanations for the large improvement of the patients. These phenomena occur not only in research but also in daily clinical practice, and are often called the non-specific effects of a treatment. They are very common and valuable in patient-care. If we had a third group with no treatment we would have been able to evaluate the non-specific effects of the treatment. But this was not the main aim of our trial and we would have needed another 75 patients. In addition, it would be very difficult to have a no treatment group because all patients who have had low back pain for 6 weeks or more want some kind of treatment.

Pilot studies

It is possible that one of the reasons for funding this research were the promising results of the pilot study (n=25),[7] although the results were not statistically significant. Trials with only small numbers of patients carry a considerable risk of failing to demonstrate a significant treatment difference when a difference is present (Type II error).[8] The power of the pilot study was too low to detect treatment differences which might be clinically relevant. With 150 patients as in our larger trial, the results in the pilot would have been statistically significant. The larger trial, however, showed that the results of the pilot study were probably due to chance.

In the larger trial we prestratificated on physiotherapy practice (n=10) participating in the study. This was done to ensure that each practice treated an equal number of patients with traction and sham traction. In this way, each practice can be regarded as a small study. Table 1 shows the rate of patients who improved in the intervention groups per physiotherapy practice. The number of patients treated per practice varies between 4 to 34. Five practices (no. 3, 5, 6, 8, and 9) treated 81% of the patients. In 5 practices the effects of traction were better than of sham traction, in 3 the sham traction showed better effects, and in 2 there were no differences between the groups. The conclusion of the 10 practices together was that there was no difference between the intervention groups. This example illustrates that small studies, such as pilot studies, are not appropriate for evaluating effects of a therapy. Positive or negative results of a pilot study should, therefore, not be used as a reason for or against funding of a research project. Pilot studies are only appropriate to evaluate whether a study is practical and feasible.

Table 1. Improved patients in the intervention groups per physiotherapy practice.*

Physiotherapy practice no.	Traction (n=77) %	Sham traction (n=74) %	Difference % traction minus sham
1 (n= 6)	50	50	0
2 (n= 5)	100	67	33
3 (n=19)	40	22	18
4 (n= 4)	100	67	33
5 (n=34)	35	59	-24
6 (n=25)	50	39	11
7 (n= 2)	100	100	0
8 (n=18)	30	63	-33
9 (n=27)	50	39	11
10 (n=11)	50	80	-30

* Ratings on a 7-point scale of patients' global perceived effect are dichotomized in improved (completely recovered and much improved); and not-improved (slightly improved, no changed, slightly worsened, much worsened and vastly worsened).

Functional status

Besides answering the primary research question about the efficacy of traction, we used (the data collected in) this trial to evaluate instruments for measuring functional status in low back pain. Several instruments for measuring functional status were used. Therefore, it was a good opportunity to compare and evaluate the instruments in the same patient group.

The choice of outcome measures is determined by the research question. Generally, there must be a clear relationship between the complaints for which the patients are included in the trial, the aim of the treatment and chosen outcome measures. In the last decade there has been a tendency to use functional status questionnaires more frequently.[9] Consensus exists that these measures should be reliable, valid, and responsive to small but clinically relevant changes.[10]

In a literature review we evaluated the methodological quality of four disease specific functional status questionnaires for patients with low back pain: Oswestry

disability questionnaire,[11] Million visual analogue scale,[12] Roland disability questionnaire,[13] and Waddell disability index.[14] The reproducibility of the questionnaires appeared to be satisfactory. The literature provided more information on the validity and responsiveness of the Oswestry and Roland questionnaires than on the Million and Waddell questionnaires.

There is no consensus about the methods to evaluate the responsiveness and a number of strategies are proposed.[10,15,16] In chapter 6 we used effect size statistics and receiver operating characteristic curves for evaluating the responsiveness of three instruments for evaluating functional status and pain. Global perceived effect was chosen as external criterium. All instruments were able to discriminate between improvement and non-improvement. In our study population the results of the Roland disability questionnaire and pain evaluated on a visual analogue scale were more responsive than severity of the main complaint and the Oswestry disability questionnaire. The two strategies for assessing responsiveness were very useful and appeared to complement each other.

In chapter 7 we evaluated a patient-specific approach for measuring functional status in low back pain. We concluded that the approach was feasible and that patients appreciated it that they could select the complaints of importance to them. The responsiveness of the patient-specific approach was comparable with more established outcome measures, such as the Roland disability questionnaire and pain evaluated on a visual analogue scale.

Several instruments have been compared and evaluated in the same patient population. It is possible that our conclusions would have been different in other patient populations: for example, in more or less disabled patients, or in more acute patients. A valid and responsive instrument may be sound for one population but not for another. As a consequence, it would be valuable to evaluate the instruments in other patient populations to try to replicate our results and provide further evaluation.

Section 2: Efficacy of lumbar traction

The conclusion of a systematic literature review on the available RCTs on the efficacy of traction was that a large trial was needed in which much attention would be paid to the methods of the study.[17] The design and results of this large trial are described in this thesis. The central question was whether continuous motorized traction is an effective treatment for patients with at least 6 weeks non-specific low back pain. The short and longer term results were very consistent: no differences between low dose (or sham) traction and high dose traction on any of the outcome measures. Although almost half of the patients improved, this improvement could not be explained by the specific effect of traction therapy.

On the basis of the mechanical rationale of traction a positive effect of high dose traction was expected compared to low dose traction. One may argue that the contrast between the intervention groups was not large enough or that the sham traction was not a real placebo. To evaluate this we plotted the traction force applied against the effect of traction. However, we saw no relation between the traction force and effect of traction. In other words, the effect of traction did not depend on the amount of traction force.

