

A Critical Appraisal of Review Articles on the Effectiveness of Conservative Treatment for Neck Pain

Jan Lucas Hoving, MSc, PT, MT,* Anita R. Gross, MSc, PT,† Diane Gasner, MA, BSc, PT,‡
Theresa Kay, BHSc, PT, MHSc,‡§ Carol Kennedy, BSc, PT,‡§ Maria A. Hondras, DC, MPH, ÷
Ted Haines, MD, PhD,¶ and Lex M. Bouter, PhD*

Study Design. A criteria-based appraisal of review articles on neck pain.

Objectives. To assess the methodologic quality, conclusions, and extent of concordance among reviews on the conservative treatment of neck disorders.

Summary of Background Data. During the past decades there has been an increasing interest in summarizing and analyzing the available evidence on the effectiveness of conservative management of neck pain. Considering the growing number and quality of reviews, consumers may question which reviews to read and believe.

Methods. Computerized bibliographic databases were searched without language restriction. The reviews assessed had been published before January 1998, included neck pain and evaluated conservative therapies, and reported at least one controlled clinical trial. Identification, selection, and quality assessment were performed independently by two investigators.

Results. Of the 108 identified articles, 25 review articles were selected, of which 12 were systematic reviews. The reviews differed in their reporting of study population, interventions, and outcomes. Statistical pooling was performed in two high-quality systematic reviews, whereas in other reviews, the investigators explicitly decided not to pool data. The results of the current study show that the concordance among reviews varied. Regarding manipulation and traction, there is inconclusive evidence among reviews. Concordance regarding the effectiveness of other conservative interventions was absent. Many of the reviews displayed major methodologic flaws.

Conclusions. Consumers should consider reports of reviews both carefully and critically, given the wide variety of review methodology, descriptive information, and final conclusions. There is a paucity of evidence from primary studies on neck pain. Therefore, more research is needed to allow systematic reviews to formulate stronger conclusions. [Key words: conservative management, discordant conclusions, methodologic quality, neck pain, review] *Spine* 2001;26:196–205

Neck pain is a common musculoskeletal symptom. It is estimated that in the general population the point prevalence for neck pain varies between 9.5% and 22%.^{2,11,17} Disorders of the cervical spine include neck pain, with or without radiation to the extremity, or headache and can be severely disabling and costly.¹⁷ Neck pain can result from many causes—for example, trauma, infections or inflammatory conditions, rheumatic diseases, and congenital diseases.⁸ Most often, however, no specific cause can be identified, and the symptoms are labeled nonspecific. Neck pain can originate from disorders in the neck, such as neural tissue, uncovertebral or intervertebral joints, discs, bones, periosteum, muscles, and ligaments. Symptoms of neck pain may often be self-limiting within a few weeks of onset, although the natural course of neck pain remains unclear.²⁵ A less favorable prognosis has been associated with high pain levels and a previous history of neck pain.¹⁰ Investigators in a recent study showed that once nonspecific neck pain becomes chronic, 44% of patients consult their general practitioners annually, with a third of these further referred to a paramedical or medical specialist. The majority of referred patients receive some form of conservative treatment. In most cases, this includes medication, some form of physical therapy, chiropractic care, or other conservative therapies.⁹

During the past decades, there has been an increasing interest in summarizing and analyzing the available evidence on conservative management of neck pain. In keeping up with the overwhelming volume of the medical literature, review articles can be a helpful tool. Consumers of review articles often experience time limitations or limited access to literature, or they may have inadequate skills in basic methodology to assess review articles appropriately. When confronted with a clinical question, the clinician, policy maker, and patient are interested in accessible and comprehensible reviews. Several types of reviews are published: These range from the more traditional narrative reviews to systematic reviews and meta-analyses. In general, narrative reviews are difficult to replicate, because the methods used are frequently unspecified or, if present, are poorly described. Systematic reviews use a more structured and rigorous approach of gathering and combining evidence, including ways to minimize bias. Sometimes attempts are made to statistically combine the results of the primary studies

From the *Faculty of Medicine, Institute for Research in Extramural Medicine, Vrije Universiteit, Amsterdam, The Netherlands; †Hamilton Health Science Corporation and ‡Clinical Epidemiology and Biostatistics, McMaster Campus, Physiotherapy Services, Hamilton, Ontario, Canada; ‡Department of Physical Therapy, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; §Sunnybrook and Women's College Health Sciences Centre, Physiotherapy, Toronto, Ontario, Canada; and the ÷Consortial Center for Chiropractic Research, Davenport, Iowa.

Supported by the College of Physiotherapists of Ontario, Toronto, Ontario; the Chedoke-McMaster Foundation, Hamilton, Ontario, and the University of Saskatchewan, Clinical Teaching and Research Award, Saskatoon, Saskatchewan, Canada.

Acknowledgment date: November 4, 1999.

Acceptance date: May 1, 2000.

Device status category: 1.

Conflict of interest category: 12,13,14,15.

into one single estimate of effect, also referred to as meta-analyses.

Discordance among reviews is not uncommon, and the sources of discordance have been studied by others.³⁵ Methodologic quality of the review article has been reported as a source of discordance regarding the final conclusions of reviews.³⁷ Systematic review methodology itself has been under scrutiny and has been investigated for its validity. A series of excellent methodologic articles have been published in the *British Medical Journal* (1997), *Annals of Internal Medicine* (1986 and 1997), and the *Journal of the American Medical Association* (1993–1998). In line with the evidence-based medicine perspective, the first place that clinicians should look when they need clinical information on treatment effectiveness is to a good review or, if available, an evidence-based treatment guideline.

