Quality assessment of randomised clinical trials

Quality Assessment of Randomised Clinical Trials

Arianne P Verhagen

ISBN: 9052782520

Subject headings: methodology / quality assessment / randomised clinical trials / review / systematic review / meta-analysis.

Lay-out: Arianne Verhagen, Cobie Martens, UM Epidemiologie, Maastricht
Cover: calligrafie van Jo Weyden, Wessem
Production: Datawyse

Quality Assessment of Randomised Clinical Trials

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof dr. A.C. Nieuwenhuijzen Kruseman
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op vrijdag 1 oktober 1999 om 14.00 uur

door

Arianne Petra Verhagen

UM
UNIVERSITAIRE PERS MAASTRICHT

Promotor:	Prof.dr.ir. P.A. van den Brandt
Co-promotor:	Dr.ir. H.C.W. de Vet
Beoordelingscommissie:	Prof.dr. J.A. Knottnerus (voorzitter) Prof.dr. L.M. Bouter (Vrije Universiteit Amsterdam) Prof.dr. J.P.M. Kleijnen (University of York, York, UK) Prof.dr. J.M.J.P. van der Linden Dr. F.E.S. Tan

This study was performed at the Maastricht Research Institute for Extramural and Transmural Health Care, which participates in the Netherlands School of Primary Care Research (CARE), acknowledged in 1995 by the Dutch Academy of Science (KNAW).

Deze uitgave is mede tot stand gekomen door subsidie van: het Nationaal Rheumafonds, de Stichting Opleiding Manuele Therapie (SOMT) en het McKenzie Instituut Benelux.

Contents

Introduction

If you realize that all things change,
there is nothing you will try to hold on to.
If you aren't afraid of dying,
there is nothing you can't achieve.

Trying to control the future
is like trying to take the master carpenter's place.
When you handle the master carpenter's tools,
chances are that you'll cut your hand.

Lao-Tse

After working as a physiotherapist in private practice for over ten years, I felt the need for a change. I enjoyed taking care of patients, but after ten years I found myself searching for a new challenge. So I sold my practice and applied to Maastricht University as an MSc student in Health Sciences. This was the first major change in my life, which evoked many more.

During my first years at the university I had no intention of going into research. I considered myself too pragmatic for such theoretical work. At that time I just liked studying and I had no career plans for the future.

After I graduated, people asked me to apply for a new PhD-project at the Department of Epidemiology at Maastricht University. This project carried the title *Development of the methodology of reviewing the literature; meta-analysis in the physiotherapy field.* I had never considered myself a researcher, much less one dealing with such abstract theoretical concepts, but I gladly accepted this new challenge. I respected and liked the other members of the project team and this was a very important factor in my decision to join the team. This was a second major change and initiated the actual start of this thesis.

Before presenting my research, I should first provide some background information. In the field of health care, Randomised Clinical Trials (RCTs), are the scientific tool for answering the question: "What is the efficacy of a specific treatment in patients with a certain disease or disability?". High quality trials may be considered a valid measure of treatment efficacy, low quality trials on the other hand, may not.

A systematic review or meta-analysis summarizes the results of the individual RCTs in a systematic way. This benefits health care providers who no longer need to read all RCTs in order to find out which treatment is best. Also patients benefit because they are more likely to receive the best treatment available at that time.

It stands to reason that one should select only RCTs of sufficient methodological quality for a systematic review. Therefore it is important to identify which are the methodological aspects that reflect on the quality of RCTs. In order to accurately assess whether an RCT is of high (or moderate or low) quality, reviewers or researchers use criteria lists, also called checklists or quality scales.

Initially I performed systematic reviews incorporating methodological issues within the field of physiotherapy. My first project focussed on the efficacy of balneotherapy (i.e. spa-therapy or hydrotherapy) in patients with arthritis. At the same time I studied the method used to assess the methodological quality of the RCTs included in this review. Chapter 1 of this thesis presents the systematic review itself.[1] In this review I used the Maastricht list[2] as criteria list to assess methodological quality. Chapter 2 comprises the results concerning the reliability of the Maastricht list as method of quality

assessment.[3]

During this first project, my fellow researchers and I realized that the Maastricht criteria list might not be a valid tool to assess the methodological quality of RCTs. The list was developed by people working at the Maastricht University, based on accepted methodological criteria as presented in textbooks. Initially, I did not question the validity of the Maastricht list as method of quality assessment. But at that time my initial belief in the validity of the Maastricht criteria list changed into uncertainty.

The Maastricht list consisted merely of a set of guidelines, which had not been rigorously tested. This prompted us to develop a new criteria list based on firmer scientific principles. National and international experts cooperated with us in this research. Using the Delphi consensus technique we reached agreement on a new criteria list.[4] Because of the consensus technique used, we called it the Delphi criteria list. Chapter 3 describes the Delphi research and presents the resulting Delphi list.

Next, we performed two systematic reviews within the field of physiotherapy, and compared the performance of the Delphi list with two other lists. Apart from the Maasticht list and the Delphi list, both described before, we also used the so-called Jadad list[5], named after AR Jadad who developed this list together with his colleagues. Chapter 4 presents the review on the efficacy of 904 nm laser therapy in musculoskeletal disorders. Overall the study quality of the studies included was low (varying from very poor to reasonable). Chapter 5 focusses on the qualitative differences between the three criteria lists. Chapter 6 presents the review of conservative treatments in acute lateral ankle sprains, and chapter 7 presents a calculation of the differences between the three criteria lists and its influence on the conclusion of the review. We evaluated the overall quality scores (which are similar to report marks) as well as how the individual criteria influenced the conclusion of the review.

During these studies our view on the impact of quality assessment changed. A commonly accepted paradigm states that studies of poor quality tend to overestimate the treatment effect. But this paradigm assumes unanimous agreement on what constitutes high or low quality. Using these three different criteria lists we seldom found that the lists produced the same conclusion on quality. Furthermore, we did not find good evidence that 'low quality' studies indeed overestimated the treatment effect.

We assumed our problems to be specific to the field of physiotherapy. Few RCTs have been conducted on physiotherapy treatments and most of them are considered to be of low to moderate quality. So, we changed and widened the scope of this PhD project and included a repetition of an existing review in the field of cardiology. The original review in 1985 of Salim Yusuf and colleagues studied the efficacy of thrombolytics in patients with acute myocardial infarction (MI).[6] This review contained no quality assessment, yet the results of this review had a major impact on the treatment of thrombolytica of people suffering from acute MI. What would have happened if some form of quality assessment had been incorporated into the final conclusions of this review? Would the conclusion of this review have been different if it had been based only on high quality studies? Therefore we set out to investigate this. We performed a quality assessment on the same set of studies as Salim Yusuf and colleagues had done. Chapter 8 presents the results of this research. We found that our review, which did include quality assessment, reached the same conclusion as the review that did not use quality assessment. Such results give one pause to think about the role of quality assessment in systematic reviews. The variety of reasons for and against quality assessment are presented in Chapter 9. During the years our view on the impact of quality on the results of trials, and quality assessment itself changed several times. New research should in our opinion focus on components of the methodological quality

and their influence on the conclusions of a review.

References

1. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Knipschild PG. Taking baths: the efficacy of balneotherapy in patients with arthritis. A systematic review. *J Rheumatol 1997*;24(10):1964-71.
2. de Vet H, de Bie R, van der Heijden G, Verhagen A, Sijpkes P, Knipschild P. Systematic Reviews on the Basis of Methodological Criteria. *Physiotherapy 1997*; 83(6):284-289.
3. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Knipschild PG. Balneotherapy and quality assessment: interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. *J Clin Epidemiol 1998*;51(4):335-41.
4. Verhagen AP, de Vet HCW, de Bie RA, et al. The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus. *J Clin Epidemiol 1998*;51(12):1235-1241.
5. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials 1996*;17:1-12.
6. Yusuf S, Collins R, Peto R, Furberg C, Stampfers MJ, Goldhaber SZ, Hennekens CH. Intravenous and intracoronary therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J 1985*;6:556-85.

1 Taking Baths.

The efficacy of balneotherapy in patients with arthritis: a systematic review.

AUTHORS:
Arianne P. Verhagen,
Henrica C.W. de Vet,
Robert A. de Bie,
Alphons G.H. Kessels,
Maarten Boers,
Paul G. Knipschild.

ABSTRACT

Objective. To review English, French, German and Dutch language studies of the effectiveness of balneotherapy. Balneotherapy (hydrotherapy or spa-therapy) is one of the oldest forms of therapy for patients with arthritis. One of the aims of balneotherapy is to relieve the pain.

Method. We performed a systematic review that included randomized and non-randomized studies. Quality scores of the studies were determined using a criteria list.

Results. Most studies report positive findings, but all studies showed methodological flaws. A quality of life measurement was never reported as an outcome measure. None of the randomized clinical trials included intention-to-treat analysis or comparison of effects between groups.

Conclusion. Because of the methodological flaws a conclusion about the efficacy of balneotherapy cannot be provided from the studies we reviewed. We conclude that most flaws found could be avoidable in future research. (*J Rheumatol 1997;24:1964-71*)

Bathing in water (balneotherapy or spa therapy) has been frequently, widely and enduringly used in classical medicine as a cure for diseases. Water from mineral and thermal springs was particularly valued.[1] In Homeric times baths were applied primarily to cleanse and refresh. By the time of Hippocrates bathing was regarded as more than a simple hygienic measure. It was considered beneficial to cure most illnesses.[1-3] The Romans too used water for the therapeutic treatment of orthopaedic conditions.[1,2,4]

After the Roman era spa-therapy fell into disuse, but in the sixteenth century baths were rediscovered.[5] Since then spa therapy has been practised continuously in the management of musculoskeletal conditions.[6-8]

Spa therapy is a very popular form of treatment for all forms of arthritis in many European countries and in Israel.[4,9,10] Spa-therapy often takes place at centres with thermal baths or sea water baths in most European countries.[11] Some consider it as a special form of physiotherapy.[6] In Israel, the main health resort area is located along the western shore of the Dead Sea. The unique environmental conditions are considered beneficial to patients with rheumatic diseases.[7]

The term balneotherapy comes from the Latin balneum (bath). The term is classically used for bathing in thermal or mineral waters, and has been distinguished from hydrotherapy, but since the beginning of this century both terms have been accepted for all forms of treatment with water.[12] We will use the term balneotherapy, since bathing for therapeutic use very often happens in spas. The water (thermal water, sea water, or tap water) is generally used at a temperature of around 34° C.[13] The hydrostatic force (Archimedes' principle) brings about a relative pain relief by reducing loading[13]; the water reduces the forces of gravity acting on painful and rheumatic joints.

The aim of balneotherapy is to improve the range of joint movements, cause muscle strengthening, relieve muscle spasm, maintain or improve functional mobility, soothe pain and as a consequence to relieve patients' suffering and let them feel well.[2,4,7,9] Sometimes bathing is combined with exercise treatment, but it is not always clearly described in textbooks whether it should be combined. Balneotherapy is most often prescribed for patients with all forms of arthritis, of which psoriatic (PsA) and rheumatoid arthritis (RA) are the most frequent. It is repeatedly noted that the development of rheumatology as a science began at the spas.[4,14]

The effectiveness of balneotherapy in the management of patients with arthritis is subject to considerable debate.[4,15] Some authors attribute the effectiveness of balneotherapy to physiological changes like increased diuresis, haemodilution and reduced rheumatoid factor.[7,16] Others think the effectiveness is due more to biomechanical changes like joint unloading, relaxation, increased muscle function and increased general condition.[8,17]

Apart from the debate about the possible physiological and biological changes there is also discussion about the indications for balneotherapy. According to Cosh[11], a patient with "most active and widespread form of rheumatoid arthritis is best treated in hospital, but when in remission a period at a thermal spa is helpful". On the other hand, Sukenik and colleagues[18-20] conclude from studies in the Dead Sea area that balneotherapy is safe and effective in patients with active RA and PsA.

We present a systematic review on the efficacy of balneotherapy in patients with arthritis. Nearly half of the studies concerned patients with RA. Based on the methodological quality of the studies we will discuss the efficacy of balneotherapy.

Materials And Methods

Identification and selection of studies. Studies were found by screening the Medline database 1966-1995 and the database from the Cochrane Field 'Rehabilitation and Therapy',

which also contains studies published in journals not covered by Medline. Reference checking and personal communication with authors was carried out to retrieve eligible studies. This search strategy was according to the recommended search strategy of the Cochrane Collaboration.[21] Using this strategy we were convinced the literature search was close to complete.

Key-words to identify the studies were: *balneotherapy, hydrotherapy, spa therapy, thalasso therapy, water therapy, arthritis, randomized clinical trial (RCT), clinical trial, experiment, evaluation study, comparative study, controlled study* and the use of the term *efficacy, effectiveness* or *effect* in the title.

To perform an adequate assessment of the methodological quality, the language of the publications had to be in English, French, German or Dutch. We did not translate studies from other languages.

Studies were eligible if they used an experimental design [randomized (RCT) or non-randomized (CCT)] and if the patients included had a form of arthritis. Balneotherapy had to be one of the interventions under study. Outcome measures also were a selection criterion. The World Health Organization (WHO) and the International League Against Rheumatism (ILAR) determined in 1992 a core set of 8 endpoints for clinical trials concerning patients with RA[22], these are listed in Table 1.

Table 1: WHO/ILAR core set of endpoints for RA clinical trials.[21]

	Endpoints
1	Pain
2	Patient global assessment
3	Physical disability
4	Swollen joints
5	Tender joints
6	Acute phase reactants
7	Physician global assessment
8	Radiographs of joints (in studies of 1 or more years' duration)

At least one of these endpoints had to be among the main outcome measures. Studies were excluded when only laboratory variables were reported as outcome measures.

Assessment of the methodological quality. The criteria to assess methodological quality were those developed at the Department of Epidemiology at the Maastricht University, The Netherlands. [23,24] (Table 2)
In this review the quality scores of the studies were determined independently by 2 authors (HdV,RdB) followed by a consensus meeting.

A total of 100 points can be obtained for the RCT, divided over 5 categories. For the CCT a maximum of 80 points is obtainable, because 15 points are given for a random treatment allocation, and a maximum of 5 points can be obtained for a blinded randomization procedure (3 points) and an adequate description of it (2 points).

RESULTS

Characteristics of the studies. The literature search resulted in 37 studies; 25 were found in Medline, 7 in the Cochrane database, 3 by reference checking and 2 by contacting authors. Two of the articles in our own database appeared in Medline also, but were not identified by the keywords used during our Medline search. For various reasons 23 studies were excluded; are listed in Table 3.

Of the studies not written in English, French, German or Dutch (n=19), 9 contained English abstracts. One of these abstracts[42] contained sufficient information about the study design and this study appeared to be not eligible.

In total, 14 experimental studies that met the eligibility criteria were found on the efficacy of balneotherapy in patients with arthritis. Studies of patients with definite or classical RA as defined by the American Rheumatism Association (ARA) criteria[48] or by Steinbrocker criteria[49] are regarded as a separate group.

Therefore, we divided the studies into two groups: Group 1, studies with patients with RA as defined by the ARA or Steinbrocker criteria, and Group 2, patients with other forms of arthritis.

Table 2: Maastricht criteria list for methodological quality assessment (the complete list and users guide is available upon request).[23]

Domain	Main items	Number of subitems	Weight
STUDY POPULATION	selection and restriction	2	2 points
	treatment allocation	3	20 points
	study size	3	10 points
	prognostic comparability	5	5 points
	drop outs	4	7 points
	loss to follow-up	3	7 points
			Total 51 points
INTERVENTION	experimental and control interventions	6	12 points
	extra treatments	2	2 points
			Total 14 points
BLINDING	blinding of patient	2	6 points
	blinding of therapist	2	6 points
	blinding of observer	2	6 points
			Total 18 points
OUTCOME	outcome measures	5	5 points
	follow-up period	3	3 points
	side effects	1	1 point
			Total 9 points
ANALYSIS	analysis and data presentation	4	8 points
			Total 8 points
TOTAL	15	47	Total 100 points

Table 3: Excluded studies.

Reasons for Exclusion	Number	Language
Language inappropriate	19	11 Russian [25-35]
		3 Czech [36-38]
		1 Romanian [39]
		1 Polish [40]
		2 Hebrew [41,42]
		1 Japanese [43]
Outcome measures inappropriate	1	German [44]
Not RCT or CCT	3	1 Dutch [45]
		1 English [46]
		1 French [47]
Total	23	

Table 4: Characteristics of RCTs of balneotherapy in patients with classical or definite RA.

Study	Sukenik et al 1990a	Sukenik et al 1990b	Elkayam et al 1991	Sukenik et al 1995
Number of patients	30	40	41	36
Interventions	I: Dead Sea salt baths (n=15) II: Sodium Chloride baths (n=15)	I: Mud packs (n=10) II: Sulphur baths (n=10) III: I and II (n=10) IV: control (n=10)	I: Mineral baths and mud packs (n=19) II: Tap water baths (n=22)	I: Dead Sea baths (n=9) II: Sulphur baths (n=9) III: I and II (n=10) IV: control (n=8)
Main outcome measures	Duration of morning stiffness 15 m walk time Hand grip strength Joint circumference Activities of daily living Patient assessment of severity of disease Number of active joints Ritchie index Laboratory variables	Duration of morning stiffness 15 m walk time Hand grip strength Joint circumference Activities of daily living Patient assessment of severity of disease Number of active joints Ritchie index Laboratory variables	Duration of morning stiffness 15 m walk time Hand grip strength Ritchie index Patient assessment of severity of disease Physician assessment of severity of disease Laboratory variables	Duration of morning stiffness 15 m walk time Hand grip strength Activities of daily living Patient assessment of severity of disease Number of active joints Ritchie index
Blinding	Patient / Observer	Observer	Patient	Observer
Follow-up	3 months	3 months	12 weeks	3 months
Efficacy according to the authors	Improvements in group I	Improvement was observed in the three treatment groups	No conclusion	Improvement was observed in the three treatment groups

Recently, Steinbrocker criteria were revised by the American College of Rheumatology (ACR)[50], but the original criteria were used in the studies found.

Group 1. Eight studies were found, of which 2 were written in German[51,52], 2 in Dutch[53,54] and 4 were of Israeli origin although written in English.[15,18,20,55] The main characteristics of the studies are summarized in Table 4 (RCT) and Table 5 (CCT).

All RCT listed in Table 4 were performed recently in Israel, 3 by the same research group. The number of patients in the intervention groups varied from 8 to 22. Two studies included a control group receiving no treatment. In all studies the intervention consisted of mineral baths, often in combination with mud packs. All patients continued their medication during balneotherapy. No extra exercise therapy is reported. All the studies used a number of outcome measures including pain and function, but no quality of life measures were used. All studies had blinded patients or observers or both, and the follow-up periods were comparable (roughly 3 months).

Table 5 shows the characteristics of the non-randomized group experimental studies (CCT). They were performed in Germany[51,52] and The Netherlands.[53,54] Two studies were published in the 1970s[51,53] and the other 2 in the 1990s.[52,54] The number of patients in each intervention group varied between 10 and 27. In the 2 Dutch studies extra exercise therapy was mentioned in both study groups. There were a smaller number of outcome measures studied in all CCT compared to the randomized studies (Table 4).

Table 5: Characteristics of CCTs of balneotherapy in patients with classical or definite RA.

Study	Günther et al 1976	Steiner et al 1979	Svarcová et al 1990	Landewé et al 1992
Number of patients	20	24	45	46
Interventions	I: Radon thermal bath (n=10) II: Tap water (n=10). Cross-over	I: hydrotherapy + exercises + electrotherapy (n=12) II: thermal bath + massage + exercises (n=12) Cross-over	I: Whirlpool (n=15) II: Low-high air pressure mass (n=15) III: Mud packs (n=15)	I: Thermal bath (n=27) II: Tap water bath (n=19)
Main outcome measures	Hand grip strength Joint circumference Range of motion Pain	Laboratory variables Duration of morning stiffness Radiographs Pain Joint circumference Number of active joints Range of motion Activities of daily living	Pain Ankle mobility Physician's assessment of treatment effect	Duration of morning stiffness Laboratory variables Ritchie index Pain Activities of daily living
Blinding	None	None	None	None
Follow-up	No	1 year	No	No
Efficacy according to the authors	Statistical improvement of both therapies, no significant difference in effect	Treatment in the thermal bath showed a marked improvement	Best improvement in groups I and III	Improvement of both therapies, no difference in effect

Blinding of patient or observer was not reported and no CCT used quality of life as an outcome measure.

Group 2. Six studies were found, of which 3 were randomized and 3 nonrandomized, all in English. The main characteristics of the studies are summarized in Table 6. Table 6 shows that the number of patients per study varied between 12 and 166, and in the intervention groups from 6 to 146. In the study by Sukenik[19] the control group is small (n=20) compared to the study group (n=146). Both groups had the regular regimen of bathing in the Dead Sea and exposure to the sun. The control group received no additional therapy. In 2 studies mineral baths were used[19,55] and in 3 studies bathing was combined with exercise treatment.[56-58] There are differences in main outcome measures among the 6 studies. In the RCT especially the patient was asked to fill in questionnaires concerning their pain or activities of daily living (ADL). In none of these studies was quality of life used as an outcome measure.

Methodological quality. The results of the assessment of methodological quality are shown in Table 7. The scores are assigned to the 5 categories and the total scores are presented for each study.

All RCT scored between 32 and 49 points, and the CCT between 10 and 29 points, indicating an overall poor methodological quality. A considerable amount of information about design or conduct of the studies was lacking or unclear. All RCT scored 15 points for a random treatment allocation; however, 2 studies[51,60] gained 2 additional points by describing the procedure.

Table 6a: Characteristics of RCTs of balneotherapy in patients with other forms of arthritis.

	Randomized studies		
Study	Nicholls et al 1990	Sylvester 1990	Green et al 1993
Form of arthritis	'Rheumatic diseases'	OA hips	OA hip
Number of patients	30	14	47
Interventions	I: Hydrotherapy (n=22) II: control (n=8)	I: Hydrotherapy + exercises (n=7) II: Electrotherapy + exercises (n=7)	I: Home exercises (n=23) II: Hydrotherapy + home exercises (n=24)
Main outcome measures	Perceived self efficacy Range of motion Pain Stiffness	Pain Function Life satisfaction Range of motion Gait	Pain Overall change scores Range of motion Muscle strength Functional tests
Blinding	None	Observer	Observer
Follow-up	6 weeks	6 weeks	18 weeks
Efficacy according to the authors	The data suggest benefits of hydrotherapy.	Best improvement in group I	Improvement of both groups, no significant difference in effect between the interventions

Table 6b: Characteristics of CCTs of balneotherapy in patients with other forms of arthritis.

	Non-randomized studies		
Study	Baldwin 1972	Szucs et al 1989	Sukenik et al 1994b
Forms of arthritis	Juvenile RA	Inflammatory arthritis	PsA
Number of patients	12	62	166
Interventions	I: Hydrotherapy (n=6) II: Home exercises (n=6)	I: Thermal baths (n=32) II: Tap water baths (n=30)	I: Mud packs, Sulphur baths (n=146) II: control (n=20)
Main outcome measures	Joint tenderness Joint swelling Muscle strength Mobility	Ritchie index Pain at movement Laboratory variables	Duration of morning stiffness Number of active joints Hand grip strength Joint circumference Patient assessment of severity of disease Ritchie index Psoriasis area
Blinding	None	Patient	Observer
Follow-up	None	18 days	3 weeks
Efficacy according to the authors	Pool therapy was the more beneficial form of therapy	Beneficial effects of the thermal water.	Addition of balneotherapy has beneficial effects.

No study scored points for 'study size' because the smallest group had to be at least larger than 25 patients. Two CCT[52,56] scored on completely describing the experimental and control interventions. Günther[51] described the interventions only superficially.

Blinding is mentioned once in the CCT.[56] In the title of his article, Szucs[56] describes his study as double-blinded, but only the blinding of the patient is mentioned. Blinding of the observer or the patients is mentioned in most RCT, except in Nicholls,[60] but success of blinding is never evaluated. Most outcome measures used in studies of Group 1 were not included in the WHO/ILAR core set.

Blinding of the analysis procedure was never described. No RCT mentioned an intention-to-treat analysis or, more important, performed a comparison of effects between groups. They all found a difference in main outcome measures between pre- and post-treatment within each group.

Effectiveness of balneotherapy. Of the CCT 5 concluded that balneotherapy was effective and 2 found improvements in both study groups but no difference between the groups. For a more reliable answer to the study question concerning the efficacy of balneotherapy in patients with arthritis, only RCT with comparisons of effects between groups are adequate. Moreover the data presentation in the reports, even after communication with authors, was too scarce to enable performance of a between group analysis. Based on the present analyses of the RCTs no conclusion about the effectiveness can be given.

Discussion

This review assessed the methodological quality of trials studying the effectiveness of balneotherapy in patients with arthritis. Unfortunately all studies showed methodological flaws. A score less than 50 points might indicate that bias in the conduct of the trial is probable, because information in the publication concerning the avoidance of bias is lacking. Therefore, a conclusion about the efficacy of balneotherapy cannot be provided because of the poor methodological quality.

A criteria list developed at the department of Epidemiology at Maastricht University is called the 'Maastricht list' because all people who have used this criteria list work (or have worked) at the department of Epidemiology, Maastricht University. With this Maastricht criteria list over 30 systematic reviews have been performed in which summarization of the results was based on methodological criteria.

In most systematic reviews in which the Maastricht criteria list is used,[23,24,63-66] few studies achieve more than 50 points. So presumably, this criteria list represents a high standard for methodological quality. A maximum score is difficult to reach, especially in studies with difficulties in blinding. For those studies a high quality score is still achievable, for example, Beurskens et al.[67] would recieve 87 points using this criteria list although in her study it was impossible to blind the therapist. It is not clear whether the low quality scores were due to real methodological flaws or poor reporting. The total score of the CCT was not only lower than the total scores of the RCT because of the absence of a random treatment allocation, but in nearly all other categories they scored fewer points compared to the RCT. By including only RCT one always risks missing a CCT of high quality. This study shows, however, that RCT score higher than CCT on almost all categories of our criteria list. While randomization is an important method to reduce bias in a clinical trial, a reason not to perform a randomized clinical trial could be based on ethical considerations. Yet we do not think ethical considerations were a major issue in these studies. As there are no ethical objections, we would advise strongly against performing non-randomized studies in future.

RA is a chronic, progressive and disabling disease and has great impact on the quality of life.

Table 7: Quality scores of all included studies.

Study	Study population Max 51 points	Interventions Max 14 points	Blinding Max 18 points	Outcome Max 9 points	Analysis Max 8 points	Total
Randomized						
Elkayam et al (1991)	25	10	2	9	3	49
Sukenik et al (1990a)	22	10	4	9	3	48
Sylvester (1990)*	26	8	2	3	6	47
Sukenik et al (1990b)	21	10	2	8	0	41
Sukenik et al (1995)	20	10	2	8	0	40
Green et al (1993)*	20	7	2	5	3	37
Nicholls et al (1990)*	17	8	0	7	0	32
Non-randomized						
Svarcová et al (1990)	8	10	0	5	6	29
Szucs et al (1989)*	6	10	2	5	3	26
Sukenik et al (1994b)*	3	8	0	7	3	21
Günther et al (1976)	7	1	0	6	6	20
Landewé et al (1992)	4	8	0	7	0	19
Baldwin (1972)*	1	6	0	3	0	10
Steiner et al (1979)	3	4	0	2	0	9

* Studies concerning patients with other forms of arthritis (Group 2).

In daily life patients are trying to deal with the pain using coping strategies and this affects their quality of life. Pain (often assessed by the patient) is reported as an outcome measure in most of the studies in Group 1, but few results about change in pain were found. Sukenik[15,18,20] and Elkayam[55] reported a patient self-assessment of the severity of disease on a 7 or 10 point scale, and reported statistically significant improvement in all groups.

A quality of life measurement is never reported as an outcome measure in the studies performed. This seems strange, because one of the aims of balneotherapy, or therapy for chronic patients in general, is improving aspects of quality of life. Measures such as hand grip strength do not adequately reflect quality of life.

At the OMERACT II conference,[61] rheumatologists pointed out that there is still little experience in using these quality of life outcome measures in rheumatology clinical trials. Of the participants, 87.5% were prepared to include quality of life measurements in future research.[61]

The spa environment is an important factor in treatment results.[7,62] Many factors may contribute positively to the reported effects,[4] like change of environment, the spa-scenery, physical and mental relaxation, the absence of (house) work duties, the non-competitive atmosphere with similarly suffering companions, the concentrated physical therapy etc.. These factors, and probably many more, can be seen as co-

interventions.

In conclusion, balneotherapy for patients with arthritis is one of the oldest forms of therapy. Most authors see it as an effective treatment of patients with arthritis. The minerals in the thermal and mineral baths, together with the increased buoyancy, may be of therapeutic value. Although for the reports we have reviewed the scientific evidence is weak because of poor methodological quality, absence of adequate statistical analysis, and the absence, for the patient, of most essential outcome measures (pain, quality of life), one cannot ignore the positive findings reported in most studies. Most methodological flaws found in these studies could be avoidable in future research. We recommend performing randomized studies using a larger study population (> 75 patients in each group). This number of study patients is rather arbitrary, but with these numbers, statistical significance comes close to clinical relevance. With a smaller number clinical relevant effects might be missed. A clear description of baseline characteristics, drop-outs/losses-to-follow-up, blinding procedures, interventions, side-effects and outcome measures (including quality of life) gives a clear understanding of possible sources of bias. Also, an intention-to-treat analysis should be performed. By intention-to-treat analysis we mean that all randomized patients (minus missing values) are included in the analysis for the most important outcome measures, and in the most important moments of effect measurement irrespective of non-compliance and co-interventions. Randomized studies performed and reported properly may give a sound answer to the question whether balneotherapy is an effective treatment for patients with arthritis. New research should use outcome measures relevant to the patients, and studies concerning patients with RA should use the WHO/ILAR core set of endpoints. We conclude that performing randomized studies with high methodological quality measuring the efficacy of balneotherapy is possible and necessary.

References

1. Jackson R. Waters and spas in the classical world. *Medical History 1990 (suppl)*;10:1-13.
2. Jagger M, Zmood D. Hydrotherapy by physiotherapists in a community health centre. *Aust Fam Phys 1984*;13:878-81.
3. Goldby LJ, Scott DL. The way forward for hydrotherapy (editorial). *Br J Rheumatol 1993*;32:771-3.
4. Fam AG. Spa treatment in arthritis: a rheumatologist's view (editorial). *J Rheumatol 1991*;18:1775-7.
5. Palmer R. "In this our lightye and learning tyme": Italian baths in the era of the renaissance. *Medical History 1990 (suppl)*;10:14-22.
6. Behrend T. The balneotherapy of rheumatoid arthritis. *Rheumatol Rehabil 1979*; (Suppl):86-7.
7. Sukenik S. Spa treatment for arthritis at the Dead Sea area (editorial). *Isr J Med Sci 1994a*;30:919-21.
8. Becker BE. The biologic aspects of hydrotherapy. *J Back Musculoskel Rehabil 1994*;4:255-64.
9. Machtey I. Dead Sea balneology in osteoarthritis. In: Machtey I, ed. Second International Seminar on Treatment of Rheumatic Diseases. Boston. John Wright PGS Inc. 1982:161-6.
10. Nicholas JJ. Physical modalities in rheumatological rehabilitation; review article. *Arch Phys Med Rehab 1994*;75:994-1001.
11. Cosh JA. The rheumatologist and the spa: a personal review. *Roy Soc Health J 1982*;102:189-92.
12. Johnson RH. Arthur Stanley Wohlman, the first government balneologist in New Zealand. *Medical History 1990 (Suppl)*;10:114-26.
13. Simon L, Blotman F. Exercise therapy and hydrotherapy in the treatment of the rheumatic diseases. *Clin Rheum Dis 1981*;7:337-47.
14. Calin A. Royal National Hospital for rheumatic diseases- Bath. A 250th birthday party (editorial). *J Rheumatol 1988*; 15:733-4.
15. Sukenik S, Neumann L, Buskila D, Kleiner--Baumgarten A, Zimlichman S, Horowitz J. Dead Sea salt baths for the treatment of rheumatoid arthritis. *Clin Exp Rheumatol 1990a*;8:353-7.
16. O'Hare JP, Heywood A, Summerhayes C. et al. Observations on the effect of immersion in bath spa water. *BMJ 1985*; 291:1747-51.
17. Golland A. Basic hydrotherapy. *Physiotherapy 1981*;67:258-62.
18. Sukenik S, Buskila D, Neumann L, Kleiner--Baumgarten A, Zimlichman S, Horowitz J. Sulphur baths and mud pack treatment for rheumatoid arthritis at the Dead Sea area. *Ann Rheum Dis 1990b*;49:99-102.
19. Sukenik S, Giryes H, Halevy S, Neumann L, Flusser D, Buskila D. Treatment of psoriatic arthritis at the Dead Sea. *J. Rheumatol 1994b*;21:1305-9.
20. Sukenik S, Neumann L, Flusser D, Kleiner--Baumgarten A, Buskila D. Balneotherapy for rheumatoid arthritis at the Dead Sea. *Isr J Med Sci 1995*;31:210-4.
21. Cochrane Collaboration Handbook 1996. Internet-address: Cochrane Home Page: http://hiru.mcmas-

ter.ca/ cochrane.
22. Boers M, Tugwell P, Felson DT, et al. World Health Organization and International League of Associations for rheumatology endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. *J Rheumatol 1994*; (suppl 41) 21:86-9.
23. Beckerman H, de Bie RA, Bouter LM, de Cuyper HJ, Oostendorp RAB. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. *Phys Ther 1992*;72:483-91.
24. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM, Knipschild PG. Spinal manipulation and mobilisation for back and neck pain: a blinded review. *BMJ 1991*;303:1298-303.
25. Kabanov SE. [Effectiveness of the treatment of dystrophic (non-infective) polyarthritis in Sochi-Matseta spa with special reference to remote results.] *Vopr Revm. 1965*;5:52-8 .(Russian)
26. Rybnikov NI, Novitskii GA. [Vascular changes as indices of the effectiveness of treatment of polyarthritis of various etiology with the radon-containing mineral waters of Chmielnik.] *Vopr Kurortol Fizioter Lech Fiz Kult 1965*;30:462-3. (Russian)
27. Pshetakovskii IL. [Effect of radon water therapy at the Khmel'nik health resort on the cardiovascular system in patients with rheumatoid arthritis.] *Vopr Kurortol Fizioter Lech Fiz Kult 1971*;36:222-6. (Russian)
28. Timofeev AV, Shautsukova LK, Shautsukova LZ. [Mathematical evaluation of the effectiveness of balneotherapy and the differential diagnosis of rheumatoid and rheumatic polyarthritis.] *Vestn Akad Med Nauk SSSR 1973*;28:85-7. (Russian)
29. Perel'muter DL, Trofimova TM, Drinevskii NP, Mel'nikov AA, Kotikov VE. [Effectiveness of step-by-step treatment of patients in early stage of rheumatoid arthritis at the AMS] USSR Institute of Rheumatism and in Eupatoria health resort. *Vopr Revm 1976*;3:27-31. (Russian)
30. Zavadiak M. [Effectiveness of treatment of patients with rheumatoid polyarthritis at the balneologic resort Siniak.] *Vrach Delo 1976*;10:105-9. (Russian)
31. Litvinenko AG, Perminov IA, Poluden' EP, Opren-ko BI, Pavlova ES. [Evaluation of efficacy of balneotherapy methods in patients with rheumatoid polyarthritis.] *Vrach Delo 1977*;5:101-4. (Russian)
32. Zaitseva VI, Bersudskaia SL, Gudilina VG, Petrova VI, Khakimdzhanov AKh. [Experience with the balneological treatment of rheumatoid arthritis at a Republic hospital.] *Vopr Kurortol Fizioter Lech Fiz Kult 1979*;1:51-3. (Russian)
33. Veinpalu-Elu, Trink RF, Veinpalu LE. [Factors affecting the results of the overall treatment of rheumatoid arthritis at mud therapy health resorts.] *Vopr Kurortol Fizioter Lech Fiz Kult 1979*;1:27-31. (Russian)
34. Taletene IP, Gaigalene BA. [Effect of combined chrisanol and balneological treatment on the clinical indicators and permeability of the synovian membrane in patients with rheumatoid arthritis.] *Vopr Kurortol Fizioter Lech Fiz Kult 1984*;2:28-31. (Russian)
35. Mar'iasis ED, Militenko SA, Shalygina IE. [Therapeutic efficacy of dry-air radon baths in the rehabilitation of patients with psoriatic arthritis.] *Vopr Kurortol Fizioter Lech Fiz Kult 1987*;6:37-9. (Russian)
36. Mackiewicz S, Sternalova L. [Evaluation of the balneological treatment of children with progressive juvenile polyarthritis.] *Fysiatr Revmatol Vestn 1966*;44: 324-7. (Czech)
37. Horvath G. [Effect of hyperthermic baths in diseases of the locomotive system.] *Fysiatr Revmatol Vestn 1967*;45:277-80. (Czech)
38. Susta A, Pavelka K, Bremova A, Salavcova V, Svandova H, Richter M. [Controlled clinical trial with Benetazone (Spofa) during spa treatment.] *Fysiatr Revmatol Vestn 1975*;53:3-11. (Czech)
39. Athanasiu P, Petrescu A, Surdan C, Moisa I. [Effect of mineral water and balneological treatment with iodated and sulfurated water on rickettsial, pararickettsial and adenoviral antibodies in patients with associated pulmonary and rheumatic chronic diseases.] *Stud Cercet Virusol 1972*;23:9-12. (Romanian)
40. Lazowski Z, Gutowska-Grzegorczyk G, Romicka A, Marchwicki I. [Tentative asses-ment of rehabilitation with particular consideration of the effects of balneotherapy on the course of rheumatoid arthritis in children.] *Reumatologia. 1966*;4:321-9. (Polish)
41. Sukenik S, Mayo A, Neumann L, Flusser D, Kleiner-Baumgarten A, Buskila D.[Dead Sea baths salts for the treatment of knee ostoearthritis.] *Harefuah 1995*;129:100-3,159,158. (Hebrew).
42. Sukenik S. [Balneological (spa)therapy for rheumatic diseases.] *Harefuah 1990c*;119:167-170. (Hebrew).
43. Nobunaga M. Balneotherapy of patients with rheumatoid arthritis. 1992; 3-8. (Japanese with English abstract, unable to trace the journal or congress it is published. The original report is in the authors possession).
44. Klemm C, Fricke R, Schattenkirchner M, Treiber W, Mathies H. [Effects and side effects of 3-chlor-4-allyloxy-phenylacetic acid (Mervan) in therapeutic study of rheumatic diseases.] *Z. Rheumaforsch 1971*;30:17-25. (German)
45. Rijswijk MH van. [A prospective study of the effectiveness of thermal bath treatments in patients with rheumatoid arthritis.]. *Ned T Geneesk 1992*;136: 163-4. (Dutch)
46. Danneskiold-Samsoe B, Lyngberg K, Risum T, Telling M. The effect of water exercise therapy given to patients with rheumatoid arthritis. *Scand J Rehab Med 1987*;19:31-5.
47. Forestier F, Augy S. [Rheumatism and thermalism. A controlled trial in 65 cases.] *Presse Thermale Climatique 1970*;107:200-5. (French)
48. Ropes MW, Bennett GA, Cobb S, Jacox R, Jessar RA. Revision of diagnostic criteria in rheumatoid arthritis. *Bull Rheum Dis 1958*;9:175-6.
49. Steinbrocker O, Traeger CH, Batterman RC.