Physiotherapists often say that traction is a part of total management and that it is not possible to evaluate the effect of traction in isolation. But it is the only way to

investigate the efficacy of traction. There might be a positive effect if traction is given in combination with exercises. It is possible that there is an interaction effect with exercises. By this we mean that the exercises might strengthen or dilute the effect of traction. Because we found no effect at all in this trial, we think that an interaction effect with exercises is unlikely.

In the study we were able to keep close to the original design and to overcome most common flaws of earlier studies on traction therapy. The prognostic comparability at baseline was good and the effect measures were clinically relevant. We were also able to blind the patients and outcome assessor for the treatment allocation. We think that this study provides a valid estimate of the ineffectiveness of lumbar traction for non-specific low back pain. In our opinion, future research on the efficacy on traction therapy studying this or other modalities of traction or other patients' groups should not have high priority any more.

References

1. Hoekstra GR. Patiënten met lage rugklachten in een huisartspraktijk. Alphen aan den Rijn: Stafleu, 1983.
2. Knipschild P, Leffers P, Feinstein A. The qualification period. J Clin Epidemiol 1991; 44: 461-464.
3. Kienle GS. Über das auftreten des Placeboeffekts. Eine Analyse des Materials von H.K. Beechers "The Powerful Placebo" (1955) und der Faktoren, die einen Placeboeffekt vortäuschen können. Inaugural-Dissertation Freien Universität Berlin, 1995.
4. Frymoyer JW. Back pain and sciatica. New Engl J Med 1988; 318: 291-300.
5. McDonald C, Mazzuca S, McCabe G. How much of the placebo effect is really statistical regression? Statistics in Medicine 1983; 2: 417-427.
6. Deyo RA. Practice variations, treatment fads, rising disability. Spine 1993; 18: 2153-2162.
7. Heijden GJMG van der, Beurskens AJHM, Dirx MJM, Bouter LM, Lindeman E. Efficacy of lumbar traction: a randomized clinical trial. Physiotherapy 1995; 81: 29-35.
8. Pocock SJ. Clinical trials. A practical approach. New York: John Wiley & Sons, 1983.
9. Beurskens AJHM, Vet HCW de, Köke AJA, Heijden GJMG van der, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease specific questionnaires. Spine 1995; 20: 1017-1028.
10. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clin Trials 1991; 12(Suppl): 142-158.
11. Fairbank JCT, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66: 271-273.
12. Million R, Hall W, Nilsen KH, Baker RD, Jayson MI. Assessment of the progress of the back pain patient. Spine 1982; 7: 204-212.
13. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low back pain. Spine 1983; 8: 141-144.
14. Waddell G, Main CJ. Assessment of severity in low back disorders. Spine 1984; 9: 204-208.
15. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chron Dis 1986; 39: 897-906.
16. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties. J Clin Epidemiol 1992; 45: 1341-1345.
17. Heijden GJMG van der, Beurskens AJHM, Koes BW, Assendelft WJJ, Vet HCW de, Bouter LM. The efficacy of traction for back and neck pain. A blinded review of randomized clinical trial methods. Physical Therapy 1995; 75: 93-104.

Summary

Although low back pain occurs frequently there is no consensus about its management. A large variety of therapeutic interventions exists, but none seems to be clearly superior to others. The efficacy of many (physio)therapeutic interventions for low back pain remains questionable. One of the treatment options is traction. The central question of the trial described in this thesis is whether traction is an effective treatment for patients with low back pain. The research question has been focused on the efficacy of continuous motorized traction for patients with at least 6 weeks non-specific low back pain.

Chapter 1 provides an introduction and guide to this thesis. The conclusion of a systematic review on the available Randomized Clinical Trials (RCTs) on the efficacy of traction was that a large trial was needed in which much attention should be paid to the methods of the study. We decided to perform this study. In the trial several instruments were used for measuring functional status in low back pain. Therefore, we were able to compare and evaluate the methodological quality of the instruments in the same patient group.

Chapter 2 describes the design of our trial in detail. The available literature is not very clear about the working mechanisms by which continuous lumbar traction could be effective. The effect of traction is mainly based on mechanical working mechanisms. Supposed mechanical effects of traction are: vertebral separation and widening of the intervertebral spaces. Various authors have reported that a certain amount of traction force is necessary to achieve separation of the vertebrae and widening of intervertebral foramina. Based on the mechanical rationale patients were randomly allocated to high dose continuous traction or low dose (sham) traction.

In the high dose traction group the force was increased until the patient indicated that the tolerance for pulling was reached, with a minimum traction force of 35% and a maximum of 50% of the total body weight. In the sham group the force was slowly increased until the patient indicated that he felt little pulling, with a maximum traction force of 20% of the total body weight. The sham traction was given with a specially developed brace which becomes tighter in the back during treatment. This was experienced as if traction were exerted. By using the special brace we hoped that the patients would feel the pulling earlier and that the traction force in the sham group could be restricted to a minimum.

Patients were selected if they had suffered for at least 6 weeks from non-specific low back pain, and had never before had any form of lumbar traction treatment. Both groups were treated 3 times per week during 4 weeks, 20 minutes per session. The patients and outcome assessor were blinded for the assigned treatment. The outcome measures were in order of importance: global perceived effect, severity of main complaints, functional status, pain, range of motion, work absence, and medical consumption. The effect measures were assessed before randomization and 5 weeks, 12 weeks and 6 months later.

Chapter 3 presents the results after 5 weeks follow-up. In total, 151 patients were randomized. Intention-to-treat analysis showed only small differences between the groups; some outcome measures favoring the traction group, some the sham group. The differences were not statistically significant; all 95% confidence intervals included the value zero. The number of withdrawals from treatment, loss to follow-up and protocol deviations were low. Consequently, the per-protocol analysis showed results similar to the intention-to-treat analysis. Subgroup analysis (e.g., duration of present episode, radiation below the knee) did not show any group for which traction might seem promising.