In the field of musculoskeletal disorders, multiple review articles are available. Most notably in low back pain and manipulation, the number of reviews outnumbers the available clinical trials.³ The growing number of reviews relating to neck pain creates new issues: How can reviews of good quality and applicability be identified? The following questions will be addressed in this study:

- What is the methodologic quality of reviews on conservative treatment of neck pain?
- What are the conclusions of these reviews, and what is the extent of concordance among them?

■ Methods

Article Identification. Computerized bibliographic databases were searched without language restriction, from 1966 (or the earliest year available, depending on the database searched) through January 1998. The citations from the search reported in Gross et al²⁸ for systematic reviews were included. Databases searched were MEDLINE, EMBASE, Manual Alternative and Natural Therapy (MANTIS), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Index to Chiropractic Literature (ICL), and the Cochrane Collaboration Trials Register (CCTR). Key words and Medical Subject Headings (MeSH) were identified on neck-related disorders or syndromes, anatomic terms, conservative treatment terms, and methodologic terms. Two search strategies for reviews by Greenhalgh²⁷ and Hunt and McKibbin³³ were used in these database searches. Additional sources included the personal files of specialists and manual searches of published textbooks. Reference tracking was performed on publications identified and judged relevant by the investigators. Two investigators (TH, CK) independently reviewed the identified titles using broad screening criteria regarding population, intervention, outcome, and design. All studies judged to be potentially relevant or of uncertain relevance were then retrieved by a research assistant. A consensus process was used when the researchers disagreed. To investigate the possibility of selection bias, the authors, journal title, and identifying features—the abstract, results, discussion, and conclusion sections—were masked by differential photocopying of these studies by a trained assistant (leaving only the unmasked text). The investigation of the effect of masking on selection will be reported separately.

Article Selection. Articles chosen to go through the selection process were independently reviewed by a randomly allocated pair of investigators (AG, TK, CK, DG, JH, and MH). For inclusion in the study, the following criteria had to be met: First, the review must have been published before the end of January 1998 in a journal, published proceedings, report, or textbook. Excluded were publications of primary studies and abstracts. Second, the review had to concern humans with neck pain or related disorders of whiplash, headache of cervical origin, or soft tissue disorders of the neck. Reviews that reported solely on systemic diseases or headache without associated neck disorder were excluded. Third, the review had to include conservative management strategies, such as manual therapies, physical medicine methods (*i.e.*, exercise, traction, acupuncture, laser, and electromagnetic therapy), drug therapies, or patient education. Invasive or surgical interventions were excluded. Finally, the review had to report on effectiveness based on an outcome from at least one randomized or clinical controlled trial (RCT and CCT, respectively). For non-English language publications, one investigator and a translator with a medical background conducted study selection in an unblinded manner. All articles were judged “included,” “unsure,” or “excluded.” A consensus process was used in case of disagreement or when the article was judged unsure by either of the investigators. The final judgment was either “included” or “excluded.” In reviews in which the features of the selection criteria (population, intervention, outcome, and design) were not clearly delineated in the article (for example, low back and neck disorder groups were both sampled) consensus was invoked to determine the selection status. A third investigator was available to resolve disagreement.

Data Extraction. As with article selection, data from each review article was extracted independently by a randomly allocated pair of investigators (AG, TK, CK, DG, TH, and JH). A standardized data-extraction form was used that contained questions on population, interventions, methodology, results, and outcome. Potential determinants of discordance were listed *a priori* and incorporated into a data extraction form. The categories of potential determinants of discordance were as follows: population (type of neck disorder, duration of disorder, recurrence rate), intervention (type of conservative treatment, type of control intervention, treatment characteristics), outcome (main outcome reported), and methodology (data pooling performed, assessment of ability to combine studies, statistical analyses). The results were compared, and each pair of investigators had to achieve consensus on each item of the data extraction form.

The reported outcomes or conclusions for the interventions of each review article were classified as “benefit,” “inconclusive,” or “no benefit” by the investigators, based on the reporting of the review article. For the interventions in a review article to be rated as beneficial, a positive effect had to be reported, either overall or in specific subgroups. Similarly, the direction of effect had to be positive. A review was rated inconclusive if no conclusions were drawn, the conclusions were unclear, or there were contrasting conclusions (both beneficial and non-beneficial outcomes reported). In cases in which the intervention was reported to have no effect or a negative effect, the review article was rated as “not beneficial.” Concordance between reviews was categorized as “no” or “yes.” Concordance between reviews with similar interventions was arbitrarily set on at least 75% agreement. If more than 25% of the review

articles disagreed on the effectiveness of a specific intervention, concordance was categorized as “no.”

Methodologic Assessment. A methodologic assessment was performed independently by two investigators (JH, DG) both experienced in performing validity assessments. Methodologic quality was assessed using a modified version of the scale reported in Oxman and Guyatt⁴⁴ and Oxman et al⁴⁵ to assess the scientific quality of research overviews. These criteria are outlined in Appendix A. This checklist was chosen as the primary checklist, because it has demonstrated acceptable validity and reliability.⁴⁴ Before formal scoring, the two investigators scored and evaluated three nonrelated review articles in a pilot study to promote uniform interpretation of the checklist. Study quality was measured using the sum of the scores of items 1 through 9, resulting in a maximum score of 18 points, indicating excellent quality. Each item ranged from a score of 0 points (“no”), 1 point (“partial” or “can’t tell”), to 2 points (“yes”). Consensus between the two investigators had to be reached on each item after all the articles were scored. A third investigator was available to resolve persisting disagreement.