Therapeutic criteria in rheumatoid arthritis. *JAMA 1949*; 140:659-62.
50. Hochberg MC, Chang RW, Dwosh I, Lindsey S, Pincus T, Wolfe F. The American College of Rheumatology 1991: Revised criteria for the classification of global functional status in rheumatoid arthritis. *Arthritis Rheum 1992*;35:498-502.
51. Günther R, Kolarz G, Thumb N, Grabner H. [The implementation of a computerized documentation system for the evaluation of spa therapy in patients with rheumatoid arthritis.] *Wiener Klinische Wochenschrift 1976*;88:84-7. (German)
52. von Svarcová J., Hofta T, Kouba A, Trnavsky K, Zvárová J. [Effects on painbehavior of the foot using physiotherapy in patients with rheumatois arthritis.] *Z Physiother 1990*;42:109-12. (German)
53. Steiner FJF, Valkenburg HA, Stadt RJ van der, Stoyanova-Drenska M, Zant J. [Balneology treatment of patients with rheumatoid arthritis.]. *Ned T Geneesk 1979*;123:661-4. (Dutch)
54. Landewé RBM, Peeters R, Verreussel RLP, Masek BA, Goei Thè HS. [No difference in effectiveness measured between treatment in a thermal bath and in an exercise bath in patients with rheumatoid arthirits.] *Ned T Geneesk 1992*;136:173-6. (Dutch)
55. Elkayam O, Wigler I, Tishler M, Rosenblum I, Caspi D, Segal R, Fishel B, Yaron M. Effect of spa therapy in Tiberias in patients with rheumatoid arthritis and ostoearthritis. *J Rheumatol 1991*;18:1799-1803.
56. Szucs L, Ratko I, Lesko T, Szoor I, Genti G, Balint G. Double-blind trial on the effectiveness of the Puspokladany thermal water on arthrosis of the knee joints. *Roy Soc Health J 1989*;109:7-9
57. Baldwin J. Pool therapy compared with individual home exercise therapy for juvenile rheumatoid arthritic patients. *Physiother 1972*;58:230-1.
58. Sylvester KL. Investigation of the effect of hydrotherapy in the treatment of osteoarthritic hips. *Clin Rehab 1989*;4:223-8.
59. Green J, McKenna F, Redfern EJ, Chamberlain MA. Home exercises are as effective as outpatient hydrotherapy for osteoarthritis of the hip. *Br J Rheumatol 1993*;32:812-5.
60. Nicholls E, Ahern M, Simionato E, Bovill I. Assessment of hydrotherapy as a therapeutic modality in rheumatic diseases. Proceedings 3rd Int Physiotherapy Congress, Hong Kong 1990: 630-5. Sydney: Link Printing Pty Ltd, 1990, 630-5.
61. Tugwell P, Boers M, Brooks P. OMERACT II conference; Outcome measures in rheumatoid arthritis clinical trials: conclusion. *J Rheumatol 1995*;22:1431-2.
62. Bálint G, Bender T, Szabó E. Spa treatment in arthritis (correspondence). *J Rheumatol 1993*;20: 1623-5.
63. van der Heijden GJMG, Beurskens AJHM, Koes BW, Assendelft WJJ, de Vet HCW, Bouter LM. The efficacy of traction for back and neck pain; a systematic, blinded review of randomized clinical trial methods. *Phys Ther 1995*;75:93-104.
64. van der Heijden GJMG, van der Windt DAWM, Kleijnen J, Koes BW, Bouter LM, Knipschild PG. Steroid injections for shoulder disorders; a systematic review of randomized clinical trials. *Br J Gen Pract 1996*;46:309-316.
65. ter Riet G, Kleijnen J, Knipschild PG. Acupuncture and chronic pain: a criteria based meta-analysis. *J Clin Epidemiol 1990*;43:1191-1199.
66. Assendelft WJJ, Koes BW, van der Heijden GJMG, Bouter LM. Efficacy of chiropractic manipulation for back pain; blinded review of relevant randomized clinical trials. *J Manipulative Physiological Therapy 1992*;15:487-494.
67. Beurskens AJHM, de Vet HCW, Köke AJA, Lindeman E, Regtop W, van der Heijden GJMG, Knipschild PG. Efficacy of traction for non-specific low back pain; 5-week results of a randomized clinical trial. *Lancet 1995*;346:1596-616.

2 Balneotherapy and Quality Assessment: Interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment.

AUTHORS:
Arianne P. Verhagen,
Henrica C.W. de Vet,
Robert A. de Bie,
Alphons G.H. Kessels,
Maarten Boers,
Paul G. Knipschild.

ABSTRACT

Objective. This study investigates aspects of the reliability of the Maastricht criteria list for quality assessment in systematic reviews, and whether blinded reviewing is necessary to prevent review bias.

Method. We used the data set of 12 articles from a systematic review concerning the efficacy of balneotherapy in patients with arthritis. Twenty reviewers participated of which two reviewers, who have been involved in developing the Maastricht criteria list, acted as reference standard. Half of all assessments were performed blindly.

Results. A high level of agreement was found between the reviewers and a high level of correlation with the reference standard. The quality scores between the blinded and unblinded assessment did not differ much.

Conclusion. Based on the results we conclude that the Maastricht criteria list is a reliable instrument in quality assessment of clinical trials. Within the limits of this study we found no evidence that blinding is necessary to prevent review bias.(*J Clin Epidemiol 1998;51:335-41*)

The aim of reviewing the literature is to summarize information from original research. Assessing the quality of clinical trials included in a review is important because variation in quality may affect the conclusion of the review.[1-4] Moreover, if the methodology of the randomized clinical trials (RCTs) included in a review is weak, the conclusions of the review cannot be very strong.[5]

In our department over 30 systematic reviews have been performed in which the summarization of the results was based on methodological criteria.[1,6-11] For these reviews we have used more or less the same criteria list, with small extensions and adaptions in order to improve the list. Because all people using this criteria list work (or have worked) at the department of Epidemiology of the Maastricht University in the Netherlands, we will call it the 'Maastricht list'.

Like every empirical study, a review needs an explicit research protocol including a description of the search strategy, quality assessment, data extraction and the analysis.[12] In our reviews the assessment of the methodological quality is always blinded for authors and institutes where the study was performed and for the journal in which the paper was published; moreover all information about the results of the intervention is deleted. Especially the latter part of the blinding procedure is very time consuming. The blinding is performed by a person not involved in the quality assessment of the studies. After blinding usually two reviewers, with knowledge of methodological principles, assess the quality of the reports using the Maastricht criteria list. In the end they compare their scorings and reach consensus on all items.

The next step in our system of quality assessment is to weigh the scores accordingly to a predesigned weighting protocol. The reason for using weights is that some criteria are regarded as more important for quality than others. The weights given to each item remain arbitrary to some extent.[7,12,13] They differ between the reviews, depending on the topic of the review, potential flaws (e.g. importance of blinding depends on the subjectivity of the specific outcome measures) and the empirical evidence available at that time. For each study an overall score for methodological quality can be calculated, by summing up the weights.

A study of the literature is by definition non-experimental and therefore open to the same forms of bias as other observational studies.[12] A specific type of bias that might occur in reviews is review bias. Reviewers have their beliefs and disbeliefs and can therefore (unconsciously) be guided into biased assessments. In order to prevent review bias some authors recommend blinded assessment of the articles.[14-17] If review bias exists, we expect that the unblinded quality scores will be higher than the blinded ones if the authors are well known, if the journal has a higher impact factor and/or when the results are favorable. This paper investigates the interobserver reliability of the Maastricht criteria list and if our method of blinding the original articles, in respect to quality assessment, is necessary to avoid review bias.

Method

Studies. We used the data set of 12 articles from a systematic review concerning the efficacy of balneotherapy in patients with arthritis.[18-29] Both RCTs (randomized clinical trials) and CCTs (controlled clinical trials) were included. Studies with a cross-over design were excluded for reasons of feasibility. In order to study the influence of review bias each article had a blinded and an unblinded version. All articles were scanned with OCR software (Optical Character Recognition) in order to facilitate the blinding procedure by a uniform lay-out. In the blinded version information was deleted about authors and institutes where the study was performed, the journal in which the paper was published and all information about the results of the intervention. After

scanning the articles it took the researcher (APV) approximately one hour to blind each article.

Reviewers. Twenty reviewers participated, of which two reviewers (RAB, HCWV), who have been involved in the development of the Maastricht criteria list[30], acted as reference standard (RS). They assessed the methodological quality and reached consensus about the coding of the articles. The other 18 reviewers assessed the quality of the papers independently. Five of them were staff members of the Department of Epidemiology of the Maastricht University (the Netherlands) and 13 were students, from an MSc Epidemiology program for physical therapists, who recently graduated (between June 1994 and June 1995) at the same University. The five staff members were seen as senior reviewers, the others as junior reviewers. Of the 18 reviewers 11 were male and seven female. All reviewers first followed a training. The object of training was to ensure that the reviewers used the forms and procedures in identical ways.[31] A computerized random table ensured that each article was scored 10 times (5 times blinded and 5 times unblinded).

Assessment of methodological quality. Table 1 presents the characteristics of the criteria list used to assess the methodological quality.

The Maastricht list consists of 5 domains divided into 15 main items with a total of 47 subitems. Each reviewer studied the articles to determine whether information of a specific item was: '+' presented and adequately done, '–' presented but not adequately done or leading to bias, '0' not presented or '?' presented but unclear. Only the items rated '+' contributed to the methodological score using weighting factors.

Table 1: Maastricht criteria list for methodological quality assessment* (the complete list and users guide is available upon request).

Domain	Main items	Number of subitems	Weight
STUDY POPULATION	selection and restriction	2	2
	treatment allocation	3	20
	study size	3	10
	prognostic comparability	5	5
	drop outs	4	7
	loss to follow-up	3	7
			Total 51 points
INTERVENTIONS	experimental and control interventions	6	12
	extra treatments	2	2
			Total 14 points
BLINDING	blinding of patient	2	6
	blinding of therapist	2	6
	blinding of observer	2	6
			Total 18 points
OUTCOME	outcome measures	5	5
	follow-up period	3	3
	side effects	1	1
			Total 9 points
ANALYSIS	analysis and data presentation	4	8
			Total 8 points
TOTAL	15	47	Total 100 points

* The weighting in the Maastricht list used for assessing the balneotherapy is somewhat different compared with other reviews using the Maastricht list.

Table 2: Mean quality scores of the articles.

Study	Reference standard (rank)	Mean reviewers (rank)	Range reviewers	Standard deviation reviewers
Randomized				
Elkayam et al (1991)	49 (1)	44.5 (3)	33-57	9.4
Sukenik et al (1990a)	48 (2)	52.4 (1)	40-62	7.5
Sylvester (1989)	47 (3)	35.0 (6)	30-39	3.3
Sukenik et al (1990b)	41 (4)	48.9 (2)	39-55	4.5
Sukenik et al (1995)	40 (5)	39.5 (5)	29-54	8.5
Green et al (1993)	37 (6)	39.9 (4)	22-54	9.6
Nichols et al (1990)	32 (7)	28.3 (7)	21-39	5.0
Non-randomized				
Svarcová et al (1990)	29 (8)	24.6 (9)	11-34	7.7
Szucs et al (1989)	26 (9)	26.5 (8)	11-41	10.1
Sukenik et al (1994)	21 (10)	24.0 (10)	13-32	5.9
Landewé et al (1992)	19 (11)	22.4 (11)	11-36	7.0
Baldwin (1972)	10 (12)	11.4 (12)	6-28	5.1

A total of 100 points could be obtained, divided over five domains. Non-randomized studies (CCTs) could only obtain a maximum of 80 points, because the lack of a random treatment allocation. In advance we determined the quality of the studies with less than 50 points as poor, between 50 and 70 points as moderate, and with more than 70 points as good.

Statistical analysis. First we calculated the mean quality scores and a Spearman rank correlation coefficient between the reviewers and the RS. Intraclass correlation coefficients (ICC), with 95% confidence intervals (95% CI), were used to measure the agreement between the raters.[34,35] Beforehand we determined ICCs and correlation coefficients greater than 0.5 to be moderate, and greater than 0.7 as a high level of agreement/correlation. We determined the correlation coefficient (r) of each reviewer with the RS using multiple regression analysis. To compare the blinded with the unblinded mean quality scores of each article we used the Student t-test. Furthermore we evaluated the influence on the quality scores of authors, the journals of publication (using the impact factors) and the results presented. Multivariate analysis was performed to assess the influence of the covariates blinding, gender and experience on the quality score.

RESULTS

Quality assessment. Table 2 shows the mean quality scores (and standard deviation) of each article and the RS scores.

The studies are ranked in the table according to the RS scores. All studies except one were regarded as of poor quality. In ranking there is not much difference between the mean scores of the reviewers and the RS (Spearman rank $r = 0.91$).

Table 3: Reviewer characteristics.

	Mean individual *r* (Range)	Mean quality scores RCTs (stand.dev)	Mean quality scores CCTs (stand.dev)
Experience			
Seniors (n=5)	0.75 (0.49 - 0.89)	37.9 (9.6)	20.2 (10.4)
Juniors (n=13)	0.84 (0.61 - 0.98)	42.4 (10.3)	21.5 (8.7)
Gender			
Male (n=11)	0.77 (0.49 - 0.98)	42.2 (10.2)	22.4 (9.8)
Female (n=7)	0.88 (0.77 - 0.93)	39.6 (10.0)	20.7 (7.8)

The level of agreement between the reviewers shows an ICC of 0.77 (0.64 - 0.89). The largest differences in scores between the reviewers and the RS scores are in the study of Sylvester[28] in the domains 'population' and 'analysis', for Sukenik et al.[23] in the domain 'population', and for Svarcová et al.[27] in the domain 'analysis'. The smallest differences between the RS and the reviewers are seen in the domain 'blinding'.

Reviewer characteristics. Each reviewer scored on average 7 articles, and to analyse the characteristics of the reviewers we calculated their individual correlation coefficient *r* with the RS. The mean correlation coefficient of the different raters with the RS was 0.82 (range 0.49 - 0.98). Only three reviewers achieved a $r < 0.70$. Table 3 presents the individual *r* of the reviewers divided into senior and junior reviewers and male or female reviewers.

One reviewer achieved a low individual *r* and decreases the mean correlation in the senior and male group. Without him, the mean scores between senior and junior reviewers were comparable (0.82 and 0.83) but there still remained a difference between male and female reviewers (0.80 and 0.87), which is not statistical significant.

Blinded study characteristics. Table 4 shows the mean quality scores (and standard deviation) of each article for the blinded and unblinded version separately. It also presents the ranking based on the mean quality scores according to the reviewers. In the right column the quality score differences were presented. Negative differences mean a higher quality score for the unblinded articles.

The quality scores between the blinded and unblinded assessment of most of the articles did not differ much. The differences were smaller for the RCTs than for the CCTs. In defined 'positive' or 'neutral' according to the authors of the articles. The highest quality score of each article is printed bold. four of the 12 studies[20,21,24,27] the blinded assessment was lower than the unblinded. The standard deviations give information about the consistency of the different raters and varied between 1.5 and 12.3 points. The standard deviations of the blinded versions were a little bit smaller than the unblinded (6.6 and 7.1 for the RCTs, and 5.9 and 6.0 for the CCTs respectively). There was a minor difference in ranking between the blinded and unblinded assessment. The ICC for agreement between the reviewers appeared to be 0.75 (0.44-0.78) for the blinded assessment and 0.76 (0.46-0.80) for the unblinded.

Table 5 presents the study characteristics that were blinded in the articles. The quality scores of each study are given, with the journal of publication, its impact factor and the authors' conclusions. The conclusions are defined 'positive' or 'neutral' according to the authors of the articles.

Table 4: Mean quality scores of the articles between blinded and unblinded quality assessment.

Study	Blinded Mean (sd)	Rank	Unblinded Mean (sd)	Rank	Difference
Randomized					
Sukenik et al (1990a)	53.8 (7.1)	1	51.0 (7.9)	1	2.8
Sukenik et al (1990b)	48.0 (2.9)	2	49.8 (6.2)	2	-1.8
Elkayam et al (1991)	44.6 (8.7)	3	44.4 (10.2)	3	0.2
Sukenik et al (1995)	40.6 (8.8)	4	38.4 (8.3)	5	2.2
Green et al (1993)	37.8 (9.4)	5	42.0 (9.8)	4	-4.2
Sylvester (1989)	35.4 (3.5)	6	34.6 (3.2)	6	0.8
Nichols et al (1990)	29.8 (5.9)	7	26.8 (4.0)	7	3
Total mean (St. dev.)	41.4 (6.6)		41.0 (7.1)		0.4 (2.6) (95%CI: -2.1,2.8)
Non-randomized					
Szucs et al (1989)	26.8 (12.3)	1	26.2 (7.9)	3	0.6
Sukenik et al (1994)	24.8 (7.5)	2	23.2 (4.4)	4	1.6
Svarcová et al (1990)	22.2 (5.4)	3	27.0 (9.6)	1	-4.8
Landewé et al (1992)	18.2 (7.5)	4	26.6 (6.4)	2	-8.4
Baldwin (1972)	12.8 (8.8)	5	10.0 (1.5)	5	2.8
Total mean (St. dev.)	20.9 (5.9)		22.6 (6.0)		-1.7 (4.7) (95%CI:-7.5,4.2)

The highest quality score of each article is printed bold.

The group of Sukenik et al. performed more than one study.[23-26] In one of these studies the unblinded quality assessment was higher than the blinded.[24] For the journals with an impact factor above 1.5 (n=5) only twice was the unblinded assessment higher than the blinded.[20,24] Also in two studies published in journals with a low (<1.5) or unknown impact factor (n=7) the unblinded assessment was higher than the blinded.[21,27]

In nine out of 12 studies a positive treatment outcome is mentioned[18,22-29], and in two of them the unblinded quality score was higher than the blinded.[24,27] Multivariate analysis did not show a statistical significant influence of the covariates blinding, gender and experience, separately or combined, on the quality sumscores.

Discussion

In this article we presented the results of a study concerning the interobserver reliability of the Maastricht criteria list for quality assessment. Also the influence of blinding the original articles when assessing the methodological quality is investigated. Although we have performed, over the years, more than 30 reviews with the Maastricht criteria list, these aspects of quality assessment were never determined.

The Maastricht list intends to measure the quality of the conducted trials. Quality assessment depends on the quality of the report. Unfortunately low reporting quality may lead to biased estimates of the quality scores. These quality scores may present an over- or underestimation of the actual trial quality.

Table 5: Study characteristics of the articles with respect to authors, journals and results.

Study	Blinded Mean	Unblinded Mean	Journal	Conclusion	Impact factor
Randomized					
Sukenik et al (1990a)	**53.8**	51.0	Clin and Exp Rheumatology	Positive	1.590
Sukenik et al (1990b)	48.0	**49.8**	Ann Rheum Dis	Positive	1.630
Elkayam et al (1991)	**44.6**	44.4	J Rheumatol	Neutral	1.869
Sukenik et al (1995)	**40.6**	38.4	Isr J Med Sci	Positive	0.440
Green et al (1993)	37.8	**42.0**	Brit J of Rheumatology	Neutral	2.331
Sylvester (1989)	**35.4**	34.6	Clin Rehab	Positive	-
Nichols et al (1990)	**29.8**	26.8	Congress Proceedings	Positive	-
Non-randomized					
Szucs et al (1989)	**26.8**	26.2	J Royal Society Health	Positive	-
Sukenik et al (1994)	**24.8**	23.2	J Rheumatol	Positive	1.869
Svarcová et al (1990)	22.2	**27.0**	Z Physiother	Positive	-
Landewé et al (1992)	18.2	**26.6**	Ned T Geneesk	Neutral	-
Baldwin (1972)	**12.8**	10.0	Physiother	Positive	0.617

The amount of non-information ('?' or '0' scores) gives an estimation of the possibility of bias. Usually two reviewers assess the methodological quality of the reports independently and compare their scorings and reach consensus. Van der Heijden et al.[7] found in his consensus meeting an initial agreement among reviewers of more than 80%. Disagreement usually meant that one coder had missed some information.[7] Sacks et al.[16] found that the two reviewers in his study agreed on more than 90% of the items scored. For feasibility reasons we chose a lot of reviewers (n=20) and a limited number of studies (n=12). We reviewed studies about the efficacy of balneotherapy in patients with arthritis. Unfortunately all studies turned out to have a low methodological quality. Our results could have been different if high quality studies were available also.

The overall scores given by the RS and the reviewers did not differ much. The ranking difference, mainly seen by the RCTs, may be due to the small differences in quality scores. We found a high level of agreement between the reviewers and a high level of correlation with the RS. This high correlation may be enhanced by the fact that all reviewers were epidemiologists, working in the same institution and all followed a training course to ensure that they used the forms and procedures the same way. Under these conditions the interobserver reliability of the Maastricht list appears to be good. Nylenna et al.[34] studied the influence of referee characteristics on their judgments on manuscript quality. She found that more experienced reviewers gave more consistent assessments and women gave higher quality scores than men. We found the interobserver reliability of both senior and junior reviewers comparable, with a higher level of agreement among the women reviewers, compared with the men, and the women gave overall lower quality scores.

There were only minor differences between blinded or unblinded quality assessment of the articles. Therefore blinding

the articles does not seem necessary and the time consuming activity of blinding the publications for its results can be saved. The ranking difference, between blinded and unblinded assessments, seen in the CCTs may be due to the fact that overall the methodological quality is rather poor and the difference in quality scores between these studies is very small. Our hypothesis that if review bias exists, the unblinded quality scores will be higher than the blinded ones if the authors were well known, the journal had a higher impact factor and/or when the results were favorable, was not confirmed. We have to take into account that all studies were published in low-impact journals. Within the limits of this study we found no relationship between the quality assessment and study characteristics such as authors, journal of publication or outcomes.

Jadad et al.[35,36] found that blinded assessment of methodological quality produced significantly lower and more consistent quality scores than unblinded assessment. In his study 14 reviewers assessed the quality of 36 studies about pain research.[35] The reviewers were researchers, clinicians and others. All researchers had participated in RCTs concerning pain relief, and all clinicians had been in-volved in managing patients with chronic pain. The articles in the research of Jadad et al.[35] were blinded for authors, journals and date of publication, sources of financial support and acknowledgements, but not for the results. Assendelft et al.[37] assessed the methodological quality of reviews. He found that a possible cause of review bias could occur when the profession of the reviewers is linked to the intervention investigated. Assendelft et al.[37] found a relationship between the review quality score, the profession of the reviewer and the conclusions of the reviewer. The reviewers in our research were all epidemiologists without a professional rela-tionship with the intervention. The reason why we cannot confirm Jadad's results may be due to all reviewers in Jadad's study being professionally involved in the intervention investigated, contrary to our research. Possibly the high level of epidemiological knowledge and the professional linkage of the reviewers might be more important characteristics in relation to review bias than study characteristics.

Conclusion

Our research shows that the Maastricht list is a reliable instrument in the assessment of the quality of clinical trials, if reviewers with a high level of epidemiological knowledge and with no professional linkage to the intervention under study are used. Under these conditions blinding the articles for quality assessment does not seem necessary. We reviewed only a small number of trials with an overall poor methodological quality. It is, however, preferable to assess the inter-observer reliability and the effects of blinding the articles with a somewhat larger data set also including articles of high quality.

Reviewers: The authors like to thank all participating reviewers for their time and efforts: Sandra Beurskens, Jeroen Borghouts, Ger Daane, Gerry Föllings, Geert v.d. Heijden, Miranda v. Hooff, Monique Hoogenkamp, Fons Kessels, Maurice de Keyser, Jan Kool, Inge v.d. Peijl, Gerben ter Riet, Thom Schambergen, Liesbeth Schoppink, Nynke Smidt, Erik Viscaal, Pieter Wolters and Marieke Zeegelaar.

References

1. ter Riet G, Kleijnen J, Knipschild PG. Acupuncture and chronic pain: a criteria based meta-analysis. *J Clin Epidemiol 1990;*43:1191-9.
2. Emerson JD, Burdick E, Hoaglin DC, Mosteller R, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trials 1990;*11:339-52.
3. Assendelft WJJ, Koes BW, van der Heijden GJMG, Bouter LM. Efficacy of chiropractic manipulation for back pain; blinded review of relevant randomized clinical trials. *JMPT 1992;*15:487-94.
4. Beckerman H, de Bie RA, Bouter LM, Cuyper HJ, Oostendorp RAB. The efficacy of laser therapy for musculoskeletal and skin disorders; a criteria based meta-analysis of randomized clinical trials. *Phys Ther 1992;*72:483-91.
5. Liberati A. Meta-analysis: statistical alchemy for the 21st century: Discussion. *J Clin Epidemiol 1995;*48:81-6.

6. van der Heijden GJMG, Bouter LM, Beckerman H, de Bie RA, Oostendorp RAB. De effectiviteit van ultrageluid bij aandoeningen van het bewegingsapparaat. *Nederlands Tijdschrift voor Fysiotherapie 1991*;101:169-77.
7. van der Heijden GJMG, Beurskens AJHM, Koes BW, Assendelft WJJ, de Vet HCW, Bouter LM. The efficacy of traction for back and neck pain; a systematic, blinded review of randomized clinical trial methods. *Phys Ther 1995*;75:93-104.
8. van der Heijden GJMG, van der Windt DAWM, Kleijnen J, Koes BW, Bouter LM, Knipschild PG. Steroid injections for shoulder disorders; a systematic review of randomized clinical trials. *Brit J General Practice 1996*;46:309-16.
9. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM, Knipschild PG. Spinal manipulation and mobilization for back and neck pain: A blinded review. *Br Med J 1991*;303:1298-303.
10. Knipschild PG. Trials and errors; alternative thoughts on the methodology of clinical trials. *Br Med J 1993*;306:1706-7.
11. Knipschild PG. Systematic reviews: some examples. *Br Med J 1994*;309:719-21.
12. Bouter LM, ter Riet G. Meta-analysis for physiotherapists; On the importance of standardization and blinding in the study of literature. Proceedings 3rd Int Physiotherapy Congress, Hong Kong. 1990.
13. van der Windt DAWM, van der Heijden GJMG, Scholten RJPM, Koes BW, Bouter LM. The efficacy of non-steroidal anti-inflammatory drugs (NSAIDS) for shoulder complaints; a systematic review. *J Clin Epidemiol 1995*;48:691-704.
14. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. *Control Clin Trials 1981*; 2:31-49.
15. Furberg CD, Morgan TM. Lessons from overviews of cardiovascular trials. *Stat Med 1987*;6:295-303.
16. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomized controlled trials. *N Engl J Med 1987*;316:450-5.
17. Gerbach ZB, Horwitz RI. Resolving conflicting clinical trials; guidelines for meta-analysis. *J Clin Epidemiol 1988*; 41:503-9.
18. Baldwin J. Pool therapy compared with individual home exercise therapy for juvenile rheumatoid arthritic patients. *Physiother 1972*;58:230-1.
19. Elkayam O, Wigler I, Tishler M, Rosenblum I, Caspi D, Segal R, Fishel B, Yaron M. Effect of spa therapy in Tiberias in patients with rheumatoid arthritis and osteoarthritis. *J. Rheumatol 1991*;18:1799-803.
20. Green J, McKenna F, Redfern EJ, Chamberlain MA. Home exercises are as effective as outpatient hydrotherapy for osteoarthritis of the hip. *Brit J Rheumatol 1993*;32:812-5.
21. Landewé RBM, Peeters R, Verreussel RLP, Masek BA, Goei Thè HS. Geen verschil in effectiviteit gemeten tussen behandeling in een thermaalbad en in een oefenbad bij patienten met rheumatoïde arthritis. *Nederlands Tijdschrift voor Geneeskunde 1992*;139:173-7.
22. Nicholls E, Ahern M, Simionato E, Bovill I. Assessment of hydrotherapy as a therapeutic modality in rheumatic diseases. Proceedings 3rd Int Physiotherapy Congress Hong Kong 1990; 630-650.
23. Sukenik S, Neumann L, Buskila D, Kleiner-Baumgarten A, Zimlichman S, Horowitz J. Dead Sea salt baths for the treatment of rheumatois arthritis. *Clinical and Experimental Rheumatology 1990a*;8:353-7.
24. Sukenik S, Buskila D, Neumann L, Kleiner-Baumgarten A, Zimlichman S, Horowitz J. Sulphur baths and mud pack treatment for rheumatoid arthritis at the Dead Sea area. *Ann Rheum Dis 1990b*;49:99-102.
25. Sukenik S, Giryes H, Halevy S, Neumann L, Flusser D, Buskila D. Treatment of psoriatic arthritis at the Dead Sea. *J. Rheumatol 1994*;21:1305-9.
26. Sukenik S, Neumann L, Flusser D, Kleiner-Baumgarten A, Buskila D. Balneotherapy for rheumatoid arthritis at the Dead Sea. *Isr J Med Sci 1995*;31:210-4.
27. Svarcová J. von, Hofta T, Kouba A, Trnavsky K, Zvárová J. Beeinflussung der Schmerzsymptomatik im Fussbereich bei Patienten mit Rheumatoid-Arthritis durch unterschiedliche Physiotherapiemittel. *Z Physiother 1990*;42:109-12.
28. Sylvester KL. Investigation of the effect of hydrotherapy in the treatment of osteoarthritic hips. *Clin Rehab 1989*;4:223-8.
29. Szucs L, Ratko I Lesko T, Szoor I, Genti G, Balint G. Double-blind trial on the effectiveness of the Puspokladany thermal water on arthrosis of the knee joints. *J Royal Society Health 1989*; 7-9.
30. de Vet HCW, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Knipschild PG. Systematic reviews on the basis of methodological criteria. *Physiotherapy 1997*;83:284-9.
31. Stock WA. Systematic coding for resarch synthesis. In: Cooper H & Hedges LV. The handbook of research synthesis. Russell Sage Foundation. New York. 1994.
32. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin 1979*;86:420-8.
33. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. *Stat Med 1994*;13:2465-76.
34. Nylenna M, Riis P, Karlsson Y. Multiple blinded reviews of the same two manuscripts. Effects of referee characteristics and publication language. *JAMA 1994*;272:149-51.
35. Jadad AR, Moore A, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials 1996*;17:1-12.
36. Jadad AR. Meta-analysis of randomised clinical trials in pain relief. PhD. Thesis. University of Oxford, 1994.
37. Assendelft WJJ, Koes BW, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. *JAMA 1995*;274:1942-8.

3 The Delphi list:
a criteria list for quality assessment of randomised clinical trials for conducting systematic reviews developed by Delphi consensus.

AUTHORS:
Arianne P. Verhagen,
Henrica C.W. de Vet,
Robert A. de Bie,
Alphons G.H. Kessels,
Maarten Boers,
Lex M. Bouter,
Paul G. Knipschild.

ABSTRACT

Objective. Most systematic reviews rely substantially on the assessment of the methodological quality of the individual trials. The aim of this study is to obtain consensus among experts about a set of generic core items for quality assessment of randomised clinical trials (RCTs).

Methods. The invited participants are experts in the field of quality assessment of RCTs. The initial item pool contained all items from existing criteria lists. Subsequently, we reduced the number of items by using the Delphi consensus technique. Each Delphi round comprised of a questionnaire, an analysis and a feedback report. The feedback report included staff team decisions made on the basis of the analysis and their justification.

Outcome. A total of 33 international experts agreed to participate of whom 21 completed all questionnaires. The initial item pool of 206 items was reduced to nine items in three Delphi rounds.

Conclusion. The final criteria list (the Delphi list) was satisfactory to all participants. It is a starting point on the way to a minimum reference standard for RCTs on many different research topics. This list is not intended to replace, but rather to be used alongside existing criteria lists. *(J Clin Epid 1998; 1235-41)*

In recent years, the number of available randomised clinical trials (RCTs) has grown exponentially. It is therefore almost impossible for clinicians to keep up with the increase of scientific information from original research.[1] An important aim of reviewing the literature in health care is to summarise the evidence that clinicians need to base their care and thus to provide the empirical basis for clinical decision making. The overall conclusions of a review often appear to depend on the quality of both the individual RCTs and the review process.[2,3] A clear description of the strategies for identifying, selecting, and integrating the information distinguishes a systematic review from the traditional narrative review.[4,5] Today, many systematic reviews rely substantially on the assessment of the methodological quality of the individual trials.[6-8]

'Quality' as a concept is not easy to define. Quality of RCTs has recently been defined as 'the likelihood of the trial design to generate unbiased results'.[9] This definition covers only the dimension of internal validity. Although most articles proposing a criteria list to assess the methodological quality of RCTs do not explicitly define the concept of quality [10], most lists measure at least three dimensions which may encompass the concept of quality in its broadest sense: internal validity, external validity and statistical analysis.[11-15] Some authors distinguish an ethical component in the concept of quality as well.[16,17]

The method to develop a quality criteria list is similar to that of other measurement instruments, for example, 'quality of life' scales.[18] Here, consensus methods are often used to select and reduce the number of items. Consensus studies are typically designed to combine the knowledge and experience of experts with the limited amount of available evidence. From the existing consensus methods, we chose the Delphi technique [19,20], because of the number of the participants we wanted to involve, the written procedure, the anonymity of the comments, and the time available (approximately 2 years) to conduct the study.

The aim of this study is to achieve consensus among experts, implicitly based on both empirical evidence and personal opinion, on how the quality of RCTs can be measured best, resulting in a quality criteria list. We have considered two approaches to reach this goal: try to achieve consensus on the definition of quality of RCTs and infer the necessary items for a criteria list, or, conversely, try to achieve consensus on items that, according to the participants, measure quality of a trial and infer from those a definition, or a description of the concept, of quality. We considered the latter approach to have a higher chance of success.

To be able to measure the quality of the design and conduct of a trial, one has to rely on the quality of the report. Our point of departure is the ideal situation, that is, that the report presents an honest, accurate, and comprehensive reflection of the conduct of the study. We regard the quality criteria list resulting from this study as a starting point for a future minimum reference standard to be used in systematic reviews. As such, it is not intended to replace existing criteria lists but to facilitate comparison of reviews more easily. This paper presents the Delphi procedure and the resulting criteria list in quality assessment of RCTs on which experts reached consensus.

METHOD

Staff team. A staff team was formed to initiate this research and consisted of all authors except L.M.B. All staff team members are epidemiologists, one of whom is also a clinician and one of whom has a statisticical background. The others are medical doctors or health scientists. The staff team was responsible for the procedures of the selection of items and the participants and was responsible for the construction of the questionnaires, the analysis of the responses and the formulation of the feedback.

Selection of the items. For the development of the initial item pool we collected all items from existing quality criteria lists for RCTs. For the search strategy four sources were used: an article by Moher *et al.*[10], the doctoral thesis of Jadad [15], information from the Methods group of the Cochrane Collaboration and a Medline search on CD-ROM using the key-words: *quality, assessment, methodology, randomised clinical trials, scales, checklists, quality scores, meta-analysis, epidemiology,* and *methods.* Papers are included when a criteria list for quality assessment of RCTs was presented. Papers were excluded when a modification of an existing list was used.

We made headings of various aspects of a design of an RCT, (e.g. aim, study question, randomization, blinding), under which all items were ordered. A total number of 17 headings (or domains) were created. On the basis of this initial item pool, we formulated the Delphi-1 questionnaire. To generate a more complete item pool, the participants were given the opportunity in Delphi-1 to add items they missed.

Selection of participants. The participants had to be epidemiologists or statisticians concerned with quality assessment in systematic reviews or meta-analyses. We tried to achieve a wide range of different points of view on quality assessment. First, we asked all first (or co-) authors of a publication of an original quality criteria list to participate, one (co-) author per original article. Next, after an extensive brainstorm of the members of the staff team, we generated a list of leading epidemiologists and statisticians in the field of quality assessment. This resulted in three groups of experts of roughly equal sizes: authors, epidemiologists, and statisticians.

Procedure. During the whole Delphi procedure, we used structured questions, for example: 'Should this item be included into the criteria list?' or 'Do you agree with the rewording of this item?'. The answer options used were 5-point Likert-scales (totally agree - totally disagree) or a 'yes/no/don't know' answer format. We invited participants to give reasons for their choices. After each Delphi round, a feedback report was made to inform the participants about opinions and arguments of the other participants. The staff team decided, on the basis of the answers and arguments of the participants, which items and questions would appear in the next questionnaire. Staff team decisions were presented and justified in the feedback report. The participants were given the opportunity to react to, or when necessary oppose to, the arguments of other participants and the decisions made by the staff team. Three or four Delphi rounds were considered sufficient to reach consensus; consensus being defined as a 'general agreement of a substantial majority'.

Analysis. The analysis of the responses from the Delphi rounds was both quantitative and qualitative. Quantitatively, we presented the mean scores on the 5-point Likert scales (strongly disagree [0 points], moderately disagree [1 point], neutral [2 points], moderately agree [3 points] and strongly agree [4 points]) as a percentage of the max. obtainable score. For example: a mean score of 1.9 is 47.5% of the max. achievable score. For questions with a 'yes/no/don't know' answer format we calculated a 'yes minus no' score from the number of participants who answered a 'yes' on a specific question minus the number of participants who answered 'no'. The necessary cut-off points were determined based on the data of each Delphi round. Qualitatively, we summarized the suggestions and comments of the participants.

Delphi-1. For every item, we asked the participants how strongly they agreed to include it in the final criteria list (5-point Likert scale). Participants were given the opportunity to suggest alternative wording and to add extra items. Some items basically asked for the same information but were formulated differently. Participants were able to choose the items in the wording they liked best.

Delphi-2. The Delphi-2 questionnaire provided opinions on the methods and results of procedural decisions made by the staff team and questions about the formulation of the items selected from Dephi-1 on which the participants agreed most. We decided, on the basis of the mixed responses in Delphi-1, to present all items not selected initially after Delphi-1 again in Delphi-2 for a second chance. Participants were able to choose the items they considered to be essential for the criteria list. Again, they were invited to give reasons for their decisions and opinions.