On the basis of the mechanical rationale of traction we expected 35-50% of the body weight to be effective and sham traction (<20%) to be not effective. To evaluate the chosen cut-off points, we plotted the traction force against the global perceived effect. We found no relation between the traction force applied and the effect of traction. In other words, the effect of traction did not depend on the amount of traction force.

Chapter 4 provides the results after 12 weeks and 6 months follow-up. The 5-week effect measurement aimed to detect short-term effects of traction and the 12-week and 6-month measurements to detect longer term effects. On the basis of the mechanical rationale of traction short term rather than long term effects are expected. The 5-week results showed no statistically significant differences between the intervention groups. As expected, after 12-week and 6-month follow-up there were also no statistically significant differences between traction and sham traction treatment on all outcome measures. The number of patients lost to follow-up was very low. This resulted in a complete data set which made the analyses very simple and clear. The types of additional treatments for low back pain during follow-up diverged and the numbers were relatively high but not remarkably different for the two groups. In conclusion, in this trial we were able to keep close to the original design of this study. Thereby we could overcome most common flaws in earlier studies on traction therapy. This trial does not support the claim that traction is efficacious for patients with at least 6 weeks non-specific low back pain.

Chapter 5 reviews the methodological quality of four disease specific functional status questionnaires for patients with low back pain: Oswestry disability questionnaire, Million visual analogue scale, Roland disability questionnaire, and Waddell disability index. The questionnaires were evaluated in terms of general description, scale structure, reliability, validity, responsiveness, and clinical research applications. There was not enough information available about the criteria of item selection used for the development of the questionnaires. The reproducibility of the questionnaires appeared to be satisfactory. The literature provided more information on the validity and responsiveness of the Oswestry and Roland questionnaires than on the Million and Waddell questionnaires. Additional research is needed to compare and improve the existing questionnaires.

Chapter 6 describes a comparison of the responsiveness of three instruments for evaluating functional status: two disease-specific questionnaires (Oswestry and Roland disability questionnaires), and a patient-specific method (severity of main

complaint). In addition, the responsiveness of pain rated on a visual analogue scale was assessed. Global perceived effect was chosen as external criterium. In a cohort of 81 patients all these measures were assessed before and after 5 weeks of treatment. Two strategies for evaluating the responsiveness in terms of sensitivity to change and specificity to change were used: effect size statistics and receiver operating characteristics curves. All instruments were able to discriminate between improvement and non-improvement. In our study population the results of the Roland disability questionnaire and pain evaluated on a visual analogue scale were more responsive than severity of the main complaint and the Oswestry disability questionnaire. The two strategies for assessing responsiveness were very useful and appeared to complement each other.

In **chapter 7** we evaluated a patient-specific approach for measuring functional status in low back pain. At baseline patients selected their main complaints in a standardized way: they selected three activities they performed frequently, which they perceived as important in day-to-day life, and which low back pain made difficult for them. A cohort of 150 patients scored the severity of the main complaints at baseline and 12 weeks later. We concluded that the approach was feasible and that patients appreciated it that they could select the complaints of importance to them. The responsiveness of the patient-specific approach was comparable with more established outcome measures, such as the Roland disability questionnaire and pain evaluated on a visual analog scale. What is more, the approach was able to detect changes on complaints that were highly relevant for the patients. On the basis of these findings it would be valuable to apply the patient-specific approach in future studies to try to replicate our results and provide further evaluation.

Chapter 8 provides a general discussion of the study described in this thesis. The chapter is divided into two sections. In the first section attention is payed to some methodological issues. Special attention is paid to the participation of patients and physiotherapists in the trial, the placebo treatment, the large improvement in both groups, the interpretation of the results of pilot studies, and the use of functional status as an outcome measure.

In the second section we looked back at the efficacy of lumbar traction. We conclude that the study provides a valid estimate of the inefficacy of lumbar traction for non-specific low back pain. In our opinion, future research on the efficacy on traction therapy studying this or other modalities of traction or other patients' groups should not have high priority any more.

Uitgebreide samenvatting

Voor geïnteresseerde lezers die in een korte tijd meer willen weten over het onderzoek naar het effect van fysiotherapeutische tractie bij mensen met lage rugklachten is deze uitgebreide Nederlandse samenvatting opgenomen. In het eerste deel wordt, na een korte beschrijving van de achtergrond van het onderzoek, de opzet, resultaten, discussie en conclusie beschreven. In het tweede deel wordt ingegaan op enkele "bijprodukten" van het onderzoek die beschreven zijn in de hoofdstukken 5, 6 en 7.

Deel 1

Achtergrond

Rugklachten komen veel voor. Ongeveer 80% van de mensen heeft tijdens hun leven wel eens last van de rug. Ondanks dat rugklachten vaak voorkomen bestaat er geen overeenstemming over de beste behandeling van deze klachten. Het effect van veel behandelingen is onbekend. Eén van de behandelmogelijkheden voor rugklachten is tractie. Bij tractie wordt er aan de rug van een patiënt getrokken.

Om na te gaan wat er eerder geschreven was over het effect van tractie hebben we op een systematische manier de kwaliteit van de 14 bestaande onderzoeken beoordeeld. Het bleek vooral dat de methodologische kwaliteit van deze onderzoeken onvoldoende was. Hierdoor was het niet mogelijk een conclusie te trekken over het effect van tractie bij lage rugklachten.

In een proefonderzoek (pilot) waaraan 25 patiënten meededen, bleek dat het mogelijk was de tekortkomingen van eerdere onderzoeken te voorkomen. De resultaten van de pilot waren veelbelovend: 64% van de mensen die echte tractie kregen herstelden tegen 34% van de mensen die placebo (nep) tractie kregen. Omdat er maar 25 patiënten meededen is het niet zeker dat tractie helpt. Het kan ook zijn dat er door toeval meer patiënten in de tractiegroep herstelden dan in de placebogroep.