Data Analyses. Data presentation and data analyses were performed on different subsets of initially selected reviews because of multiple publications (duplicate publications and updates). Reviews were labeled as belonging to one review group if they had at least one author in common and dealt with similar populations or interventions. For the methodologic quality assessment, the full set of reviews was used, including all (multiple) publications. This was meant to expose possible differences in reporting within these review groups, based on publications in different journals. In case of multiple publications by the same review group, the best quality review was used for further analyses and reporting in the current study. The rationale for the classification of interventions in this study was based on the descriptions of the interventions provided by the authors of the review articles. Four primary intervention groups were identified: manual therapies, physical medicine methods, drug therapies and education. In the absence of adequate information in the review article, the investigators classified the described interventions in these groups and their respective subcategories (such as mobilization, manipulation or acupuncture).

Agreement between the two investigators for selection and methodologic assessment was calculated using the software program Agree, developed at Maastricht University, (Maastricht, The Netherlands). The linear weighted κ was used as a measure of agreement. Agreement on the methodologic assessment was also performed on each item (a total of nine items) of the methodologic checklist. Agreement statistics are based on the complete set of review articles selected.

■ Results

Article Identification and Selection

One hundred eight potentially relevant review articles were identified. Most of the articles could be retrieved by more than one search method. Of the 108 articles 28% were identified on MEDLINE, 25% on EMBASE, 20% on MANTIS, 2% on ICL, 3% on CINHAL, and 1 on the CCTR. Furthermore, 43% (of 108 articles) were retrieved by manual searches and consulting personal

files of experts. Thirty-eight reviews fulfilled all four criteria for selection. Thirty reviews were published as journal articles (including one in an on-line journal), two as book chapters, four as reports, and two in one PhD thesis.^{1,5-7,12,13,15,18-24,26,28-32,34,38-43,46-55}

Most reviews were excluded, because they did not qualify as a review (70% of all excluded articles), did not report on neck pain or related disorders (34%), or did not include relevant interventions or outcomes (both 6%). Some review articles were excluded for more than one reason. The only non-English articles were evaluated by just one evaluator. One German²² and three Dutch language articles^{7,38,51} were selected.

Many of the publications originated from the same review groups. Two review groups are responsible for five publications each: the reviews by Hurwitz and Coulter et al^{18,19,32,34,48} and Gross et al.^{1,28-31} One of the reviews by Gross et al³⁰ is the only Cochrane Review in the current study. Twenty-five unique reviews are evaluated and discussed, excluding duplicate or previous versions of reviews written by the same review group with lower methodologic scores (13 reviews). The initial overall agreement between the two investigators regarding selection was 83%; the linear weighted κ value was 0.66. Descriptive statistics were used to answer each research question. Because of the small number of reviews on each intervention, the use of statistical analyses was limited. Therefore, the relation between the review article's conclusions and other review characteristics could not be studied.

Assessment of Methodologic Quality

Table 1 outlines the results of the quality assessment for each of the 38 identified eligible review articles. Potential sources of bias during identification, selection, validity assessment, and synthesis are presented. All review articles, including updates and duplicate publications by the same review group, are reported separately (25 reviews and 13 duplicate versions or updates). The articles are ordered by their methodologic scores, according to the checklist. Only the review articles with the highest methodologic scores or most recent publication date (numbered 1-25) were used for subsequent analysis.

The methodologic quality of the review articles was variable; the average quality score was 8.5 points and ranged from 0 points (lowest score) to 18 points (highest score). Most common flaws were a noncomprehensive search (Item 2), no inclusion or exclusion criteria reported (Item 3), bias in the selection of articles (Item 4), and inadequate or absent validity assessment (Items 5 and 6). The scores indicated that the full range of the checklist was used and differentiated between high- and low-quality review articles. Only four review articles fully satisfied Item 4: was biased in the selection of articles avoided. These four publications all belong to the same review group^{1,28-30} that also achieved the highest overall quality score.

Table 1. Methodological Scores of Review Articles on the Effectiveness of Conservative Treatment of Neck Pain