Delphi-3. We reworded the initial items based on the arguments given in Delphi-2, and presented them in the Delphi-3 questionnaire. We asked whether the participants preferred the rewording or the original phrasing. Furthermore, we presented the items that received a second chance (based on the answers in Delphi-2) to be included into the criteria list. The participants were able to state which of these items should be added into the final list of items. Subsequently, we asked whether they agreed with the omission of domains not chosen in previous rounds (Delphi-1 and Delphi-2).

Definition of quality. After Delphi-1, at the 3rd Cochrane Colloquium in Oslo in 1995 in a meeting with some of the participants, the issue was raised of whether we should continue talking about the 'quality of RCTs' or whether we should limit ourselves to identifying a set of 'parameters which may be related to effect sizes', which implies a restriction to internal validity. Therefore, in Delphi-2 we asked the participants whether they had problems with using the word 'quality' related to this criteria list. On the basis of their answers, we generated two possible definitions about quality, and the participants were asked in Delphi-3 which of the two different definitions they considered to be most accurate.

RESULTS

Participants. We were able to locate 15 of 17 identified authors (or co-authors) of original criteria lists. One of them refused to participate, and three did not respond. We located 13 of 19 epidemiologists, of whom two refused to respond and two did not respond. Of the 15 statisticians we located, one refused to respond and one did not respond. Potential participants declined mostly because they were too busy, only one declined because he did not like the Delphi-method for this purpose. We started with 33 persons who agreed to participate, of whom 26 returned the first questionnaire and 21 the second and third questionnaires. One participant returned the second and third questionnaire. Reasons mostly mentionend for nonresponse was lack of time.

Delphi-1. A total number of 24 papers were found presenting an criteria list.[9,10,13,14,16,21-40] Several articles used the same criteria list, namely the 'Maastricht list'[33-39] or the list developed by Chalmers.[13,40] Once, a double publication of the same criteria list was found.[27,28] We started with 17 articles[9,10,13,14,16,21-33] after excluding articles in which a modification of the 'Chalmers list' or the 'Maastricht list' was used. From these criteria lists, we generated a large initial item pool of 206 items ordered under 17 domains. Of the 33 Delphi-1 questionnaires, 26 were returned and analyzed.

The initial item list generated intense disagreement: on 25% of the items (n=52) five or more participants scored 'strongly agree' to include this item, whereas five or more other participants scored 'strongly disagree' to include that item (see Table 1). The disagreement was in part due to different formulations of the items but also to the different priorities of the statisticians and the epidemiologists regarding the inclusion of statistical items.

Table 1: Some examples of items (Delphi-1) on which the participants (n= 26) showed strong disagreement.

Items	Number of participants that answered 'strongly agree': this item must be included in the list	Number of participants that answered 'strongly disagree': this item must not be included in the list
The study design is: a. Poor (e.g. no comparative groups). b. Inadequate (e.g. comparative, single blind or open). c. Adequate (e.g. comparative, double blind).	9	10
Is the method described used to conceal the intervention assignment schedule from participants and clinicians until recruitment was complete and irrevocable?	10	7
Was the study described as randomized (this includes the use of words such as randomly, random and randomization)?	7	6
Dates of starting and ending accession?	6	6

Epidemiologists stated repeatedly that items concerning the statistical analysis had nothing to do with the quality of RCTs, whereas the statisticians consider, for example, the performance of an *a priori* sample size calculation to be of importance to quality. Table 1 shows examples of items on which the participants disagreed strongly.

We saw no obvious difference in scoring between the authors and the epidemiologists, but observed a difference when we divided the participants in statisticians on the one hand and epidemiologists + authors on the other. The statisticians scored 31 items greater than 70% of the maximum obtainable score, of which five items concerned statistical analysis and seven items concerned withdrawals or drop-outs.

Staff team decisions. The aim of the staff team was a short final criteria list. On the basis of the data, we chose a rather high cut-off point of 70%, resulting in a preliminary list of seven items, to which items could be added during the procedure. The feedback report of the Delphi-1 presented all items with their scores in percentage and all comments made by the participants (anonymously). We decided to present all items of Delphi-1 again in Delphi-2 so that participants were able to reconsider their first decisions, before any definite decision on inclusion or exclusion was taken.

Delphi-2. Of the 33 Delphi-2 questionnaires sent to all initial participants, 21 were returned and analysed. Non response was mainly in the authors/epidemiologists group. The most reported reason was lack of time, and one participant was on maternity leave. Eight participants agreed with the cut-off point of 70%, whereas nine participants answered 'don't know'. The majority of the participants (n=15) accepted the seven initial items to be included, but all considered rephrasing of most items necessary. Most participants chose some of the items from Delphi-1 that had a score below the 70% (second chance items) to be included also.

Staff team decisions. The data showed a large group of 'second chance items' that were never chosen or were chosen by only one or two participants; that is, most participants did not feel those items were essential.

We decided to give the items chosen at least four times a final chance to be included. Table 2 presents the reworded preliminary items and the extra items receiving a final chance to be included.

Delphi-3. All 21 Delphi-3 questionnaires sent to the participants of Delphi-2 were returned and analysed. The majority of the participants accepted the rewording of the initial items. One second-chance item was added to the final criteria list because 19 participants regarded this item as essential. On the other items, the opinion on whether or not to include was divided with roughly equal 'yes' and 'no' responses. We decided these items to be important but not essential and, thus, did not include them in the final criteria list.

Table 2: All items selected for the definitive criteria list.

Items selected and reworded in Delphi-2
1. Treatment allocation
a) Was a method of randomisation performed?
b) Was the treatment allocation blinded?
2. Are the groups similar at baseline regarding the most important prognostic indicators?
3. Eligibility criteria:
a) Are inclusion criteria operationalised?
b) Are exclusion criteria operationalised?
4. Was the outcome assessor blinded?
5. Was the therapist/care provider blinded?
6. Is the numerical information regarding the primary endpoint sufficient to enable statistical pooling?
7. Does the analysis include an intention-to-treat analysis?
Items receiving a final chance in Delphi-3 to be included also.
1. Is the withdrawal/drop-out rate unlikely to cause bias?
2. Are therapeutic and control regimens/interventions operationalised?
3. Is the compliance rate (in each group) unlikely to cause bias?
4. Is controlled for co-interventions which could explain the results?
5. Was the patient blinded?
6. Is a sample size justification described?

Table 3: Final Delphi list after three Delphi rounds.

1. Treatment allocation	
a) Was a method of randomisation performed?	Yes / No / Don't know
b) Was the treatment allocation concealed?	Yes / No / Don't know
2. Were the groups similar at baseline regarding the most important prognostic indicators?	Yes / No / Don't know
3. Were the eligibility criteria specified?	Yes / No / Don't know
4. Was the outcome assessor blinded?	Yes / No / Don't know
5. Was the care provider blinded?	Yes / No / Don't know
6. Was the patient blinded?	Yes / No / Don't know
7. Were point estimates and measures of variability presented for the primary outcome measures?	Yes / No / Don't know
8. Did the analysis include an intention-to-treat analysis?	Yes / No / Don't know

Table 4: Items and domains per Delphi round.

Domains	Number of items in Delphi-1 & 2	Number of items in Delphi-3	Number of items in final Delphi list
1. Study question	2	-	-
2. Population	15	1	1
3. Sample size and power calculations a priori	9	1	-
4. Treatment allocation	12	1	1
5. Study design	2	-	-
6. Ethics	4	-	-
7. Intervention	19	1	-
8. Outcome measures	21	-	-
9. Follow-up/withdrawals	14	1	-
10. Blinding	28	3	3
11 Co-intervention	5	1	-
12. Side-effects	5	-	-
13. Compliance	6	1	-
14. Prognostic comparability	6	1	1
15. Analysis	41	2	2
16. Conclusion	10	-	-
17. Presentation	7	-	-

This final list is called the Delphi list (Table 3) and includes a description about the interpretation of the items as well (available upon request from the first author). In Table 4 we present in detail the items and domains per Delphi round.

Definition of quality. According to the majority of the participants, restriction of 'quality' to 'internal validity' does not capture the concept of 'quality' and, consequently, a definition of quality should also contain elements of external validity and the statistical analysis. But during the process, we noticed inconsistencies, even within participants within one Delphi round. For example, a participant stated explicitly on one page that quality was only concerned with internal validity. But on the next page, the same participant suggested the inclusion of three items into the final criteria list that clearly concerned the external validity. Therefore, the staff team generated two different definitions based on the answers of Delphi-2. The first definition was: *'Quality is a set of parameters in the design and conduct of a study related to effect sizes.'* This definition had emerged from a workshop with some of the participants at the 3rd Cochrane Colloquium in Oslo. The second definition was generated from the remarks in the Delphi-2 questionnaire: *'Quality is a set of parameters in the design and conduct of a study that reflects the validity of the outcome, related to the external and internal validity and the statistical model used'.*

The majority (n =17) of the participants in Delphi-3 were in favor of the second definition of quality, but most of them did not like the phrasing. Only two participants preferred the first definition, and two participants answered 'don't know'. The participants achieved consensus on quality being more than internal validity alone, but the staff team was not able to capture this consensus into an acceptable definition.

DISCUSSION

After three Delphi rounds, the participants achieved consensus on a generic core set of items for quality assessment in RCTs. Because of the chosen Delphi consensus procedure, we will call this list: the Delphi list. In our effort to develop a criteria list, we chose not to define the word 'quality' beforehand because a well-accepted definition does not exist. We assumed that the participants (all experts in

the field of quality assessment) would have their own clear picture of what quality is. The advantage of a consensus method such as the Delphi approach is that the different ideas of the concept of quality integrate in the resulting criteria list, thus determining the content validity. During the process, most participants appeared to have difficulties with this approach, and we decided to try to formulate a definition of quality.

In a consensus procedure the choice of the participants is crucial.[19,20] In the process of selecting the participants, our aim was to achieve a broad representation of all different points of view on quality assessment using three different groups of roughly equal sizes.

In a Delphi consensus procedure, the staff team has to decide about the procedural steps.[19,20] Their decisions can vary from fully autocrative to fully democratic. Because of the expected fundamental differences, we assumed that a too-directive role would be ineffective. Therefore, we decided to allow all Delphi-1 items for a second chance. The data of Delphi-2 showed much more agreement, and we considered that a consensus could be achieved. After Delphi-3, the participants seemed satisfied with the resulting criteria list, and we believe no new arguments were given, so a fourth round would probably not add new or different information.

Based on the comments and remarks of the participants during the whole procedure, an Appendix has been constructed on the interpretation of the items. The reviewers have to decide, depending on the topic of the review, whether enough information is provided to score a 'yes' on certain items. As long as these decisions are stated explicitly in the review, it will be clear for the reader how the items are scored and a comparison with reviews on other topics using the same criteria list can be made.

Empirical research concerning assessment of the methodological quality of RCTs is relatively new. Awaiting of empirical research, we think it is usefull to prioritize items using a group of experts. All different opinions in this field of research should be respected at this stage. Starting this research, we were well aware of the different views on quality and quality assessment of RCTs. Despite this knowledge, we were surprised by the many initial differences between the participants. Notwithstanding these differences, the participants achieved consensus on the final Delphi list. New in the ongoing discussion about quality is that we achieved broad consensus concerning the need for inclusion of three dimensions of quality into any definition of the 'concept' of quality: 'internal validity', 'external validity' and 'statistical considerations'. In the feedback of Delphi-3, in which we presented the Delphi list to the participants as a result of this research project, we asked participants to react to the final result. No negative, and four positive reactions or comments were received.

When a consensus concerning the content of a criteria list is reached, the following issue of what to do with the results of quality assessment has to be addressed. A quality criteria list can be used in different ways.[33,38,40,41] It can provide a quality score as an estimate of the methodological quality. These quality scores can be used as a 'threshold score' for inclusion of the article in a review, as a 'weighting factor' in the statistical analysis [40,42,43], or as the input sequence in a cumulative meta-analysis.[43-45] Sometimes a visual plot of the effect size against a quality score is presented.[40,42,43] The next step will be to achieve consensus (based on empirical evidence) about how to incorporate quality into the final conclusions of a systematic review or meta-analysis.

CONCLUSION

The participants in this Delphi process achieved consensus on a generic criteria list for quality assessment in RCTs: The Delphi list. The adoption of this core set, by the participants and other researchers may be the first step towards a minimum reference standard of quality measures for all RCTs. It is not our intention to replace existing criteria lists, but

suggest it should be used alongside these lists. The validity of this criteria list will have to be measured and evaluated over time.

PARTICIPANTS: The authors like to thank the following persons for their participation: DG Altman, E Andrew, J Berlin, LM Bouter, SA Brown, MK Cho, M Clarke, K Dickersin, M Evans (& AV Pollock), C Friedenreich, PC Gøtzsche, S Greenland, J van Houwelingen, TF Imperiale, J Lau, C Mulrow, M Nurmohamed, I Olkin, P Onghena, G ter Riet, H Sacks, KF Schultz, K Smith, P Tugwell, S Yusuf. Their participation in this project does not necesarily mean that they fully agree with the final criteria list, but the criteria list is the result of a 'communis opinio'.

References

1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Boston: Little-Brown,1991:359-78.
2. Haynes RB. Clinical review articles: should be as scientific as the articles they review (editorial). *Br Med J 1992*;304:330-1.
3. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol 1991*; 44:1271-8.
4. Dickersin K, Berlin JA. Meta-analysis: state of the science. *Epidemiologic Reviews 1992*;14:154-76.
5. Mulrow CD. Rationale for systematic reviews. *Br Med J 1994*;309:597-9.
6. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *Can Med Assoc J 1988*;138:697-703.
7. Mulrow CD. The medical review article: state of the science. *Ann Intern Med 1987*;106:485-8.
8. Mulrow CD, Oxman AD (eds). Cochrane Collaboration Handbook [updated 1 March 1997]. In: The Cochrane Library [database on disk and CDROM]. The Cochrane Collaboration. Oxford: Update Software; 1996-. Updated quarterly.
9. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clin Trials 1996*;17: 1-12.
10. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clin Trials 1995*;16:62-73.
11. The Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials: special communication. *JAMA 1994*;272: 1926-31.
12. Assendelft WJJ, Koes BW, Heijden GJMG van der, Bouter LM. The efficacy of chiropractic for back pain: blinded review of the relevant randomized clinical trials. *J Manipulative Physiol Ther 1992*;15: 487-94.
13. Chalmers TC, Smith H.jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. *Controlled Clin Trials 1981*;2:31-49.
14. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: medical. *Stat Med 1989*;8:441-54.
15. Jadad AR. Meta-analysis of randomised clinical trials in pain relief. (PhD Thesis) Oxford: University of Oxford, 1994.
16. Andrew E. Method for assessment of the reporting standard of clinical trials with Röntgen contrast media. *Acta Radiologica Diagnosis 1984*;25:55-8.
17. Lumley J, Bastian H. Competing or complementary? Ethical considerations and the quality of randomized trials. *Int J Technology Assessment in Health Care 1996*;12;247-63.
18. Jaeschke R, Guyatt GH. How to develop and validate a new quality of life instrument. In: Spilker B ed. Quality of life assessments in clinical trials. New York: Raven Press, 1990.
19. Delbecq AL, Ven AH van de, Gustafson DH. Group techniques for program planning; a guide to nominal group and Delphi processes. Glenview: Scott, Foresman, 1975.
20. Dalkey NC, Helmer O. An experimental application of the Delphi-method to the use of experts. Management Science, 1963.
21. Evans M, Pollock AV. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *Br J Surg 1985*; 72:256-60.
22. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. *Controlled Clin Trials 1989*;10:31-56.
23. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. *Pediatrics 1989*;84:815-27.
24. Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C, Tonascia S, Chalmers TC. A cohort study of summary reports of controlled trials. *JAMA 1990*;263:1401-5.
25. Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? *Ann Intern Med 1990*;113:299-307.
26. Spitzer WO, Lawrence V, Dales R, et al. Links between passive smoking and disease: a best evidence synthesis. *Clin Investigative Medicine 1990*;13:17-42.
27. Brown SA. Measurement of quality of primary studies for meta-analysis. *Nursing Research 1991*;40: 352-5
28. Brown SA. Meta-analysis of diabetes patient education research. *Research in Nursing and Health 1992*; 15:409-19.
29. Nurmohamed M, Rosendaal F, Buller H, Dekker E, Hommes D, Vandenbroucke J, Briët E. Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. *Lancet 1992*;340:152-6.
30. Onghena P, Houdenhove B van. Antidepressant-induced analgesia in chronic non-malignant pain: a meta-analysis of 39 placebo-controlled studies. *Pain 1992*;49:205-19.

31. Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation: a meta-analysis. *Am Rev Respir Dis 1992*; 145:533-9.
32. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA 1994*;272:101-4.
33. Koes BW, Assendelft WJJ, van der Heijden GJMG , Bouter LM, Knipschild PG. Spinal manipulation and mobilization for back and neck pain: A blinded review. *Br Med J 1991*;303:1298-303.
34. Beckerman H, de Bie RA, Bouter LM, De Cuyper HJ, Oostendorp RAB. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. *Phys Ther 1992*;72:483-91.
35. van der Heijden GJMG, Bouter LM, Beckerman H, de Bie RA, Oostendorp RAB. De effectiviteit van ultra geluid bij aandoeningen van het bewegings apparaat. *Ned T Fysiother 1991*;101:169-77.
36. Kleijnen J, Knipschild P, ter Riet G. Clinical trials of homeopathy. *Br Med J 1991*;302:316-23.
37. Knipschild PG. Trials and errors; alternative thoughts on the methodology of clinical trials. *Br Med J 1993*;306:1706-7.
38. Knipschild PG. Systematic reviews: Some examples. *Br Med J 1994*;309:719-21.
39. ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. *J Clin Epidemiol 1990*;43:1191-9.
40. Detsky AS, Naylor CD, Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta analysis. *J Clin Epidemiol 1992*;45:255-65.
41. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA 1995*;273:408-12
42. Jenicek M. Meta-analysis in medicine: Where we are and where we want to go. *J Clin Epidemiol 1989*; 42:35-44.
43. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technology Assessment 1996*;12;195-208.
44. Koes BW, van Tulder MW, van der Windt DAWM, Bouter LM. The efficacy of back schools: a review of randomized clinical trials. *J Clin Epidemiol 1994*;47: 851-62.
45. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM. Spinal manipulation for low back pain; an updated systematic review of randomized clinical trials. *Spine 1996*;221:2860-2873.

4 Efficacy of 904 nm laser therapy in musculoskeletal disorders: a systematic review

AUTHORS:
Robert A. de Bie,
Arianne P. Verhagen,
Anton F. Lenssen,
Henrica C.W. de Vet,
Frans A.J.M. van den Wildenberg,
Gauke Kootstra,
Paul G. Knipschild,

ABSTRACT

Objective: This systematic review was undertaken to assess the effectiveness of 904 nm low level laser therapy (LT) in musculoskeletal disorders.

Method: In order to retrieve randomised trials, computer-aided searches of databases and of bibliographic indexes were performed. Furthermore, congress reports, reviews and handbooks were all checked for relevant citations. Subsequently, all retrieved studies were scored on methodological quality.

Results: This review found 25 studies that investigated the effects of 904 nm LT versus placebo or any other intervention, in subjects with a condition for which LT was thought a feasible intervention. Of these, 21 fulfilled the entry criteria for this review, and were assessed in a blinded manner on methodological criteria. Overall, study quality ranged from 'poor' to 'reasonable'. In a classification of the material into diseases studied, no clear evidence was found for the effectiveness of LT, except perhaps for knee problems and myofascial pain.

Conclusion: It is conclude that 904 nm LT does not seem to be effective in the treatment of musculoskeletal disorders, but that further and improved research is needed to shed more light on its effectiveness. *(Based on: RA de Bie, AP Verhagen, AF Lenssen, HCW de Vet, FAJM van den Wildenberg, G Kootstra, PG Knipschild. Efficacy of 904 nm laser therapy in the management of musculoskeletal disorders: a systematic review. Phys Ther Rev 1998; 3: 59-72.)*

Laser therapy (LT) is a relatively novel treatment modality based upon the application of relatively low intensity laser light to treat soft tissue injuries, pain and wounds of various aetiologies. Medical lasers were first introduced in the early 1960's; favourable results in surgery on humans were reported around 1965.[1,2] About 15 years later the first reports on LT in physiotherapy[3] and in acupuncture[4] appeared. The first randomised clinical trials were published in 1981 by Gallachi et al.[5] and Lewith et al.[6] Gallachi described the effects of LT in the treatment of cervical and lumbar pain, whereas Lewith conducted a randomised trial to evaluate the effect of infrared stimulation of local trigger points on the pain caused by cervical osteo-arthrosis.

In the treatment of musculoskeletal disorders, lasers with varying wavelengths are used, mostly ranging from 632 to 904 nm (i.e. from visible red to near infrared). The clinical use of LT has increased rapidly over the last few years, despite the lack of adequate insights into the underlying biological mechanisms of action and the extent of its clinical effects.

The interaction between (low intensity) laser light and tissue can be characterized or classified as two processes; absorption and scattering. Absorption can be seen as the transformation of light energy into another form of energy, ultimately resulting in the dissipation of heat. Absorption of laser light occurs mainly at a molecular level, where three underlying mechanisms can be distinguished; either i) atoms are excited to higher modes of oscillation; or ii) electron bonds are excited within the biomolecule; or iii) rotation of (parts of) the biomolecule takes place.[7] Scattering, due to differing relative refractive indices of the various cellular substances and molecules, may be defined as a change in the direction of light propagation. Both absorption and scattering are wavelength dependent, and result in a dissipation of laser power as the light beam penetrates the irradiated tissue.

The reported clinical effects of low intensity laser irradiation (as outlined above), have led to a series of theories of mechanisms of action, in which the term 'biostimulation' tipically occurs; indeed, during the past 15 years so called biostimulative effects of LT have been described by several authors. Biostimulation refers to the application of electromagnetic energy by LT to body tissues, which supposedly influences a wide variety of cell functions.[8-10] These effects are thought to consist of both stimulation or inhibition of biochemical, physiological and proliferative activities; the magnitude of such effects is reported to be dependent on wavelength, dosage and dose-intensity of LT.[11]

The 'cellular communication' theory, which is proposed to explain the bioeffectivity of LT, claims that there is an impairment or disorder, the energy state of a cell is changed, consequently altering the electromagnetic communication between cells. LT is thought to influence this communication in a positive way.[7,12] The 'photochemical theory' offers an alternative explanation; in this case action of laser light is explained in terms of its absorption by tissue chromophores (photo acceptors). These chromophores may either be enzymes, membrane molecules or other cellular or extracellular substances; activation of these by LT is considered responsible for the postulated bioeffects.[13]

Neither theory has been thoroughly confirmed in research, and the supposed underlying mechanism remains unclear. Moreover, recent research with 904 nm laser on tissue samples has failed to show any effects on cell metabolism, and hence has not provided corroborative evidence.[14] The lack of a convincing biological basis that might explain the clinical effects induced by LT serves to maintain the controversy surrounding optimal dosage and treatment indications. Musculoskeletal disorders are thought to be influenced positively by LT in five main areas: i.e. analgesic effects, anti-inflammatory effects, nerve regeneration, regeneration of muscular tissues and of bone tissues.[12,13] Thus an analysis of the potential effects of LT on each of these areas seems

imperative. It was therefore decided to perform a systematic review to evaluate the effectiveness of 904 nm LT.

METHODS

Trials on 904 nm low level LT were identified by searches in Medline and Embase (both up to 1996) and by checking the Database of the Cochrane Field 'Rehabilitation & Related Therapies' at Maastricht University, the Netherlands. Keywords used for the intervention under study were: *laser, low level laser therapy, infra red, light therapy* in combination with *rehabilitation, exercise therapy, physiotherapy* and *physical therapy*. Keywords used to describe the design were: *controlled trials, experiments, clinical trials, randomised* and *evaluation studies*. Additionally, Current Contents and Physiotherapy Index, reviews, congress reports, and handbooks on LT were checked. Retrieved references were followed-up by citation tracking. Papers published in English, French, German, Dutch, Spanish, Italian, Norwegian, Swedish and Danish were eligible for inclusion. Abstracts and unpublished studies were not included.

The selected studies had to fulfil the following criteria for inclusion in the review. Firstly, the subjects in the study had to have a condition for which LT was thought a feasible treatment, and this had to be compared with placebo, no treatment or other interventions. The treatment regimen had to consist of 904 nm LT, and the study design had to be a randomised clinical trial.

Table 1: Criteria for assessing methodological quality in RCTs of low level laser therapy.

Criterion*		Weights
Study population (total points=49)		
A	Homogeneity	4
B	Randomisation procedure mentioned	10
	Concealed method of randomisation	10
C	Comparability of relevant baseline characteristics	6
D	Numbers of patients	9
E	Dropouts described for each study group separataly	7
F	Loss to follow up not leading to bias	3
Intervention (total points=14)		
G	Intervention adequately described and performed	12
H	Co-interventions avoided or equal in study groups	2
Blinding (total points=18)		
I	Patients blinded	6
J	Therapist blinded	6
K	Observer blinded	6
Outcome (total points=12)		
L	Adequate outcome measures	5
M	Adequate follow up period	5
N	Description of side effects	2
Data presentation and analysis (total points=7)		
O	Mean or frequencies of most important outcome measures presented for each group	3
P	Intention to treat analysis	1
Q	Adequate correction for base-line differences of dropouts	3
		100

This design is considered the optimal paradigm for intervention studies, because of its potential to provide a valid assessment of the efficacy of an intervention.[15,16] The papers eligible for reviewing were blinded for author(s), journal and references, and the lay-out of the original papers was changed. They were distributed to two reviewers (APV and AFL) who independently assessed the quality of the studies; the reviewers attempted to reach consensus in a subsequent meeting on the items on which there was disagreement. Where consensus could not be reached, a third (not blinded) reviewer (RAB) made the final decision.

Table 1 shows the criteria used for assessing the methodological quality of the trials. The list was originally designed by Ter Riet et al.,[17] and subsequently modified over a number of years by Koes et al.,[18] Van der Heijden et al.[19] and Assendelft et al..[20] It is based on generally accepted principles of intervention research.[15,16] The criteria list was adapted for LT, with respect to the intervention, laser parameters, and relevant outcome measures. Studies could earn points for methodological quality in five categories; these dealt with study population, interventions, blinding, outcome and data presentation and analysis. A maximum score of 100 points could be obtained (see Table 1).

With respect to the evidence on effectiveness of LT, a study was considered 'positive' if its author(s) concluded that laser treatment was more effective than the reference treatment. This usually corresponded with a statistical significant difference between treatments. However, a number of studies also reached this conclusion on pre-post group comparisons; we corrected for this in the result section. A study was labelled 'negative' if no difference between the study treatments was reported or if the reference treatment showed better results.

Results

Twenty-five trials[21-45] were identified that used some form of randomization and compared 904 nm LT with a different intervention or placebo therapy. Of these, four papers were excluded from further reviewing; three were controlled clinical trials but not randomized[42,43,44] and one trial used healthy individuals, instead of patients.[45] One author[40,41] presented different arms of the same trial in separate articles. Since the methodology was reported differently as well, we describe the results of each arm (article) separately. The methodological characteristics of the remaining 21 trials are presented in Table 2, ranked according to their methodological score. The methodological score ranged from 6 to 65 points. Seven out of the 21 trials used a cross-over design.[21,26-29,36,38] The descriptions of inclusion and exclusion criteria (A), the intervention (G) and blinding of the patient, therapist and/or observer (I, J, K) were in general, satisfactory. The descriptions of prognostic comparability (C), dropouts (E), co-interventions (H) and side effects (N) were rather disappointing.

Quite a number of studies mentioned that the allocation procedure was randomized, but failed to mention how this was done or if the method of randomization was concealed. Some studies showed obvious flaws in their design leading to bias. Vasseljen[41] mentioned that patients and physiotherapists were 'fully aware of the treatment being given', whereas in the study of England et al.[25] the therapist was not blinded 'for reasons of safety and practicality'. Rather large dropout rates were reported in the studies of Flöter et al.[26] and Lucas et al.,[32] where more than 15% and 27% of the patients dropped out, respectively. Incorrect allocation procedures were suspected in the studies by Longo et al.[31] and Dolan et al.[24] In the latter study the person who was responsible for the randomization had experienced 'serious ethical problems' when he had randomized patients to the control group.

Table 2: Methodological item scores per study

maximal score per item	A 4	B 20	C 6	D 9	E 7	F 3	G 12	H 2	I 6	J 6	K 6	L 5	M 5	N 5	O 3	P 1	Q 3	Tot 100
Olavi, 1988 [36]							5								1			6
Beard, 1990 [21]							8						2					10
Gobelet, 1986 [27]							2		4	4					1			11
Dolan, 1988 [24]							8			4					1			13
Meier, 1988 [34]	4		2	3			6											15
Bihari, 1989 [22]							6	2	4		4			2				18
Longo, 1988 [31]	2		2				4	2			4		5					19
Seichert, 1987 [38]							9		4					2	3	1		19
Jensen, 1987 [29]	2				5		10			4								25
England, 1989 [25]	2			3			9		4		4	2			1			25
Ceccherelli, 1989 [23]	4			6			6		4	4			5		1			30
Flöter, 1988 [26]		10	2				10		4	4		1						31
Nivbrant, 1989 [35]	2	10		3	1		10		4	4		3	2		1			40
Lucas, 1995 [32]	2			9	2		8	2	4	4	4	2			3	1		41
Hansen, 1990 [28]	4	10					7		6	6	6	3		2				44
Rogvi, 1991 [37]	4	10		9	1		7	2		4	4	2	5					48
Siebert, 1986 [39]	2	10	2	9			10	2	4		4	5	5					53
Lundeberg, 1987 [33]	4	20					8		4	4	4	4	5		1			54
Vasseljen, 1992 [40]	4	10		9	7	3	10	2					5		3	1		54
Klein, 1990 [30]	4	10		9		3	10	2	4	4	4	5		2	3	1		58
Vasseljen, 1992 [41]	4	20		9	5		10	2	4			5	5		3	1		65

This study also allowed for many co-interventions and groups were prognostically not comparable. Taken together the reviewed studies used 70 measures of effect in total, of which 30 were related to pain and 3 to the use of analgesics. Five studies reported on range of motion (ROM) and five on influence of LT on activities of daily life. Additional measures of outcome were very diverse, and made an estimate of the overall effect of LT difficult. Quite a number of outcome measures per study has been used. However, comparison of outcomes between studies was hampered by incomparability of patient groups or diseases studied.

904 nm laser was used for the treatment of various diseases. Three studies reported on the efficacy of LT in rheumatoid arthritis,[21,27,38] of which two studies reported no effect and only Gobelet et al.[21] found a positive effect using a pre-post comparison of the data. No between groups effect was present. Two out of three studies[23,26] reported positive effects regarding the efficacy of LT on myofascial pain; the other study[28] reported negative effects. The positive study by Floter et al.[26] was hampered by a dropout rate of 15%.

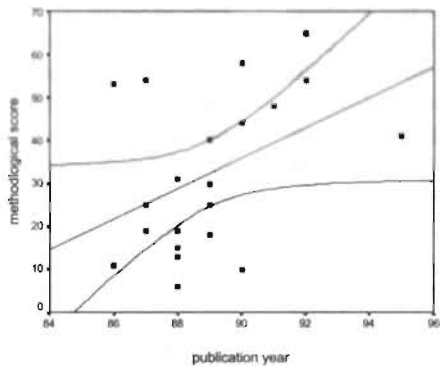
Four studies [24,34,35,37] reported the efficacy of LT in knee problems. Nivbrant et al.[35] and Rogvi et al.[37] found a positive, although not significant, trend towards effectiveness, whereas Jenssen et al.[24] found no effect. The positive results from Meier et al.[34] were due to a pre-post comparison of the data, and disappeared in a between groups analysis. Only the study by Olavi et al.[36] reported on the efficacy of LT on trigger points and showed favourable results, but this study was hampered by a rather large amount of non-informativeness (see Table 2). Three studies [22,24,32] reported on efficacy of LT in the treatment of (pressure) ulcers. Although all three studies were positive, no firm conclusions about efficacy could be drawn here. The positive trend noted by Bihari et al.[22] was due to an incorrect analysis procedure, whereas the positive study by Dolan et al.[24] was hampered by an incorrect allocation procedure, many co-interventions, and

Table 4: Summary of the results of the review per disease

Disease (number of studies)	Outcome according to author(s)	Outcome according to reviewers
rheumatoid arthritis (n=3)	$+^{23}$, $-^{27}$, $-^{36}$	$-^{\Phi}$, $-^{\Phi}$, $-^{\Phi}$
myofascial pain (n=3)	$+^{25}$, $+^{26}$, $-^{29}$	$+^{\Phi}$, $+^{\Phi}$, $-$
knee problems (n=4)	$+^{35}$, $+^{37}$, $+^{34}$, $-^{24}$	$+$, $+$, $-^{\Phi}$, $-^{\Phi}$
trigger points (n=1)	$+^{36}$	
pressure ulcers (n=3)	$+^{22}$, $+^{24}$, $+^{32}$	$-$, $+^{\Phi}$, $+^{\Phi}$
tennis elbow (n=4)	$-^{33}$, $-^{39}$, $-^{40}$, $-^{41}$	$-$, $-$, $-$, $-$
tendonitis (n=1)	$+^{25}$	$+^{\Phi}$
low back pain (n=2)	$+^{30}$, $-^{31}$	$-$, $-^{\Phi}$

Pre-post comparisons are recalculated as between comparisons.
Superscript numbers identify the study, where; - = not effective, + = effective and ? = undecided.
$^{\Phi}$ identifies studies with a quality score of less than 40 points.

Figure 1: Methodological score versus publication year



prognostically incomparable groups; and the positive study by Lucas et al.[32] was flawed by a dropout rate of 27%. Five studies reported on the efficacy of LT in tennis elbow [33,39-41] or tendinitis.[25] Only the study on tendonitis [25] reported positive results, whereas in all studies on tennis elbow, no effect was noted. Finally, two studies reported on the efficacy of LT in low back pain.[30,31] Klein et al.[30] observed no effect but used a pre-post comparison analysis procedure, whereas Longo et al.[31] found an effect but an incorrect allocation procedure was suspected in this study.

Table 4 summarizes the available evidence per disease; as can be seen from this, there is no single disease in which LT obviously excels in effectiveness. Studies reporting efficacy (according to the authors) had used, on average dose of 1.3 J/cm^2, whereas studies reporting no results had used an average dose of 2.1 J/cm^2. The use of different dosages per study was not apparently related to the year of publication of the study.

However, it is interesting to note that the methodological quality of the studies seems to be increasing over time (see Figure 1). It is also noteworthy that studies with a negative outcome (according to the authors) have on average a significant better methodological score (p=0.002). The average methodological score for positive studies is 24 (sd=13), and for negative studies is 47 (sd=16). This phenomenon may provide an indication of publication bias, showing that positive

studies with low methodological quality get published more easily.

The methodological score of the studies is based on the items that are reported and well performed. Figure 2 shows that the missing points are mainly due to lack of information and only to a minor extent to bias.

Discussion

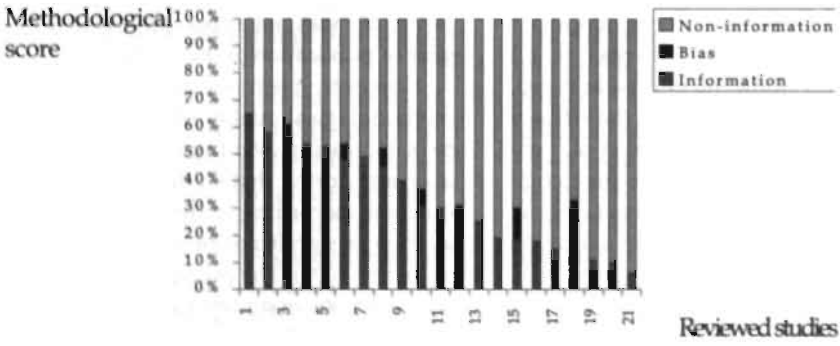
The value of a literature review depends on the success in obtaining the results of all trials (RCTs) which have been conducted on the issue of interest. It is possible that relevant studies reported in fora not accessible to us or in languages incomprehensible to us were omitted from this review. There are also indications that (especially) small clinical trials with negative results are not as easily published as small positive trials.[46] Thus, publication bias could form a threat to the validity of the results presented here.

There is a considerable number of randomised clinical trials that study the effect of LT. However, many of the reviewed studies showed serious methodological flaws, and much information was lacking in the publications. When assessing methodological quality, the primary goal is to achieve an estimate of both the effects and the chance of bias in the results of the studies performed.[47,48] One method of assessing quality is by using a criteria list which tries to measure internal validity, precision of the study and relevance of the choices with respect to population, interventions and measures of effect. Assigning weights to these criteria anticipates the argument that some errors in trial design are more crucial than others. Although controversial, weighting does give some insight into the quality of the studies performed and provides an overview of the credibility of the results. It also enables the reviewed studies to be ranked to some extent, according to their methodological quality.

The fact that systematic reviews use reported material to judge the quality of the trials under consideration is a consequence of this type of research, but is also cause for concern. As seen in this review, the methodological score is more influenced by lack of information than by methodological flaws (Figure 2). This is a pity, since lack of information could have been easily avoided by authors of studies or journal editors.

Of the 21 trials in our review, seven reported results that were not based upon an appropriate data-analysis procedure. When these positive studies were corrected by appropriate analysis (that is, analysis between groups instead of within groups), no effects of LT were found.

Figure 2: Informativeness, non informativeness and bias per study



A study of the dosage in these trials did not reveal firm relationships between the dosage and the outcome of the study. In cases where no dosages were provided, but enough parameters were at hand to calculate the dose at skin level, this was done. Where necessary, specifications from the laser industry were obtained to be able to calculate the applied dosages as accurate as possible. It should be noted that the actual laser dose is often lower than the one cited by the authors, or the laser industry. Second, 904 nm lasers are notorious for problems with cooling the diode, since the output critically depends on the rise in temperature of the laser resonator. The higher the temperature, the lower the output; as a consequence, 904 nm lasers lose (on average) about 20% of their output in the first half hour.[49] A third factor which influences output negatively is the optical system of the laser apparatus; mirrors, fibre-optics and lenses each reduce the output power by around 25%. For instance a three mirror system, often used in laser devices capable of scanning an area, lose 75% in output owing to these mirrors. Fourth, divergence of the laser beam, which can amount to up to 35^0 in 904 nm lasers, results in loss of power density.[50] Finally, reflection, refraction and absorption play an important role in diminishing output power, especially in laser scanning devices. Although these mechanisms are also claimed to be essential parts of the working mechanism, when output is lowered before reaching the target tissue these phenomena work as barriers, preventing adequate dose delivery in the target tissue.