Op basis van het proefonderzoek en het literatuuronderzoek concludeerden wij dat een groot onderzoek nodig was, waarbij veel aandacht wordt besteed aan de kwaliteit van het onderzoek.

In de literatuur bestaat geen duidelijkheid over de manier waarop tractie zou werken. Vaak wordt geschreven dat door aan de rug te trekken de ruimtes tussen de wervels van de wervelkolom groter zouden worden. Mogelijke effecten die daardoor optreden zijn het verminderen van een uitpuiling van een tussenwervelschijf (discus) en het vergroten van de ruimte voor zenuwen of gewrichtkapsels. Een zekere tractiekracht is nodig om deze effecten te krijgen. Een gedeelte van de toegediende tractiekracht gaat verloren aan wrijving van het tractieapparaat en wrijving tussen de patiënt en behandelbank, spierspanning en ligamentaire weerstand te overwinnen. Als er gebruik wordt gemaakt van een behandelbank met een verrolbaar bovenblad beschouwen wij een lumbale

tractiekracht van minder dan 20% van het lichaamsgewicht als placebo (nep) tractie. Om voldoende contrast tussen de groepen te hebben en ongunstige effecten te vermijden, hebben we gekozen om als echte tractie een tractiekracht tussen de 35% en 50% te gebruiken.

Onderzoeksopzet

Onderzoeksvraag

De vraagstelling van het onderzoek is: "Wat is het effect van fysiotherapeutisch hard trekken met elektrische continue tractie ten opzichte van zacht trekken (placebo) op de mate van het herstel bij patiënten met minimaal 6 weken a-specifieke lage rugklachten?"

Selectie van patiënten

Patiënten mochten meedoen aan het onderzoek als ze minstens 6 weken last hadden van hun lage rug zonder dat er een duidelijke oorzaak voor de klachten bekend was (a-specifieke klachten). Ze moesten minstens 18 jaar zijn en mochten nooit eerder met tractie behandeld zijn. Verder werden patiënten bij wie de rugklachten in de afgelopen twee weken duidelijk verbeterd waren, uitgesloten. De patiënten werden behandeld in fysiotherapiepraktijken en poliklinieken van ziekenhuizen in Midden- en Zuid-Limburg die aan het onderzoek meewerkten. Een ervaren onderzoeks-fysiotherapeut ging bij alle patiënten na of ze voldeden aan onze criteria. Ook gaf hij de patiënten informatie over het onderzoek en vroeg schriftelijke toestemming (informed consent) aan iedere patiënt.

Loting en blindering

Door loting (randomisatie) werd bepaald welke van de twee behandelingen een patiënt kreeg: echte (hoge dosis) tractie of placebo (lage dosis) tractie. Een persoon die niets met het onderzoek te maken had, maakte enveloppen klaar met daarin de behandelcode "echt" of "placebo".

De behandelend fysiotherapeut kreeg voor iedere patiënt zo'n gesloten envelop. Hij opende de envelop tijdens de eerste behandeling en wist daardoor welke behandeling de patiënt kreeg (hij was niet geblindeerd voor de toegewezen behandeling). De patiënten en degene die het effect beoordeelde (de onderzoeks-fysiotherapeut) wisten niet of de behandeling echte of placebo tractie was (zij waren geblindeerd).

Behandelingen

De fysiotherapeuten behandelden de patiënten volgens een vast voorschrift (protocol). Op die manier wisten we precies wat er tijdens een behandeling gebeurde. De tractie werd uitgevoerd bij de patiënt in rugligging waarbij de onderbenen op een bankje lagen. In de tractiegroep verhoogde de fysiotherapeut de tractiekracht totdat de patiënt aangaf dat de tractie nog net comfortabel aanvoelde. De minimale kracht was 35% van het lichaamsgewicht en de maximale 50%. In de placebogroep verhoogde de fysiotherapeut de kracht vanaf 0 totdat de patiënt aangaf de tractie net te voelen. De maximale kracht was 20% van het lichaamsgewicht. De placebo tractie werd gegeven met behulp van een speciaal ontwikkelde gordel om het bekken waarbij er tijdens de tractie aan de rugzijde een

insnoereffect ontstond. Hierdoor hoopten we dat de patiënt eerder aangaf iets te voelen waardoor de tractiekracht in de placebogroep zo laag mogelijk kon blijven. In beide groepen bleef de ingestelde tractiekracht gedurende 20 minuten hetzelfde.

Naast de tractiebehandeling kregen beide groepen een boekje met voorlichting over rugklachten en tractie en instructies over houdingen en bewegingen. De fysiotherapeuten gaven aan de hand van dit boekje voorlichting over het omgaan met de rug. De patiënten werden gedurende 4 weken 3 keer per week behandeld.

Meten van het effect

De twee belangrijkste (primaire) effectmaten in dit onderzoek waren [1] het oordeel van de patiënt over het algeheel ervaren herstel van de rugklachten en [2] de ervaren hinder bij de 3 belangrijkste klachten van de patiënt. Bij het begin van het onderzoek koos elke patiënt zijn of haar belangrijkste klachten als gevolg van de rugklachten. Belangrijkste klachten zijn dagelijks terugkomende en niet te vermijden pijnlijke activiteiten of bewegingen.

Andere (secundaire) effectmaten waren: [3] functionele status gemeten met een vragenlijst (Roland disability questionnaire); [4] pijn tijdens de meting en gedurende de afgelopen week; [5] ernst van de rugklachten bepaald door de onderzoeks-fysiotherapeut; [6] beweeglijkheid van de rug; [7] werkverzuim en [8] extra behandelingen voor de rugklachten (medische consumptie). Het effect van de behandelingen werd 5 en 12 weken en 6 maanden na de loting gemeten.