Review Group Number and First Author (Year)	Methodological Items*									Total
	1	2	3	4	5	6	7	8	9	
1. Gross (1998) ²⁹	2	2	2	2	2	2	2	2	2	18
2. Bronfort (1997) ¹²	2	2	2	0	2	2	2	2	2	16
3. Bronfort (1997) ¹³	2	2	2	0	2	2	2	2	2	16
4. Hurwitz (1996) ³⁴	2	2	2	0	2	2	2	2	2	16
5. V.D. Heijden (1995) ⁵²	2	2	2	0	2	2	2	2	2	16
6. Beckerman (1990) ⁷	2	2	2	0	2	2	1	2	2	15
7. Koes (1991) ⁴²	2	1	2	0	2	2	2	1	2	14
8. Magee (1997) ⁴³	2	1	2	1	2	1	2	1	1	13
9. Spitzer (1995) ⁴⁹	1	2	1	1	2	1	2	2	1	13
10. Ter Riet (1989) ⁵¹	2	2	1	0	2	2	1	1	1	12
11. Di Fabio (1992) ²⁰	1	1	2	0	0	1	2	1	1	9
12. Jans (1996) ³⁸	2	1	1	0	1	1	1	1	1	9
13. Florian (1991) ²⁴	1	1	1	0	2	2	0	1	1	9
14. Conlon (1992) ¹⁵	1	1	0	0	2	1	1	1	1	8
15. Patel (1989) ⁴⁶	1	2	2	0	0	0	1	1	1	8
16. Winkel (1996) ⁵⁵	0	0	0	0	0	0	1	1	1	3
17. Jordan (1996) ⁴¹	1	1	1	0	0	0	0	0	0	3
18. Ernst (1993) ²²	0	0	0	0	0	1	0	1	1	3
19. Ellenberg (1994) ²¹	1	1	0	0	0	0	0	0	0	2
20. Gebhart (1994) ²⁶	0	0	0	0	0	0	0	0	1	1
21. Vernon (1995) ⁵⁴	0	0	0	0	0	0	0	0	1	1
22. Richardson (1986) ⁴⁷	0	0	1	0	0	0	0	0	0	1
23. Evans (1992) ²³	0	0	0	0	0	0	0	0	0	0
24. Johnson (1996) ³⁹	0	0	0	0	0	0	0	0	0	0
25. Swezey (1996) ⁵⁰	0	0	0	0	0	0	0	0	0	0
Duplicates or Updates of the Same Review Group Listed Above										
1. Gross (1996) ²⁸	2	2	2	2	2	2	2	2	2	18
Gross (1998) ³⁰	1	2	2	2	2	2	2	2	2	17
Aker (1996) ¹	1	2	1	2	2	2	2	2	2	16
Gross (1996) ³¹	0	0	0	0	0	0	1	2	2	5
4. Coulter (1996) ¹⁹	2	1	1	1	2	2	2	2	1	14
Coulter (1996) ¹⁸	1	1	1	0	1	0	0	1	1	6
Haldeman (1997) ³²	0	0	0	0	0	0	1	0	1	2
Shekelle (1997) ⁴⁸	0	0	0	0	0	0	0	0	0	0
6. Beckerman (1992) ⁶	2	2	2	0	2	2	1	2	1	14
7. Beckerman (1993) ⁵	2	2	1	0	2	2	1	1	1	12
9. Spitzer (1995) ⁴⁹	2	2	1	1	0	0	2	2	1	11
17. Jordan (1996) ⁴⁰	0	0	0	0	0	0	0	0	0	0
21. Vernon (1991) ⁵³	0	0	0	0	0	0	0	0	1	1
% of items fulfilled	53	53	49	16	50	47	50	53	54	
Total of 25 reviews and 13 duplicates or updates.										

* See Appendix A for detailed criteria list.

Overall agreement between the two investigators (JH, DG) regarding quality assessment was 84%, indicating good initial agreement. The overall linear weighted κ was 0.83. The individual linear weighted κ for Items 1–9 on the checklist ranged from 0.66 (Item 9) to 0.94 (Item 2). Disagreement between investigators was primarily a result of reading errors. Both investigators reached full agreement on all items, and thus a third investigator was not consulted.

■ Conclusions of Reviews

Table 2 outlines the main characteristics of the 25 review articles including year of publication, type of neck disorder, and reported conclusions per intervention. The review articles are ordered by their quality score. Each intervention-specific conclusion reported in a review article was classified as “benefit” (+ score), “inconclusive” (? score) or “no benefit” (– score). The absence of a score indicates that the intervention was not studied in the review article at issue.

Review articles frequently reported on more than one intervention, which resulted in multiple conclusions within a review article. Manipulation was studied most frequently (10 reviews), followed by traction therapy (7 reviews), a combination of manipulation-mobilization and other treatments (6 reviews), and electromagnetic therapy and acupuncture (5 reviews each). Other interventions, such as electrical stimulation, infrared light, and spray and stretch were included in the reviews by Gross et al,^{28–30} which covered all interventions listed in this study.

■ Concordance Among Reviews

Table 3 outlines the extent of concordance among reviews and is organized according to intervention (see also Table 2). Concordance was achieved if 75% of the review articles within an intervention category reported similar conclusions. Manipulation was studied in 10 reviews, 4 of which included patients with headache of cervical origin, and most of which reported mixed re-

Table 2. Population and Intervention Characteristics of Selected Review Articles

No. First Author (Year)	No. of Studies	Type of Neck Disorder			Direction of Effect by Intervention													
		Neck Pain	Headache	Whiplash	Manual Therapies				Physical Medicine Methods						DT			
					Man	Mob	Com	Trac	Imm	Ex	Lasr	EMT	Acu	Mag	Oth	Med	Inject	ED
1. Gross (1996, 1998) ²⁸⁻³⁰	23	x	x	x	?	?	+	?		?	?	+	+		?	?		?
2. Bronfort (1997) ¹²	4	x					+											
3. Bronfort (1997) ¹³	6		x				+											
4. Hurwitz (1996) ³⁴	14	x	x		?	+												
5. V.D. Heijden (1995) ⁵²	3	x						?										
6. Beckerman (1990) ⁷	3	x										?						
7. Koes (1991) ⁴²	5	x			?													
8. Magee (1997) ⁴³	5			x				+		?			+			+		
9. Spitzer (1995) ⁴⁹	9			x	?		+	?		?	+		+		?	+	+	+
10. Ter Riet (1989) ⁵²	3	x											?					
11. Di Fabio (1992) ²⁰	1	x			?													
12. Jans (1996) ³⁸	5			x												?		
13. Florian (1991) ²⁴	11	x			?	+		?					+	?	?			
14. Conlon (1992) ¹⁵	3	x			?													
15. Patel (1989) ⁴⁶	3	x	x										+					
16. Winkel (1996) ⁵⁵	1	x						?										
17. Jordan (1996) ⁴¹	4	x								?								
18. Ernst (1993) ²²	6	x			?													
19. Ellenberg (1994) ²¹	1	x						?	?							?		
20. Gebhart (1994) ²⁶	5	x		x					?	?								
21. Vernon (1995) ⁵⁴	2		x		+													
22. Richardson (1986) ⁴⁷	3	x	x										+					
23. Evans (1992) ²³	4			x				?										?
24. Johnson (1996) ³⁹	6			x				?	?				?					?
25. Swezey (1996) ⁵⁰	2	x														?		