The 21 RCTs studied can be considered the best available evidence when studying the efficacy of 904 nm LT in musculoskeletal disorders. Because of their use of random allocation of the patients and the use of control treatments, their potential to supply valid ans-wers is much larger than that of uncontrolled or nonrandomized controlled studies. Nevertheless, the observed study quality ranged from very poor (6 points) to only reasonable (65 points). Many (avoidable) errors in design and data-analysis were noted. Therefore, we suggest that in the future more attention should be paid to larger sample sizes, improved prognostic comparability of the groups, and avoiding drop-outs and co-interventions. Furthermore, mentioning side-effects and the use of between group comparisons could help answer the question of whether 904 nm LT has favourable effects in musculoskeletal disorders. Last but not least, better reporting of future studies is imperative.

We did not pool the results of the trials because patient characteristics, illnesses studied and treatments given were not similar enough to allow for pooling, neither as a total nor in subgroups. In addition, the methodological quality of the studies was low. Furthermore, sixteen out of 21 studies did not present data that allowed pooling.

In conclusion, the results clearly show that the efficacy of LT in musculoskeletal disorders is questionable. In none of the studied diseases in which LT is supposed to be effective could firm evidence be provided that LT was superior to placebo, sham or other treatment modalities. Also, outcome measures such as general improvement or change in range of movement, failed to show advantageous effects of LT.

We conclude that there is little evidence that 904 nm LT is effective in musculoskeletal disorders. Larger trials with better methodological quality could provide more definite and convincing answers.

References

1. Flock M, Zweng HC. Laser coagulation of ocular tissues. *Archives of Ophtalmology 1964*;72:604.
2. Goldman L. Wilson R, Hornby P et al. Laser radiation of malignancy in man. *Cancer 1965*;1:93-101.
3. Goldman JA. Investigative studies of laser technology in rheumatology and immunology. Biomedical Laser Technology Clinical Applications New York: Springer Verlag, 1981.
4. Wei X. Laser treatment of common diseases in surgery and acupuncture in the people's republic of China: preliminary report. *Acupuncture Electro 1981*; 6:19-31.
5. Gallachi G, M uller W. Akupunktur und Laserstrahlenbe handlung beim Cervical- und

Lum-balsyndrom. *Z Phys Med 1981;2:95-102.*
6. Lewith GT, Machin DA. A randomized trial to evaluate the effect of infra-red stimulation of local trigger points, versus placebo, on the pain caused by cervical osteoarthrosis. *Acupuncture Electro 1981; 6:277-84.*
7. Diamanthopoulos C. Bioenergetics and tissue optics. In: GD Baxter. Therapeutic lasers: theory and practice. Churchill Livingstone, Edinburgh, 1994.
8. Belkin M, Schwartz M. New biological phenomena associated with laser radiation. *Health Phys 1989;4:* 141-50.
9. Karu TI.Molecular mechanisms of the therapeutic effect of low intensity laser irradiation. *Laser in Life Sciences 1988;2:53-74.*
10. Basford JR. Low energy laser therapy: controversies and new research findings. *Lasers in Surgery and Medicine 1989;9:1-5.*
11. Keijzer M, Jacques SL, Prahl SA et al. Light distributions in artery tissue: Monte Carlo simulations for finite laser beams. *Lasers in Surgery and Medicine 1989;9:148-54*
12. King PR. Low level laser therapy: a review. *Physiotherapy Theory and Practice 1990;6:127-38.*
13. Kitchen SS, Partridge CJ. A review of low level laser therapy. *Physiotherapy, 1991;77:161-168.*
14. Bouma MG, Buurman WA, van den Wildenberg FAJM. Low energy laser irradiation fails to modulate the inflammatory function of human monocytes and endothelial cells. *Lasers in Surgery and Medicine 1996;19:207-215.*
15. Pocock SJ. *Clinical trials; a practical approach.* John Wiley & Sons, Chichester, 1991.
16. Meinert CL. Clinical trials; design, conduct and analysis. Oxford University Press, New York, 1986.
17. ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain. A criteria based meta-analysis. *J Clin Epid. 1990;43:1191-99.*
18. Koes BW, Bouter LM. van der Heijden GJMG, Knipschild PG. Physiotherapy exercises and back pain. *Br Med J 1991;302:1572-76.*
19. van der Heijden GJMG, Beurskens AJHM, Koes BW, Assendelft WJJ, de Vet HCW, Bouter LM. Traction for back and neck pain: a blinded review. *Phys Ther 1995;75:93-104.*
20. Assendelft WJJ, Koes BW, Knipschild PG, Bouter LM. The relation between methodological quality and conclusions in reviews of spinal manipulation. *JAMA 1995;274:1942-48.*
21. Beard M. Third international physiotherapy congress, Hong Kong. Link Printing Pty. Ltd, Sydney, N.S.W. 1990.
22. Bihari I, Mester AR. The biostimulative effects of low level laser therapy of longstanding crural ulcers using helium neon laser, helium neon plus infrared lasers, and noncoherent light: preliminary report of a randomized double blind comparative study. *Laser Ther 1989;1:75-8.*
23. Ceccherelli F, Altafini L, Lo Castro G, Avila A, Ambrosio F, Giron GP. Diode laser in cervical myofascial pain: a double blind study versus placebo. *Clin J Pain 1989;5:301-4.*
24. Dolan M, Spiker T, Valkenburg P, Sterenborg HJCM. Infra-rood softlaser behandeling van decubitus. *Tijdschrift voor Fysiotherapie 1989;7:* 124-40.
25. England S, Farrell AJ, Coppock JS, Struthers G, Bacon PA. Low power laser therapy of shoulder tendonitis. *Scand J Rheumatol 1989;18:* 427-31.
26. Flöter Th, Rehfisch HP. Schmerzbehandlung mit laser. Eine doppelblind-studie. *Top Medizin 1990;4:* 53-7.
27. Gobelet C, Meier JL, Volken H. Mid ou mythe laser et pathologie abarticulaire. In: Simon L. Actualite's reeducation fonctionelle readaptation. Paris: Masson, 1986 (IIe serie).
28. Hansen HJ, Thoroe U. Low power laser biostimulation of chronic oro-facial pain. A double-blind placebo controlled cross-over study in 40 patients. *Pain 1990;43:169-79*
29. Jensen H, Harreby M, Kjer J. Is infrared laser effective in painful arthrosis of the knee ? {Infrarod laser - effekt ved smertende knaeartose?} *Ugeskr Laeger 1987;149:3104-6.*
30. Klein RG, Eek BC. Low energy laser treatment and exercise for chronic low back pain: double blind controlled trial. *Arch Phys Med Rehab 1990;71:34-7.*
31. Longo L, Tamburini A, Monti A, Cattaneo L, Sesti AG. Treatment with 904 nm and 10600 nm laser of acute lumbago: double blind control. *Laser. Clinical research. 1988;16-20.*
32. Lucas C, Moll WAW, Coenen CHM Low level laser therapy bij decubitus stadium III. Een dubbelblind, placebo gecontroleerd effectonderzoek. Afdeling Contractactiviteiten, Faculteit Gezondheidszorg, Hogeschool van Amsterdam; 1-53
33. Lundeberg T, Haker E, Thomas M. Effect of laser versus placebo in tennis elbow. *Scand J Rehabil Med 1987;19:135-8.*
34. Meier JL, Kerkour K. Traitement laser de la tendinite. *Med Hyg 1988;46:907-11.*
35. Nivbrant B, Friberg S. Laser tycks ha effekt på knäledsartros men vetenskapligt bevis saknas. *Läkartidningen 1992;89:859-61.*
36. Olavi A, Pekka R, Kolari Pertti J Effect of the infrared laser therapy at treated and non-treated trigger points. *Int J Acupuncture & Electro-Therapeutics Res. 1989;14:9-14*
37. Rogvi-Hansen B, Ellitsgaard N, Funch M, Dall-Jensen M, Prieske J. Low level laser treatment of chondromalacia patellae. *International Orthopaedics 1991;15:359-61.*
38. Seichert N, Sch`ps P, Siebert W, Schnizer W, Liebmeister R. Wirkung einer Infrarot-Laser-Therapie bei weichteilrheumatischen Beschwerden im Doppelblindversuch. *Therapiewoche 1987;37:1375-9.*
39. Siebert W, Seichert N, Siebert B, Wirth JC. What is the efficacy of 'soft' and 'mid' lasers in therapy of tendinopathies? *Arch Orthop Traum Su 1987;106:* 358-63.
40. Vasseljen O, Hoeg N, Kjeldstad B, Johnsson A, Larsen S. Low level laser versus placebo in the

treatment of tennis elbow. *Scan J Rehab Med 1992*; 24:37-42.
41. Vasseljen O. Low-level laser versus traditional physiotherapy in the treatment of tennis elbow. *Physiotherapy 1992*;78:329-43.
42. Seichert N, Siebert, B, Schöps P. Die Soft und Mid Lasertherapie in der Physikalischen Medizin. Eine kritische Discussion. *Z Phys Med Baln Med Klim 1986*;15:400-4.
43. Gudmundsen J, Vikne J. Laserbehandling av epicondylitis humeri og rotatorcuffsyndrom. *Nor Tidskr Idrettsmed. 1987*;2:6-15.
44. Flöter T. Laser in the management of chronic pain. *Scand J Acupuncture Electrotherapy 1987*;2:18-21.
45. Greathouse DG, Currier DP, Gilmore RL. Effects of clinical infrared laser on superficial radial nerve conduction. *Phys Ther 1985*;65:1184-7.
46. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith Jr H. Publication bias and clinical trials. *Contr Clin Trials 1987*;8:343-53.
47. Lipsey MW. Identifying potentially interesting variables and analysis opportunities. In: The handbook of research synthesis; eds. Cooper H & Hedges LV. Russell Sage Foundation, New York,1994.
48. de Bie RA. Methodology of systematic reviews: an introduction. *Phys Ther Rev 1996*;1:47-51.
49. Dolan M, Spiker T, Valkenburg P. Laseronderzoek. Deel I: een experimenteel placebo onderzoek naar de invloed van een infrarood laserbehandeling op de wondgenezing bij decubitus. Den Haag: Haagse Hogeschool, 1988.
50. Medisch Technologische Dienst TNO. MTD-markt-overzicht Laserapparatuur, fysiotherapie. Leiden, MTD-TNO, 1989.

5 Quality assessment of trials: a comparison of three criteria lists.

AUTHORS:
Arianne P. Verhagen,
Robert A. de Bie,
Anton F. Lenssen,
Henrica C.W. de Vet,
Alphons G.H. Kessels,
Maarten Boers,
Piet A. van den Brandt.

ABSTRACT

Objective. The conclusion of a systematic review depends on the quality of the individual studies included. This article presents the results of a comparison of three different methods of quality assessment.

Method. A data set of 21 randomized clinical trials (RCTs) from a systematic review concerning the efficacy of laser therapy in patients with musculoskeletal disorders is used. The criteria lists to assess the methodological quality were the 'Maastricht' list, the 'Jadad' list and the 'Delphi' list.

Results. The three criteria lists show moderate to good correlation. Major differences between the lists are the number of items, and differences in wording of the items. The latter seem to affect the ranking of the studies.

Conclusion. Based on our results we conclude that the Delphi list seems a practical and satisfactory instrument for quality assessment of RCTs. *(Submitted)*

Many systematic reviews rely on measurement of the methodological quality of the individual trials. Various ways of assessing the quality of randomized clinical trials (RCTs) are used, such as quality scales, criteria lists and checklists.[1] A criteria list consists of items concerning different methodological aspects of RCTs. It can provide a quality score (QS) as an estimation of the methodological quality of the design and conduct of the trials by summation of the various items fulfilled.[2] These quality scores result in a hierarchical list in which higher scores indicate studies with a better methodological quality.[2,3] They can also be used as a 'threshold score' for inclusion of the article in a review, or as a 'weighting factor' in the statistical analysis.[4-6] Sometimes a visual plot of the effect size against a QS is presented.[7] Apart from providing quality scores a criteria list can also be used as a list to provide (peer review) guidelines for investigators about the content of the trial report.[3,5,6]

The first criteria list was developed in 1981 by Chalmers et al.[8] Fifteen years later at least 16 more lists have been developed.[6] These criteria lists were designed to measure the quality of the trials or the trial reports and were often designed for specific research areas. The number of items in the criteria lists varies between 3 and 47. These items usually constitute 'accepted criteria', as listed in textbooks on clinical trials as aspects of importance for the quality of a trial. Only a few criteria lists were developed by formal scale development techniques.[9,10]

Different criteria lists, applied to the same set of trials, do not always provide similar results.[6] Detsky et al.[5] assessed the quality of 18 trials using three quality measuring instruments. They reported slight differences in the absolute scores but found no substantial difference in the ranking of the trials.[5] Contrary to the findings of Detsky et al.,[5] Moher et al.[6] detected considerable variance in both absolute scores and ranking of the studies when comparing six quality scales assessing the quality of 12 trials.

Differences in ranking, as a result of using different criteria lists, may cause problems when quality scores are incorporated into a systematic review or meta-analysis. Therefore empirical evidence is needed to establish whether methods in quality assessment are valid and reliable. This study compares three different methods of quality assessment. For this study we chose the two criteria lists which were developed using scale developing techniques: the 'Delphi' list and the 'Jadad' list.[9-11] The third list used is the one frequently used at our department: the 'Maastricht' list.[12] We investigated whether item choice, the number of items or the wording of items cause a difference in absolute quality scores or ranking.

Description Of The Criteria Lists

The Delphi list. This list has recently been developed using scale developing techniques.[9] A pool of items was constructed from existing criteria lists and narrowed down by means of the Delphi Consensus Technique, by an international panel of more than 25 experts in the field of quality assessment in RCTs (statisticians and epidemiologists). The Delphi list contains nine items and measures three dimensions of quality: internal validity, external validity and statistical considerations. All items have a 'yes/no/don't know' answer option. No formal description of the calculation of a QS is presented.

The Jadad list. This is a criteria list developed by Jadad et al.[10,11] For this list a pool of items was generated by a multidisciplinary panel of six 'judges' and narrowed down by means of the Nominal Group Consensus Technique.[11] The result was a set of 3 items, directly related to the reduction of bias (internal validity). All questions have 'yes/no' answers options. For the QS a maximum of 5 points can be earned: three times one point for each 'yes', and two additional points for a proper method of randomization and blinding.

The Maastricht list. This list, developed at the Department of Epidemiology of Maastricht University, is called the 'Maastricht' list, because of its origin.[12] This list is used in over 30 systematic reviews.[7,13-18] The Maastricht list consists of 15 main items based on methodological criteria. The 15 main items are divided into 47 subitems measuring three dimensions of quality: internal validity and external validity and statistical considerations. The list contains four answer options in order to determine whether information on a specific item was: '+' presented and adequately done, '?' presented but unclear, '–' presented but not adequately done or leading to bias, or '0' not presented in the publication. The items rating '+' contribute to the QS. Furthermore, weights are assigned to all items to reflect relative importance. Based on empirical evidence,[19] items on treatment allocation and blinding of patients, therapists and observers are weighted heavily in this list. Summing up the weights of the '+' rated items results in an overall QS. A methodologically perfect study receives a maximum of 100 points.

Method

Studies. A data set of 21 RCTs[22-42] from a systematic review concerning the efficacy of 904 nm laser therapy in patients with muskulosceletal disorders is used.[20] Studies with healthy subjects are excluded, and seven RCTs used a cross-over design.[26-29,35,36,39]

Assessment of the methodological quality. We combined the three criteria lists into one list including all items from the 'Delphi', the 'Jadad' and the 'Maastricht' list (see Appendix). The original guidelines for assessment of the individual lists were used. The assessment of the studies was performed independently by two of the authors (APV, AFL), followed by a consensus meeting. We compared the three lists about inclusion and wording of items and the overall QS.

Statistical methods. We calculated the QSs for each list according to the original weighting. For the Delphi list we used an equal weight (of one point) for each item. To compare the different lists with each other we present QSs as percentages of the maximum score. In order to assess ranking differences we calculated Spearman rank correlation coefficients between the three lists. To study the influence of a different wording we also calculated Spearman rank correlation coefficients for the sum score on each list on items concerning randomization, blinding and withdrawals only. Prior to analyzing the data, we defined a correlation coefficient of $r \geq 0.7$ as good, $0.7 > r \geq 0.5$ as moderate and $r < 0.5$ as poor.

Results

Quality scores. The quality scores obtained with the different criteria lists are presented, as a percentage of the maximum score in Table 1. The Delphi scores vary between 1 and 6 points out of 9 points (11%-66%), the Jadad scores between 1 and 4 points out of 5 points (20%-80%) and the Maastricht scores vary between 6 and 65 points out of 100 points (6%-65%). None of the studies achieved the maximum score on any of the used criteria lists. Mean QSs on each criteria list vary from 32.3% on the Maastricht list, 45.1% on the Delphi lists, to 52% on the Jadad list. RCTs with a cross-over design had relatively low QS on all criteria lists compared to the parallel RCTs. Studies are ranked in decreasing order according to the Delphi list. The QSs of the three lists are comparable, especially the ranking of the Maastricht and Delphi lists correspond well. Overall, the Jadad quality scores are higher (mean 52%) than the quality scores on both other lists, while the overall Maastricht quality scores are lower (mean 32.3%). In 14 out of 21 studies the Jadad quality score is the highest, once (number 11)[42] the Maastricht QS is higher than the Delphi score and once (number 10)[25] the Maastricht QS is higher than the Jadad score.

Table 1: Quality scores for the RCTs with a cross-over design and the concurrent RCTs, ranked according to the Delphi quality scores.

		Quality Scores (rank)		
	Author	Delphi list	Jadad list	Maastricht list
1	Vasseljen (1992a)	66(1)	80(1)	65(1)
2	Klein (1990)	66(1)	80(1)	58(2)
3	Lundeberg (1987)	66(1)	60(2)	54(3)
4	Siebert (1986)	66(1)	60(2)	53(4)
5	Hansen (1990)*	66(1)	60(2)	44(6)
6	Lucas (1995)	66(1)	80(1)	41(7)
7	Nivbrant (1989)	66(1)	80(1)	40(8)
8	Rogvi (1991)	55(2)	80(1)	48(5)
9	Cecchelli (1989)	55(2)	40(3)	30(10)
10	England (1989)	55(2)	20(4)	25(11)
11	Vasseljen (1992b)	44(3)	60(2)	54(3)
12	Flöter (1988)*	44(3)	40(3)	31(9)
13	Jensen (1987)*	44(3)	60(2)	25(11)
14	Longo (1988)	33(4)	40(3)	19(12)
15	Bihari (1989)	33(4)	40(3)	18(13)
16	Gobelet (1986)*	33(4)	20(4)	11(16)
17	Seichert (1987)*	22(5)	40(3)	19(12)
18	Meier (1988)	22(5)	40(3)	15(14)
19	Dolan (1988)	22(5)	40(3)	13(15)
20	Beard (1990)*	11(6)	40(3)	10(17)
21	Olavi (1988)*	11(6)	40(3)	6(18)
	Mean QS	45.1%	52%	32.3%

* RCT with cross-over design

In the study of Vasseljen (number 11)[42] the higher Maastricht QS compared with the Delphi score is because of the proper description of dropouts, losses to follow-up, the intervention and the use of a blinded data analist. Two studies (numbers 10 and 16)[25,27] achieved the lowest QS on the Jadad list, a comparable low Maastricht score, but a much higher Delphi score. This discrepancy is especially large in the study of England et al. (number 12).[27] On the Jadad list England et al.[27] received only one point for being randomized. On the Maastricht and Delphi list they also received points for blinding the patient and observer and for the presentation of the data. On most other items on the Maastricht list they did not receive additional points resulting in an overall low Maastricht QS.

Table 2: Spearman correlation coefficients of the overall sumscores between the three criteria lists (In parentheses: Spearman correlation coefficients on the sum scores of items concerning randomization, blinding and withdrawals only).

	Delphi	Jadad	Maastricht
Delphi	1	0.71 (0.41)	0.87 (0.73)
Jadad		1	0.78 (0.72)
Maastricht			1

Correlation coefficient. The Spearman rank correlation coefficients of the overall QSs are presented in Table 2. In parenthesis the Spearman correlation coefficients are presented, calculated on the items about randomization, blinding and withdrawals only.

The correlation coefficients on the overall QSs between the three lists are good (≥ 0.7). The difference in wording between the Maastricht & Delphi lists does not seem to affect the ranking of the studies. Obviously the difference in wording between the Jadad & Delphi list creates a major ranking difference (Spearman of 0.41). This difference is mainly due to the difference in scoring on the blinding item. Most studies describe themselves as double blinded in the title or abstract, but often in the report no information about blinding was presented.

Main Elements Of The Criteria Lists.

The three criteria lists all aim to measure the methodological quality of RCTs, but include different items in their list. In this paragraph we describe differences and similarities between the criteria lists, when applied to the same study set.

Treatment allocation. All criteria lists contain items whether randomization is performed and items about the randomization procedure, although the wording of these items slightly differs (see Appendix items D.1a/1b;J.1;M.2). To judge whether a method of randomization is performed requires more information in the original report than to judge if the study is described as randomized. The term 'random' is sometimes presented only in the title or summary of the report. All criteria lists ask for additional information about the randomization procedure. Four studies performed a blinded treatment allocation.[26,29,33,40] In the two cross-over studies which mentioned blinded treatment allocation,[26,28] no information about the procedure is presented. Concerning the items on treatment allocation only small differences in scoring appeared between the criteria lists.

Blinding. All criteria lists contain one or more items about blinding. The Jadad list does not discriminate between the different levels of blinding (patient, therapist or observer). An extra point could be earned when the method of 'double blinding' was regarded 'appropriate'. Nearly all studies (n=18) obtained at least one point because they described their study as 'double blind', however sometimes only in the title of the report. Moreover, although studies described themselves as 'double blind', some studies did not describe whether the patient, observer or therapist was blinded.[34,36,37] The Maastricht list contains an item to determine if the blinding procedure was evaluated and successful, and one report[28] provided this information. Twice the use of a blinded data-analist (asked only by the Maastricht list) is mentioned.[30,32]

Withdrawals. The Maastricht and the Jadad lists contain one or more items concerning withdrawals, the Delphi list does not. On this topic both the Maastricht and Jadad lists show major differences. The Maastricht list contains items about whether there are withdrawals and whether it may cause bias when the withdrawal rate is too high [meaning > 5% dropouts (= withdrawal during the treatment period) and > 20% loss to follow-up (= withdrawal after the treatment period)]. The Jadad list asks for 'a description of withdrawals and dropouts'. Six studies described withdrawals and dropouts,[29,32,35,38,41,42] but in three the dropout rate was less than 5%.[29,41,42] Studies with a high withdrawal rate can receive a 'yes' on the Jadad list and a '--' on the Maastricht list, i.e. 'leading to bias'.[26,38] In the Jadad list it contributes to the QS while according to the Maastricht list it can be an important source of bias.

Analysis. The Maastricht and the Delphi list both contain items concerning the analysis while the Jadad list does not. Both lists check if an 'intention-to-treat' analysis is performed. This item can also receive a 'yes' score when there were no dropouts and the compliance was good: in such a case the analysis is intention-to-treat by default. There was no study without dropouts, and no study mentioned the use of or performed an 'intention-to-treat' analysis. Also information about the presentation of 'point estimates and measures of variability' was asked by both lists. In four studies means and confidence intervals were presented.[25,35,41,42] Concerning the analysis issues there is hardly any difference between the Maastricht and the Delphi list.

Bias. Only the Maastricht list contains the answer option '--' meaning: 'not adequately done or leading to bias'. In 7 out of 21 studies this answer option was used in one or more items. For instance: reading the report of Longo et al.[31] the reviewers had serious doubts whether the allocation procedure was adequate and not leading to bias. Although the word 'randomly' was used, the sentence '.... the second doctor, who put him (the patient, APV) in one of the three groups....' made the reviewers suspicious. Vasseljen[42] mentioned the patients and physiotherapists to be 'fully aware of the treatment being given' (active laser versus traditional physiotherapy). According to the reviewers (AFL,APV) there is a difference between 'not blinded' and 'fully aware'. In studies with difficulties to blind patients they could try to keep the patients 'naive' for the treatment, meaning that the patients do not exactly know what the alternative treatment is. It was certainly not done in this study. In the study of England et al.[25] the therapist was not blinded 'for reasons of safety and practicality'. The researchers doubted the validity of these reasons.

Rogvi et al.[38] mentioned a 10% dropout rate, most in the sham treatment group, and Flöter and Rehfisch[26] mentioned a selective dropout rate of more than 15%. Lucas et al.[32] did not describe a dropout rate but, after calculations of the reviewers the dropout rate in this study appeared to be more than 29% of which 8 patients (= 21%) having 'data which are not representative'. According to the Maastricht list any dropout rate of more than 5% without presenting reasons for dropouts, is possibly leading to bias. Description of number and reason of dropout in each group is required to judge whether bias is (un)likely. Dolan et al.[24] received a '--' answer on more than one item. Firstly, when selecting the patients a non-homogeneous group was formed of patients with all kinds of pressure ulcers. The person performing the randomization first divided the patients in different strata (of place and severity of the ulcers) and then he experienced ethical problems randomizing the patients to the control group. Therefore the groups were not comparable for important prognostic characteristics. Secondly, many co-interventions were allowed which did not enhance comparability between the groups.

Discussion

This study compares three quality criteria lists. Unfortunately no 'gold standard' for quality assessment exists. No major differences in overall quality score or ranking of the studies are found using different lists. It is, on the other hand, possible that combining the three lists into one combined criteria list might have affected our findings. Overall the Maastricht list gives a lower estimation of the quality compared with the Delphi list, and the Jadad list a higher one. Knowing these systematic differences, we consider, based on the correlation coefficients between the lists, all three lists equally valid instruments in quality assessment.

In the development of the Jadad list the methodological quality was defined as the internal validity of the trial.[11] The items in the Jadad list are therefore directly related to the internal validity, while the Maastricht and Delphi lists also contain items related to the external validity and the statistical analysis of a report. Both the Maastricht and Delphi list do not define the methodological quality of an RCT explicitly. The question can be raised: 'Does the methodological quality of a study only relate to the domain of internal validity (Jadad list) or is there more (Maastricht and Delphi lists) to relate to?' No conclusion in this ongoing debate about the correctness of the definition of quality can yet be drawn.

Another difference between the three criteria lists, apart from the difference in domains of quality, is the number of items. All items from the Delphi list are in the Maastricht list. Furthermore other differences between the Maastricht and the Delphi list on the one hand and the Jadad list on the other, are the wording of the items. In the Jadad list only a description of randomization, blinding and withdrawals is asked. For the Maastricht and Delphi lists ask more detailed information about the procedure and are more focussed on the performance of the study rather than on the description of certain elements. This difference in wording seems to affect the ranking of the studies. The wording of the Jadad items may well be chosen pragmatically based on the conclusions of the SORT-group (Standard of Reporting Trials),[21] in which is stated that the reporting of a trial should first improve before the quality (here defined as internal validity) can be assessed. The shortness of the Jadad list (3 items) as well as the difference in wording may have affected the correlation coefficients more than the difference in view on quality.

Based on empirical evidence [19] we consider a blinded (or concealed) treatment allocation and blinding of the patient, therapist or observer important in preventing bias. According to the randomization procedure, additional information about the procedure is gathered by all three criteria lists. We consider information about the blinding procedure and its succesfulness, as important as information about the randomization procedure, yet hardly any report provides it.

Withdrawals in a study can be a source of bias especially when there is selective withdrawal. It seems illogical that studies receive a point on the Jadad list for a description of the withdrawal rate. A high withdrawal rate can be a serious threat on the internal validity of the trial.

With the '--' answer option (used in the Maastricht list) the reviewer is, to some extent, able to distinguish between information not given (the '0' answer option) and possible souces of bias ('--'). Information not provided in a report might hide a possible cause of bias, but when information deserves the '--' answer option it is considered clearly a bias in the conduct of the trial. To determine if something is 'not adequately done or leading to bias' reviewers need to have an epidemiological background. A '--' score on one or more items may give the reviewer the opportunity to subtract points from the quality score because of clear causes of bias, or the reviewer can present the amount of biased information graphically.[20]

Despite the differences in scoring between the lists, they show a rather high correlation. Issues as external validity and statistical analysis appeared to have no major influence

on the absolute quality scores and the ranking. What is the optimal number of items necessary to provide enough information about the trial quality? The high correlation between the Maastricht and Delphi lists indicates that using the long Maastricht list did not provide important additional information.

CONCLUSION

In choosing a criteria list for quality assessment two issues are important. Essential is whether the reviewer is able to get a good picture of the validity of the trial using a specific criteria list. Another issue is whether the criteria list is easy to handle. Concerning this question the answer is: the shorter the criteria list is the better. In this data set, the short Jadad list appeared to be less sensitive for differences in methodological quality, while the extensive Maastricht list is less practical and does not provide important additional information compared with the Delphi list. Based on these results we conclude that the Delphi list is a satisfactory and the most practical instrument for quality assessment.

References

1. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Control Clin Trials 1995*;16:62-73.
2. Assendelft WJJ, Koes BW, van der Heijden GJMG, Bouter LM. The efficacy of chiropractic for back pain: blinded review of the relevant randomized clinical trials. *J Manipulative Physiol Ther 1992*; 15:487-94.
3. de Bie RA. Methodology of systematic reviews: an introduction. *Phys Ther Rev 1996*;1:47-51.
4. Jenicek M. Meta-analysis in medicine: Where we are and where we want to go. *J Clin Epidemiol 1989*; 42:35-44.
5. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L' Abbé JKA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol 1992*;45:255-65.
6. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technology Assessment 1996*;12; 195-208.
7. Koes BW, van Tulder MW, van der Windt DAWM, Bouter LM. The efficacy of back schools: a review of randomized clinical trials. *J Clin Epidemiol 1994*; 47: 851-62.
8. Chalmers TC, Smith H.jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. *Controlled Clin Trials 1981*;2:31-49.
9. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG. The Delphi list; a criteria list for quality assessment of Randomized Clinical Trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epid 1998*;51: 1235-41.
10. Jadad AR. Meta-analysis of randomised clinical trials in pain relief. PhD. Thesis. University of Oxford, 1994.
11. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials 1996*; 17:1-12.
12. de Vet HCW, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Knipschild PG. Systematic reviews on the basis of methodological criteria. *Physiotherapy 1997*; 83:284-9.
13. ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. *J Clin Epidemiol 1990*;43:1191-9.
14. Kleijnen J, ter Riet G, Knipschild P. Vitamin B6 in the treatment of the premenstrual syndrome: a review. *Br J Obstet Gynaecol 1990*;97:847-52.
15. van der Heijden GJMG, van der Windt DAWM, Kleijnen J, Koes BW, Bouter LM. Steroid injections for shoulder disorders: a systematic review of randomized clinical trials. *Brit J Gen Pract 1996*;46:309-16.
16. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM, Knipschild PG. Spinal manipulation and mobilization for back and neck pain: A blinded review. *Br Med J 1991*;303:1298-303.
17. Knipschild PG. Trials and errors; alternative thoughts on the methodology of clinical trials. *Br Med J 1993*; 306:1706-7.
18. Knipschild PG. Systematic reviews: some examples. *Br Med J 1994*;309:719-21.
19. Schultz KF, Chalmers I, Hayes R, Altman DG. Empirical evidence of bias; Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA 1995*;273:408-12.
20. de Bie RA, Verhagen AP, Lenssen AF, de Vet HCW, van den Wildenberg FAJM, Kootstra G, Knipschild PG. Efficacy of 904 nm laser therapy in musculoskeletal disorders: a systematic review. *Phys Ther Rev 1998*;3:59-72.
21. The Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials; special communication. *JAMA 1994*; 272:1926-31.

References of the systematic review

22. Bihari I, Mester AR. The biostimulative effects of low level laser therapy of longstanding crural ulcers using helium neon laser, helium neon plus infrared lasers, and noncoherent light: preliminary report of a randomized double blind comparative study. *Laser Ther 1989*;1:75-8.
23. Ceccherelli F, Altafini L, Lo Castro G, Avila A, Ambrosio F, Giron GP. Diode laser in cervical myofascial pain: a double blind study versus placebo. *Clin J Pain 1989*;5:301-4.
24. Dolan M, Spiker T, Valkenburg P, Sterenborg HJCM. [The infrared softlaser treatment of decubitus]. *Nederlands Tijdschrift voor Fysiotherapie 1989*;7: 124-40.(Dutch)
25. England S, Farrell AJ, Coppock JS, Struthers G, Bacon PA. Low power laser therapy of shoulder tendinitis. *Scand J Rheumatol 1989*;18:427-31.
26. Flöter Th, Rehfisch HP. [Pain treatment with laser; a double-blind trial]. *Top Medizin 1990*;4:53-7. (German)
27. Gobelet C, Meier JL, Volken H. [Mid- or myth laser and abarticular rheumatism]. :In: Simon L. Actualite's reeducation fonctionelle readaptation. Paris: Masson, (Ile serie). (French) 1986
28. Hansen HJ, Thoroe U. Low power laser biostimulation of chronic oro-facial pain. A double-blind placebo controlled cross-over study in 40 patients. *Pain 1990*; 43:169-79.
29. Jensen H, Harreby M, Kjer J. [Infrared laser: effect in painful arthrosis of the knee?] *Ugeskr Laeger 1987*; 149:3104-6. (Danish)
30. Klein RG, Eek BC. Low energy laser treatment and exercise for chronic low back pain: double blind controlled trial. *Arch Phys Med Rehab 1990*;71:34-7.
31. Longo L, Tamburini A, Monti A, Cattaneo L, Sesti AG. Treatment with 904 nm and 106 nm laser of acute lumbago: double blind control. *Laser. Clinical Research 1988*;16-20.
32. Lucas C, Moll WAW, Coenen CHM. [Low level laser therapy in decubitus stage III. A double blind, placebo-controlled trial]. International report Dept Contract activities, Faculty of Health Care, Hogeschool van Amsterdam;1-53. (Dutch).
33. Lundeberg T, Haker E, Thomas M. Effect of laser versus placebo in tennis elbow. *Scan J Rehab Med 1987*;19:135-8.
34. Meier JL, Kerkour K. [Laser treatment of tendinitis]. *Med Hyg 1988*;46:907-11. (French)
35. Nivbrant B, Friberg S. [Laser treatment of knee joint arthrosis sems to be effective but scientific evidence is lacking]. *Lkartidningen 1992*;89:859-61. (Swedish)
36. Olavi A, Pekka R, Kolari-Pertti J. Effect of the infrared laser therapy at treated and non-treated trigger points. *Int J Acupuncture & Electro-Therapeutics Res 1989*;14:9-14.
37. Beard M, Hansel N, Furnis A. The treatment of rheumatoid arthritis with low power laser. Third International Physiotherapy Congress, Hong Kong. Link Printing Pty. Ltd, Sydney, N.S.W. 1990
38. Rogvi-Hansen B, Ellitsgaard N, Funch M, Dall-Jensen M, Prieske J. Low level laser treatent of chondromalacia patellae. *Intern Orthopaedics 1991*; 15:359-61.
39. Seichert N, Schups P, Siebert W, Schnizer W, Liebmeister R. [A double blind, cross over trial about the efficacy of low power infrared laser therapy in rheumatic complaints]. *Therapiewoche 1987*; 37:1375-9. (German)
40. Siebert W, Seichert N, Siebert B, Wirth CJ.What is the efficacy of 'soft' and 'mid' lasers in therapy of tendinopathies? *Arch Orthop Traum Su 1987*; 106: 358-63.
41. Vasseljen O, Hoeg N, Kjeldstad B, Johnsson A, Larsen S. Low level laser versus placebo in the treatment of tennis elbow. *Scan J Rehab Med 1992a*; 24:37-42.
42. Vasseljen O. Low-level laser versus traditional physiotherapy in the treatment of tennis elbow. *Physiotherapy 1992b*; 78: 329-34

APPENDIX: combined quality assessment form

Codes with an M means: from the Maastricht list; codes with a D: Delphi list and codes with a J: Jadad list

STUDY POPULATION

M.1	Selection and restriction:		
	1 description of in- and exclusion criteria	0 ? +	
	2 restriction to a homogeneous study population	0 ? +	-
D.3	Were both inclusion and exclusion criteria specified?	Yes / no / Don't know	
J.1	Was the study described as randomised (this includes the use of words such as randomly, random and randomisation)?	Yes / no / Don't know	
	Is the method appropriate?	Yes / no / Don't know	
D.1a	Was the method of randomization performed?	Yes / no / Don't know	
D.1b	If subjects were randomly allocated to treatment groups, was the method of random allocation concealed?	Yes / no / Don't know	
M.2	Treatment allocation:		
	1 randomization	yes / no	
	2 allocation procedure adequate	0 ? +	-
	3 blinded allocation procedure	0 ? +	-
M.3	Study size		
	0 smallest group smaller than 25 subjects	0 ? +	
	1 smallest group larger than 25 subjects	0 ? +	
	2 smallest group larger than 50 subjects	0 ? +	
	3 smallest group larger than 75 subjects	0 ? +	
M.4	Prognostic comparability		
	1 duration of the complaint	0 ? +	-
	2 baselinescores for outcome measures	0 ? +	-
	3 age 0	? +	-
D.2	Were the groups similar at baseline regarding the most important prognostic characteristics?	Yes / no / Don't know	
M.5	Drop outs		
	1 no drop outs	0 ? +	-
	or		
	2 number of drop outs given in each group	0 ? +	
	3 reasons for withdrawal (of drop outs) given in each group	0 ? +	
	4 drop outs not leading to bias (less than 5%)	0 ? +	-
J.3	Was there a description of withdrawels and drop-outs?	Yes / no / Don't know	
M.6	Loss to follow-up		
	1 less than 20 % loss to follow up in all groups	0 ? +	
	2 less than 10 % loss to follow up in all groups	0 ? +	
	3 loss to follow up not leading to bias	0 ? +	-

INTERVENTIONS

M.7.1	Intervention #1 = experimental; name:		
	1 type of intervention	0 ? +	
	2 intensity of intervention parameters	0 ? +	
	3 duration of each treatment session	0 ? +	
	4 treatment frequency	0 ? +	
	5 number of treatment sessions	0 ? +	
	6 compliance presented	0 ? +	-
M.7.2	Intervention #2 = experimental / control; name:		
	IF control group: placebo control group?	YES / NO	

1	type of intervention	0 ? +	
2	intensity of intervention parameters	0 ? +	
3	duration of each treatment session	0 ? +	
4	treatment frequency	0 ? +	
5	number of treatment sessions	0 ? +	
6	compliance presented	0 ? +	-

M.8 Extra treatments

1	no co-interventions	0 ? +	-
or			
2	comparable co-interventions between groups	0 ? +	-

BLINDING

J.2 Was the study described as double blind? Yes / no / Don't know
Is the method was appropriate? Yes / no / Don't know

M.9 Blinding of patient

1	(attempt for) Blinding or naive patient	0 ? +	-
2	Blinding evaluated and successful?	0 ? +	-

D.6 Was the patient blinded? Yes / no / Don't know

M.10 Blinding of therapist

1	(attempt for) Blinding or naive therapist	0 ? +	-
2	Blinding evaluated and successful?	0 ? +	-

D.5 Was the care providor blinded? Yes / no / Don't know

M.11 Blinding of observer

1	(attempt for) Blinding or naive observer	0 ? +	-
2	Blinding evaluated and successful?	0 ? +	-

D.4 Was the outcome assessor blinded? Yes / no / Don't know

OUTCOME

M.12 Outcome measures

		measured		measured by pat/the/obs/ ?	blinded	
1	. . .	yes	no	| | | | |	0 ? +	-
2	. . .	yes	no	| | | | |	0 ? +	-
3	. ..	yes	no	| | | | |	0 ? +	-
4	. . .	yes	no	| | | | |	0 ? +	-
5	. . .	yes	no	| | | | |	0 ? +	-

M.13 Follow-up period

1	measured	0 ? +	-
2	relevant	0 ? +	-

M.14 Side effects

1	description of side effects in each group	0 ? +	-

ANALYSIS

M.15 Analysis and presentation of data

1	Use of blinded data-analist	0 ? +	-
2	frequencies	0 ? +	-
or mean & standard deviation			
or median & quartiles (for most important measurements)			
3	intention to treat analysis	0 ? +	-
or			
4	adequate corrections for base-line differences or drop outs	0 ? +	-

D.8 Did the analysis include an 'intention-to-treat' analysis? Yes / no / Don't know

D.7 Were point estimates and measures of variability presented for primary outcome measure(s)? Yes / no / Don't know

6 Efficacy of conservative interventions in the treatment of acute lateral ankle sprains: a systematic review

AUTHORS:
Robert A. de Bie,
Arianne P. Verhagen,
Anton F. Lenssen,
Henrica C.W. de Vet,
Frans A.J.M. van den Wildenberg,
Gauke Kootstra,
Paul G. Knipschild.