Onderzoeksresultaten

Patiënten populatie

Tijdens de periode juni 1993 - december 1994 ging de onderzoeks-fysiotherapeut bij 243 patiënten na of ze voldeden aan de selectiecriteria. 92 patiënten (38%) konden niet meedoen. De belangrijkste redenen waarom ze niet mee konden doen waren: onvoldoende gemotiveerd (hieronder valt o.a.: geen tijd, vakantie, weigering in verband met de kans op placebo tractie) en minder dan 6 weken lage rugklachten. Uiteindelijk tekenden 151 patiënten het informed consent formulier en ondergingen de loting: 77 patiënten kregen echte tractie en 74 placebo tractie. Alle vooraf gemeten kenmerken (zoals leeftijd, geslacht en duur van de klachten) waren goed vergelijkbaar voor beide groepen. De loting was dus goed gelukt.

Tractiekracht en blindering

De gemiddelde tractiekracht werd berekend op basis van de individuele gemiddelde tractiekrachten (percentage van het lichaamsgewicht) over alle tractiebehandelingen. Dit was 42% voor de tractiegroep en 15% voor de placebogroep. De blindering van de patiënten voor de echtheid van hoge of lage dosis tractie was goed gelukt. In de placebogroep dachten slechts 4 patiënten (6%) dat ze placebo tractie hadden gehad en in de tractiegroep slechts 1 patiënt (1%). Van de patiënten dacht 74% van de patiënten in de placebo groep en 71% in de tractie groep dat ze echte tractie hadden gehad. En 25% van de patiënten in de placebo groep en 24% in de tractiegroep wist niet welke therapie ze gehad hadden.

Resultaten

Voor het beoordelen van de resultaten hebben we twee soorten analyses uitgevoerd. Ten eerste was er de "intention to treat" analyse, waarbij alle patiënten geanalyseerd werden in de groep waar ze door loting aan toegewezen waren, dus inclusief uitvallers en patiënten waarbij de behandeling niet volgens het voorschrift verlopen was. In tabel 1 staan de resultaten van deze analyse voor de belangrijkste effectmaten. Het bleek dat in beide behandelgroepen de rugklachten minder werden, maar voor alle effectmaten waren de verschillen tussen de groepen erg klein. Alle 95% betrouwbaarheidsintervallen voor de verschillen tussen de groepen bevatten de waarde 0. Dit betekent dat er geen statistisch significante verschillen waren tussen de groepen.

Tabel 1. Intention-to-treat analyse na 5 weken: verbetering en verschil tussen de behandelgroepen met 95% betrouwbaarheidsintervallen (BI).

Effectmaat	Tractie (n=77)	Placebo (n=74)	Verschil (95% BI) Tractie minus Placebo	P waarde
Algeheel ervaren herstel**	34 (44%)	37 (51%)	-7% (-23%; 9%)	0.42
Eerste belangrijkste klacht †	28.5	28.4	0.1 (- 9.0; 9.2)	0.99
Tweede belangrijkste klacht †	27.0	24.6	2.4 (- 6.8; 11.5)	0.61

* Lost to follow-up: 1 in placebo groep.

** Scores op de 7-puntsschaal werden gesplitst in verbeterd (geheel hersteld en veel verbeterd) en niet verbeterd (weinig verbeterd, niet veranderd, weinig verslechterd, veel verslechterd, slechter dan ooit).

† Gescoord op een 100 mm. visual analog schaal (VAS) (beste score 0, slechtste score 100).

De resultaten van het ervaren herstel gemeten op een 7-puntsschaal werden gesplitst in twee groepen: verbeterd (geheel hersteld en veel verbeterd) en niet verbeterd (iets verbeterd, niet veranderd, iets verslechterd, veel verslechterd en slechter dan ooit). Vijf van de patiënten (7%) in de tractiegroep en 1 patiënt (1%) in de placebogroep vulde in dat de rugklachten veel verslechterd waren of slechter dan ooit. Ongeveer de helft van de patiënten zei dat de rugklachten geheel of veel verbeterde: 44% in de tractiegroep en 51% in de placebogroep.

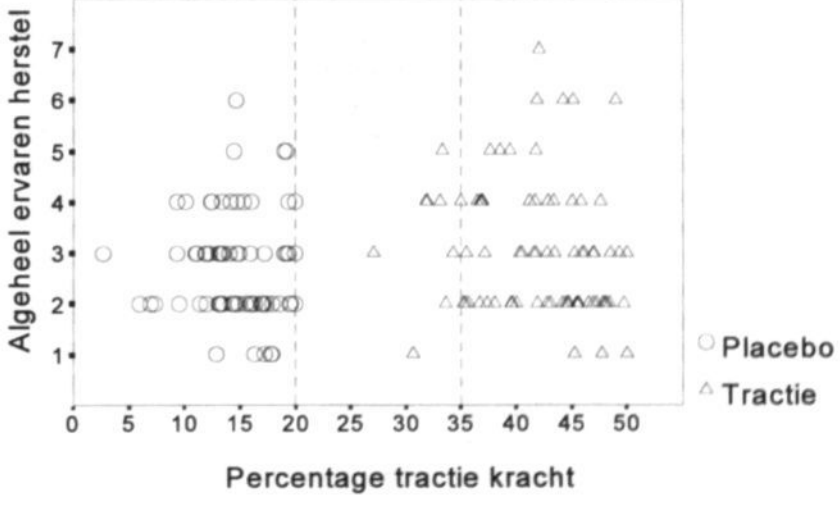
Bij het begin van het onderzoek vroegen we de patiënten om hun 3 belangrijkste klachten te selecteren. Een aantal patiënten kon maar 2 klachten selecteren. Daarom werd alleen het effect van tractie op de 2 belangrijkste klachten nagegaan. De moeite met het uitvoeren van de belangrijkste klachten was in beide groepen minder, maar er was geen verschil tussen de twee groepen.