The review articles are numbered according to their methodologic score. Scores: the conclusions in review articles were categorized in a "+", "?", or a "-" score, indicating "benefit," "inconclusive," or "not beneficial" effect.

DT—drug treatment; ED—education; Manual therapies—Man = manipulation, Mob = mobilization; Com = combination of mobilization and manipulation with other treatments. Physical medicine methods—Trac = tractions; Imm = immobilization such as a neck collar; Ex = exercise; Lasr = laser; EMT = electro-magnetic therapy; Acu = acupuncture; Mag = magnetic necklace; Other = infrared therapy, spray and stretch, therapy, transcutaneous electrical nerve stimulation (Gross); multimodal therapy/physical therapy (Magee, Jans and Swezey). Drug treatment—Med = medication such as NSAID's and analgesics; Inject = trigger point injections (Spitzer, Evans and Johnson), and intra-articular injection (Johnson). Education = individual or group education (Gross) and education/advice to promote function (Spitzer).

sults. Therefore, there is concordance among reviews regarding the inconclusiveness of evidence for manipulation in treating neck pain. There was no concordance in conclusions about effectiveness of mobilization and a combination of manipulation and mobilization in conjunction with other conservative treatments. It is of interest that all four review articles in the benefit category were in fact higher quality systematic reviews. Traction, in the physical medicine methods category, was reviewed most frequently. All seven reviews determined that the evidence for traction was inconclusive. Similarly, inconclusive evidence was reported concordantly for immobilization and exercise laser therapies. Acupuncture was rated both beneficial and inconclusive. Electromagnetic therapy was the only intervention reported to be consistently beneficial by most review articles. It should be noted, however, that only two small trials were responsible for these results. Drug therapies were the least frequently reported on and were rated as either beneficial or inconclusive. Similar results were found for injection therapy. Finally, mixed results were found for education, reported by two reviews.

■ Review Characteristics

Apart from criteria related to bias (see Table 1), the reviews differ in their reporting of basic descriptive char-

acteristics—that is, study population, intervention, and outcomes. Table 4 describes the number of reviews that report on a specific characteristic. No reviews published before 1986 were selected. More than half the reviews were published in 1995 or later. No correlation was found between the year of publication and the methodologic quality; the Spearman correlation was 0.2 ($P = 0.23$). Until 1998, lower quality reviews continued to be published. In many cases, a target intervention was not defined, or there was an absence of intervention characteristics for each primary study listed in the review article. Frequently, there was no explanation for the selection of specific intervention categories.

A specific outcome was reported in 21 of 25 reviews. The outcome measure used varied: Pain and range of motion were most commonly reported (21 reviews and 14 reviews, respectively). Outcomes such as function (condition-specific and general outcomes: 7 reviews), general health status (7 reviews), and costs (5 reviews) were less frequently reported on. In 4 reviews on manual therapies, the issue of harm was addressed.^{28,34,43,54} Although review articles addressed complications in general, none reported that manipulation should be discouraged for this reason.

Reviewers were hesitant to attempt statistical pooling and, if feasible, the data pooled were from an arbitrary

Table 3. Concordance Among Review Articles About the Efficacy of Conservative Treatments for Neck Pain

Treatment Group	Specific Intervention	No. of Reviews*	Conclusion Review Article			Consensus
			Benefit	Inconclusive	No Benefit	
Manual Therapies	Manipulation	N = 10	2	8	0	Yes
	Mobilization	N = 3	2	1	0	No
	Combination Mob/Man + other treatments	N = 6	4	2	0	No
Physical Medicine Modalities	Traction	N = 7	0	7	0	Yes
	Immobilization	N = 4	0	4	0	Yes
	Exercise	N = 4	1	3	0	Yes
	Laser	N = 2	0	2	0	Yes
	Electromagnetic therapy	N = 5	4	1	0	Yes
	Acupuncture	N = 5	3	2	0	No
	Magnetic necklace	N = 2	0	2	0	Yes
	Infrared light	N = 1	0	1	0	†
	Spray + stretch	N = 1	0	1	0	†
	TENS	N = 1	0	1	0	†
Physical therapy	N = 3	1	2	0	No	
Drug Treatment	Drug therapies	N = 3	1	2	0	No
	Injections	N = 3	1	2	0	No
	Intra-articular injection	N = 1	0	1	0	†
Education	Education + other interventions	N = 2	1	1	0	No

* A few reviews reported on more than one intervention category (for example, manipulation, exercise, and traction); these figures are based on 25 original reviews.

† No comparison possible: only one review for this intervention.