Abstract

Objective. This systematic review summarises the efficacy of conservative interventions in acute lateral ankle sprains.

Method. We performed computer aided searches of databases and of bibliographic indexes. Furthermore, we checked congress reports, reviews and relevant citations. Subsequently, all retrieved empirical studies were scored on methodological quality and effect sizes were calculated for days of sick leave, pain and swelling.

Results. We found 81 studies that investigated the effects of physiotherapy interventions versus other interventions or placebo interventions, in subjects with acute lateral ankle sprain. Of these, 44 fulfilled our entry criteria. Study quality ranged from poor (9 points) to rather good (70 points). Only two studies scored more than 60 points (on pulsed shortwave therapy) both and they showed no effect. Tape was found to be superior over other types of treatment, in effect shortening the duration of sick-leave, while plaster of Paris treatment seemed to prolong sick-leave. Placebo therapies delivered no positive results. Further and improved research is needed to shed more light on efficacy of other treatment interventions. *(Based on: De Bie RA, Verhagen AP, Lenssen AF, de Vet HCW, van den Wildenberg FAJM, Kootstra G, Knipschild PG. Efficacy of conservative interventions in the treatment of acute lateral ankle sprains: a systematic review. Submitted).*

Ankle sprains are one of the most common injuries of the ankle, and are most often reported in relation to sports participation.[1] They relate to the lateral ligament complex of the ankle, and are reported on as sprains, strains, inversion injuries, lateral ankle injuries and lateral ligament injuries.

There is still much debate about which therapy is most effective in the treatment of ankle sprains. There are many treatment options available, but there is little consensus which treatment is the most efficacious. Many choices regarding therapy seem to be driven by tradition or fashion rather than efficacy.

Adequate initial treatment for ankle sprains is thought to consist of rest, ice, compression and elevation (RICE)[2,3] while treatment after the initial phase can consist of plaster of Paris, taping, braces or special orthoses, often combined with all kinds of adjunct therapies. In the case where the ankle is being surgically treated, one often chooses for a plaster of Paris approach afterwards.

To shed some light on the efficacy of conservative treatment approaches for ankle sprains, we performed a systematic review to evaluate the effects of the provided therapies on the outcome measures pain, swelling and sick-leave.

Method

Trials on interventions for lateral ankle sprains were identified by searches in Medline and Embase (both up to 1996) and by checking the Database of the Cochrane Field 'Rehabilitation & Therapy' at Maastricht University, the Netherlands. Additionally, we checked Current Contents, Physiotherapy Index, reviews, congress reports and handbooks. Retrieved references were followed-up by citation tracking. Papers published in English, French, German, Dutch, Spanish, Italian, Norwegian, Swedish and Danish were eligible for inclusion. Languages outside the above mentioned range, as well as abstracts and unpublished studies were not included. The search strategy was adapted from the search strategy described by Dickersin [4] which is now widely being used by Cochrane reviewers. Keywords used to describe the design were: *randomised controlled trials, controlled clinical trials, random allocation, double blind, single blind, experiments* and *evaluation studies*. Keywords used to identify the illness were: *ankle, sprain, inversion, lateral ankle sprain, lateral ankle ligament injury* and *strain*. Keywords to identify the interventions were: *therapy, exercise, rehabilitation, bracing, taping, cast, plaster of Paris, orthosis* and all *physical therapy modalities* known to us [(pulsed) ultrasound therapy, laser therapy, (pulsed) short wave therapy, electro therapy, thermo therapy and cryo therapy].

The selected studies had to fulfil the following criteria for inclusion in the review. The subjects in the study had to suffer from a lateral ankle sprain. The therapy should consist of conservative treatment approaches and had to be contrasted with placebo, no treatment, physiotherapeutical or other interventions, while studies comparing various surgical techniques were excluded. The study design had to be a randomised clinical trial.

The papers eligible for reviewing were given to two reviewers (APV and AFL) who independently assessed the quality of all retrieved studies. In a consensus meeting they tried to reach agreement on items on which they had different opinions. If consensus could not be reached, a third reviewer (RAB) made the final decision. Table 1 shows the criteria for assessing the methodological quality of the trials. The list was originally designed by Ter Riet et al.[5], and modified over the years by Koes et al.[6], De Vet et al.[7] and Assendelft et al.[8] It is based on generally accepted principles of intervention research.[9,10] The criteria list was adapted for ankle sprains with respect to the intervention and relevant outcome measures. Studies could obtain points for methodological quality in five categories. These categories consisted of study population, interventions, blinding, outcome, and data presentation and analysis.

Table 1: Criteria for assessing methodological quality in randomised clinical trials of low level laser therapy.

	Criterion	Weights
Study population (total points=48)		
A	Homogeneity	4
B1	Randomisation procedure mentioned	10
B2	Concealed method of randomisation	10
C	Comparability of relevant baseline characteristics	6
D	Numbers of patients	8
E	Dropouts described for each study group separately	7
F	Loss to follow up not leading to bias	3
Intervention (total points=12)		
G	Intervention adequately described and performed	10
H	Co-interventions avoided or equal in study groups	2
Blinding (total points=21)		
I	Patients blinded	7
J	Therapist blinded	7
K	Observer blinded	7
Outcome (total points=12)		
L	Adequate outcome measures	5
M	Adequate follow up period	5
N	Description of side effects	2
Data presentation and analysis (total=7)		
O	Mean or frequencies of most important outcome measures presented for each	1
P	group	3
Q	Intention to treat analysis	3
	Adequate correction for base-line differences or drop outs	

A maximum score of 100 points could be obtained. To synthesise the data we pooled the data using a random effects model.[11] All outcomes are reported as effect sizes with 99% confidence intervals. Effect sizes allow comparisons among studies that address the same research hypotheses but use somewhat different manipulations and/or outcome measures. The thus obtained effects are measured in terms of their own standard deviations. Effect sizes were calculated by using Cohen's *d* or Hedges's *g*.[11,12] Tests for heterogeneity in pooled estimates were done by using chi-square statistics.[11]

Results

81 Trial reports were identified that use some form of randomisation and study various interventions for the treatment of ankle sprains and contrast them with a different intervention, placebo therapy or no therapy. Of these, 27 were excluded from further reviewing (references can be obtained from the authors): ten studies use medication versus placebo, eight describe preventive effects, six study biomechanical aspects of ankle orthosis and braces, one incorporates multiple foot injuries, one studies healthy subjects and one reports only 2-years results. Of the remaining 54 studies[1,2,3,14-58], four were reported more than once[19,24,27,30] and two 3 or 5 times.[54,58] Of duplicate trials the most

complete descriptions were used in this review.

The methodological characteristics of 44 trials are presented in Table 2. They are ranked according to their methodological score. The methodological score ranged from 9 to 70 points. Only 2 studies scored more than 60 points, implying the methodological inadequacy of 95% of the material. There was no apparent relationship between the year of publication and the methodological quality. The average methodological score for negative trials was 35 points (SD 14), and the average methodological score for positive trials was 29 points (SD 10). The blanks in the table give insight in methodological aspects of the reviewed trials that were not reported. Only three studies reported to have used a concealed randomisation method (item B).[18,36,42] Many studies were hampered by large numbers of dropouts and losses to follow-up (items E and F). A few studies even lost over 50% of the population during follow-up.[21,22,33] Blinding of patient, therapist or observer occurred seldom (items I, J, K). Side effects were scarcely reported upon (item N). Finally, presentation of data and analyses was inadequate in many cases (items P and Q).

The 44 reviewed trials contained in total 4646 patients with ankle sprains. In total 35 interventions were studied, among which 5 larger subgroups could be detected. 13 Trials reported on plaster of Paris (cast) versus another intervention, tape was studied in 19 trials, whereas a form of bracing or bandage was studied in 10 and 7 trials respectively. 11 Studies used placebo therapy as a contrast to another intervention, that mostly consisted of a physical therapy modality such as laser, ultrasound or pulsed shortwave therapy.

The two best studies are the ones by Barker et al[18] (70 points) and McGill et al[42] (63 points). Both study pulsed shortwave therapy (PST) versus placebo. Barker et al.[18] randomise patients with lateral ankle sprains over 2 groups of 34 and 39 patients. They administer PST on 3 consecutive days for 45 minutes to both groups, either in a verum or a placebo fashion. Additionally, all patients receive tubigrip, elbow crutches and analgesics. Assessments on 1, 2, 3, 8 and 15 days show no efficacy of PST over placebo on range of motion, walking velocity, step length, pain, swelling or other walking indices. Methodological points were lost by the lack of an intention-to-treat analysis and a 10% dropout rate.

McGill et al.[42] randomise patients with lateral ankle sprains over 2 groups of 18 patients each. They administer PST on three consecutive days for 15 minutes in a verum and a placebo fashion. All patients receive additional tubigrip, crutches and analgesics. Assessments at 1, 2, 3, 8 and 15 days show no effects on number of analgesics taken, pain, time to weight bearing on injured foot or swelling. Also here no intention-to-treat analysis was performed and on average 13.5% of patients were lost during follow-up. Outcome measures that were most prominently reported upon were pain, swelling and days of sick leave. Since the format of most outcome measures differed, we used an effect size calculation to make the outcomes comparable over studies. Most of the studies failed to provide adequate data to calculate effect sizes. For these studies results are expressed as positive (+), negative (-) or undecided (?).

In table 4 the outcomes on pain, swelling and sick leave during the first six weeks are reported for the five previously detected subgroups. From both the qualitative as well as the quantitative analysis brace therapy seems to provide some pain relief, although not statistically significant (quantitatively) nor consistent (qualitatively). The qualitative analysis regarding tape shows no pain relief. However, quantitative data supporting this claim are lacking. The other interventions do not show clear benefits. Placebo therapy (placebo laser-, short wave- or ultrasound therapy) shows no pain relief, but both quantitative as qualitative data are scarce.

Both quantitative as qualitative analyses show no clear evidence in reduction of

Table 2: Methodological scores of the reviewed trials.

Author	Year	A† 4	B 20	C 8	D 8	E 7	F 3	G 10	H 2	I 7	J 7	K 7	L 5	M 5	N 2	O 1	P 3	Q 3	Total 100	Outcome †
barker	85	4	20	2	6	2		10	2	7	7		4	5		1			70	negative
mcgill	88	4	20		8	6	1	10	2		3		3	5		1			63	negative
zeegers	95	4	10	4	4	5		6	2				5	5	2	1			48	negative
o'hara	92	4	10	6	4	7		6	2				4				3		46	positive
bie	88	4			6	7		10	2	3	3	3	2			1	3		44	positive
klein	91	2	20	2	4			6					1	5		1			41	negative
sloan	89	4	10	4	6			6	2			3	5						40	positive
pennington	93	4		2		5		10	2	3		3	5			1	3		38	positive
oostendorp	87	4			4	7	3	6				3	4	5		1			37	positive
wester	96	4	10	2	2			8					5	5		1			37	positive
pasila	78	2	10	6	6			8					4			1			37	positive
konradsen	91	2		2	8	7	2	6					2	5	2	1			37	positive
johannes	93	4	10	4	4	2		6	2				1		2	1			36	negative
williamson	86	2	10	4	4	2		6	2			3	1			1			35	negative
scotece	92	4		4	2	7	3	6	2				1	2		1	3		35	positive
nilsson	83	4		4	2	6	1	6					5	5		1			34	negative
axelsen	93	4	10		6	1		10					2			1			34	positive
bradnock	95	4				7		10	2	3			1	2		1	3		33	negative
cetti	84	4		4		7	3	6						5		1	3		33	positive
dettori	94	4			8	1	1	6				3	4	5		1			33	positive
sommer	93	4		2	4	7	1	6					2	5		1			32	positive
clark	91	2			4			10	2	3	3		2	5					31	negative
eiff	94	4		2			2	10	2				5	5		1			31	positive
hedges	80	4		4	8			6					3	5		1			31	negative
holmer	91	4		6	4	1		6	2				5	2		1			31	negative
cotee	88	2			6	7		10					1			1	3		30	positive
leanderson	95	4		2	6			6	2				3	5		1			29	positive
korkala	87	4		4	2			6					4	5	2	1			28	negative
moller-larson	88	4		4	4		1	6					2	5		1			27	positive
brakenbury	83	2		6	6			6	2				4			1			27	positive
rucinski	91	2			2	7		8	2				1			1	3		26	positive
michlovitz	88				2	7		10					3			1	3		26	negative
jongen	92	4		4	2	1	1	6					1	5		1			25	negative
gronmark	80			2		7	3	6					1	2		1	3		25	negative
muwanga	86	4		2	8		2	6	2				1						25	positive
airaksinen	90				6			8					4	5		1			24	positive
lelieveld	79				2	5		10	2				1						20	positive
zwipp	92			2	2			6	2				1	5		1			19	positive
allen	85	4			4			6					2			1			17	negative
wilkerson	93	4			2			6	2				1			1			16	negative
freeman	65							6					3	2					14	positive
caro	64			2				6	2				2			1			13	positive
makulowu	77			2				6					1						9	positive
brooks	81	2						6					1						9	positive

† See Table 1 for explanation of categories A to Q

Table 4: Effects of interventions in acute ankle sprains in the first six weeks.

Intervention	Qualitiative analysis outcome according to reviewer *by study*	*conclusion*	Quantitative analysis Outcome according to data *difference (95% CI)*	*conclusion*	*combined outcome*
Brace vs other					
pain	+[27], ?[29], ?[36], +[54], -[57], +[57]	pain relief	1.03 (-0.40/2.46)[27,29,54]	no pain relief	undecided
swelling (in ml)	+[27], ?[29], -[57], +[57]	reduction	no data		reduction of swelling
days sick leave	+[27], ?[57], +[36], ?[54], -[57]	undecided	0.55 (-0.64/1.74)[36,54]	no reduction	undecided
Tape vs other					
pain	-[16], -[19], -[48], -[57]	no pain relief	no data		no pain relief
swelling (in ml)	+[57]	reduction	no data		reduction of swelling
days sick leave	?[14], ?[48], +[52], +[57]	reduction	-2.97 (-4.31/1.63)[48,52]	reduction	reduction
Bandage vs other					
pain	-[15], +[27], -[33], +[34], -[46], +[47]	undecided	0.72 (-295/4.38)[3,27,33,34]	no pain relief	undecided
swelling (in ml)	+[27], +[46], -[57], -[51]	undecided	-14.8 (20/-9.58)[51]	no reduction	undecided
days sick leave	+[21], -[23], +[27], +[34], -[39], -[46], +[59]	undecided	-0.38 (-0.59/-0.17)[21,23,34,39]	reduction	reduction
Cast vs other					
pain	-[27], ?[29], +[33], ?[36], -[54]	undecided		no pain relief	undecided
swelling (in ml)	?[21], -[27], ?[29], +[33]	undecided	-0.29 (-1.64/1.06)[27,33,36,54]	no reduction	undecided
days sick leave	-[21], -[23], -[27], -[29], ?[37], -[52], -[54]	increase	0.17 (-0.06/0.27)[33] 4.52 (3.94/5.1)[21,23,54]	increase	increase
Placebo vs other					
pain	-[17], ?[18], -[19], ?[25], ?[40], ?[42], -[50], ?[53]	no pain relief		no pain relief	no pain relief?
swelling (in ml)		no reduction	-3.59 (-1.16/6.00)[17,42]	no reduction	no reduction
days sick leave	?[17], ?[18], ?[21], ?[40], -[42], -[49], -[50], -[17]	no reduction	-3.38ml(-3.679/3.081)[17,50] no data		undecided

+ reduction campared to other intervention(s); - no reduction; ? no difference

swelling by the studied interventions. Reduction of sick leave is best accomplished by taping. Both quantitative as qualitative analyses show this. Cast therapy seems to prolong sick leave significantly (quantitatively). For the other interventions the efficacy of the studied therapies remains undecided. Few studies reported side effects. Complications due to surgical therapy were reported as deep venous thrombosis in four patients[32, 38], loss of sensation in six cases[58] or infection (2 cases).[58] Deep venous thrombosis also occurred in two patients treated with plaster of Paris[38], while severe dermal lesions were found in three patients treated with tape and in one patient treated with a bandage.[35] Tape was also mentioned to be more painful in some cases.[45]

Discussion

The value of a literature review depends on the success in obtaining the results of all trials that have been conducted on the issue of interest. Despite the extensive search strategy it is possible that relevant studies reported in fora not accessible to us or in languages incomprehensible to us were missed in this review.

When one assesses the methodological quality, the primary goal is to achieve an estimate of both the effects and the chance of bias in the results of the performed studies. One method of assessing quality is by using a criteria list which tries to measure internal validity, precision of the study and relevance of the choices with respect to population, interventions and measures of effect.

Assigning weights to the criteria anticipates the argument that some errors in trial design are more crucial than others. Although controversial, weighting does give some insight into the quality of the performed studies and provides an overview of the credibility of the results. It also enables the reviewed studies to be ranked to some extent, according to their methodological quality. Another advantage of our scoring system is that it is already well known[5-8,60] and transparent. It permits the reviewer to allocate the distribution of methodological points elsewhere or to adjust scores if this is felt necessary.

The fact that systematic reviews use reported material to judge the quality of the trials under consideration is a consequence of this type of research, but is also cause for concern. It might very well be that the quality of the reported material does not reflect the quality of the primary research. Especially in trials where more reports of the same research question were generated, one tends to find differences, which in the end translates to differences in methodological scores.

We did not blind the reviewing procedure. From previous investigations[59] we found that blinding is not likely to bias the findings or results of the methodological score when articles are evaluated by skilled and trained people.

The 44 trials in this review can be considered the best available evidence when one studies the efficacy of conservative treatment regimens in ankle sprains. Because of the use of random allocation of the patients and the use of control treatments, their potential to supply valid answers is much larger than that of uncontrolled or non-randomised controlled studies. Nevertheless, the observed study quality ranged from very poor (9 points) to rather good (70 points).

Many (avoidable) errors in design and data-analysis were noted. In the studied trials, also the more recent ones, there is still a reluctance (or sloppiness) to assure proper blinding. Of course, some interventions are hard to blind due to the nature of the intervention. However, in the here studied trials many interventions that used apparatuses could have been blinded and in all trials at least attempts could have been made to blind the assessors and data-analysts.

A considerable number of randomised clinical trials studies the efficacy of conservative treatment regimens in ankle sprains. However, many of the reviewed studies showed serious methodological flaws, and much information was lacking in the publications. In fact only two studies score above 60 points.

There seems to be no relationship between methodological quality and year of publication, although trials reporting positive effects were of lower methodological quality. Apparently one continues to repeat methodological flaws over time, and low quality trials are still accepted by indexed (and peer reviewed) journals. Therefore, we suggest that in the future more attention should be paid to appropriate blinding procedures (where feasible), avoiding dropouts, to avoid or standardise co-interventions and to ensure better data representation.

We did try to pool the results of the trials on outcome measures as pain, swelling and days of sick leave. However the pooled results should be interpreted carefully. As can be seen from table 4, in many of the findings rather a small amount of trials contribute to the pooled measure of effect, since a large part of the studies did not present data that allowed pooling. Therefore, also a qualitative assessment of the data was performed, resulting in an overall assessment combining both quantitative and qualitative methods. As can be seen there is much uncertainty about the efficacy of the studied interventions. Pooling of the two best trials on pulsed shortwave therapy (PST) was impossible; qualitative assessment showed no results of PST. Larger and methodologically more adequate trials are called for in future to provide more definite and satisfying answers.

References

1. Kannus P, Renstr¨m P. Treatment for acute tears of the lateral ligaments of the ankle. *J Bone Joint Surg 1991;*73A:305-12.
2. Cote DJ, Prentice WE, Hooker DN, Shields EW. Comparison of three treatment procedures for minimizing ankle sprain swelling. *Phys Ther 1988;*68: 1072-6.
3. Wilkerson GB, Horn-Kingery HM. Treatment of the inversion ankle sprain: comparison of different modes of compression and cryotherapy. *JOSPT 1993;* 17(5):240-6.
4. Dickersin K, Berlin JA. Meta-analysis: State of the science. *Epid Rev 1992;*14:154 -76.
5. ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain. A criteria based meta-analysis. *J Clin Epid 1990;*43:1191-9.
6. Koes BW, Bouter LM. van der Heijden GJMG, Knipschild PG. Physiotherapy exercises and back pain. *Br Med J 1991;*302:1572-6.
7. de Vet HCW, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Knip schild PG. Systematic reviews on the basis of methodological criteria. *Physiotherapy 1997;*83:284-9.
8. Assendelft WJJ, Koes BW, Knipschild PG, Bouter LM. The relation between methodological quality and conclusions in reviews of spinal manipulation. *JAMA 1995;*274:1942-8.
9. Pocock SJ. Clinical trials; a practical approach. John Wiley & Sons, Chichester, 1991.
10. Meinert CL. Clinical trials; design, conduct and analysis. Oxford University Press, New York, 1986.
11. Cooper H, Hedges LV eds. The handbook of research synthesis. Russell Sgae Foundation, New York, 1994.
12. Rothman KJ. Modern Epidemiology. Little, Brown and Company, Boston/Toronto, 1984.
13. Wester U, Jespersen SM, Nielsen KD, Neumann L. Wobble board training after partial ankle sprains of lateral ligaments of the ankle: a prospective randomized study. *JOSPT 1996;*5:332-6.
14. Jongen SJM, Pot JH, Dunki-Jacobs PB. De behandeling van de verzwikte enkel. *Geneeskunde en Sport 1992;*25(3):98-101.
15. Airaksinen O, Kolari PJ, Miettinen H. Elastic bandages and intermittent pneumatic compression for treatment of acute ankle sprains. *Arch Phys Med Rehab 1990;*71(6):380-3.
16. Allen MJ, McShane M. Inversion injuries to the lateral ligament of the ankle joint. A pilot study of treatment. *Br J Clin Practice 1985;*July:282-6.
17. Axelsen SM, Bjerno T. Laserbehandling af fodledsdistorsion. (Laser treatment in ankle sprains) *Ugeskr Laeger 1993;*155(48):3908-11.
18. Barker AT, Barlow PS, Porter J, Smith ME, Clifton S, Andrews L, O'Dowd WJ. A double-blind clinical trial of low power pulsed shortwave therapy in the treatment of a soft tissue injury. *Physiotherapy 1985;*71(12): 500-4.
19. de Bie RA, Steenbruggen RA, Bouter LM. Effect of laser therapy on ankle sprains. *Ned T Fysiotherapie (English edition) 1989;*99:4-7.
20. Bradnock B, Law HT, Roscoe K. A quantitative comparative assessment of the immediate response to high frequency ultrasound and low frequency ultrasound ("longwave therapy") in the treatment of acute ankle sprains. *Physiotherapy 1995;*81(7):78-84.
21. Brakenbury PH, Kotowski J. A comparative study of the management of ankle sprains. *Br J Clin Practice 1983;*37(5):181-5.
22. Brooks SC, Potter BT, Rainey JB. Treatment for partial tears of the lateral ligament of the ankle: a prospective trial. *Br Med J 1981;*282:606-7.
23. Caro D, Craft IL, Howells JB, Shaw PC. Diagnosis and treatment of injury of lateral ligament of the ankle joint. *Lancet 1964;* 720-3.
24. Cetti R, Christensen SE, Corfitzen MT. Ruptured fibular ankle ligament: plaster or pliton brace? *Brit J Sports Med 1984;*18(2):104-9.
25. Clark A. A double blind trial of low level laser therapy for soft tissue injury. World Confed. Phys Therapy 11th Intern Congress Barbican Centre London 28 July-2 Aug 1991:807-809.
26. Cote DJ, Prentice WE, Hooker DN, Shields EW. Comparison of three treatment procedures for minimizing ankle sprain swelling. *Physical Ther 1988;*68(7):1072-6.
27. Dettori JR, Pearson BD, Basmania CJ, Lednar WM. Early ankle mobilization, part I: The immediate effect on acute, lateral ankle sprains. A randomized clinical trial. *Military Medicine 1994;*159:15-20.
28. Dettori JR, Basmania CJ. Early ankle mobilization, Part II: One-year follow-up of acute, lateral ankle sprains. A randomized clinical trial. *Military Medicine 1994;*159:20-4.
29. Eiff MP, Smith AT, Smith GE .Early mobilization versus immobilization in the treatment of lateral ankle sprains. *American Journal Sports Medicine 1994;*22(1):83-8.
30. Freeman MAR. Treatment of ruptures of the lateral ligament of the ankle. *J Bone Joint Surg 1965;*47B(4):661-8.
31. Freeman MAR. Instability of the foot after injuries to the lateral ligament of the ankle. *J Bone Joint Surgery 1965;*47B(4):669-677.
32. Gronmark T, Johnsen O, Kogstad O. Rupture of the lateral ligaments of the ankle: a controlled clinical trial. Injury: *British Journal of Accident Surgery 1980;*11:215-8.
33. Hedges JR, Anwar RAH. Management of ankle sprains. *Ann Emerg Med 1980;*9(6):296-302.
34. Holmer P, Carstensen ND, Merrild UB. Stottestromper kontra stottbind i behandlingen af akutte ankel-distorsioner. (Compression stocking versus bandage in acute ankle sprains) *Ugeskr Laeger 1991;*153(6):430-2.
35. Johannes EJ, Kaulesar Sukul DMKS, Spruit PJ, Putters JLM. Controlled trial of a semi-rigid bandage ("scotchwrap") in patients with ankle ligament lesions. *Curr Med Res Opin 1993;*13:154-62.
36. Klein J, Rixen D, Albring Th, Tiling Th. Funktionelle versus gipsbehandlung bei der frischen aussen-bandruptur des oberen sprunggelenks. (Functional treatment versus plaster of Paris treatment in acute ankle sprains) *Unfallchirurg 1991:*94:99-104

37. Konradsen L, Holmer P, Sondergaard L. Early mobilizing treatment for grade III ankle ligament injuries. *Foot Ankle 1991*;12(2):69-73.
38. Korkala O, Rusanen M, Jokipii P, Kytomaa J, Avikainen V. A prospective study of the treatment of severe tears of the lateral ligament of the ankle. *International Orthopaedics 1987*;11:13-7.
39. Leanderson J, Wredmark T. Treatment of acute ankle sprain. Comparison of a semi-rigid ankle brace and compression bandage in 73 patients. *Acta Orthop Scand 1995*;66(6):529-31.
40. Lelieveld van DW. Vaerdien af ultralyd og el-stimulation ved behandling af distorsioner. (Ultrasound or electrical stimulation in the treatment of ankle sprains) *Ugeskr Laeger 1979*;141:1077-80.
41. Makuloluwe RTB, Mouzas GL. Ultrasound in the treatment of sprained ankles. *Practitioner 1977*;218: 586-588.
42. McGill SN. The effects of pulsed shortwave therapy on lateral ligament sprain of the ankle. *NZ J Physiotherapy 1988*; December:21-4.
43. Michlovitz S, Smith W, Watkins M. Ice and high voltage pulsed stimulation in treatment of acute lateral ankle sprains. *J Orthop Sports Phys Ther 1988*; 9:301-4.
44. Muller-Larsen F, Whetelund JO, Jurik AG, Carvalho A, Lucht U. Comparison of three different treatments for ruptured lateral ankle ligaments. *Acta Orthop Scand 1988*;59:564-6.
45. Muwanga CL, Quinton DN, Sloan JP, Gillies P, Dove AF. A new treatment of stable lateral ligament injuries of the ankle joint. *Injury 1986*;17:380-2.
46. Nilsson S. Sprains of the lateral ankle ligaments. An epidemiological and clinical study with reference to different forms of conservative treatment. *J Oslo City Hosp 1983*;33:13-36.
47. O'Hara J, Valle-Jones JC, Walsh H, O'Hara H, Davey NB, Hopkin-Richards H, Butcher R. Controlled trial of an ankle support (Malleotrain) in acute ankle injuries. *Br J Sp Med 1992*;26:139-42.
48. Oostendorp RAB .Functionele instabiliteit na het inversietrauma van de enkel en voet: een effect onderzoek pleisterbandage versus pleisterbandage gecombineerd met fysiotherapie. (Functional instability after ankle sprains; a trial of taping versus taping and exercise) *Geneeskunde en Sport 1987*;20:45-55.
49. Pasila M, Visuri T, Sundholm A. Pulsating short-wave diathermy: value in treatment of recent ankle and foot sprains. *Arch Phys Med Rehab 1978*;59:383-6.
50. Pennington GM, Danley DL, Sumko MH, Bucknell A, Nelson JH. Pulsed, non-thermal, high-frequency electromagnetic energy (diapulse) in the treatment of grade I and grade II ankle sprains. *Military Medicine 1993*;2:101-4.
51. Rucinski TJ, Hooker DN, Prentice WE, Shields EW, Cote-Murray DJ. The effects of intermittent compression on oedema in post acute ankle sprains. *JOSPT 1991*;14:65-9.
52. Scotece GG, Guthrie MR. Comparison of three treatment approaches for grade I and II ankle sprains in active duty soldiers. *JOSPT 1992*;15:19-23.
53. Sloan JP, Hain R, Pownall R. Clinical benefits of early cold therapy in accident and emergency following ankle sprain. *Arch Emergency Med 1989*;6:1-6.
54. Sommer von HM, Schreiber H. Die fruh konservative Therapie der frischen fibularen Kapsel-band-ruptur aus sozial-"konomischer sicht. (Early functional conservative therapy of a fresh fibular rupture of the capsular ligament from a socioeconomical viewpoint). *Sportverl Sportschad 1993*;7:40-6.
55. Wester U, Jespersen SM, Nielsen KD, Neumann L. Wobble board training after partial ankle sprains of lateral ligaments of the ankle: a prospective randomized study. *JOSPT 1996*;5:332-6.
56. Williamson JB, George TK, Simpson DC, Hannah B, Bradbury E. Ultrasound in the treatent of ankle sprains. Injury 1986;17:176-8.
57. Zeegers AVCM. Het supinatieletsel van de enkel. Thesis University of Rotterdam, 1995.
58. Zwipp H, Schievink B. Primary orthotic treatment of ruptured ankle ligaments: a recommended procedure. Prosthetics and Orthotics International 1992;16:49-56.
59. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Knipschild PG. Balneotherapy and Quality Assessment. The interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. *J Clin Epidemiol 1998*;51:335-41.
60. van der Heijden GJMG, van der Windt DAWM, Kleijnen J, Koes BW, Bouter LM, Knipschild PG. The efficacy of corticosteroid injections for shoulder complaints: a blinded systematic review. *J Clin Epidemiol 1995*;48:691-704.

7 Impact of quality items on study outcome: treatments in acute lateral ankle sprains

AUTHORS:
Arianne P. Verhagen,
Robert A. de Bie,
Anton F. Lenssen,
Henrica C.W. de Vet,
Alphons G.H. Kessels,
Maarten Boers,
Piet A. van den Brandt.

ABSTRACT

Objective. This study investigates the influence of different aspects of methodological quality on the conclusions of a systematic review concerning treatments of acute lateral ankle sprains.

Method. A data set of 44 trials was used studying the efficacy of conservative interventions in patients with an acute lateral ankle sprain. Quality assessment of the individual studies was performed using the Delphi list. We calculated effect sizes of the main outcome measures in each study in order to evaluate the relationship between overall quality scores and outcome. Next, we set out to investigate the impact of design attributes on pooled effect sizes by subgroup analysis according to the design attributes.

Results. Quality scores vary from rather low to reasonably good. Only 23 of the 44 studies allowed calculation of effect sizes, for one or more outcome measures. We determined an estimate of an 'overall average of effects'. Studies with a proper randomisation procedure and those which report a form of blinding both produce a slightly higher estimate of this average effect size.

Conclusion. Previous research has suggested that methodologically poor designed studies tend to overestimate the effect. Our study does not confirm these conclusions. *(Submitted)*

Are the conclusions of a systematic review influenced by the methodological quality of the included studies? Some researchers have not found any difference in results between studies of good and poor quality.[1-4] Others have found that the methodologically sound studies showed positive treatment effects while the studies with a poor quality did not,[5,6] or vice versa.[7,8]

Some research on the relationship between design attributes, such as randomisation and blinding, has already been performed. Shapiro & Shapiro[9] studied, among others, treatment effects and 'features of the experimental design' in 143 trials, published over a five year period in the field of psychotherapy. Regarding the assignment of patients they divided the studies in four categories: non-random, random without matching and two different groups with matched randomisation. Furthermore, the studies were divided into three groups of 'blindness of the person obtaining the outcome data' (observer): 'single blind', 'knew group composition' and 'acted as therapist'. They found no relationship between patient assignment and effect sizes and a negative relation between blindness of the observer and effect sizes, meaning that when the observer was blinded, the effect sizes were smaller compared to studies with an unblinded observer.

Colditz, Miller and Mosteller found conflicting results in their studies.[10,11] They studied 113 reports published in 1980 in a sample of medical journals[10], and 221 reports published in six leading surgery journals in 1983.[11] In both studies no differences were found in effect sizes between randomised and non randomised parallel group comparisons. Colditz et al.[10] found a lower effect size in double blinded studies compared with not double blinded ones, while Miller et al.[11] found that 'double blind comparisons produced the largest average gains (effect sizes), significantly larger than the average for comparisons that involved no blinding'.

Ottenbacher[12] investigated the influence of random assignment on outcome in 30 randomised clinical trials (RCTs) and 30 controlled clinical trials (CCTs) published in the Journal of the American Medical Association (JAMA) and the New England Journal of Medicine (NEJM). He found no influence of random assignment on study outcome.

In contrast to previous studies, Schultz et al.[13,14] and Moher et al.[15] differentiated between 'randomised', 'randomised adequately' or a 'concealed randomisation' procedure. Schultz et al. performed their studies in the field of obstetrics and gynaecology, while Moher et al. randomly selected eleven meta-analysis of different interventions and diseases. Both showed in their studies that methodologically poor designed trials (focussing on randomisation and blinding) tend to exaggerate treatment effects. Kunz et al.[16], on the other hand, concluded in their methodological review that 'failure to use adequate concealed random allocation can distort the apparent effects of care in either direction'.

Most studies about design characteristics were done outside the context of a specific therapeutic research question.[13-15] We chose to place our research within a specific research question, i.e. the efficacy of conservative treatments in acute lateral ankle sprains. The purpose of this study is to evaluate whether overall trial quality and trial design attributes, such as the randomisation procedure and blinding, have an impact on outcomes of RCTs.

METHOD

Studies. We used a data set from a systematic review[17] of 44 randomised clinical trials (RCTs) on the efficacy of conservative interventions in the treatment of acute lateral ankle sprains. All studies were randomised and compared conservative treatment with either no treatment, a placebo or non surgical treatment. Studies were excluded from the analysis when they presented a withdrawal

rate > 50%, or when no effect sizes could be calculated.

Assessment of methodological quality. For the assessment of the methodological quality of individual studies we used the Delphi list.[18] This quality criteria list contains nine items and measures three dimensions of quality: internal validity, external validity and statistical considerations. The quality score consists of the number of items satisfied and ranges from 0 - 9. The assessment of the studies was performed independently by two of the authors (APV, AFL) followed by a consensus meeting. For the component analysis studies were divided into several categories according to items concerning the randomisation procedure, blinding and the analysis used. With an appropriate randomisation we mean that reports present information about a proper randomisation procedure instead of just using the word 'random'. With a concealed randomisation we mean that a random (unpredictable) assignment sequence is generated by an independent person not responsible for determining eligibility of the patients, and this sequence is concealed until allocation occurs.[13] For blinding we divided the studies into two main categories: blinding reported or not reported. When blinding is reported we note whether the observer was blinded or the term 'double blind' was used. For the statistical analysis items we divided the studies into two categories: performance of an intention-to-treat (ITT) analysis or not.

Statistical methods. For the primary outcome measures we calculated the effect sizes and their 95% confidence intervals (CI) according to the methods described in Cooper & Hedges.[19] These effect sizes transform the results of continuous data from any parallel group comparison into a standardized metric. Pooled effect sizes were calculated according to a random effects model,[19] using one effect zise (of the main outcome measure) out of each study. In a funnelplot we evaluated the possibility of publication bias in this review. If there is publication bias in a meta-analysis, the funnelplot will often be skewed and asymmetrical.[20,21]

Next we calculated overall quality scores (QS) for the individual studies. A cut-off point between 'high' and 'low' quality studies is set at 50% of the maximum achievable score of 9 points, meaning 'high' quality studies scored ≥ 5 points and 'low' quality studies ≤ 4 points. For the analysis of major components of quality, i.e. randomisation, blinding and an intention-to-treat analysis (ITT) we performed component analysis. We based our decision about the impact of these design characteristics on the overall pooled results, or the 'overall average of effects', instead of on the results of the methodological best studies.