Ook op de andere effectmaten verbeterden de patiënten in beide groepen veel, behalve de beweeglijkheid van de rug, hierop veranderden de patiënten niet. Het aantal bijwerkingen (bijvoorbeeld: hoofdpijn, 6 keer genoemd; 'raar' gevoel in benen, 9 keer genoemd) was gelijk verdeeld over de groepen. De resultaten van de metingen 12 weken en 6 maanden na de loting waren gelijk aan de resultaten na 5 weken. Er was geen statistisch significant verschil tussen de echte en placebo tractie.

De tweede analyse die we gedaan hebben is de "per protocol" analyse. In een per protocol analyse worden alleen patiënten opgenomen waarbij alles gegaan is zoals van tevoren gepland. Patiënten werden voor deze analyse uitgesloten als: [1] ze tijdens de behandelperiode uitvielen; [2] ze andere behandelingen dan tractie hadden gehad; [3] ze een tractiekracht lager dan 35% van het lichaamsgewicht in de tractiegroep hadden gehad; [4] er te veel tijd tussen de behandeldagen was geweest.

Bij de effectmeting 5 weken na de loting werd de per protocol analyse gedaan bij 135 patiënten: 66 in de tractie en 69 in de placebogroep. Een aantal patiënten werd om meer dan 1 van bovengenoemde redenen voor deze analyse uitgesloten. De resultaten van de per protocol analyse waren gelijk aan de resultaten van de intention to treat analyse: op alle effectmaten was er geen statistisch significant verschil tussen de echte en placebo tractie. Ook nu waren de resultaten van de per protocol analyses van de resultaten 12 weken en 6 maanden na de loting weer gelijk aan de resultaten na 5 weken. Er was geen verschil tussen de echte en placebo tractie.

Op basis van het mechanisch werkingsmechanisme van tractie werd een positief effect van hoge dosis tractie (hoger dan 35% van het lichaamsgewicht) verwacht vergeleken met lage dosis tractie (lager dan 20% van het lichaamsgewicht). Om na te gaan of de keuze van de afkappunten invloed had op het resultaat hebben we de tractiekracht uitgezet tegen het algeheel ervaren herstel. In de figuur is te zien dat er geen relatie bestaat tussen de tractiekracht als percentage van het lichaamsgewicht en ervaren herstel.



Figuur. Percentage tractie kracht van het lichaamsgewicht (20% = maximum voor placebo; 35% = minimum voor tractie) versus algeheel ervaren herstel (1=geheel hersteld; 7= slechter dan ooit). Er is geen relatie tussen de gegeven tractie kracht en effect.

Subgroep analyse
Om na te gaan of er groepen patiënten waren waarbij tractie meer effect had dan placebo, voerden we bij de gegevens van de meting 5 weken na de loting analyses uit in 8 kleinere groepen (subgroepen). Voor de analyse werden alleen de belangrijkste effectmaten gebruikt. Subgroepen werden gevormd op basis van de volgende kenmerken, allen gesplitst in twee groepen: [1] leeftijd (afkappunt 40 jaar); [2] geslacht; [3] duur van de huidige klachten periode (afkappunt 6 maanden); [4] uitstraling in het onderbeen (ja/nee); [5] score bij het begin van het onderzoek op een algemene gezondheidsvragenlijst (General Health questionnaire) (afkappunt 11 punten); [6] ernst van de belangrijkste klacht (afkappunt 70 mm op de visual analog schaal); [7] geschiktheid voor tractie volgens de behandelend fysiotherapeut (ja/nee); [8] behandeling in fysiotherapiepraktijk of polikliniek van ziekenhuis. In geen van de subgroepen vonden we een statistisch significant effect van tractie vergeleken met placebo.

Discussie en conclusie

De centrale vraagstelling van dit onderzoek was of tractie een effectieve behandeling is voor patiënten met minstens 6 weken a-specifieke lage rugklachten. De resultaten waren erg duidelijk: op alle effectmaten waren de verschillen tussen tractie en placebo tractie erg klein en niet statistisch significant. In beide behandelgroepen ging ongeveer de helft van de patiënten vooruit, maar deze vooruitgang kon niet verklaard worden door de gegeven tractiekracht (specifieke effect van tractie).

De grote vooruitgang van de patiënten is opmerkelijk: de helft van de patiënten zei dat de rugklachten veel verbeterd of geheel hersteld waren. De patiënten hadden minimaal 6 weken last van hun rug. Dit is een groep patiënten bij wie we zonder effectieve behandeling weinig vooruitgang verwachtten. Behalve dat deze vooruitgang een placebo effect zou kunnen zijn, zijn ook nog andere redenen waarom mensen beter worden na een ineffectieve behandeling, zoals spontaan herstel.

In dit onderzoek konden we de tekortkomingen van eerder onderzoek voorkomen. Bijvoorbeeld: de vergelijkbaarheid van de groepen bij het begin van het onderzoek (baseline) was goed, de patiënten en de onderzoeks-fysiotherapeut waren geblindeerd voor de behandeling en de effectmaten waren relevant. Wij denken dat op basis van deze studie een geldige uitspraak over het effect van tractie mogelijk is. We kunnen ons afvragen waarom we geen verschil vonden tussen tractie en placebo tractie. Een aantal verklaringen is mogelijk.

Een reden zou kunnen zijn dat de keuze van de afkappunten voor de tractie-krachten verkeerd is. Diverse auteurs beschreven dat een bepaalde tractiekracht nodig is om mechanische effecten te krijgen. We dachten dat een tractiekracht van 35-50% van het lichaamsgewicht effectief zou zijn en een kracht van minder dan 20% niet. Het had kunnen zijn dat het contrast tussen de groepen niet groot genoeg was of dat de lage dosis tractie geen echte placebo was. Maar we zagen dat er helemaal geen relatie was tussen de toegediende tractiekracht en het effect van tractie (figuur). Met andere woorden, het effect van tractie was niet afhankelijk van de hoeveelheid tractiekracht.