Mob/Man = mobilization or manipulation.

selection of conservative treatment methods available. Illustrative of the heterogeneity of the primary studies were the reviews in which investigators explicitly decided not to pool data, based on the decision that the population, outcomes, interventions, or follow-up were too dissimilar. Nonetheless, statistical pooling was performed in two high-quality systematic reviews on manual therapies and in one low-quality meta-analysis on acupuncture.^{29,34,46}

In one systematic review, data pooling was performed on three RCTs on manipulation for subacute (duration of 3–13 weeks) and chronic neck pain (>13 weeks), reporting an effect size of -0.42 (95% confidence interval [CI] $-0.005, 0.850$).³⁴ The outcomes pooled were subjective pain improvement, 100-mm visual analog scale (VAS) for pain, and improvement of the main symptom. In the other systematic review, pooling of data was used on a combination of mobilization and manipulation that investigated the effect on neck pain of 1–4 weeks' duration (five RCTs) and of 6–8 weeks' duration (three RCTs), reporting effect sizes of -0.6 (95% CI $-0.9, -0.4$) and -0.5 ($-0.8, -0.2$), respectively.²⁸ The VAS for pain was used as the outcome data for pooling.

Regarding the set of clinical trials used for data pooling, only one clinical trial was used in both review arti-

cles. Table 4 shows additional descriptive information about the selected review articles. The number of primary studies per intervention category varied considerably among reviews. In general, there is a paucity of evidence from primary studies especially regarding the effectiveness of conservative treatment for acute neck pain.

■ Discussion

Reviewers from all over the world have been involved in summarizing and analyzing the available evidence on conservative management of neck pain. Recently, this has been supported by the expanding interest of the Cochrane Collaboration. The growing number and the quality of reviews are creating new issues. The question arises of which review consumers should read and believe. In several recent articles investigators have studied discrepancies between reviews in other health care fields and have tried to resolve them.^{4,16,37}

The reviews identified in the current study addressed a wide variety of interventions and neck-related disorders. Unfortunately, current classifications for neck disorders have no clinimetric standards.¹⁴ Many review articles have no basic descriptive information or rationale for classifying disorders, interventions, and outcomes. The insufficiency in information hinders to a large extent the

Table 4. Characteristics of Review Articles

Review Article Characteristics	No. of Reviews (Total n = 25)	Numbers of Review groups*
Population		
Specific neck disorder type described?	n = 15	1, 4, 8–10, 12, 15–17, 19–21, 23–25
Duration complaints specified?	n = 11	2–4, 8–10, 12, 13, 15, 18, 22
Interventions		
Type of intervention described?	n = 21	1–16, 18, 20–22, 24
Frequency or number of treatments specified?	n = 7	1–4, 6, 8, 9
Duration treatment period specified?	n = 10	1–4, 6, 8, 9, 12, 13, 20
Outcomes		
Outcomes specified?	n = 21	1–16, 20–22, 24–25
Type of review		
Systematic review	n = 12	1–10, 13, 14
Nonsystematic review	n = 13	11, 12, 15–25
Meta-analyses		
Pooling performed?	n = 3	1, 4, 15
Decided not to pool	n = 6	2, 3, 5–8, 14
Power		
Discussed or calculated per study?	n = 5	2, 3, 5, 6, 11
Year of publication		
1995–1998	n = 13	1–5, 8, 9, 12, 16, 17, 21, 24, 25
before 1995	n = 12	6, 7, 10, 11, 13–15, 18–20, 22, 23
Country of study		
North American	n = 14	1–4, 8, 9, 11, 13, 14, 19–21, 23, 25
European	n = 11	5–7, 10, 12, 15–18, 22, 24

No. = number of reviews; n = number of reviews in sample fulfilling this item. See Table 1 for review groups.

application of the results of review articles for the consumer, making it more difficult to extrapolate the findings presented in the review article to clinical practice. It is, however, not clear whether certain subgroups of specific disorders or interventions are associated with larger treatment effects than others. Some more extensive high-quality reviews, however, provide valid information, but are also limited by the information provided in publications on primary studies (controlled clinical trials).

The identification of reviews on neck pain through sources other than computerized databases turned out to be very important. Of the 38 review articles selected 50% were not identified through computerized databases; some of these were systematic reviews. The findings in the current study show that few methodologically sound reviews (of RCTs or CCTs) are available using a comprehensive search strategy.^{27,33} The selection criteria used in the review articles (if listed) often seem rea-

sonable but do not have further clarification. Terms such as manipulation and mobilization seem unambiguous, but the exact meaning of manipulative therapy, manual therapy, or chiropractic manipulation remains open to interpretation. In many cases, the only way to find out is to retrieve the primary studies. Unfortunately, the number of primary studies discussed in these review articles is typically low, and using very strict selection criteria would limit the number of available clinical trials even more.