RESULTS

Studies. In total eight studies compared an intervention such as 'short wave' or 'laser-therapy' with a placebo and 23 studies compared 'brace', 'tape' or 'bandage' with 'cast' or 'plaster'. In 25 studies pain was reported as the main outcome measure, and in 18 studies swelling was reported as an outcome measure. All studies included patients with acute ankle sprains (< 48 hours). Only 23 of the 44 studies allowed for calculation of effect sizes for one or more outcome measures and were included. One study[67] is excluded from the analysis because of a high withdrawal rate: over 60% loss to follow-up after three months and over 80% after one year. The sample of excluded studies was comparable with the included ones concerning patient characteristics, randomisation schedule, blinding, interventions and outcome measures.

Effect sizes. In the 22 included studies we were able to calculate in 27 effect sizes. Of these 27 effect sizes 8 (27.6%) were negative, suggesting an effect in favor of the control group. In Table 1 we present the characteristics of all studies and their main outcome measures, and the effect sizes for 22 studies.

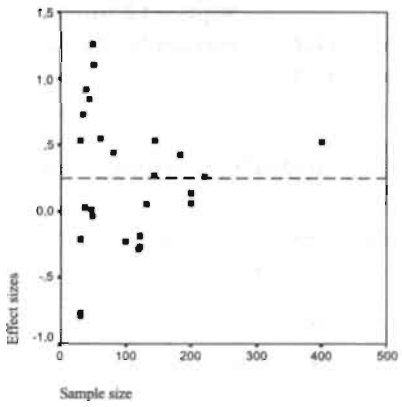
Table 1: Characteristics of the studies, ranked according to the Delphi quality score.

Study	Sample size	Randomisation	Intervention	Blinding	Outcome measures	Intention-to-treat	Effect size (95% CI)	Delphi QS
Barker '85	n = 82	Concealed	Pulsed short wave vs placebo	Care provider and patient	Pain Swelling	No	– –	7
McGill '88	n = 37	Concealed	Pulsed short wave vs placebo	Care provider	Pain Swelling	No	0.03 (-0.61, 0.67) –	6
Bie '88	n = 38	Unknown	Laser vs placebo (vs tape)	Care provider and patient	Pain	Yes	0.92 (0.05, 1.78)	6
Pennington '93	n = 50	Unknown	Diapulse vs placebo	Observer and patient	Swelling	Yes	1.1 (0.52, 1.68)	6
Bradnock '95	n = 47	Unknown	Low freq ultrasound vs placebo (vs long wave)	Patient	Gait	Yes	0.01 (-0.65, 0.67)	5
Hedges '80	n = 121	Unknown	Bandage vs plaster	Unknown	Pain Swelling	No	-0.19 (-0.59, 0.21) -0.27 (-0.67, 0.13)	4
Williamson '86	n = 154	Adequate	Physiother. + ultrasound vs physiother. alone	Observer	Clinical score	No		4
Oostendorp '87	n = 24	Unknown	Physiother. + bandage vs bandage alone	Observer	Pain	Yes	--	4
Coté '88	n = 30	Unknown	Cold vs heat (vs bath)	Unknown	Swelling	Yes	-0.78 (-1.68, 0.12)	4
Sloan '89	n = 143	Adequate	Cold vs stimulated ther.	Observer	Swelling	No	0.28 (-0.03, 0.59)	4
Rucinski '91	n = 30	Unknown	Wrap vs elevation (vs intermitt. compression)	Unknown	Swelling	Yes	-0.77 (-1.68, 0.14)	4
O'Hara '92	n = 220	Adequate	Malleotrain vs tubigrip	Unknown	Pain	Yes	0.26 (0.01, 0.51)	4
Axelsen '93	n = 48	Adequate	Laser vs placebo	Unknown	Pain Swelling	No	1.26 (-0.36, 2.15) -0.04 (-0.87, 0.79)	4
Sommer '93	n = 120	Unknown	Aircast vs plaster (vs cast)	Unknown	Loss of working hours	No	-0.29 (-0.72, 0.14)	4
Dettori '94	n = 64	Unknown	Wrap vs plaster (vs cast)	Observer	Pain	No	--	4
Pasila '78	n = 321	Adequate	Diapulse vs placebo (vs curapulse)	Unknown	Pain	No	--	3
Brakenbury '83	n = 400	Unknown	Bandage (+ chymoral or placebo) vs cast (+ chymoral or placebo)	Unknown	Range of motion	No	0.52 (0.25, 0.79)	3

Study	Sample size	Randomisation	Intervention	Blinding	Outcome measures	Intention-to-treat	Effect size (95% CI)	Delphi QS
Ceti '84	n = 130	Unknown	Brace vs plaster	Unknown	Pain/-swelling	Yes	--	3
Muwanga '86	n = 144	Unknown	Nothingham ankle support vs tubigrip (vs strapping)	Unknown	Range of motion	No	0.53 (0.12, 0.94)	3
Michlovitz '88	n = 30	Unknown	Ice vs ice + (placebo or real) pulsed stimulation	Unknown	Pain Swelling	Yes	0.53 (-0.35, 1.41) -0.21 (-1.07, 0.65)	3
Airaksinen '90	n = 44	Unknown	Bandage + compression vs bandage alone	Unknown	Pain Swelling	No	0.85 (0.23, 1.46) 0.85 (0.23, 1.46)	3
Klein '91	n = 60	Concealed	Bandage vs cast	Unknown	Health index	No	0.56 (0.02, 1.1)	3
Clark '91	n = 55	Unknown	(low or high) Laser vs placebo	Care provider and patient	Pain/-swelling	No	--	3
Hølmer '91	n = 200	Unknown	Wrap vs bandage	Unknown	Pain Swelling	No	0.13 (-0.14, 0.40) 0.06 (-0.21, 0.33)	3
Scotece '92	n = 184	Unknown	Gel cast vs *daily strapping* (vs tape)	Unknown	Return to *duty*	Yes	0.42 *(-0.07, 0.77)*	3
Jongen '92	n = 100	Unknown	Tape vs malleotrain	Unknown	Health index	No	-0.3 (-0.69, 0.09)	3
Zeegers '95	n = 264	Adequate	Coumans bandage vs aircast (vs shoe vs sock)	Unknown	Pain Swelling	No	2.22 (1.72, 2.71) 2.1 (1.6, 2.6)	3
Leanderson '95	n = 73	Unknown	Bandage vs brace	Unknown	Pain/-swelling	No	--	3
Grønmark '80	n = 95	Unknown	Tape vs cast (vs surgery)	Unknown	Symptoms	Yes	--	2
Brooks '81	n = 205	Unknown	Physiother. vs tubigrip (vs cast vs no treatment)	Unknown	Health index	No	--	2
Nilsson '83	n = 180	Unknown	Tape vs ice (with or without xylocain injection)	Unknown	Pain	No	--	2
Allen '85	n = 57	Unknown	Strapping (weave or stirrup) vs no treatment	Unknown	Pain	No	--	2
Korkola '87	n = 150	Unknown	Bandage vs cast (vs surgery)	Unknown	Pain	No	--	2
Möller-Larsen '88	n = 200	Unknown	Tape vs cast (vs surgery)	Unknown	Satisfaction	No	--	2
Konradsen '91	n = 80	Unknown	Brace vs aircast	Unknown	Pain	No	0.44 (0, 0.88)	2

Study	Sample size	Randomisation	Intervention	Blinding	Outcome measures	Intention-to-treat	Effect size (95% CI)	Delphi QS
Johannes '93	n = 136	Adequate	Tape vs semi rigid bandage	Unknown	Pain/-swelling	No	--	2
Wilkerson '93	n = 34	Unknown	Tape vs compression (with or without ice)	Unknown	Function	No	0.73 (0.09, 1.55)	2
Eiff '94	n = 82	Unknown	Mobilisation vs immob.	Unknown	Pain/weight-bearing	No	--	2
Wester '96	n = 61	Adequate	Training vs no-training	Unknown	Pain/-swelling	No	--	2
Caro '64	n = 132	Unknown	Tape vs cast (vs hydrocortisone injection)	Unknown	Time to cure	No	0.05 (-0.42, 0.52)	1
Freeman '65	n = 45	Unknown	Tape vs cast (vs surgery)	Unknown	Pain	No	--	1
Makulowuwe '77	n = 80	Unknown	Ultrasound vs plaster	Unknown	Recovery rate	No	--	1
Lelieveld '79	n = 60	Unknown	Ultrasound (real or placebo) vs electrotherapy	Unknown	Pain/-swelling	No	--	1
Zwipp '92	n = 100	Unknown	Brace vs cast (vs two kinds of surgery)	Unknown	Health index	No		1

Figure 1: Funnelplot of sample size against effect sizes of the individual studies.



In order to assess potential publication bias, Figure 1 presents the funnelplot of the effect sizes, as presented in Table 1, against the sample size. The sample size is presented horizontally and the effect sizes vertically. The funnelplot shows no asymmetry, therefore, we assume that our meta-analysis is probably not biased. The pooled estimate or the 'overall average of effects' is 0.25 (95%CI: 0.07-0.43).

Quality scores. Table 1 presents the characteristics of the studies included in this research. The Delphi quality scores vary from 1 to 7 points. The mean QS is 3.1, which is low compared to the maximum achievable score of 9 points. 75% of the studies reported no information about the method of randomisation and 77% of the reports presented no information about blinding procedures.

The two reviewers, who assessed the articles independently, had an initial agreement on the Delphi criteria list of approximately 95%. The 5% disagreement occurred mostly because one reviewer had missed some information (4%) but rarely because of a difference in interpretation of the information (1%).

Relationship between overall quality and outcome. We made a scatter plot between the overall quality scores (QS) and effect sizes (Figure 2). The scatterplot shows no relation between the QSs and the effect sizes (intercept = 0.217; slope = 0.045). The pooled effect size of 'high' quality studies (n = 4) is 0.53 (95% CI: -0.21-1.27) and of 'low' quality studies (n = 18) is 0.19 (95% CI: 0.005-0.38). This difference is not statistically significant.

Component analysis. The effect sizes, pooled for subgroups according to the various design attributes, are presented in Table 2.

Randomisation. Of all 44 studies three reported a concealed randomisation procedure,[28,46,51] and seven an adequate method of randomisation.[27,41,44,57,61,64,66] Because of the small numbers of studies in both categories we combined them in the component analysis (n=5). When the randomisation method is unknown the pooled effect size is lower than when the method is appropriate or concealed (Table 2).

Blinding. When 'double blinding' is mentioned[28,29,35,51,58,66] all studies described at least one level of blinding, and five described the method of blinding. Four studies[37,56,61,66] described blinding of the outcome measurement (observer), and four studies[28,29,35,57] described blinding of two different levels (three times patient and therapist, once patient and observer). One study[29] evaluated whether the blinding procedure was succesfull. The pooled effect size for the category 'blinding not reported' is lower than when blinding is reported and is comparably low with the pooled effect size in the category randomisation procedure unknown (Table 2).

Analysis. 37 of the 44 reports presented frequencies or point estimates of the main outcome measures, and seven reports supplied hardly any information about the main outcome measures.

Figure 2: Plot of the Delphi QSs of the individual studies against effect size.

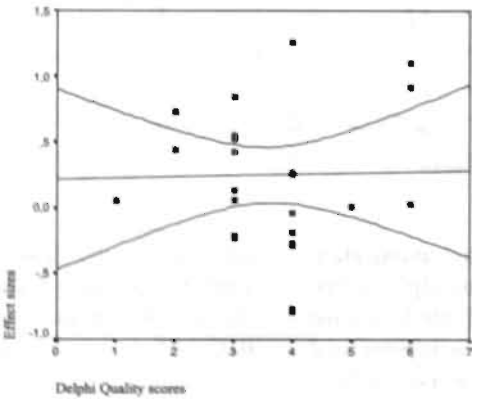


Table 2: Pooled effect size of all studies according to various design attributes.

Design attributes/Components 44 studies	*Pooled effect size 22 studies*	*95% CI*
Randomisation		
method unknown (n=33)	0.21 (n=17)	0.001-0.44
method appropriate or concealed (n=11)	0.34 (n=5)	0.017-0.68
Blinding		
blinding not reported or unknown (n=34)	0.19 (n=17)	0.008-0.40
blinding reported (n=10)	0.46 (n=5)	-0.06-0.99
blinding observer (n=5)	0.70 (n=2)	-0.31-1.72
'double blind' (n=6)	0.70 (n=3)	-0.19-1.60
Intention-to-treat analysis		
not performed (n=33)	0.24 (n=13)	0.04-0.45
performed (n=11)	0.23 (n=9)	-0.17-0.64

In another 14 studies we were unable to calculate effect sizes from the presented information. In 11 studies the performance of the analysis was carried out according to the intention-to-treat (ITT) principle. The difference in pooled effect size is small in studies with an ITT and no ITT analysis (Table 2).

Discussion

In general, design factors, such as proper randomisation and blinding procedures, do influence the interpretation of the results of individual clinical trials. The use of design factors in the interpretation of aggregated research in systematic reviews or meta-analysis is more difficult. In our study the overall methodological quality scores varied between poor and reasonably good, but most studies scored less than half of the maximum available score. Our findings support the conclusion of other researchers that just a few clinical trials meet the minimum standards of methodological rigor to be validly interpretable from a scientific point of view.[22-24]

A leading paradigm in empirical research is that clinical trials which do not meet some design criteria, such as concealed randomisation or 'double blinding', will be biased in favor of the intervention, and therefore produce more likely positive treatment effects. Concerning design factors, we found a lower estimate of effect in trials with an unknown randomisation procedure, or where blinding was not reported, compared to the ones using a proper randomisation and blinding schedule. There are several possible explanations why our findings do not confirm this paradigm.

The validity of our investigation is limited by the small number of trials, the small number of patients involved and the quality of the data presented. Our results could also be affected by the fact that we had to exclude almost half of our studies, because data to be able to calculate effect sizes was not presented. However, the included and excluded studies were similar with regard to the most important design characteristics. Contrary to other studies about design characteristics,[13-15] we chose to place our research within a specific research question. According to Kunz et al.[16] evidence about the influence of randomisation is less clear in comparisons across interventions, compared to empirical studies using studies with more or less the same intervention. Combining trials concerning varying interventions in varying diseases or disorders there is such a large heterogeneity, an estimation of an 'overall average of true effects' cannot be given. In that case the assumption that non-concealed randomised trials, or not blinded trials, provide an overestimation of the treatment effect cannot be tested.

The problem still is that we do not know what the 'true' treatment effect is, we can only estimate it. We based our estimate of the 'overall average of true effects' on the pooled results of 0.25 (95%CI: 0.07-0.43). Stating this, the pooled estimates of the studies with a concealed or appropriate randomisation or blinding reported, provide a slightly higher estimate, and the pooled effect estimate of the high quality studies (0.53) provides a much higher estimate compared to the 'overall average of effects', although not statistically significant.

In conclusion, this study confirms that trial design attributes can modify outcome, but the direction and magnitude of this effect is unpredictable, and may depend on the research question. Quality assessment is seen as an important part of a meta-analysis, but the influence of quality on outcome is yet unclear and needs further research.

References

1. Emerson JD, Burdick E, Hoaglin DC, Mosteller R, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trials 1990*;11:339-52.
2. Kleijnen J, ter Riet G, Knipschild PG. Vitamin B in the treatment of the premenstrual syndrome; a review. *Britt J of Obstetrics and Gynaecology 1990*;97:847-52.
3. Kleijnen J, Knipschild PG, ter Riet G. Clinical trials of homeopathy. *Br Med J 1991*;302:316-23.
4. Aker PD, Gross AR, Goldsmith CH, Peloso P. Conservative management of mechanical neck pain: systematic overview and meta-analysis. *Br Med J 1996*;313: 1291-6.
5. Assendelft WJJ, Koes BW, van der Heijden GJMG, Bouter LM. Efficacy of chiropractic manipulation for back pain; blinded review of relevant randomized clinical trials. *JMPT 1992*;15:487-94.
6. Van der Heijden GJMG, van der Windt DAWM, Kleijnen J, Koes BW, Bouter LM, Knipschild PG. The efficacy of corticosteroid injections for shoulder complaints: a blinded systematic review. *J Clin Epid 1995*;48:691-704.
7. Kleijnen J, ter Riet G, Knipschild PG. Acupuncture and asthma: a review of controlled trials. *Thorax 1991*; 46:799-802.
8. Beckerman H, de Bie RA, Bouter LM, Cuyper HJ, Oostendorp RAB. The efficacy of laser therapy for musculoskeletal and skin disorders; a criteria based meta-analysis of randomized clinical trials. *Phys Ther 1992*;72:483-91.
9. Shapiro DA, Shapiro D. Meta-Analysis of Comparative Therapy Outcome Studies: A Replication and Refinement. *Psychological Bulletin 1982*;92(3):581-604.
10. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med 1989*;8:441-54.
11. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: Surgical. *Stat Med 1989*;8:455-66.
12. Ottenbacher K. Impact of random assignment on study outcome: an empirical examination. *Control Clin Trials 1992*;13:50-61.
13. Schultz KF, Chalmers I, Hayes R, Altman DG. Empirical evidence of bias; Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA 1995*;273:408-12.
14. Schultz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaegology. *Br Med J 1996*;312:742-4.
15. Moher D, Pham Ba', Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analysis? *Lancet 1998*;352:609-13.
16. Kunz R, Oxman AD. The unpredictability paradox: re-view of emppirical comparisons of randomised and non-randomised clinical trials. *Br Med J1998*;317:1185-90.
17. de Bie RA, Verhagen AP, Lenssen TF, de Vet HCW, van den Wildenberg FAJM, Kootstra G, Knipschild PG. Efficacy of conservative interventions in the treatment of acute lateral ankle sprains; a systematic review. *Submitted*
18. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG. The Delphi list: a criteria list for quality assessment of Randomised Clinical Trials for conducting systematic reviews developed by Delphi consensus. J Clin Epid 1998;51:-1235-41.
19. Cooper H, Hedges LV. The handbook of research synthesis. New York. *Russell Sage Foundation. 1994.*
20. Egger M, Smith GD, Sneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *Br Med J 1997*;315:629-34.
21. Egger M, Smith GD. Misleading meta-analysis: lessons from 'an effective, safe, simple' intervention that wasn't. *Br Med J 1995*;310:752-4.
22. Tyson JE, Furzan JA, Reisch JS, Mize SG. An evaluation of the quality of therapeutic studies in perinatal medicine. *J Pediatr 1983*;102:10-3.
23. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM. Spinal manipulation for low back pain; an updated systematic review of randomized clinical trials. *Spine 1996*;21:2860-73.
24. Assendelft WJJ, Koes BW, van der Heijden GJMG, Bouter LM. The effectiveness of chiropractic for

treatment of low back pain; an update and attempt at statistical pooling. *JMPT 1996*;19:499-507.

References of the systematic review

25. Airaksinen O, Kolari PJ, Miettinen H. Elastic bandages and intermittent pneumatic compression for treatment of acute ankle sprains. *Arch Phys Med Rehabil 1990*;71:380-3.
26. Allen MJ, McShane M. Inversion injuries to the lateral ligament of the ankle joint. A pilot study of treatment. *Br J Clin Pract 1985* jul 39(7):282-6.
27. Axelsen SM, Bjerno T. Laserbehandling af fodledsdistorsion. *Ugesk Laeger 1993*;155(48):3908-11. (scand)
28. Barker AT, Barlow PS, Porter J, Smith ME, Clifton S, Andrews L, ODowd WJ. A double blind clinical trial of low power pulsed shortwave therapy in the treatment of a soft tissue injury. *Physiotherapy 1985*;71:500-4.
29. De Bie RA, Steenbrugge RA, Bouter LM. Effect van lasertherapie op inversietraumata van de enkel. *Nederlands Tijdschrift voor Fysiotherapie 1988*;98:108-12. (dutch)
30. Bradnock B, Law HT, Roscoe K. A quantitative comparative assessment of the immediate respons to high frequency ultrasound and low frequency ultrasound in the treatment od acute ankle sprains. *Physiotherapy 1995*;81:78-84.
31. Brakenbury PH, Kotowski J. A comparative study of the management of ankle sprains. *Br J Clin Pract 1983*; 37(5):181-5.
32. Brooks SC, Potter BT, Rainey JB. Treatment for partial tears of the lateral ligament of the ankle: a prospective trial. *Br Med J 1981*;282:606-7.
33. Caro D, Craft IL, Howells JB, Shaw PC. Diagnosis and treatment of injury of lateral ligament of the ankle joint. *Lancet 1964*;i:720-3.
34. Cetti R, Christensen SE, Corfitzen MT. Ruptured fibular ankle ligament: plaster or pliton brace? *Brit J Sports Med 1984*;18:104-9.
35. Clark A. A double blind trial of low level laser therapy for soft tissue injury. World Confed. Phys Therapy 11th Intern Congress Barbican Centre London. July 28th - Aug 2nd 1991:807-9.
36. Cote DJ, Prentice WE, Hooker DN, Shields EW. Comparison of three treatment procedures for minimizing ankle sprain swelling. *Phys Ther 1988*;68:1072-6.
37. Dettori JR, Pearson B, Basmania CJ, Lednar WM. Early ankle mobilisation, Part I: the immediate effect on acute, lateral ankle sprains (a randomized clinical trial). *Military Medicine 1994*;159:15-9.
38. Eiff MP, Smith AT, Smith GE .Early mobilization versus immobilization in the treatment of lateral ankle sprains. *American Journal Sports Medicine 1994*;22(1): 83-8.
39. Freeman MAR. Treatment of ruptures of the lateral ligament of the ankle. *Journal of Bone and Joint Surgery 1965*;47b:661-8.
40. Gronmark T, Johnsen O, Kogstad O. Rupture of the lateral ligaments of the ankle: a controlled clinical trial. *Injury 1980*;11:215-8.
41. O'Hara J, Valle-Jones JC, Walsch H et. al. Controlled trial of an ankle support (Malleotrain) in acute ankle injuries. *Br J Sp Med 1992*;26:139-42.
42. Hedges JR. Management of ankle sprains. *Ann Emerg Med 1980*;9:296-302.
43. Holmer P, carstensen NC, Merrild UB. Stottestromper kontra stottebind i behandlingen af akutte ankeldistorsioner. En prospektiv randomiseret undersogelse. *Ugesk Laeger 1991*;153(6):430-2. (scand)
44. Johannes EJ, Kaulesar Sukul DMKS, Spruit PJ, Putters JLM. Controoled trial of a semi-rigid bandage (Scotchrap) in patients with ankle ligament lesions. *Current Medical Research and Opinions 1993*;13:154-62.
45. Jongen SJM, Pot JH, Dunki Jacobs PB. De behandeling van de verzwikte enkel: tape versus malleotrain. *Geneeskunde en Sport 1992*;25:98-101. (dutch)
46. Klein J, Rixen D, Albring Th, Tiling Th. Funktionelle versus Gipsbehandlung bei der frischen Aussenbandruptur des oberen Sprunggelenks; eine randomisierte klinische Studie. *Unfallchirurg 1991*;94:99-104. (german)
47. Konradsen L, Holmer P, Sondergaard L. Early mobilizing treatment for grade III ankle ligament injuries. *Foot & Ankle 1991*;12:69-73.
48. Korkola O, Rusanen M, Jokipii P, Kytomaa J, Avikainen V. A prospective study of the treatment of severe tears of the lateral ligament of the ankle. *International Orthopaedics 1987*;11:13-17.
49. Leanderson J, Wredmark T. Treatment of acute ankle sprain; comparison of a semi-rigid ankle brace and compression bandage in73 patients. *Acta Orthop Scand 1995*;66:529-31.
50. Makulowuwe RTB, Mouzas GL. Ultrasound in the treatment of sprained ankles. *Practitioner 1977*;218: 586-8.
51. McGill SN. The effects of pulsed shortwave therapy on lateral ligament sprain of the ankle. *NZ Journal of Physiother 1988*: 21-4.
52. Michlovitz S, Smith W, Watkins M. Ice and high voltage pulsed stimulation in treatment of acute lateral ankle sprains. *Journal of Orthopaedic and Sports Physical Therapy 1988*;9:301-4.
53. Moller-Larsen F, Wethelund JO, Jurik AG, Carvalho A de, Lucht U. Comparison of three different treatments for ruptured lateral ankle ligaments. *Acta Orthop Scand 1988*;59:564-6.
54. Muwanga CL, Quinton DN, Sloan JP, Gillies P, Dove AF. A new treatment of stable lateral ligament injuries of the ankle joint. *Injury 1986*;17:380-2.
55. Nilsson S. Sprains of the lateral ankle ligaments. *J Oslo City Hosp 1983*;3:16-36.
56. Oostendorp RAB. Functionale instabiliteit na het inversietrauma van enkel en voet: een effectonderzoek pleisterbandage versus pleisterbandage gecombineerd met fysiotherapie. *Geneeskunde en Sport 1987*; 20:45-55. (dutch)
57. Pasila M, Visuri T, Sundholm A. Pulsating shortwave diathermy: value in treatment of recent ankle and foot sprains. *Arch Phys Med Rehabil*

1978;59:383-6.
58. Pennington GM, Danley DL, Sumko MH, Bucknell A, Nelson JH. Pulsed, non-thermal, high-frequency electromagnetic energy (diapulse) in the treatment of grade I and grade II ankle sprains. *Military Medicine 1993;*158:101-4.
59. Rucinski TJ, Hooker DN, Prentice WE, Shields EW, Cote-Murray D. The effects of intermittent compression on edema in postacute ankle sprains. *JOSPT 1991;* 14:65-9.
60. Scotece GG, Guthrie MR. Comparison of three treatment approaches for grade I and II ankle sprains in active duty soldiers. *JOSPT 1992;*15:19-23.
61. Sloan JP, Hain R, Pownall R. Clinical benefits of early cold therapy in accident and emergency following ankle sprain. *Arch Emerg Med 1989;*6:1-6.
62. Sommer HM, Schreiber H. Die fruh funktionelle konservative Therapie der frischen fibularen Kapsel-Band Ruptur aus Sozial ekonomischer Sicht. *Sportverl Sportschad 1993;7*:40-6. (german)
63. Van Lelieveld DW. Vaerdien af ultralyd og el-stimulation ved behandling af distorsioner. En kontrolleret undersogelse. *Ugesk Laeger 1979* Apr 16: 41(6): 1077-80. (scand)
64. Wester JU, Jespersen SM, Nielsen KD, Neumann L. Wobble board training after partial sprains of the lateral ligaments of the ankle: a prospective randomized study. *JOSPT 1996;*23:332-6
65. Wilkerson GB, Horn-Kingerly HM. Treatment of the inversion ankle sprain: comparison of different modes of compression and Cryotherapy. *JOSPT 1993;*17:240-6.
66. Williamson JB, George DC, Simpson DC, Hannah B, Bradbury E. Ultrasound in the treatment of ankle sprains. *Injury 1986;*17:176-8.
67. Zeegers AVCM. Supinatie letsel van de enkel. Academisch proefschrift, Thesis 1995. Utrecht. (dutch)
67. Zwipp H, Schievink B. Primary orthotic treatment of ruptured ankle ligaments: a recommended procedure. *Orthitics International 1992;*16:49-56.

8 The influence of methodological quality on the conclusion of a landmark meta-analysis on thrombolytic therapy

AUTHORS:
Arianne P. Verhagen,
Henrica C.W. de Vet,
Frank Vermeer,
Jos WMG. Widdershoven,
Robert A. de Bie,
Alphons G.H. Kessels,
Maarten Boers,
Piet A van den Brandt.

ABSTRACT

Objective. To study the influence of the methodological quality of individual trials on the outcome of a landmark meta-analysis on thrombolytic therapy in acute myocardial infarction.

Data source. Studies included in a meta-analysis of Yusuf et al. (Eur Heart J. 1985;6:556-85).

Data extraction. From each study we extracted the number of patients in the treated and control groups who died in hospital or during follow-up. Methodological quality was assessed using the Delphi list.

Data synthesis. We first recalculated pooled Odds Ratios (ORs), and their 95% confidence intervals (CIs), on the studies found and compared them with the original results of Yusuf et al. We incorporated the results of the quality assessment in several ways in the calculation of the pooled ORs: a) component analysis of the methodological items or components: randomization, blinding, withdrawals and analysis; b) visual plot of the individual ORs against the quality score; c) the quality score used as a 'threshold score' for inclusion of the article in the pooling; d) quality score used as a 'weighting factor'; e) cumulative pooling using quality scores as the input sequence.

Results & Conclusion. No correlation between overall quality scores and ORs was found. Studies with a proper description of the different quality components provided a good estimate of the true treatment effect. No major differences were found between the results of the five different methods of incorporating the quality scores into the final outcome or conclusions. (*Submitted*)

Scientific guidelines for reviewing the literature often include assessment of *the methodological quality of trials.*[1,2] The value of the conclusion of a meta-analysis not only depends on the quality of the review process itself, but also on the methodological quality of the randomized clinical trials (RCTs) included.[3] A leading paradigm in empirical research is that clinical trials which do not meet certain design criteria, such as concealed randomization and 'double blinding', will be usually biased in favor of the intervention, and are therefore more likely to produce positive treatment effects.[4-7]

Assessment of the quality of clinical trials by criteria lists, provides an estimation of the possibility of biased results of a trial. One approach in assessing quality is to focus on components such as randomization, blinding etc. in trial reports.[5,8] Furthermore, a criteria list can provide a quality score as an estimation of the overall methodological quality of the design and conduct of the trial.[9] These quality scores result in a hierarchical list in which higher scores indicate studies with a better methodological quality.[10] Quality scores can be used as a 'threshold score' for inclusion of the article in a review, as a 'weighting factor in the statistical analysis,[6,11,12] or as the input sequence in a cumulative meta-analysis.[6,13,14] Finally, a visual plot of the effect size against a quality score can be presented.[6,13,14]

Historically, the effectiveness of thrombolytic therapy for acute myocardial infarction (MI) was long disputed. In the years before 1980 intravenous streptokinase (SK) was tested in RCTs, but the results were not unequivocally in favor of this therapy. Later, intracoronary application of SK became in use because of the angiographically documented recanalisation of the occluded coronary artery. Streptokinase became licenced for use in MI after positive results of a meta-analysis in 1985 of Yusuf et al.[15] and two very large trials[16,17] in which the benefit of intravenous thrombolysis in acute MI was confirmed. However, cumulative meta-analysis showed (in retrospect) that there already was a clear evidence of the benefit in 1973.[18,19] Yusuf et al.[15] did not perform quality assessment in their meta-analysis. They closed their discussion with a comment on the general validity of their overview because some trials were "undoubtly less well executed than others".

We consider the meta-analysis of Yusuf et al.[15] as a landmark because the results had great impact on health care, and large trials studying the same intervention[16,17] confirmed their conclusions. In a study of Egger et al.[20] the meta-analysis of Yusuf et al.[15] was regarded as valid, because the funnelplot derived from it was not skewed.

In this research quality will be measured by the Delphi list.[21] This list has recently been developed using scale developing techniques. A pool of items is constructed from existing criteria lists and narrowed down by means of the Delphi Consensus Technique, using the cooperation of an international panel of more than 25 experts in the field of quality assessment in RCTs (statisticians and epidemiologists). We set out to investigate, whether and in which way quality can affect the overall conclusions of Yusuf et al.s'[15] meta-analysis. In this study the conclusions of the authors are regarded as the 'gold standard', namely the overall average of true effects.

METHOD

Selection of studies. All full reports presenting mortality data included in the meta-analysis of Yusuf et al.[15] are included in this study. Where they used an abstract or personal communication, we required full reports. Therefore we searched in MEDLINE and EMBASE and consulted leading cardiologists in the trombolytic field. Studies only available as abstracts or personal communications were excluded, because quality assessment could not be performed.

Table 1: The Delphi list for quality assessment

Items	Answer option
1. Treatment allocation	
a) Was a method of randomisation performed?	Yes / No / Don't know
b) Was the treatment allocation concealed?	Yes / No / Don't know
2. Were the groups similar at baseline regarding the most important prognostic indicators?	Yes / No / Don't know
3. Were the eligibility criteria specified?	Yes / No / Don't know
4. Was the outcome assessor blinded?	Yes / No / Don't know
5. Was the care provider blinded?	Yes / No / Don't know
6. Was the patient blinded?	Yes / No / Don't know
7. Were point estimates and measures of variability presented for the primary outcomemeasures?	Yes / No / Don't know
8. Did the analysis include an intention-to-treat analysis?	Yes / No / Don't know

Quality assessment. For the quality assessment of individual studies we applied the Delphi list.[21] The Delphi list contains 9 items, measuring three dimensions of quality: internal validity, external validity and statistical considerations (Table 1). The assessment of the studies was performed independently by two epidemiologists and two cardiologists. The assessors reached a final score during a consensus meeting resulting in an overall quality score (QS). We studied the relationship between QS and pooled ORs first by analysing the effects of the main components of the QS (component analysis). We then studied the relationship between the overall QS and ORs in several ways.

For the component analysis we divided the studies into three categories of *randomization*: method concealed, method appropiate but not concealed, and method unknown. Concealed randomization implies that Aa random (unpredictable) assignment sequence is generated by an independent person not responsible for determining eligibility of the patients, and this sequence is concealed until allocation occurs". With appropriate we mean that reports present additional information about a proper randomization procedure.
For *blinding* we divided the studies into two categories: blinding reported or not reported. Concerning *withdrawals* we divided the studies into three categories: a) no withdrawals or a withdrawal rate not leading to bias, b) withdrawal rate unknown, and c) a withdrawal rate possibly leading to bias (meaning >5% dropouts, >20% loss to follow-up). For the *statistical analysis items* we divided the studies into two categories: performance of an intention-to-treat (ITT) analysis or not.

For the overall analysis we first construct a scatterplot of the QS against the individual ORs. In order to evaluate possible selection bias we also made a scatterplot including the studies we were unable to find or did not include. These studies all recieved a QS of 2.

To use the QS as a 'threshold score' we followed Chalmers et al.'[22] suggestion and did a restricted analysis on studies with a mean QS and above mean QS. Further we also used the QS as a 'weight': we weighted each individual study estmate by their achieved Delphi QS, thereby deriving more impact from higher quality studies on the overall pooled results.[12] Finally, to achieve cumulative pooling we started the pooling with the study with the highest QS and subsequently added the others, rankordered by decreasing QS.

Data extraction. For the main outcome measure we extracted from each report the number of patients in the treated and control groups who died in hospital or during follow-up as mentioned in the report.

Analysis. We first recalculated pooled ORs, and their 95% confidence intervals (CIs), using a Peto fixed effects model, and compared it with the results of Yusuf et al.[15] Also a funnelplot was made according to Egger et al.[20] to evaluate possible publication/selection bias of the studies found. We calculated Spearman rank correlation coefficients between the QSs of the epidemiologists and the cardiologists separately, and to evaluate the relationship between quality and effect estimate. Next the quality scores (QS) were incorporated in the pooling in the 5 different ways mentioned above.

Table 2: Characteristics of the studies.

Study	Participants	Method	Intervention	Odds Ratio (95% CI)	Delphi QS
Gormson '73	acute MI < 24 hours; age < 80 , n = 28	concealed R; blinding of patient, therapist and observer	UK vs placebo	0.61 (0.09-4.37)	8
Schreiber '86	acute MI < 6 hours; age < 76 , n = 38	unknown R; blinding of patient and therapist	SK + heparin vs placebo + heparin	0.20 (0.02-2.07)	7
Italian (Dioguardi) '71 [30]	acute MI < 12 hours; no age limits , n = 321	concealed R; no blinding; multicenter	SK + anticoag. vs glucose + anticoag	1.01 (0.51-2.01)	6
2nd Frankfurt (Breddin) '73	acute MI < 12 hours; age < 70 , n = 206	unknown R; blinding of patient and therapist; multicenter	SK + heparin vs placebo + heparin	0.38 (0.18-0.77)	6
Lippschutz '65 [45]	acute MI < 48 hours; no age limits , n = 84	concealed R; blinding of patient and observer	UK + heparin vs placebo + heparin	0.79 (0.24-2.58)	6
Rentrop '84 [51]	acute MI < 12 hours; age < 72 , n = 124	unknown R; blinding of observer	IC-SK or IC-SK + NTG vs NTG or control	2.38 (0.84-6.75)	6
Kennedy '85	acute MI < 12 hours; age < 75 , n = 250	concealed R; no blinding; multicenter	IC-SK + heparin vs heparin	0.52 (0.23-1.16)	6
2nd European '71 [34]	acute MI < 24 hours; no age limits , n = 730	concealed R; no blinding; multicenter	SK vs heparin	0.64 (0.45-0.9)	5
Australian (Bett) '73 [28]	acute MI < 24 hours; age < 65 , n = 517	adequate R; no blinding; multicenter	SK + heparin + anticoag. vs heparin + anticoag	0.75 (0.43-1.31)	5
UK Collab. (Aber) '76 [23]	acute MI < 24 hours; no age limits , n = 595	concealed R; no blinding; multicenter	SK vs control	0.88 (0.57-1.35)	5

Study	Participants	Method	Intervention	Odds Ratio (95% CI)	Delphi QS
3rd European '81 [32]	acute MI < 12 hours; age < 80 , n = 315	concealed R; no blinding; multicenter	SK + coumarin vs glucose + coumarin	0.41 (0.24-0.72	5
Olson '86 [48]	acute MI < 12 hours; no age limits; males only, n=52	unknown R; no blinding	SK + heparin vs saline + heparin	0.83 (0.20-3.29)	5
European Cooperative '75 [31]	acute MI < 12 hours; age < 80 , n = 341	concealed R; no blinding; multicenter	UK + heparin + anticoag. vs glucose + heparin + anticoag.	1.09 (0.64-1.87)	5
Khaja '83 [41]	acute MI < 6 hours; no age limits , n = 40	adequate R; no blinding	IC-SK vs dextrose placebo	0.21 (0.02-2.08)	5
Anderson '84 [28]	acute MI < 4 hours; age < 80 , n = 50	unknown R; probably blinding of observer	IC-SK + heparin vs heparin	0.38 (0.06-2.19)	5
Simoons '85 [54]	acute MI < 4 hours; age < 70 , n = 533	adequate R; no blinding; Zeelen design	IC-SK + conventional vs conventional treatment	0.51 (0.30-0.88)	5
2nd German (Poliwoda)'77 [49]	acute MI < 12 hours; no age limits , n = 492	concealed R; no blinding; multicenter	SK + heparin + marcoumar vs heparin + standard	1.28 (0.84-1.94)	5
1st European (Amery) '69 [34]	acute MI < 72 hours; no age limits, n = 167	adequate R; no blinding	SK + coumarin vs heparin + coumarin	1.46 (0.69-3.1)	4
Heikinheimo '71 [37]	acute MI < 72 hours; no age limits, n = 426	adequate R; no blinding; multicenter	SK + anticoag. vs glucose + antocoag.	1.25 (0.64-2.42)	4
Witchitz '77 [58]	acute MI < 24 hours; age < 75, n = 58	adequate R; no blinding	SK vs heparin	0.77 (0.20-3.04)	4
Leiboff '84 [44]	acute MI < 4 hours; age < 75, n = 40	unknown R; no blinding	SK + heparin vs NTG + heparin	1.78 (0.29-11.04)	4
Raizner '85 [50]	acute MI < 6 hours; age < 70, n = 64	concealed R; no blinding;	IC-SK + NTG vs NTG or control	2.8 (0.48-16.5)	4
Austrian (Benda) '77 [27]	acute MI < 12 hours; no age limits, n = 728	adequate R; no blinding; multicenter	SK vs control	0.56 (0.36-0.87)	3
Lasierra '77 [43]	acute MI < 48 hours; no age limits, n = 24	adequate R; no blinding;	SK + heparin vs heparin	0.22 (0.02-2.53)	2

SK = streptokinase IC-SK = intracoronair streptokinase UK = Urokinase
NTG = nytroglycerin R = randomization procedure

RESULTS

Trials included. In the original meta-analysis of Yusuf et al.[15] 33 trials were included. Of four studies two reports (short term and long term results) of the same trial were available.[25,26,32,33,39,40,50,57] For quality assessment we chose the report in which the method of research was most clearly described.[25,33,39,50] We identified 27 of these 33 references.[23-25,27-31,33-37,39,41-47,49,50,52,56,58] The language of most publications was in English, three in German,[27,29,49] one in French[58] and one in Spanish.[43] The Spanish report was translated into English to facilitate quality assessment. Of the original 33 references two were based on 'personal communications'.[38,55] We were able to trace one of them[38] as full report[54] and this paper was also included.