Een andere reden zou kunnen zijn dat we alleen patiënten geselecteerd hebben met a-specifieke lage rugklachten. Omdat er veel onenigheid en onduidelijkheid bestaat bij het stellen van een juiste diagnose van lage rugklachten, hebben we geen onderscheid gemaakt tussen patiënten met problemen van de discus, tussenwervelgewrichten of spieren. Om na te gaan of bepaalde subgroepen baat hadden bij tractie, hebben we een aantal subgroep analyses gedaan. We beperkten deze subgroepen tot brede patiënten categorieën, en niet tot patiënten met een bepaalde mogelijke diagnose. Maar de subgroep analyse liet geen groep zien die baat had bij tractie. We hebben geen patiënten geselecteerd met acute klachten of met een hernia. Op basis van dit onderzoek kunnen we niet zeggen dat tractie niet helpt bij deze groepen, maar de kans is klein omdat hetzelfde werkingsmechanisme verondersteld wordt.

Het kan zijn dat er een positief effect was geweest als tractie samen met oefeningen was gegeven. De oefeningen zouden dan het effect van tractie kunnen versterken. Omdat we geen enkel effect van tractie vonden is het onwaarschijnlijk dat tractie wel helpt in combinatie met oefeningen.

Samenvattend kunnen we zeggen dat op basis van dit onderzoek een duidelijke uitspraak mogelijk is: continue elektrische tractie is niet effectief bij patiënten met minimaal 6 weken a-specifieke lage rugklachten. Ongeveer de helft van de patiënten ging vooruit ongeacht welke behandeling ze kregen. Maar deze vooruitgang kon niet verklaard worden door het specifieke effect van tractie. Volgens ons heeft toekomstig onderzoek naar het effect van deze vorm van tractie of naar andere vormen of bij andere patiëntengroepen geen hoge prioriteit meer.

Deel 2

De hoofdstukken 5, 6 en 7 gaan niet over het effect van tractie. In deze hoofdstukken wordt ingegaan op "bijprodukten" van het onderzoek. Ze gaan over het meten van de functionele status. Met functionele status bedoelen we het vermogen van een persoon om zijn dagelijkse activiteiten uit te kunnen voeren. Het doel van veel therapieën is het verbeteren van het functioneren van patiënten. Functionele status is daarom een belangrijke effectmaat voor onderzoek op het gebied van rugklachten.

Om na te gaan wat er in de literatuur bekend was over vier bekende vragenlijsten die functionele status bij rugklachten meten, hebben we een literatuuronderzoek uitgevoerd. De vragenlijsten zijn op diverse punten beoordeeld o. a. op de bruikbaarheid, of ze bij herhaalde metingen dezelfde resultaten opleveren (betrouwbaarheid), of ze meten wat ze beogen te meten (validiteit) en of ze in staat zijn klinisch relevante veranderingen in de tijd te meten (responsiviteit). De resultaten van dit literatuuronderzoek staan beschreven in hoofdstuk 5.

In het onderzoek naar het effect van tractie zijn verschillende vragenlijsten voor het meten van functionele status gebruikt. Hierdoor was het mogelijk de kwaliteit van deze vragenlijsten te beoordelen en te vergelijken in dezelfde patiëntengroep. In hoofdstuk 6 beoordelen en vergelijken we de responsiviteit van drie vragenlijsten om de functionele status te meten en de effectmaat pijn. In hoofdstuk 7 presenteren we een patiënt-specifieke vragenlijst om functionele status bij rugklachten te bepalen. Hierbij selecteert elke patiënt op een gestandaardiseerde wijze zijn of haar belangrijkste klachten. We evalueerden de bruikbaarheid en responsiviteit van deze

vragenlijst. Het bleek dat deze vragenlijst het even goed deed als andere veel gebruikte lijsten.

Tot slot staat in hoofdstuk 8 een algemene discussie over het onderzoek dat in dit proefschrift is beschreven. Het hoofdstuk is verdeeld in twee delen. Het eerste deel beschrijft een aantal methodologische aspecten. Speciale aandacht is besteed aan de deelname van fysiotherapeuten en patiënten aan het onderzoek, de placebo behandeling, de grootte van de vooruitgang in beide groepen, de interpretatie van resultaten van pilot onderzoeken en het gebruik van de functionele status als effectmaat. In het tweede deel kijken we terug op het effect van tractie.

Dankwoord

Op 1 november 1992 ging het driejarig onderzoek naar het effect van tractie van start. Drie jaar leek een lange tijd. Maar ze zijn omgevlogen. Het waren drie leuke en leerzame jaren, wat met name kwam door de medewerking van veel mensen.

Zonder patiënten kan effectonderzoek niet uitgevoerd worden. Allereerst wil ik daarom alle 151 patiënten bedanken voor hun medewerking aan de metingen. Zonder hun geduld om ongeveer 3 uur vragen te beantwoorden en onderzocht te worden, was van dit onderzoek weinig terechtgekomen.

Een belangrijke factor in het onderzoek was de goede samenwerking met de aan het onderzoek deelnemende fysiotherapeuten. Frans Philippens, Wilma Simonis, Willy van Baal, Thijs Belgers, Richard Hoen, Jan van Beurden, Colinda Gulikers, Jos Coenjaerts, Tjeu Dols, Roger Maessen, Wim Lahaye, Wim Schmetz, Rob Valkenburg, Henk Creusen, Caroline van der Velde en Armond Hamelers bedankt voor jullie inzet! Jullie dachten mee over de opzet van het onderzoek en hebben de patiënten geworven en behandeld. Ik respecteer het dat jullie, in deze voor de fysiotherapie woelige tijden, een onderdeel van jullie vak durfden te evalueren.