What is the methodologic quality of reviews on conservative treatment of neck disorders? Many of the reviews evaluated in this study display major methodologic flaws. Remarkably, publications from the same review group differ in their reporting of methodologic characteristics, resulting in different methodologic quality scores. Indeed, the methodologic quality of reviews has been shown to be variable, certainly between systematic reviews.⁴ One reason for this could be the different submission requirements of the journals (such as number of words, methodologic requirements) or the target audience for which the article is written. Some findings, however, are encouraging; at least five systematic reviews^{12,13,29,34,52} show adequate measures to minimize bias according to our methodologic assessment. As would be expected, the Cochrane review received the highest quality rating. This is consistent with other research showing that Cochrane reviews appear to have greater methodologic rigor and are more frequently updated than systematic reviews or meta-analyses published in paper-based journals.³⁶

What are the results of these reviews, and what is the extent of concordance among them? The distinction between statements based on personal belief and those based on systematic evidence supported by data is often obscure, even among systematic reviews in the current study. Results of this study show that the concordance among reviews varied across the interventions. In more than half the intervention groups listed, concordance was absent. Regarding manipulation and traction, there was good concordance regarding the inconclusive evidence reported by review articles. However, reviews reporting on manipulation and mobilization in combination with other conservative therapies were not in agreement. Although there was concordance among reviews that electromagnetic therapy is beneficial, the results are based on just two primary studies on electromagnetic therapy. The issue of publication bias could be raised, because none of the reviews concluded that any of the 14 interventions had no benefit. The current authors attribute the reporting of inconclusive evidence to the small number of primary studies available, which is especially true for acute neck pain.²⁹

Some reviews were more generic or discussed just one modality or specific syndrome, such as whiplash. These reviews were included, because some statement was made about the effectiveness of neck-related disorders. Unfortunately, the sources of heterogeneity in these review articles could not be investigated. The inhomoge-

neity and the limited number of review articles per intervention category made it difficult to study which factors were associated with discordant conclusions among reviews. Within the group of conservative treatment, multiple interventions were found. Similarly, the neck disorders in this review covered many different subtypes of syndromes. The large differences in methodologic quality were an additional source of heterogeneity. Although statistical analysis was not feasible, a qualitative evaluation including a quality assessment has been included, so the reader can draw his or her own conclusions. Although there are no strong quantitative results in this review, some methodologic issues were exposed among review articles. Resolving these problems will improve future review articles.

It is striking to find that the quality of some recently published reviews is so variable, despite efforts to improve the methodology of reviews during the past decade. Improved reporting and the use of internationally accepted classifications in combination with high methodologic standards should minimize discordance between review articles or at least make those differences more transparent. Some high-quality reviews for neck disorders are available, with adequate descriptive information. Authors of review articles should consider how and by whom their results will be used. Only good systematic reviews can provide the unbiased information needed for clinical decision making. There is a paucity of evidence from primary studies on neck pain and conservative treatment, and therefore much more research is needed to allow reviews to form strong conclusions. Given the wide variety of review methodology, descriptive information, and final conclusions, consumers should consider reports of reviews both carefully and critically.

■ Key Points

- Review articles should
- avoid bias in the selection of articles;
- explicitly describe the population and symptoms reviewed;
- detail the number of treatments and their specific characteristics;
- use accepted classifications if possible; and
- use systematic techniques in conducting the review.

Acknowledgments

The authors thank Bart W. Koes for help in supervision and planning of the study, Charlie H. Goldsmith for statistical consultation, Caroline Quartly for assisting in selection, Kay Yardley for data extraction, and Angela Eady for lending her expertise in the computerized database literature searches.

References

1. Aker PD, Gross AR, Goldsmith CH, et al. Conservative management of mechanical neck pain: systematic overview and meta-analysis. *BMJ* 1996;313:1291-6.
2. Andersson HI, Ejlertsson G, Leden I, et al. Chronic pain in a geographically defined general population: studies of differences in age, gender, social class, and pain localization. *Clin J Pain* 1993;9:174-82.
3. Assendelft WJ, Koes BW, Knipschild PG, et al. The relationship between methodological quality and conclusions in reviews of spinal manipulation. *JAMA* 1995;274:1942-8.
4. Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. *JAMA* 1998;279:1566-70.
5. Beckerman H, Bouter LM, Van Der Heljden GJMG, et al. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? *Br J Gen Pract* 1993;43:73-7.
6. Beckerman H, de Bie RA, Bouter LM, et al. 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. Beckerman H, de Bie RA, Bouter LM, et al. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials (in Dutch). *Ned Tijdschr Fysiother* 1990;100:306-16.
8. Bogduk N. Neck pain: an update. *Aust Fam Physician* 1988;17:75-80.
9. Borghouts AJ, Janssen HJ, Koes BW, et al. The management of chronic neck pain in general practice: a retrospective study. *Scan J Rehabil* 1999;17:215-20.
10. Borghouts AJ, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. *Pain* 1998;77:1-13.
11. Brattberg G, Thorslund M, Wikman A. The prevalence of pain in a general population: the results of a postal survey in a county of Sweden. *Pain* 1989;37:215-22.
12. Bronfort G. Efficacy of spinal manipulation and mobilisation for low back and neck pain: a systematic review and best evidence synthesis. In: *Efficacy of manual therapies of the spine* (thesis). Amsterdam: Thesis Publishers; 1997.
13. Bronfort G. Efficacy of spinal manipulation for headache and conditions other than spinal related pain: a systematic review and best evidence synthesis. In: *Efficacy of manual therapies of the spine* (thesis). Amsterdam: Thesis Publishers; 1997.
14. Buchbinder R, Goel V, Bombardier C, et al. Classification systems of soft tissue disorders of the neck and upper limb: do they satisfy methodological guidelines? *J Clin Epidemiol* 1996;49:141-9.
15. Conlon J. Chiropractic treatment of neck and back disorders: a review of selected studies. Report of the Canadian Coordinating Office for Health Technology Assessment (CCOHTA). Ottawa, Ontario, Canada: CCOHTA 1992.
16. Cook DJ, Reeve BK, Guyatt GH, et al. Stress ulcer prophylaxis in critically ill patients: resolving discordant meta-analyses. *JAMA* 1996;275:308-14.
17. Cote P, Cassidy JD, Carroll L. The Saskatchewan health and back pain survey: the prevalence of neck pain and related disorders in Saskatchewan adults. *Spine* 1998;23:1689-98.
18. Coulter I. Manipulation and mobilization of the cervical spine: the results of a literature survey and consensus panel. *Journal of Musculoskeletal Pain* 1996;4:113-23.
19. Coulter IDC, Hurwitz EL, Adams AH, et al. The appropriateness of manipulation and mobilization of the cervical spine. Santa Monica, CA: RAND Corporation, Report 96, 1995.
20. Di Fabio RP. Efficacy of manual therapy. *Phys Ther* 1992;72:853-64.
21. Ellenberg MR, Honet JC, Treanor WJ. Cervical radiculopathy. *Arch Phys Med Rehabil* 1994;75:342-52.
22. Ernst E. An analysis of the clinical effectiveness of manipulations as a treatment for cervical pain. *Wien Med Wochenschr* 1993;143:129-30.
23. Evans RW. Some observations on whiplash injuries. *Neurol Clin* 1992;10:975-97.
24. Florian T. Conservative treatment for neck pain: distinguishing useful from useless therapy. *J Back Muscle Rehabil* 1991;1:55-66.
25. Gargan M, Bannister G. The rate of recovery following whiplash injury. *Eur Spine J* 1994;3:162-4.
26. Gebhart JS, Donaldson DH, Brown CW. Soft-tissue injuries of the cervical spine. *Orthop Rev*. 1994;May (Suppl):9-16.
27. Greenhalgh T. How to read a paper: the Medline database. *BMJ* 1997;315:180-3.
28. Gross AR, Aker PD, Goldsmith CH, et al. Conservative management of mechanical neck disorders: a systematic overview and meta-analysis. *Online J Curr Clin Trials* 1996.
29. Gross AR, Aker PD, Goldsmith CH, et al. Conservative management of mechanical neck disorders: physical medicine modalities for mechanical neck disorders (Cochrane review). In: *The Cochrane Library*, Issue 1, 200. Oxford: Updated software.
30. Gross AR, Aker PD, Goldsmith CH, et al. Conservative management of