After detailed reading two of the 28 references did not meet the selection criteria (i.e. full report, presenting mortality data). One appeared to be a case series instead of an RCT.[35] Although the paper used the terms: 'RCT' and 'control group', it reported the results of a case series (n = 23), later on compared with a group of controls (n =11).
Another report focussed on the complications of streptokinase treatment, based on data derived from an RCT, but no data about mortality were presented.[46] Both studies were excluded from the analysis.

Of the remaining 26 references, four were abstracts.[42,47,52,56] We found two full reports of the same group of authors, about the same trial[48,53] and included them in this research. The main characteristics of the remaining 24 trials and their Delphi QSs are presented in Table 2 (see Appendix). In our calculations we used the data of the longest follow-up period available, found in the reports.

Pooled OR. Following Yusufs'[15] approach, we divided the reports into two categories: intravenous (streptokinase or urokinase) and intracoronary (streptokinase) treatment.

Regarding intravenous treatment we included 17 of the original 24 references. Our pooled OR was: 0.78 (0.67-0.90), equal to that of Yusuf et al. [0.78 (0.68-0.89)].

Regarding intracoronary treatment we included seven of the nine original references of Yusuf et al.[15] We calculated an OR of 0.68 (0.48-0.98). Contrary to Yusuf et al. we found a significant pooled OR, suggesting that intracoronary treatment with SK is benificial for patients with acute MI. The pooled OR found by Yusuf was only presented in a figure (OR = 0.8; 95% CI = 0.55-1.1).

To study the relationship between quality of an RCT and outcome we restricted ourselves to the studies concerning intravenous treatment, because we reproduced the same pooled OR as Yusuf et al. did with the studies found. The funnelplot we made of these 17 studies (Figure 1) showed no apparent publication/selection bias.

Quality assessment. The quality score (QS) of the Delphi list is presented in Table 2. The mean QS was 5 (max = 9). The Spearman rank correlation coefficient between both epidemiologists was 0.65, and between the cardiologists 0.59. The Spearman of the QSs versus the ORs was -0.21.

INCORPORATION OF QUALITY

Component analysis. Of the included 17 studies, nine[23,28,30-32,34,36,45,49] mentioned a concealed randomization, in five studies[24,27,37,43,58] the method was appropiate but not concealed, and in three[29,48,53] the method of randomization was unknown. Four studies[29,36,45,53] mentioned a form of blinding and all four used the term 'double blind'. In six studies[27,31,33,34,49] no withdrawals or a withdrawal rate not leading to bias was found, in seven the withdrawal rate was unknown[24,36,37,43,58,53,58] and in two studies[23,45] a withdrawal rate 'possibly leading to bias' was found. In five studies[30,31,48,53,58] an ITT analysis is performed. In Table 3 we present the pooled ORs and 95% CIs for each component in the different categories.

Figure 1: Funnelplot of the studies concerning intravenous treatment of MI.

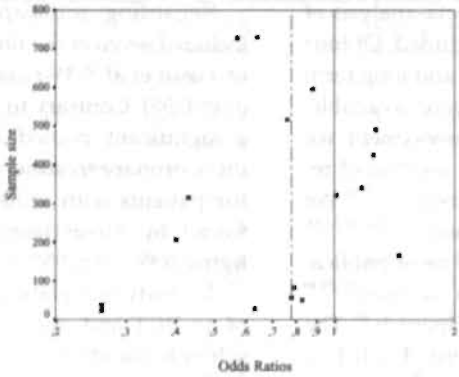


Table 3: Pooled ORs (95% CI) for subgroup of studies according to the individual components.

Component	*Pooled OR*	*95% CI*
Randomization		
concealed (n=9)	0.81	0.68-0.97
appropriate (n=5)	0.79	0.58-1.08
unknown (n=3)	0.43*	0.24-0.77
Blinding		
blinding reported (n=4)	0.46*	0.27-0.78
blinding not reported (n=13)	0.81	0.69-0.94
Withdrawals		
no withdrawals or not leading to bias (n=8)	0.73	0.61-0.86
withdrawal rate unclear or unknown (n=7)	1.02	0.68-1.54
withdrawal rate possibly leading to bias (n=2)	0.87	0.58-1.30
Intention-to-treat (ITT) analysis		
ITT analysis used/mentioned (n=5)	0.74	0.63-0.87
no ITT analysis used/mentioned (n=12)	1.00	0.69-1.45

* significantly different from the 'golden standard' pooled OR (OR = 0.78).

When the randomization method is unknown or when blinding is reported the pooled OR significantly differs from the overall average of true effects (the 'gold standard' pooled OR). Overall there is a tendency to underestimate the effect in the categories unknown except for randomization. When the different components (except blinding) are properly described, the pooled estimate of effect is close to the 'golden standard' OR.

Visual plot. To get insight in the relationship between overall quality and effect sizes we constructed a scatterplot of the QS against the individual ORs. Figure 2a shows the Delphi QS (vertically) against the effect size (horizontally) of all 17 full reports. Figure 2b presents the QS and the ORs of all 24 original studies; the studies of which we were unable to find received a QS of 2. Both plots show no correlation between overall quality score and effect.

Threshold score. We found a mean QS of 5 and included studies with at least a QS of 5. The pooled OR was 0.77 (0.65-0.91).

Weighting factor. We weighted each individual study estimate by their achieved Delphi QS

(Table 1). The pooled estimate was 0.78 (0.73-0.83). There is no difference in pooled OR using quality scores as a "threshold" or as a "weight". The confidence interval using a weighted analysis is smaller because all studies received a weight above 1.

Cumulative pooling. We started the cumulative pooling with the highest scoring study, and subsequently added the ones with lower QS. When 100% of the studies is included the original pooled estimate of effect (OR = 0.78) is reached. Figure 3 shows the cumulative pooling. The number of pooled studies in the category of high quality studies is small and the confidence intervals are wide. When the best 40% of the studies is included the pooled OR becomes significantly lower than 1,0.

Figure 2a: scatterplot between Delphi QS and individual Odds Ratios of the 17 full reports.

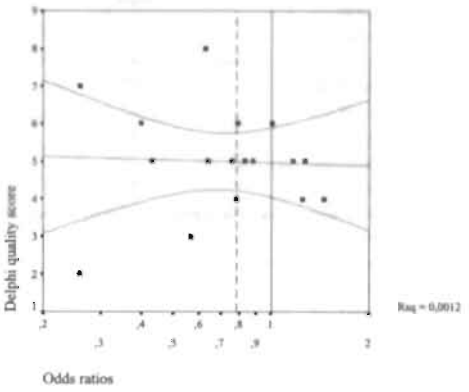


Figure 2b: Scatterplot between Delphi QSs and individual Odds Ratios of all 24 original studies.

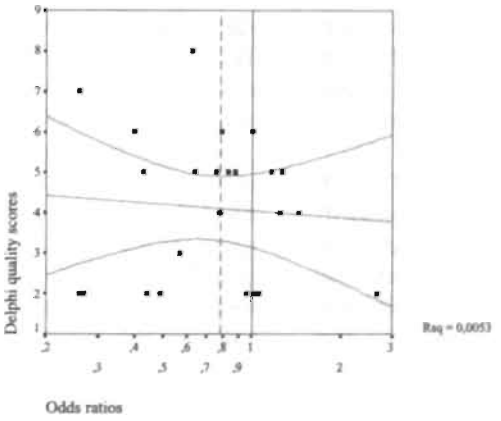
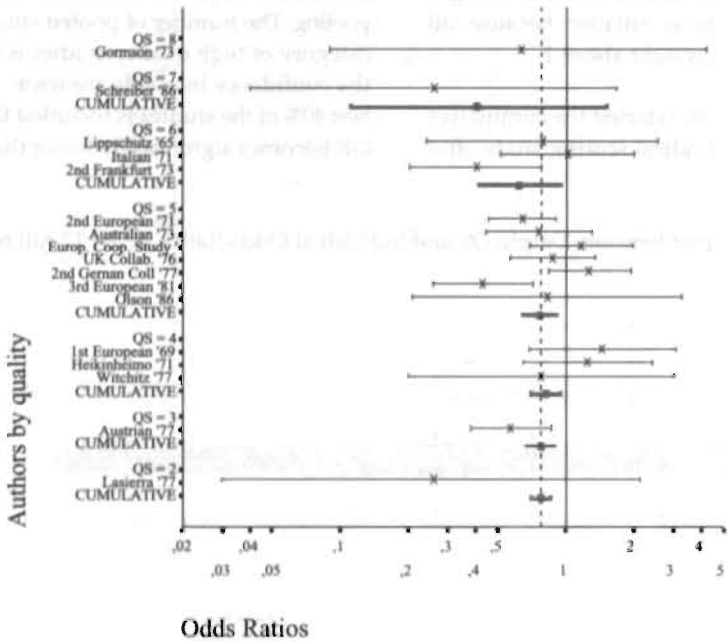


Figure 3: Cumulative Odds Ratio's.



Discussion

We found no apparent influence of methodological quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Therefore the results were also consistent over the five different methods of incorporating the quality into a final conclusion. Studies with a proper description of different quality components seem to provide a good estimate of the overall average of 'true' treatment effect (pooled OR).

We decided to accept the pooled OR as found by Yusuf et al.[15] as our 'gold standard', providing a 'true' treatment effect. This means that the Yusuf pooled OR is an overall average score. When the leading paradigm in empirical research is true, meaning that clinical trials with a poor methodological quality will usually be biased in favor of the intervention, we would expect to find a pooled OR of high quality studies closer to 1, and a pooled OR of the lower quality studies lower than the average Yusuf OR of 0.78. However, our research does not confirm this paradigm in empirical research. We must acknowledge that the number of studies is small and only a few are high quality studies. The high-powered studies (i.e. large sample size) were only of low to intermediate quality. Thus our study is not strong enough to confidently reject the paradigm, more empirical studies are needed.

When a reviewer performs a systematic review or meta-analysis, and quality assessment is a part of the review process, two decisions are important. First, the choice of a criteria list as a valid and reliable measuring instrument. We regard the Delphi list as a valid one, because of the way it is developed. Secondly, the reviewer has to decide about the way quality will be incorporated into the final conclusion. In the literature five different ways of incorporating the quality are des-

cribed.[6,11-14] Overall our results of incorporating quality into a final conclusion is consistent, but component analysis on 'blinding' provides a strange result we cannot explain.

Unfortunately, our data set was rather small to be able to draw firm conclusions. Because the studies we used were all published before the rise of empirical evidence of the importance of some design characteristics, our data set might be biased one way or the other. It is possible that authors may not have provided information needed to assess the quality, resulting in a lower quality score. On the other hand authors did probably not provide information just in order to be regarded as a high quality study when it is not.

In conclusion, quality assessment is seen as an important part of a meta-analysis, but the influence of quality on outcome is yet unclear and needs further research. In this research the pooled effect estimate is reached irrespective of the way quality is incorporated into the final conclusion.

References

1. Mulrow CD. The medical review article; state of the science. *Ann Intern Med 1987*;106:485-8.
2. Oxman A, ed. Preparing and maintaining systematic reviews: The Cochrane Collaboration Toolkit 1996. Available by Internet. Internet-address: Cochrane Home Page: http://hiru.mcmaster.ca/cochrane.
3. Haynes RB. Clinical review articles should be as scientific as the articles they review. *Br Med J 1992*; 304:330-1.
4. Schultz KF, Chalmers I, Hayes R, Altman DG. Empirical evidence of bias; Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA 1995*;273:408-12.
5. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technology Assessment 1996*;12; 195-208.
6. Schultz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaegology. *Br Med J 1996*; 312:742-4.
7. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet 1998*;352:609-13.
8. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *Br Med J 1998*; 317:1185-90.
9. Assendelft WJJ, Koes BW, van der Heijden GJMG, Bouter LM. Efficacy of chiropractic manipulation for back pain; blinded review of relevant randomized clinical trials. *JMPT 1992*;15:487-94.
10. de Bie RA. Methodology of systematic reviews: an introduction. *Phys Ther Rev 1996*;1:47-51.
11. Jenicek M. Meta-analysis in medicine: Where we are and where we want to go. *J Clin Epidemiol 1989*;42:35-44.
12. Detsky AS, Naylor CD, ORourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol 1992*;45:255-65.
13. Koes BW, van Tulder MW, van der Windt DAWM, Bouter LM. The efficacy of back schools: a review of randomized clinical trials. *J Clin Epidemiol 1994*;47: 851-62.
14. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM. Spinal manipulation for low back pain; an updated systematic review of randomized clinical trials. *Spine 1996*;221:2860-73.
15. Yusuf S, Collins R, Peto R, Furberg C, Stampfers MJ, Goldhaber SZ, Hennekens CH. Intravenous and intracoronary therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J 1985*;6:556-85.
16. GISSI-1. (Gruppo Italiano per lo Studio della Streptochinasi nell Infarcto miocardico). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet 1986*;1:397-401
17. ISIS-2 (Second International Study of Infarct Survival) Collaborative group. Randomized trial of intravenous streptokinase, oral aspirin, both or neither among 17,187 cases of suspected acute myocardial infarction. *Lancet 1988*;2:349-60.
18. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recomendations of clinical experts. *JAMA 1992*;268: 240-8.
19. Egger M, Smith GD, Phillips AN. Meta-analysis. Principles and procedures. *Br Med J 1997*;315:1533-7.
20. Egger M, Smith GD, Sneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *Br Med J 1997*;315:629-34.
21. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG. The Delphi list: a criteria list for quality assessment of Randomised Clinical Trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epid 1998*; 51: 1235-41.
22. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. *Control Clin Trials 1981*;2:31-49.

References of the meta-analysis

23. Aber CP, Bass NM, Berry CL et al. Streptokinase in acute myocardial infarction: a controlled multicentre study in the United Kingdom. *Br Med J 1976;*2:1100-4. [UK Collaborative].
24. Amery A, Roeber G, Vermeulen HJ, Verstraete M. Single-blind randomised multicentre trial comparing heparin and streptokinase treatment in recent myocardial infarction. *Acta Med Scand Suppl 1969;*505:5-35. [1st European].
25. Anderson JL, Marshall HW, Bray BE et al. A randomized trial of intracoronary streptokinase in the treatment of acute myocardial infarction. *New Eng J Med 1983;*308:1312-18.
26. Anderson JL, McIlvaine PM, Marshall HW et al. Long-term follow-up after intracoronary streptokinase for myocardial infarction. A randomized controlled study. *Am Heart J 1984;*108:1402-8.
27. Benda L, Haider M, Ambrosch F. [Results of the austrian myocardial infarction study on the effects of streptokinase.] *Wein Klin Wochenschr 1977;*89:779-83. [Austrian] (German).
28. Bett JHN, Biggs JC, Castaldi PA et al. Australian multicentre trial of streptokinase in acute myocardial infarction. *Lancet 1973;*1:57-60. [Australian].
29. Breddin K, Ehrly AM, Fechler L et al. [Short-term fibrinolytic treatment in acute myocardial infarction.] *Dtsch Med Wochenschr 1973;*98:861-73. [2nd Frankfurt] (German).
30. Dioguardi N, Marnicci PM, Lotto A. Controlled trial of streptokinase and heparin in acute myocardial infarction. *Lancet 1971;*2:891-5. [Italian].
31. European Cooperative Study. Controlled trial of urokinase in myocardial infarction. *Lancet 1975;*2:624-6.
32. European Cooperative Study Group for Streptokinase Treatment in Acute Myocardial Infarction. Streptokinase in acute myocardial infarction. *New Eng J Med 1979;*301:797-802. [3rd European].
33. The European Cooperative Study Group. Streptokinase in acute myocardial infarction extended report of the European Cooperative trial. *Acta Med Scand Suppl 1981;*648:7-57. [3rd European].
34. European Working Party. Streptokinase in recent myocardial infarction: a controlled multicentre trial. *Br Med J 1971;*3:325-31. [2nd European].
35. Fletcher AP, Sherry S, Alkjaersig N, Smyrniotis FE, Jick S. The maintainance of a sustained thrombolytic state in man. II. Clinical observations on patients with myocardial infarction and other thromboembolic disorders. *J Clin Invest 1959;*38:1111-9.
36. Gørmsen J, Tidstrom B, Fedderson C, Ploug J. Biochemical evaluation of low dose of urokinase in acute myocardial infarction. A double blind study. *Acta Med Scand 1973;*194:191-8.
37. Heikinheimo R, Ahrenberg P, Honkapohja H et al. Fibrinolytic treatment in myocardial infarction. *Acta Med Scand 1971;*189:7-13.
38. Hugenholtz PG, Serruys PW, Simoons ML et al. Randomized trial of intracoronary thrombolysis in acute myocardial infarction. Personal Communication 1985.
39. Kennedy JW, Ritchie JL, Davis KB, Fritz JF. Western Washington randomized trial of intracoronary streptokinase in acute myocardial infarction. *New Eng J Med 1983;*309:1477-82.
40. Kennedy JW, Ritchie JL, Davis KB et al. Western Washington randomized trial of intracoronary strep-tokinase in acute myocardial infarction. A 12-month follow-up report *New Eng J Med 1985;*312:1073-8.
41. Khaja F, Walton JA, Brymer JF et al. Intracoronary fibrinolytic therapy in acute myocardial infarction. Report of a prospective randomized trial. *New Eng J Med 1983;*308:1305-11.
42. Kolibash AJ, Magorien RD, Bashore TM, Bush CA. Does acute thrombolytic recanalization favorably alter segmental left ventricular function? *Circulation 1984;70*(Suppl II)258.(Abstract).
43. Lasierra C, Vilades J, Fernandez C et al. Estreptoquinasa en el infarcto agudo de miocardio. *Revista Clinica Espaniola 1977;*144:251-7. (Spanish).
44. Leiboff RH, Katz RJ, Wasserman AG et al. A randomized, angiographically controlled trial of intracoronary streptokinase in acute myocardial infarction. *Am J Cardiol 1984;*53:404-7.
45. Lippschutz EJ, Ambrus JL, Ambrus CM et al. Controlled studie on the treatment of coronairy occlusion with urokinase-activated human plasmin. *Am J Cardiol 1965;*16:93-8.
46. Ness PM, Simon TL, Cole C, Walston A. A pilot study of streptokinase therapy in acute myocardial infarction: observations on complications and relation to trial design. *Am Heart J 1974;*88:705-12. [NHLBI].
47. Olson HG, Lyons KP, Butman S et al. A randomized controlled trial of intravenous streptokinase in acute myocardial infarction. *Circulation 1984;* 70 (Suppl II); 155 (Abstract).
48. Olson HG, Butman SM, Piters KM et al. A randomized controlled trial of intravenous streptokinase in evolving acute myocardial infarction. *Am Heart J. 1986;*111:1021-9.
49. Poliwoda H, Schneider B, Avenarius HJ. [Investigations of the clinical course of acute myocardial infarction. I. The fibrinolytic treatment od acute myocardial infaction with streptokinase.] *Med Klin 1977;*72:451-63. [2nd German Collaborative] (German).
50. Raizner AE, Tortoledo FA, Verani MS et al. Intracoronary thrombolytic therapy in acute myocardial infarction. A prospective randomized trial. *Am J Cardiol 1985;*55:301-5.
51. Rentrop KP, Feit F, Blanke H et al. Effects of intracoronary streptokinase and intracoronary nitroglycerin infusion on coronairy angiographic patterns and mortality in patients with acute myocardial infarction. *New Eng J Med 1984;*311:1457-63.
52. Schreiber TL, Miller DH, Borer JS et al. Efficacy of intravenous heparin vs streptokinase in acute

myocardial infarction. *Circulation 1984*; 70 (Suppl II); 27 (Abstract).
53. Schreiber TL, Miller DH, Silvasi DA et al. Randomized double blind trial of intravenous streptokinase for acute myocardial infarction. *Am J Cardiol 1986*;58:47-52.
54. Simoons ML, Serruys PW, van der Brand M, Br F, de Zwaan C, Res J, Verheugt FWA, Krauss XH, Remme WJ, Vermeer F, Lubsen J. Improved survival after early thrombolysis in acute myocardial infarction; a randomized trial by the interuniversity cardiology institute in The Netherlands. *Lancet 1985*;i:578-81.
55. Theroux P. A trial of intracoronary streptokinase in acute myocardial infarction. (Personal communication to C Furberg).
56. Valre PE, Gurot C, Castillio-Fenoy A et al. L'infarctus myocardique. Traitement randomise par la streptokinase. *La Nouvelle Presse medicale 1975*;4:190. (Abstract) (French).
57. Verani MS, Tortoledo FA, van Reet RE, Young JB, Poliner LR, Raizner AE. Intracoronary thrombolysis in acute myocardial infarction: effects on function of myocardium at risk. Preliminary findings of a randomized trial. *J Am Coll Cardiol 1983*;1:592. (Abstract) [Baylor/Methodist].
58. Witchitz S, Kolsky H, Moisson P, Chiche P. Streptokinase et infarctus du myocarde aigu. La fibrinolyse peut-elle limiter la necrose? *Ann Cardiol Anglol 1977*;26:53-6. (French)

9 Discussion. The art of quality assessment of RCTs.

AUTHORS:
Arianne P. Verhagen,
Henrica C.W. de Vet,
Robert A. de Bie,
Maarten Boers,
Piet A van den Brandt.

Based on: Verhagen AP, de Vet HCW, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs. *Submitted.*

Well conducted randomized clinical trials (RCTs) provide the best evidence on the efficacy of medical interventions. In 1998 the British Medical Journal celebrated the 50th anniversary of the RCT in the medical field. During the past 50 years the number of RCTs published each year increased immensely. According to MEDLINE over 29,000 new RCTs were published in 1997. For the ordinary clinician it has become impossible to keep up with the latest evidence.

Systematic reviews include a comprehensive search strategy and a predetermined and explicit method to appraise and synthesise the information from individual studies.[1,2] Obviously the validity of the conclusions of a systematic review will depend on the quality of the included primary studies, based on the dictum: garbage in, garbage out. Therefore, assessment of trial quality is often a part of the process of a systematic review.

Over the years the Department of Epidemiology of the Maastricht University in The Netherlands, has gained experience in performing systematic reviews. The protocol for these reviews emphasizes the assessment of the methodological quality of individual trials. Ter Riet et al.[3], Kleijnen et al.[4], Koes et al.[5] and Beckermann et al.[6] can be credited with the publication of one of the first quality criteria lists of this Department. In the course of time additional items have been added to form what is now called: The Maastricht list.[7]

During the development of the Maastricht list, people felt the need to scientifically ground the process of quality assessment of RCTs to be used in systematic reviews. These efforts resulted in a series of studies presented in this thesis. We developed a new criteria list: The Delphi list,[8] in a consensus exercise using formal scale development techniques (see Chapter 3). Furthermore we performed four reviews using different criteria lists for quality assessment. In this chapter we review the data from these methodological studies to discuss the validity and reliability of quality assessment in randomized clinical trials (RCTs). What is the empirical evidence to support quality assessment of RCTs?

What Is Quality ?

Most criteria lists proposed to assess the methodological quality of RCTs do not explicitly define the concept of quality.[9] These lists usually include at least three dimensions that may encompass the concept of quality: internal validity, external validity and statistical analysis.[10-12] Quality of RCTs has recently been defined as: 'the likelihood of the trial design to generate unbiased results.[13] This definition covers only the dimension of internal validity. During our development of the "Delphi list" for quality assessment (Chapter 3), the participants, all experts in the field of RCTs, failed to agree on a specific definition, but did agree that the concept quality should comprise more than internal validity alone.[8] From this context we propose the following definition of quality: 'the likelihood of the trial design to generate unbiased results, that are sufficiently precise to allow application in clinical practice."

How Can Quality Be Assessed ?

The last decade has shown efforts to develop appropriate tools for quality assessment. This includes the 1996 CONSORT guidelines[14] that aim to set the standard for the written report of an RCT. When assessing trial quality one has to rely on the information retrieved from these reports.

Mainly, a criteria list is used to provide a quality score as an estimation of the overall methodological quality of the design and conduct of the trial. Higher quality scores should indicate studies with a better methodological quality.[15,16] Another approach in quality assessment is to focus on components such as randomization, blinding etc. in trial reports.[17,18]

Moher reported in 1995 that at least 25

criteria lists have been developed.[9] We currently estimate the number of quality scales at 50 or 60, and the number is still increasing. Most tools to assess trial quality are regarded as valid because they comprise a selection of 'accepted criteria', such as those listed in textbooks on clinical trials as aspects of importance for the quality of a trial.[9] The Maastricht list is an example of such a criteria list.[7]

We feel that the application of formal scale development techniques, including consensus can increase the validity (face and content) of the resulting quality scale. As far as we are aware the only criteria lists developed by such techniques are the list developed by Jadad et al.[13] and the Delphi list[8] (Chapter 3).

Although consensus provides some validity it is not a paradigm. Consensus is always achieved within a theoretical model, which can be proven wrong in time. A well known example of wrong consensus occurred in the beginning of this century. Not long after Einstein developed his relativity theory, a hundred physicists reached consensus against it, some say based on anti-Semitic emotions. Einsteins' famous answer was: you need just one physicist to prove that this relativity theory is wrong. To date, no one has succeeded in falsifying this theory.

Notably, consensus did not prevent marked differences between the Jadad and the Delphi list. Whether quality assessment is of surplus value in the process of a systematic review partly depends on the validity and reliability of the criteria list used.

What Is The Evidence For The Validity And Reliability Of Quality Assessment?

The validity of quality assessment itself needs further study. Such studies should answer the question: Does this criteria list measure what it is supposed to measure, namely the methodological quality of the trial? To know how close a criteria list measures the true state, we need a gold standard, or external criterion to compare it against (*criterion validity*). However, a gold standard of quality assessment does not, and probably will never, exist. Sensitivity and specificity (to important differences in quality) requires an understanding of what such a difference is; this understanding is currently not available.

When a gold standard is lacking one has to fall back on a theoretical model. This means assessing the *content validity*: does the method of measurement include all dimensions of the theoretical framework? Almost all existing criteria lists are based on accepted methodological criteria as presented in textbooks. In the development of the Jadad list, the participants used a definition that covers only the domain of internal validity.[13] During the development of the Delphi criteria list[8], each participant had their own 'clear picture' (or definition) of what quality comprises, but the picture varied between participants (Chapter 3). We selected the participants to achieve a broad representation of all different points of view on quality assessment. The advantage of this approach, in our opinion, was that the different ideas of the concept of quality were integrated in the resulting criteria list, enhancing content validity.

Another aspect of validity is *construct validity*: is the measurement consistent with other measurements of quality assessment? This means comparing the results of different quality criteria lists with each other. In our studies in which we compared the Maastricht, Jadad and Delphi lists we found sometimes that they gave conflicting results (Chapter 5, 7 and 8). On the one hand we found high Spearman rank correlations between the criteria lists (min $r = 0.71$, max $r = 0.87$) in the Laser-review (Chapter 5). On the other hand we found lower correlations (min $r = 0.47$, max $r = 0.82$) in the Yusuf-review (Chapter 8). This indicates variability in ranking of the trials. Other studies confirm our conclusion that different criteria lists, applied to the same set of trials, do not always provide similar results.[17,19] In conclusion, the validity of quality assessment is unclear.

In order to discriminate between high and

low quality trials, quality assessment must, apart from being valid, also be reliable. The reliability of most criteria lists is unknown. In reviews where more than one reviewer assesses the trial quality, reviewers discuss their differences and reach a consensus score. The interrater agreement of the Jadad list, expressed with Intra Class Correlation coefficients (ICC), was reported as 0.56 (0.36-0.75)[13], and as 0.85 [20]. We found a Spearman rank correlation on the Jadad list between two trained reviewers (epidemiologists) of $r = 0.45$ in the Yusuf-review (Chapter 8).

Concerning the interrater agreement of the Maastricht list, we found an ICC of 0.77 (0.64 - 0.89) in the Balneo-review (Chapter 2).[21] In the Yusuf-review (Chapter 8) we reported a Spearman rank correlation of $r = 0.72$. Other studies performed with the Maastricht list reported an overall agreement between two reviewers of approximately 80%[22] up to 95% in the Ankle-review (Chapter 7).

Concerning the interrater reliability of the Delpi list, we found an overall agreement between the two reviewers of 95% in the Ankle review (Chapter 7), and in the Yusuf review a Spearman rank correlation coefficient $r = 0.65$, when reviewed by two epidemiologists.

We conclude that, when studied, the interrater agreement of these three criteria lists vary from moderate to good.[13,20,21,23,24]

What Are The Major Potential Biases In The Application Of Quality Assessment?

Reporting bias is a specific form of information bias: Are published RCTs a true reflection of what went on in the trial? Reporting of trials may be flawed in a way that it provides a misleading impression of methodological quality, one way or the other.[25]

On the one hand, poorly reported trials could be judged as having low quality, while it may not be so. This is a problem of 'underreporting'. A flawed report, i.e. lacking the necessary information about trial design and conduct, does not necessarily mean that the underlying study was flawed. It may reflect a lack of understanding of the reporting requirements for such studies.[26] On the other hand quality can also be 'overreported'. Between 5% and 30% of 'randomized' studies, may actually not have performed a method of randomization.[27,28] With the growing emphasis on the quality of trials, the problem of overreporting may be increasing.

The only way to prevent 'underreporting' is to educate investigators in the methodology and reporting requirements of trials, which are now well established.[29,30] But is there a way to prevent 'overreporting', other than being a spy or a fly on the wall when researchers discuss their reports? We consider reporting bias at this moment still a major threat to the validity of quality assessment, despite current efforts to improve the quality of trial reports.[14]

Review bias is a type of information bias specifically for reviews. Beliefs and disbeliefs can (subconsciously) guide reviewers into biased assessments. To prevent it, it has been suggested to perform quality assessment under masked conditions, i.e. authors, institutes, sponsorships, journals of publication, or study results should be unknown to the reviewer.[11,13,31-33]

In some studies blinded assessment of methodological quality produced significantly lower and more consistent quality scores than unblinded assessment.[13,34] The overall quality scores between the blinded and unblinded assessment in our Balneo-review (Chapter 1 and 2) did not differ much.[21] The reviewers in the latter review were all epidemiologists without a professional relationship with the intervention. Thus we surmise that the level of epidemiological knowledge and the professional linkage of the reviewers might be more important in relation to review bias than study characteristics.[21]

We set out to investigate the influence of professional linkage and the need for blinding in the Yusuf-review (Chapter 8). Two cardiologists and two epidemiologists reviewed the same set of trials using the

Delphi list. Unfortunately, the cardiologists involved saw no need to mask the articles, because they felt able to recognize the separate studies based on the number of patients randomized and the intervention under study. Therefore we refrained from blinding the articles. The interrater reliability of the Delphi list between epidemiologists was r=0.65 (Spearman), and between cardiologists r=0.59. The epidemiologists' final quality score was based on consensus. The Spearman correlation coefficient of each cardiologist independently with this consensus quality score was r=0.68 and r=0.37 resp. One cardiologist interpreted the items on the Delphi list in a personal way. Assendelft et al.[35] stated that a possible cause of review bias could occur when the profession of the reviewers is linked to the intervention investigated. He found a relationship between the review quality score, the profession of the reviewer and the conclusions.

In conclusion, research concerning the influence of masked quality assessment shows inconsistent results.[13,21,34] Reviewers should, for the moment, state explicitly in their reviews if quality was assessed by experts in the field or not.

Bias due to *misclassification* may result when overall quality is used to determine a cut-off point between high versus low quality studies. When the validity and reliability of the quality criteria list is poor or unknown, there is a real chance of bias due to misclassification. Using only quality components might decrease the problem of misclassification, but does not fully overcome it.

Can We Incorporate Quality Into The Conclusion?

Whether or how the results of quality assessment should be incorporated into the conclusion of a review is under debate, especially when quality scores are used.[36] Several strategies are available to do this. First of all, a visual plot of the effect size against an overall quality score can be presented.[17,37,38] Further, quality scores can be used as a 'threshold score' for inclusion of the article in a review, as a 'weighting factor' in the statistical analysis,[17,19,39] or as the input sequence in a cumulative meta-analysis.[17,37,38] Hardly any empirical evidence concerning the different ways of incorporating quality in a review is available.[19] The same conclusion was reached in our Yusuf-review (Chapter 8) irrespective of the way the overall quality scores were incorporated. From these methods we prefer to visually plot the effect size against a quality score. This way one can evaluate whether and how quality influences the final conclusion. Finally, a recent suggestion is to focus on components on quality, and to use meta-regression techniques to evaluate the possibility and the direction of bias.[36] This is, on our opinion, for the moment the best way to evaluate the influence of design characteristics on outcome.

Which Way Ahead: Quality Components Or Quality Scores?

The leading paradigm in the field of quality assessment of RCTs, is that low quality studies tend to overestimate effect estimates. This paradigm is based on the assumption that investigators are (subconciously) biased in favor of the intervention.[18] The advantage of using an overall quality score is its simplicity, but methodologically it is arguable. Shortcomings on some methodological criteria can be compensated by positive scores on other criteria. On the basis of our own research we might state that low study quality can both underestimate and overestimate the true effect (Chapter 7 and 8).

Empirical research has show that components of quality can indeed influence the effect estimates, however, the direction of this influence is not consistent.[18,25,34,40] In our opinion research should primarily focus on components of quality, measured using a criteria list. Apart from randomization and blinding items, this criteria list should contain

items concerning other design characteristics that possibly influence the results (e.g. the Delphi list).[8] Empirical research should determine the relevance of these and other items. When their influence is clear, summation of the important components of quality into an overall quality score can be investigated. Such research should then guide the improvement of existing quality criteria lists.

Plans For The Future.

The Cochrane Collaboration publishes guidelines on how to perform systematic reviews. The Methods Working Group on Empirical Methodological Studies (EMS MWG) of the Cochrane Collaboration reviews methodological studies to evaluate the relationship between overall quality and design characteristics and effect sizes. Clearly reviews are useful, but individual studies are still needed. As stated before such studies should primarily focus on design characteristics. Emphasis should be put on two or three generic criteria lists for quality assessment, and to study the impact of quality components. Review groups can always add specific items to one of these generic criteria lists, relevant for the aim of the review.

We assume that the relevance and importance of several quality components will strongly depend on the topic of research. Therefore we recommend to evaluate quality items within a series of specific research question, and to eventually qualitatively summarise these studies into a methodological review. Qualitative summarization, instead of statistical pooling, is important in order to understant the relevance of quality components and why this relevance differs between different research questions.

We consider quality assessment a valuable tool to differentiate between studies, in order to find a clinically relevant effect estimate in systematic reviews. The task is to generate a valid quality criteria list to provide more insight in the still hazy relationship between quality and outcome.

In this thesis I described five years of research in the field of quality assessment. During those five years our view on the impact of quality on the results of trials, and quality assessment itself, changed several times. The aim of research in methodological quality is to promote more valid trial results through improved methodology. These results allow health care providers to provide better care for their patients. This in itself is reason enough to continue studies such as these, and I hope to be able to make further contributions on a research level in the day-to-day reality of patient care.