Alle belangrijke beslissingen tijdens de voorbereidingsfase en later bij het schrijven van de artikelen werden genomen in de projectgroep. In de voorbereidingsfase van het onderzoek kwam deze groep bijna wekelijks bij elkaar. Onze vaste vergadertijd was 17.15 uur. Stukken werden steeds zorgvuldig besproken en van commentaar voorzien. Als beginnend onderzoeker heb ik hier veel van geleerd. Door jullie medewerking aan de strakke tijdsplanning verliep alles volgens schema.
Wiel Regtop, als man uit de praktijk wist je altijd zeer relevante vragen te stellen. Eline Lindeman, ondanks dat je druk bezig was met het afronden van je promotie heb je altijd tijd gevonden om commentaar te geven en naar de projectgroep bijeenkomsten te komen. Geert van der Heijden, jij stond aan het begin van dit onderzoek. Ik vind het leuk dat je nu ook het einde van dichtbij meemaakt. Geert Walenkamp en Mark de Krom, als adviseurs van de projectgroep gaven jullie tijdens de voorbereidingsfase op deskundige wijze advies over cruciale onderdelen van het onderzoek.

Ook de leden van de begeleidingscommissie (bestaande uit: Johan Peeters, André Knottnerus, Lex Bouter en Gerard Engel) wil ik bedanken voor het bewaken van de voortgang van het onderzoek.

Paul Knipschild, mijn promotor, je was op de achtergrond aanwezig. Van jou heb ik geleerd dat logisch nadenken in de epidemiologie erg belangrijk is.

De bijdrage van Riekie de Vet, mijn co-promotor, was groot. Van begin tot eind was je de ideale "werkbaas". Ik ben de eerste promovendus die je "aflevert". Ik hoop dat

er nog veel mogen volgen. De enthousiaste en deskundige begeleiding is mij prima bevallen. Je was heel nauw betrokken bij de dagelijkse gang van zaken. Je liet me voldoende zelfstandig werken maar wist ook op tijd in te grijpen om, indien nodig, bij te sturen. Supersnel heb je telkens weer mijn artikelen van goed en bruikbaar commentaar voorzien. Het werk hebben we regelmatig afgewisseld met onze ontspannende sportieve bezigheden: fietsen en schaatsen. Riekie, ik hoop dat we in de toekomst nog veel samen kunnen werken.

Albère Köke, mijn maatje in onderzoeksland. Ik wil je bedanken voor de gezellige uren. Als onderzoeksassistent heb je de metingen bij bijna alle patiënten uitgevoerd. Hiervoor heb je heel wat kilometers afgelegd. Het was steeds weer spannend of nieuw aangemelde patiënten mee wilden doen aan het onderzoek. Het zat niet altijd mee. Dat begon al met de eerste patiënt, die hebben we na één behandeling nooit meer teruggezien. De uitdrukking "een goed begin is het halve werk" ging bij ons dus niet op.
In alle fasen van het onderzoek heb je kritisch meegedacht en bijgedragen aan de voortgang van het project. Vooral jouw praktijkervaring was van grote waarde. Albère, ik hoop dat het einde van dit project niet het einde van onze samenwerking is.

Met veel plezier heb ik de afgelopen jaren bij de vakgroep Epidemiologie gewerkt. Collega's, met name Ingeborg Poorterman en Jeanne van Loon, hebben gedurende anderhalf jaar bijna dagelijks moeten aanhoren hoeveel patiënten er nu weer waren aangemeld, afgewezen of geselecteerd. Bedankt voor jullie luisterend oor.
Met mijn fysiotherapie-collega's: Rob de Bie, Petra Sijpkes, Geert van der Heijden, Arianne Verhagen en Pieter Wolters had ik vaak boeiende gesprekken over onderzoek op het terrein van de fysiotherapie. Mede dankzij het uitwisselen van ervaringen is het leuk om onderzoek te doen.
Cobie Martens was als administratief medewerker betrokken bij het project. Bedankt voor de zeer secure invoer van de gegevens in de computer. Bij Annemie Mordant kon ik altijd terecht voor vragen over SPSS. Thum Aarts ben ik dankbaar voor de deskundige wijze waarop hij de lay-out van het proefschrift heeft verzorgd. Van Bob Wilkinson heb ik geleerd hoe ik, op een voor iedereen duidelijke manier, in het Engels moet schrijven en presenteren.

Mijn ouders hebben me altijd onvoorwaardelijk gesteund in wat ik wilde doen. Bedankt voor jullie steun, interesse en belangstelling. Last but not least, Jan, ik weet dat je hier niet genoemd wilt worden. Maar zonder jou te noemen is het dankwoord niet compleet!

Curriculum Vitae

Sandra Beurskens werd geboren op 22 april 1966 te Roermond. Zij volgde de MAVO aan De Kreppel te Heythuysen en de HAVO aan de Scholengemeenschap St. Ursula te Horn.
Van 1984 tot 1988 studeerde zij Fysiotherapie te Heerlen. Daarna startte zij in 1988 met de studie Gezondheidswetenschappen aan de Rijksuniversiteit Limburg, waar zij in 1992 afstudeerde met als afstudeerrichting Gezondheidsvoorlichting en -opvoeding.
Tijdens haar studie Gezondheidswetenschappen was zij part-time werkzaam in een praktijk voor fysiotherapie en vervulde zij een student-assistentschap bij de vakgroep Epidemiologie van de Rijksuniversiteit Limburg te Maastricht. Sinds november 1992 is zij werkzaam als toegevoegd onderzoeker bij de vakgroep Epidemiologie, waar zij werkte aan dit proefschrift.