References

1. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the potsdam consultation on meta-analysis. *J Clin Epidemiol 1995;*48:167-171.
2. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol 1991;*44: 1271-8.
3. Ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. *J Clin Epidemiol 1990;*43:1191-1199.
4. Kleijnen J, Knipschild P, ter Riet G. Clinical trials of homeopathy. *Br Med J 1991;*302:316-23.
5. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM, Knipschild PG. Spinal manipulation and mobilization for back and neck pain: A blinded review. *Br Med J 1991;*303:1298-303.
6. Beckerman H, de Bie RA, Bouter LM, de Cuyper HJ, Oostendorp RAB. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. *Phys Ther 1992;*72:483-91.
7. de Vet H, de Bie R, van der Heijden G, Verhagen A, Sijpkes P, Knipschild P. Systematic Reviews on the Basis of Methodological Criteria. *Physiotherapy 1997;* 83:284-9.
8. Verhagen AP, de Vet HCW, de Bie RA, et al. The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus. *J Clin Epidemiol 1998;*51:1235-41.
9. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials 1995;*16:62-73.
10. The SORTG. A proposal for structured reporting of randomized controlled trials. *Jama 1994;*272:1926-31.
11. Chalmers TC, Smith Jr H, Blackburn B, Silverman B, Schoeder B, Reitman D. A method for assessing the

quality of a randomized control trial. *Control Clinical Trials 1980;2*:31-49.
12. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy:I: Medical. *Stat Med 1989;8*:441-54.
13. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials 1996;*17:1-12.
14. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement [see comments]. *JAMA 1996;276*:637-9.
15. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. *Intern J Technology Assesment Health Care 1996;*12:195-208.
16. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA 1995;*273:408-12.
17. Assendelft WJ, Koes BW, van der Heijden GJ, Bouter LM. The efficacy of chiropractic manipulation for back pain: blinded review of relevant randomized clinical trials. *J Manipulative Physiol Ther 1992;*15:487-94.
18. Bie RAd. Methodology of systematic reviews: an introduction. *Phys Ther Rev 1996;*1:47-51.
19. Detsky AS, Naylor CD, Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol 1992;*45:255-65.
20. Bender JS, Halpern SH, Thangaroopan M, Jadad AR, Ohlsson A. Quality and retrieval of obstetrical anaesthesia randomized controlled trials. *Can J Anaesth 1997;*44:14-8.
21. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Knipschild PG. Balneotherapy and quality assessment: interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. *J Clin Epidemiol 1998;*51:335-41.
22. van der Heijden GJ, Beurskens AJ, Koes BW, Assendelft WJ, de Vet HC, Bouter LM. The efficacy of traction for back and neck pain: a systematic, blinded review of randomized clinical trial methods. *Phys Ther 1995;*75:93-104.
23. Brown SA. Measurement of quality of primary studies for meta-analysis. *Nursing Research 1991;*40:352-5.
24. Emerson JD, Burdick E, Hoaglin DC, Mosteller R, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clinical Trials 1990;*11:339-52.
25. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *Br Med J 1996;* 312:742-4.
26. Grimes DA, Schulz KF. Methodology citations and the quality of randomized controlled trials in obstetrics and gynecology. *Am J Obstet Gynecol 1996;*174: 1312-5.
27. Evans M, Pollock AV. Trials on Trial. A Review of Trials of Antibiotic Prophylaxis. *Arch Surg 1984;*119: 109-13.
28. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet 1990;*335:149-53.
29. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials. *Control Clinical Trials 1980;*1:37-58.
30. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. *New Eng J Med 1982;*3:1332-7.
31. Gerbarg ZB, Horwitz RI. Resolving conflicting clinical trials: guidelines for meta analysis. *J Clin Epidemiol 1988;*41:503-9.
32. Sacks HS, Berrier J, Reitman D, Ancona BVA, Chalmers TC. Special article: Meta-analysis of randomized controlled trials. *New Eng J Med 1987;*316:450-5.
33. Jadad AR, Moher D, Klassen TP. Guides for reading and interpreting systematic reviews: II. How did the authors find the studies and assess their quality? *Arch Pediatr Adolesc Med 1998;*152:812-7.
34. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet 1998;*352:609-13.
35. Assendelft WJJ, Koes BW, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. *JAMA 1995;*274:1942-8.
36. Greenland S. Invited commentatory: a critical look at some popular meta-analytic methods. *Am J Epidemiol 1994;*140:290-6.
37. Koes BW, van Tulder MW, van der Windt WM, Bouter LM. The efficacy of back schools: a review of random-ized clinical trials. *J Clin Epidemiol 1994;*47: 851-62.
38. Koes BW, Assendelft WJ, van der Heijden GJ, Bouter LM. Spinal manipulation for low back pain. An updated systematic review of randomized clinical trials. *Spine 1996;*21:2860-71.
39. Jenicek M. Meta-analysis in medicine: where we are and where we want to go. *J Clin Epidemiol 1989;*42:35-44.
40. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *Br Med J 1998;*317:1185-90.

Summary

Quality of care depends, among others, on the efficacy of health care interventions. A causal relation between health care interventions and their effects can only be determined by randomised clinical trials (RCTs). These RCTs are the scientific tool for answering the question: What is the efficacy of a specific treatment in patients with a certain disease or disability, compared to a placebo treatment or no treatment? High quality trials may be considered a valid measure of treatment efficacy, low quality trials on the other hand, may not.

The aim of systematic reviews or meta-analyses is to summarize the results of the individual RCTs in a systematic way. This benefits health care providers who no longer need to read all RCTs in order to find out which treatment is best. Also patients benefit because they are more likely to receive the best treatment available at that time. In a systematic review the literature is searched and judged in a systematic way. Part of the judgement is the assessment of the methodological quality. It stands to reason that one should base the conclusion of a systematic review on RCTs of sufficient methodological quality. In order to assess the quality, reviewers or researchers use criteria lists, also called checklists or quality scales. The aim of this thesis is to identify which are the methodological aspects that reflect on the quality of RCTs. Apart from validating existing criteria lists, also a new criteria list is developed and tested.

Chapter 1. We started with a systematic review on the effectiveness of balneotherapy, including English, French, German and Dutch language studies. Balneotherapy (hydrotherapy or spa therapy) is one of the oldest forms of therapy for patients with arthritis. One of the aims of balneotherapy is to relieve the pain. This systematic review included randomized and non-randomized clinical trials. We performed computer-aided searches of databases and of bibliographic indexes and congress reports; also reviews and handbooks were all checked for relevant citations. Quality scores of the studies were determined using the 'Maastricht' criteria list.

14 studies fulfilled the entry criteria for this review. In eight studies patients with rheumatoid arthritis were included, in the other studies patients with other forms of arthritis were included, such as osteoarthritis of the hip, juvenile arthritis and psoriatic arthritis. Most studies reported positive findings, but all studies showed methodological flaws. A quality of life measurement was never reported as an outcome measure. None of the randomized clinical trials included an intention-to-treat analysis or a comparison of effects between groups.

Because of these methodological flaws a conclusion about the efficacy of balneotherapy cannot be provided from the studies we reviewed. We concluded that most flaws found can be avoided in future research.

Chapter 2. In this study we investigated aspects of the reliability of the 'Maastricht' criteria list for quality assessment in systematic reviews, and whether blinded reviewing is necessary to prevent review bias. We used the data set of 14 articles from the systematic review concerning the efficacy of balneotherapy in patients with arthritis as presented in Chapter 1. Two studies were excluded because of the cross-over design they used. Twenty reviewers participated of which two reviewers, who have been involved in developing the 'Maastricht'

criteria list, acted as reference standard. Half of all assessments were performed blindly. Studies were blinded for author and affiliations of the study, journal and year of publication and results.

A high level of agreement was found between the reviewers and a high level of correlation with the reference standard. The quality scores between the blinded and unblinded assessment did not differ much. Based on these results we concluded that the 'Maastricht' criteria list is a reliable instrument in quality assessment of clinical trials. Within the limits of this study we found no evidence that blinding is necessary to prevent review bias.

Chapter 3. Most systematic reviews rely substantially on the assessment of the methodological quality of the individual trials. The aim of the study presented in this Chapter, is to obtain consensus among experts about a set of generic core items for quality assessment of randomised clinical trials (RCTs). The invited participants are experts in the field of quality assessment of RCTs. The initial item pool contained all items from existing criteria lists. Subsequently, we reduced the number of items by using the Delphi consensus technique. Each Delphi round comprised of a questionnaire, an analysis and a feedback report. The feedback report included staff team decisions made on the basis of the analysis and their justification.

A total of 33 international experts agreed to participate of whom 21 completed all questionnaires. The initial item pool of 206 items was reduced to nine items in three Delphi rounds. The final criteria list (the 'Delphi' list) was satisfactory to all participants. It is a starting point on the way to a minimum reference standard for RCTs on many different research topics. This list is not intended to replace, but rather to be used alongside or incorporated in existing criteria lists.

Chapter 4 presents a systematic review to assess the effectiveness of 904 nm low level laser therapy (LT) in musculoskeletal disorders. In order to retrieve randomised trials, computer-aided searches of databases and bibliographic indexes were performed. Furthermore, congress reports, reviews and handbooks were all checked for relevant citations. Subsequently, all retrieved studies were scored on methodological quality using the 'Maastricht' criteria list. We found 25 studies that investigated the effects of 904 nm LT versus placebo or any other intervention, in subjects with musculoskeletal disorders for which LT was thought a feasible intervention. Of these 25 studies, 21 fulfilled the entry criteria for this review, and quality was assessed in a blinded manner.

Overall, study quality ranged from 'poor' to 'reasonable'. In a classification of the material into diseases studied, no clear evidence was found for the effectiveness of LT, except perhaps for knee problems and myofascial pain. In conclusion, 904 nm LT does not seem to be effective in the treatment of musculoskeletal disorders, but further and improved research is needed to shed more light on its effectiveness.

Chapter 5. The conclusion of a systematic review depends on the quality of the individual studies included. This article presents the results of a comparison of three different methods of quality assessment. The data set of 21 randomized clinical trials (RCTs) from a systematic review concerning the efficacy of laser therapy in patients with musculoskeletal disorders is used (Chapter 4). The criteria lists to assess the methodological quality were the 'Maastricht' list, the 'Jadad' list and the 'Delphi' list. The three criteria lists show moderate to good correlation. Major differences between the lists are the number of items, and differences in wording of the items. The latter seemed to affect the ranking of the studies. Based on our results we concluded that the 'Delphi' list seems the most practical and satisfactory instrument for quality assessment of RCTs.

Chapter 6 consists of a systematic review that summarises the efficacy of conservative interventions in acute lateral ankle sprains. We performed computer aided searches of databases and of bibliographic indexes. Furthermore, we checked congress reports, reviews and relevant citations. Subsequently, all retrieved empirical studies were scored on methodological quality and effect sizes were calculated for days of sick leave, pain and swelling.

We found 81 studies that investigated the effects of physiotherapy interventions versus other interventions or placebo interventions, in subjects with acute lateral ankle sprain. Of these, 44 fulfilled our entry criteria. Study quality was assessed using the 'Maastricht' list and ranged from poor to rather good. Only two studies, both on pulsed shortwave therapy, scored more than half of the maximum available score and they showed no effect compared to placebo therapy. Tape was found to be superior over other types of treatment, in effect shortening the duration of sick-leave, while plaster of Paris treatment seemed to prolong sick-leave. Further and improved research is needed to shed more light on efficacy of other treatment interventions.

Chapter 7. This study investigates the influence of different aspects of methodological quality on the conclusions of a systematic review concerning treatments of acute lateral ankle sprains. A data set of 44 trials was used studying the efficacy of conservative interventions in patients with an acute lateral ankle sprain, as presented in the previous Chapter. Quality assessment of the individual studies was performed using the 'Delphi' list. We calculated effect sizes of the main outcome measures in each study in order to evaluate the relationship between overall quality scores and outcome. Next, we set out to investigate the impact of design attributes on pooled effect sizes by subgroup analysis according to the design attributes.

Quality scores vary from rather low to reasonably good. Only 23 of the 44 studies allowed calculation of effect sizes, for one or more outcome measures. We determined an estimate of an 'overall average of effects'. Studies with a proper randomization procedure and those which report a form of blinding both produce a slightly higher estimate of this average effect size. When the randomization and blinding procedures are unknown we found an underestimation of the effect rather than an overestimation. Design factors such as randomization and blinding have an impact on outcome. Previous research has suggested that methodologically poor designed studies tend to overestimate the effect. Our study does not confirm these suggestions.

Chapter 8. To study the influence of the methodological quality of individual trials on the outcome we performed a remake of a landmark meta-analysis on thrombolytic therapy in acute myocardial infarction. The aim was to included the same studies as in the meta-analysis of Yusuf et al. (Eur Heart J. 1985;6:556-85). From each study we extracted the number of patients in the treated and control groups who died in hospital or during follow-up. The methodological quality was assessed using the 'Delphi' list.

Unfortunately we were unable to trace all studies, so we first recalculated the pooled Odds Ratios (ORs), and their 95% confidence intervals (CIs), on the studies found and our results were similar compared with the original results of Yusuf et al. We incorporated the results of the quality assessment in several ways in the calculation of the pooled ORs: a) component analysis of the methodological items or components: randomization, blinding, withdrawals and analysis; b) visual plot of the individual ORs against the quality score; c) the quality score used as a 'threshold score' for inclusion of the article in the pooling; d) quality score used as a 'weighting factor'; e) cumulative pooling using quality scores as the input sequence. No correlation between overall quality scores and effect size was found. Studies with a proper description of the different quality compo-

nents provided a good estimate of the true treatment effect. The other studies gave an overestimation or an underestimation of treatment effect.

Chapter 9. In this chapter we discuss quality assessment in general: what is quality, its' validity and reliability. In the literature no satisfying definition of 'quality' can be found. At this moment the number of criteria lists is estimated at 60. The only two criteria lists developed using formal scale developing techniques are; the 'Jadad list' and the 'Delphi list'. We assume these criteria lists to have 'expert' validity, because of the way they are developed. Overall no conclusions about the validity of criteria lists can be drawn.

The reliability of the three criteria lists used in this thesis seem to be 'reasonable' to 'good'. In this Chapter some of the most important fallacies of quality will be discussed. One of the important fallacies is 'reporting' bias, meaning: is the information presented in the report an adequate reflection of what happened during the study? For the assessment of the quality we use the information presented in the written report.

The leading paradigm in the field of quality assessment of RCTs, is that low quality studies tend to overestimate effect estimates. This paradigm is based on the assumption that investigators are (subconciously) biased in favor of the intervention. On the basis of our own research we might state that low study quality can both underestimate and overestimate the effect.

In our opinion research should primarily focus on components of quality, measured using a criteria list, and determine the relevance of different items. We assume that the relevance and importance of several quality components will strongly depend on the topic of research. Therefore we recommend to evaluate quality items within a series of specific research question.

In conclusion, we consider quality assessment a valuable tool to differentiate between studies, in order to find a clinically relevant effect estimate in systematic reviews. The task is to generate a valid quality criteria list to provide more insight in the still hazy relationship between quality and outcome.

Samenvatting

De kwaliteit van zorg wordt onder andere bepaald door de effectiviteit van de behandeling die wordt gegeven. Bewijzen voor die effectiviteit worden geleverd door gerandomiseerd effectonderzoek onderzoek (*randomised clinical trial* = RCT). Het RCT is de vorm van onderzoek dat een antwoord kan geven op de vraag: Wat is de effectiviteit van die behandeling bij patiënten met die bepaalde ziekte of aandoening, ten opzichte van een andere behandeling of niets doen? RCT's met een hoge methodologische kwaliteit geven waarschijnlijk een meer valide antwoord op deze vraag, dan studies met een lage methodologische kwaliteit. In de loop der jaren zijn soms meerdere RCT's gedaan met vergelijkbare onderzoeksvragen. Deze onderzoeken hadden niet altijd dezelfde uitkomsten. Hierdoor is het voor de persoon in de dagelijkse praktijk moeilijk te bepalen welke uitkomst ze nu moeten geloven of niet.

Het doel van een literatuuronderzoek (review), is om een samenvatting te geven van de aanwezige kennis op een bepaald vakgebied. Het voordeel is dat de behandelaar niet zelf alle bestaande onderzoeken hoeft te lezen en beoordelen, voordat men een conclusie kan trekken over welke behandeling nu het beste is. Het voordeel voor de patiënt is dat de kans om de beste behandeling te krijgen groter wordt.

Voor een systematische review geldt dat de literatuur op een systematische wijze is verzameld en beoordeeld. Onderdeel van de beoordeling is dat de methodologische kwaliteit van de onderzoeken wordt bepaald. Het klinkt logisch om, bijvoorbeeld, alleen de resultaten van die RCT's in de conclusie van een systematische review te betrekken waarvan de methodologische kwaliteit voldoende is. Die beoordeling van de kwaliteit gebeurt met behulp van een criteria lijst. Het doel van dit promotieonderzoek was om meer inzicht te krijgen in de manier van kwaliteitsmeting bij RCT's.

In **Hoofdstuk 1** wordt een systematische review beschreven met de vraagstelling: Wat is de effectiviteit van balneotherapie (dit is spa- of kuuroord therapie) bij patiënten met een vorm van arthritis? Balneotherapie is een van de oudste vormen van therapie bij deze patiënten. In de review zijn effectonderzoeken opgenomen waarbij de patiënten zijn gerandomiseerd (=door het lot verdeeld) of door de onderzoeker zijn verdeeld over de interventie groep of de controle groep. Er zijn alleen Engels-, Duits-, Frans- en Nederlandstalige artikelen in de review opgenomen. Na een intensieve zoekactie met behulp van gecomputeriseerde literatuurbestanden en referenties, zijn 14 studies gevonden die voldeden aan de in- en exclusie criteria voor deze review. Acht studies waren uitge-voerd met patiënten met rheumatische arthritis. In de andere studies waren patiënten opgenomen met osteoartritis van de heup, juveniele arthritis of psoriatrische arthritis. In de helft van alle studies is een vorm van randomisatie uitgevoerd. Opvallend was dat in geen van de onderzoeken de 'kwaliteit van leven' van deze patiënten als effectmaat is gemeten.

De methodologische kwaliteit werd gemeten met behulp van een al bestaande criteria lijst, de zogenoemde 'Maastricht criteria list'. De meeste onderzoeken in deze review rapporteerden positieve resultaten, maar de meeste studies vertoonden methodologische mankementen. Vanwege deze mankementen kan geen antwoord worden gegeven op de onderzoeksvraag of balneotherapie effectief is bij deze groep patiënten.

Hoofdstuk 2. In dit hoofdstuk beschrijven we een studie waarin verschillende methodologische aspecten van de 'Maastricht criteria list' zijn onderzocht. We hebben gebruik gemaakt van dezelfde studies als in de review naar het effect van balneotherapie (zie Hoofdstuk 1). Van de 14 originele studies zijn 2 uitgehaald omdat deze een zgn 'cross-over design' hadden. Onafhankelijk van elkaar hebben 18 reviewers, elk ongeveer 8 artikelen met behulp van de 'Maastricht list' beoordeeld. Twee andere reviewers beoordeelden alle artikelen en hun oordeel, ofwel kwaliteits score, werd verheven tot 'gouden standaard'. De interbeoordelaars betrouwbaarheid werd bepaald, maar ook de verschillen in beoordeling tussen de 18 reviewers en de 'gouden standaard'. Tevens werd onderzocht of het zin had de artikelen, voordat ze werden beoordeeld, te blinderen voor de auteur van het artikel, het tijdschrift waarin het is gepubliceerd en voor de resultaten. Dit omdat informatie hierover wel eens vertekenend zou kunnen werken op de methodologische beoordeling.

Er werd een hoge mate van interbeoordelaars betrouwbaarheid gevonden en een hoge mate van overeenstemming tussen de 18 individuele reviewers met de gouden standaard. De conclusie was dat 'Maastricht list' een betrouwbaar instrument was om de methodologische kwaliteit van effectonderzoek te meten. Tevens werden geen aanwijzingen gevonden waaruit bleek dat de blindering van de artikelen in deze studie nuttig was.

Hoofdstuk 3. Steeds vaker wordt in systematische reviews de kwaliteit van de geïncludeerde RCT's gemeten. De meeste criteria lijsten zijn niet ontwikkeld volgens algemeen geldende wetenschappelijke criteria, maar op basis van criteria hoe een goede RCT moet worden uitgevoerd. Deze criteria staan vermeld in theorieboeken over de methodologie van RCT's. Door een meetinstrument, zoals een criteria lijst, op een wetenschappelijk verantwoorde manier te ontwikkelen wordt de validiteit (= meet men echt de methodologische kwaliteit?) van zo'n meetinstrument vergroot. Het doel van het onderzoek dat is beschreven in dit hoofdstuk was om door middel van internationale consensus tussen experts een criteria lijst te ontwikkelen waarmee de kwaliteit van RCT's kan worden gemeten. Alle deelnemers die waren uitgenodigd waren experts op het gebied van kwaliteitsmeting bij RCT's. De oorspronkelijke 'item pool' waarmee we dit onderzoek begonnen bestond uit alle items uit al bestaande criteria lijsten. Door middel van de 'Delphi' consensus technique is het aantal items gereduceerd. Elke Delphi ronde bestond uit een vragenlijst en een feedback rapport met de analyse van de voorgaande vragenlijst.

Meer dan 30 experts op dit vakgebied werden uitgenodigd om mee te doen aan dit onderzoek, waarvan er 21 het onderzoek hebben voltooid. De oorspronkelijke item pool bestond uit 206 items en deze is in drie Delphi rondes gereduceerd tot 9 items. Uiteindelijk is consensus bereikt over de definitieve criteria lijst: de 'Delphi list'. De bedoeling is dat deze lijst niet zozeer andere criteria lijsten vervangt, maar naast bestaande criteria lijsten wordt gebruikt.

Hoofdstuk 4. Hier wordt een systematische review beschreven met de vraag: Wat is de effectiviteit van 904 nm lasertherapie bij patiënten met aandoeningen van het bewegingsapparaat? De methodologische kwaliteit werd gemeten met behulp van de 'Maastricht list'. De studies waren geblindeerd voor de auteurs van de publicatie, het tijdschrift van publicatie en de gebruikte referenties.

We vonden 25 studies die het effect van 904 nm lasertherapie vergelijken met een placebo behandeling, een andere controle behandeling of geen behandeling. In totaal voldeden 21 RCT's aan de selectie criteria en zijn opgenomen in de review. De kwaliteit van de studies varieerde van 'mager' tot 'redelijk'. De RCT's werden opgedeeld in subgroepen naar aanleiding van de aandoening die ze bestudeerden. Er werd geen bewijs gevonden voor de effectiviteit

van 904 nm lasertherapie, behalve misschien voor knieproblemen en myofaciale pijnklachten waar een positieve trend kon worden vastgesteld.

Hoofdstuk 5 beschrijft de review uit Hoofdstuk 4 waarin nu de kwaliteit is gemeten met drie verschillende criteria lijsten: de 'Maastricht list', de 'Delphi list', en de 'Jadad list', genoemd naar A.R. Jadad die deze lijst heeft ontwikkeld. De overeenkomst tussen de drie criteria lijsten blijkt redelijk tot goed: studies die met behulp van de ene lijst werden geclassificeerd als van goede kwaliteit werden in veel gevallen met behulp van de andere twee lijsten ook als zodanig geclassificeerd. De grote verschillen tussen de drie lijsten zijn: het aantal items per criteria lijst (variërend van 3 tot 47) en de formulering van de verschillende items. Met name dit laatste had invloed op de rangordening van de studies (van goede kwaliteit tot slechte kwaliteit).

De conclusie was dat de 'Delphi list' de voorkeur had boven de andere twee onderzochte criteria lijsten vanwege de lengte (9 items) en de goede formulering van de items.

Hoofdstuk 6. Deze systematische review probeert antwoord te geven op de vraag: Wat is de effectiviteit van verschillende conservatieve fysiotherapeutische therapieën bij patiënten met acute enkelband laesies. De review is op dezelfde manier uitgevoerd als de review beschreven in Hoofdstuk 4. De belangrijkste uitkomstmaten waren: pijn, zwelling, en periode ziekteverlof.

In totaal voldeden 44 RCT's aan de in- en exclusie criteria voor deze review. De kwaliteit van de studies varieerde van 'mager' tot 'goed'. Twee studies over de effectiviteit van 'pulsed short wave therapy' versus placebo therapy scoorden meer dan de helft van het aantal te behalen punten (= een voldoende). Beide studies vermeldden geen effect van 'pulsed short wave therapy' vergeleken met de placebo. Wel bleek dat tape een positief effect had in vergelijking met ander therapieën, met name in het verkorten van het 'ziekte verlof' (sick-leave) en dat het gebruik van plaster of Paris daarop een negatief effect had.

Hoofdstuk 7. In de bovenbeschreven review (zie Hoofdstuk 6) is de kwaliteit ook gemeten met de 'Delphi list', en de studies kregen zo een algemene kwaliteitsscore (soort rapportcijfer). De grootte van het behandel effect (effect size) van de individuele studies is berekend met het doel de relatie tussen de kwaliteit en de grootte van het behandeleffect te onderzoeken. Hiervoor hebben we subgroep analyses naar de verschillen tussen sommige kwaliteits componenten. De effect sizes van individuele studies werden per subgroep statistisch bij elkaar opgeteld (gepooled). Helaas kon maar van 23 van de 44 oorspronkelijke studies een effect size worden berekend op basis van de gegevens die waren gepresenteerd in de publicaties.

Als de manier van randomisatie onbekend was of als er (waarschijnlijk) geen blindering had plaatsgevonden, dan bleek de gepoolde effect size veel lager dan wanneer er in de studie een goede randomisatie procedure was toegepast, en/of een vorm van blindering had plaatsgevonden. Deze uitkomst kwam niet overeen met onze begin aanname, namelijk dat in studies van slechte kwaliteit (met name met een onbekende randomisatie procedure en waarschijnlijk geen blindering) er een overschatting van het effect wordt vastgesteld.

Hoofdstuk 8. Om meer inzicht te krijgen in de invloed van verschillende kwaliteits componenten van studies op het eindresultaat hebben we een meta-analyse uit 1985 opnieuw uitgevoerd. Deze meta-analyse ging over de vraag: Wat is het effect van thrombolytica op de overleving van patiënten met een acuut myocard infarct (MI). De conclusie van deze meta-analyse was dat als men na een MI aan patiënten thrombolitica voorschrijft hun kans om alsnog te overlijden met 20% afneemt. Deze studie heeft indertijd veel impact gehad, maar de kwaliteit van de studies was niet gemeten. Dit eindresultaat van deze

studie is later bevestigd door hele grote RCT's, en is daarom als gouden standaard gebruikt.

Helaas konden we niet alle publicaties terug vinden, zodat we allereerst van de gevonden studies de resultaten hebben gepoold om te weten of we op hetzelfde resultaat uitkwamen. De kwaliteit van de individuele studies is gemeten met de 'Delphi list'. Vervolgens zijn vijf verschillende manieren bestudeerd waarmee de kwaliteit kon worden meegewogen in de eindconclusie van de review: a) component analysis: de artikelen zijn in subgroepen verdeeld op basis van kwaliteits componenten; b) plot van de individuele effectschattingen en de kwaliteits score; c) de kwaliteitsscore als een inclusie criterium; d) de kwaliteitsscore als wegings factor; e) cumulatieve pooling met de kwaliteitsscore als criterium voor de invoer volgorde.

Er is geen correlatie tussen kwaliteit en effect gevonden. Wanneer studies een heldere beschrijving gaven van een aantal kwaliteits kenmerken resulteerde dat in een redelijk goede schatting van het uiteindelijke effect. De andere studies gaven ófwel een overschatting ófwel een onderschatting van het behandeleffect.

Hoofdstuk 9. Dit hoofdstuk bevat een algemene discussie over de vraag: wat is nu eigenlijk kwaliteit van een RCT, hoe kun je dat meten en is dat meten eigenlijk wel een valide en betrouwbare exercitie. Uit de literatuur blijkt dat pogingen tot het definiëren van het begrip kwaliteit geen eenduidige definitie oplevert. Op dit moment wordt het aantal bestaande criteria lijsten geschat op ongeveer 60, en er komen er vast nog bij. De twee criteria lijsten die volgens de wetenschappelijke spelregels zijn ontwikkeld, zijn in onze onderzoeken gebruikt, namelijk: de 'Delphi list' en de 'Jadad list'. We nemen aan dat deze criteria lijsten, vanwege de manier waarop ze zijn ontwikkeld, een iets betere 'expert' validiteit hebben, dan de andere criteria lijsten, maar meer kun je nauwelijks zeggen over de validiteit van deze meetmethode.

De betrouwbaarheid (inter- en intra beoordelaars betrouwbaarheid) van de in dit proefschrift onderzochte criteria lijsten blijkt 'redelijk' tot 'goed' te zijn. Verder worden een aantal mogelijke valkuilen wat betreft de meting van de methodologische kwaliteit besproken in deze discussie. Eén van de belangrijkste valkuilen is 'reporting bias', dat wil zeggen: is de informatie zoals die is gegeven in de publicatie over de opzet en uitvoering van het RCT wel optimaal en volledig? De meting van de kwaliteit gebeurt met name aan de hand van de rapportage van het onderzoek in de publicatie.

Een bekend paradigma op dit terrein luidt dat RCTs van lage kwaliteit waarschijnlijk een overschatting van het behandel effect geven. Dit paradigma is gebasseerd op de aanname dat onderzoekers onbewust de interventie bevoordelen. Uit eigen onderzoek kunnen we de voorlopige conclusie trekken dat onderzoek van lage kwaliteit het behandel effect zowel kan onderschatten als overschatten. Eventueel vervolg onderzoek zou zich moeten richten op de verschillende componenten van kwaliteit. De relevantie van de diverse items is waarschijnlijk afhankelijk van de context of het onderwerp van de review. Vervolg onderzoek zal vooral moeten plaatsvinden in een serie reviews met specifieke vraagstellingen.

Het meten van de methodologsche kwaliteit zien we als een belangrijk middel om te kunnen differentieren tussen studies van lage en hoge kwaliteit. Doelstelling moet zijn om uiteindelijk te komen tot een valide criteria lijst, waardoor men meer inzicht krijgt in de relatie tussen kwaliteit en gevonden onderzoeksresultaten.

Dankwoord

De bijdrage van de mensen aan wie in dit dankwoord woorden van dank worden gericht was essentieel voor het ontstaan van dit proefschrift en was bepalend voor een belangrijk deel van mijn werk- en leefplezier. In dat licht gezien zou dit proefschrift met het dankwoord moeten beginnen in plaats van eindigen. Gelukkig voor hen is het dankwoord vaak het eerst gelezen hoofdstuk.

Dit dankwoord kan niet beginnen zonder me allereerst te richten op het thuisfront. Ton, Leanne en Nander, het klinkt clichématig, maar dit alles had niet kunnen ontstaan zonder jullie steun. Tijdens allerlei huishoudelijke en organisatorische beslommeringen de *afgelopen jaren, heeft niemand ooit geroepen:* maar waarom moet jij ook zo nodig?!?

Onder het thuisfront in de breedste zin des woords versta ik ook iedereen die heeft geholpen met de opvang van de kinderen. Allereerst de au-pairs Lydia, Marina en Sharon, later de verschillende gastouders: André, Ingrid en Gerry. Jullie steun vormde en vormt nog steeds voor mij de essentiële randvoorwaarde om te kunnen werken zoals ik heb gedaan. Nogmaals, iedereen heel erg bedankt.

Verder wil ik de leden van de projectgroep bedanken voor de prettige samenwerking, goede ideeën en adviezen.

Riekie de Vet, co-promotor, dagelijkse begeleider, en reisgenoot. Je wist me vriendelijk en adequaat op de goede weg te houden en me voor ernstige slordigheden en oppervlakkigheid te behoeden. De reisjes samen met jou ter promotie van (in random volgorde) onszelf, ons werk en onze werkgever, zal ik me altijd met veel plezier blijven herinneren. Jammer dat we als lustrum niet nog één keer samen naar het Cochrane Colloquium in Rome kunnen.

Paul Knipschild als promotor in de startfase. Je plastische voorbeelden, kritische en heldere kijk op de methodologie en je nonconformistische manier van denken gedurende de opzet van het onderzoek heb ik erg gewaardeerd.

Piet van den Brandt, je had de moeilijke taak om als promotor pas halverwege het traject op te treden. Ondanks het feit dat je tegelijkertijd ook begon als hoogleraar aan de capaciteitsgroep Epidemiologie, met alle werkdruk van dien, heb je deskundig en op een prettige manier bijgedragen aan dit eindproduct.

Maarten Boers, als creatief denker heb je aan dit proefschrift en het onderhavige project, naast je vakinhoudelijke kennis een wezenlijke bijdrage geleverd aan de originaliteit.

Fons Kessels, volgens mij en sommige anderen ben je een geniaal denker. Maar je weet hoe dat gaat met genieën: het duurt een tijdje voordat ze worden begrepen. Mijn eigenwijze aard heeft een goed begrip van jouw ideeën soms wat lang in de weg gestaan. Door de adviezen van jou en Jos ben ik in staat geweest de vereiste analyses te kiezen en te doorgronden. Heren, hiervoor mijn hartelijke en oprechte dank!

Last but not least, Rob de Bie. Ik vond het een eer je paranimf te zijn vorig jaar. In je dankwoord noemde je mij één van je eerste studenten van eigen kweek; namelijk het door jou opgezetten studieprogramma voor fysiotherapeuten (KBW, Kort Bewegings Wetenschappen). Na mij zullen binnenkort nog enkele van die studenten promoveren. Ik denk dan ook dat je met recht trots mag zijn

op al die eerste studenten van je kweekje. Dank voor al je hulp en vaderlijke zorgen de afgelopen jaren.

Veel dank gaat ook naar iedereen die heeft meegeholpen de studies te reviewen en heeft meegedaan als deelnemer om de Delphi lijst te ontwikkelen.

Alle (ex-) collega's van de capaciteitsgroep Epidemiologie in Maastricht wil ik hartelijk bedanken voor de bijzonder prettige samenwerking en de gezelligheid. Met heel veel plezier heb ik de afgelopen jaren met iedereen in meer of mindere mate samengewerkt.

Miranda, na jaren samen een kamer te hebben gedeeld en na vele diepe en minder diepe gesprekken verder ben ik blij in jou een vriendin gevonden te hebben.

Raymond, mede-student KBW en later ook collega op de capaciteitsgroep. Onze wegen kruisen elkaar steeds. Op die kruisingen ontstaan steeds zeer plezierige ontmoetingen en gesprekken. Ik hoop dat onze wegen elkaar met enige regelmaat blijven kruisen.

Sandra, als collega fysiotherapeut en onderzoeker was je mijn grote voorbeeld. Ik heb in de loop der jaren veel van je geleerd. Helaas, een Lancet publicatie kan ik je niet nadoen.

Harry en Jos, als onverbeterlijke digibeet heb ik jullie voor menig probleem geplaatst. Heel veel dank voor al het geduld en jullie bereidheid me elke keer weer uit de computer technische brand te helpen.

Ute, Maurice, Erik en alle anderen van de capaciteitsgroep Epidemiologie, nogmaals dank voor de prettige werkomgeving en ik hoop met jullie ook in de toekomst contact te blijven onderhouden.

Nynke en Ilja, vriendinnen ieder uit een verschillende periode in mijn leven en nu de paranimfen. Nynke, ook jou ken ik sinds onze KBW tijd. Je enthousiasme is aanstekelijk, je perfectionisme en je organisatietalent maken voor mij het samenzijn en -werken met jou tot een feest.

Ilja, jou ken ik uit de tijd dat we oefenden voor de opleiding tot manueel therapeut. We delen de liefde voor het vak, maar gingen ieder een eigen kant op. Jij bent dé persoon om mij uit de ivoren wetenschappelijke toren te halen op het moment dat ik neig op te stijgen.

Guus Panken en Pierre Graus. Dit dankwoord zou niet volledig zijn zonder jullie te bedanken voor het feit dat ik in jullie beider praktijken nog een tijd mijn oude vak heb kunnen uitoefenen. Jullie waren plezierige en open 'werkgevers'. Heel jammer dat er niet meer uren in een dag en meer dagen in de week zitten, want dan was ik zeker ook bij jullie blijven werken.

Mijn laatste woord van dank gaan naar Kees en Margriet. Zonder jullie liefde voor elkaar bestond ik niet, zonder jullie liefde voor mij was ik niet geworden wie ik nu ben. Mij keuze om weer te gaan studeren, zal jullie vast een hartverzakking hebben bezorgd. Nogmaals mijn dank voor jullie onvoorwaardelijke steun; het is me veel waard.

Authors and affiliations

Department of Epidemiology,
Maastricht University
Maastricht, The Netherlands:
Robert A. de Bie, PhD, PT
Piet A. van den Brandt, PhD
Arianne P. Verhagen, MSc, MT, PT
Henrica C.W. de Vet, PhD

Department of Clinical Epidemiology,
Vrije Universiteit,
University Hospital, Amsterdam,
The Netherlands:
Maarten Boers, MD, MSc, PhD

Research Unit Patient Care,
Maastricht University Hospital,
Maastricht, The Netherlands:
Alphons G.H. Kessels, MD

Department of Primary Care,
Maastricht University
Maastricht, The Netherlands:
Paul G. Knipschild MD, PhD

Department of Physiotherapy,
Maastricht University Hospital,
Maastricht, The Netherlands:
Anton F. Lenssen, PT

Department of Epidemiology and
Biostatistics,
Institute for Research in Extramural
Medicine, Vrije Universiteit, Amsterdam,
The Netherlands:
Lex M. Bouter PhD

Department of Cardiology,
Maastricht University Hospital,
Maastricht, The Netherlands:
Frank Vermeer, MD, PhD

Department of Cardiology,
Twee Steden Hospital
Tilburg, The Netherlands:
Jos WMG. Widdershoven, MD, PhD

Department of General Surgery,
Maastricht University Hospital,
Maastricht, The Netherlands:
Gauke Kootstra, MD, PhD

VVAA,
Utrecht,The Netherlands:
Frans A.J.M. v/d Wildenberg, MD, PhD

Curriculum Vitae

- Geboren op 28 maart 1959 te Uithoorn.
- VWO-b diploma gehaald in 1978 aan het Fivelcollege te Delfzijl.
- Fysiotherapie opleiding aan de SAFA (Stichting Academie voor Fysiotherapie Amsterdam) te Amsterdam; diploma in 1982.
- Opleiding manuele therapie aan de SOMT (Stichting Opleiding Manuele Therapie) te Eindhoven; diploma in 1987.
- Als fysiotherapeut/manuele therapeut gewerkt in een praktijk voor fysiotherapie in Dordrecht; van 1983 tot 1992.
- Student 'Kort Bewegings Wetenschappen' (KBW) aan de Universiteit Maastricht vanaf 1992; *cum laude* afgestudeerd in 1994.
- Sinds 1995 status Epidemioloog A.
- Als fysiotherapeut/manuele therapeut gewerkt in praktijken voor fysiotherapie in Roermond en Urmond van 1992 tot 1997.
- AiO (Assistent in Opleiding) aan de Capaciteitsgroep Epidemiologie van de Universiteit Maastricht van 1 januari 1995 tot 1 juni 1999.
- Sinds 1 juni 1999 werkzaam als post-doc bij het Instituut Huisartsgeneeskunde aan de Erasmus Universiteit te Rotterdam.

'Water'. Dit staat voor de wilskracht, synomiem voor de geboorte van een project.

'Hout'. Dit staat voor de creatieve energie, synomiem hier voor het proces om te komen tot de definitieve vormgeving van een project.

'Vuur'. Dit staat voor het denken, iets tot bloei laten komen. Synomiem hier voor de uitvoer van een project.

'Aarde'. Dit staat voor bundeling, de harmoniserende energie die de andere energieën bundelt.

'Metaal'. Dit staat voor de ervaring en evaluatie, synomiem hier voor de evaluatie van de uitkomsten van het project én hoe het de volgende keer anders, beter kan.

Deze karakters symboliseren een 'empirische cyclus' van energie die de auteur gedurende het hele promotie proces een heel aantal keren heeft doorlopen.