mechanical neck disorders. Patient education for mechanical neck disorders (Cochrane review). In: The Cochrane Library, Issue 1, 200. Oxford: Update software.

31. Gross AR, Aker PD, Quartly C. Manual therapy in the treatment of neck pain. *Rheum Dis Clin North Am* 1996;22:579–98.
32. Haldeman S, Hooper PD, Phillips RB, et al. Spinal manipulative therapy: Clinical efficacy. In: Frymoyer JW, ed. *The Adult Spine: Principles and Practice*. 2nd ed. Philadelphia: Lippincott–Raven Publishers; 1997:1837–61.
33. Hunt DL, McKibbin KA. Locating and appraising systematic reviews. *Ann Intern Med* 1997;126:532–8.
34. Hurwitz EL, Aker PD, Adams AH, et al. Manipulation and mobilization of the cervical spine: a systematic review of the literature. *Spine* 1996;21:1746–60.
35. Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *Can Med Assoc J* 1997;156:1411–6.
36. Jadad AR, Cook DJ, Jones A, et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *JAMA* 1998;280:278–80.
37. Jadad AR, McQuay HJ. Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology. *J Clin Epidemiol* 1996;49:235–43.
38. Jans MP, Oostendorp RAB. Physical therapy for patients after a whiplash injury: an overview of the literature (in Dutch). *Ned Tijdschr Fysiother* 1996;106:112–8.
39. Johnson G. Hyperextension soft tissue injuries of the cervical spine: a review. *J Accid Emerg Med* 1996;13:3–8.
40. Jordan A, Manniche C. Rehabilitation and spinal pain. *J Neuromuscular Syst* 1996;4:89–93.
41. Jordan A, Ostergaard K. Implementation of neck/shoulder rehabilitation in primary health care clinics. *J Manipulative Physiol Ther* 1996;19:36–40.
42. Koes BW, Assendelft WJ, van der Heijden GJ, et al. Spinal manipulation and mobilisation for back and neck pain: a blinded review. *BMJ* 1991;303:1298–303.
43. Magee DJ, Oborn-Barret E, Turner S, et al. A systematic overview of the current research evidence on selected treatment interventions on soft tissue neck injury following trauma. 1997. Alberta, British Columbia: University of Alberta; 1997.
44. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991;44:1271–8.
45. Oxman AD, Guyatt GH, Singer J, et al. Agreement among reviewers of review articles. *J Clin Epidemiol* 1991;44:91–8.
46. Patel M, Gutzwiller F, Paccaud F, et al. A meta-analysis of acupuncture for chronic pain. *Int J Epidemiol* 1989;18:900–6.
47. Richardson PH, Vincent CA. Acupuncture for the treatment of pain: a review of evaluative research. *Pain* 1986;24:15–40.
48. Shekelle PG, Coulter I. Cervical spine manipulation: summary report of a systematic review of the literature and a multidisciplinary expert panel. *J Spinal Disord* 1997;10:223–8.
49. Spitzer WO, Skovron ML, Salmi LR, et al. Scientific monograph of the Quebec Task Force on Whiplash-Associated Disorders: redefining “whiplash” and its management. *Spine* 1995;20:1S–73S.
50. Swezey RL. Chronic neck pain. *Rheum Dis Clin North Am* 1996;22:411–37.
51. Ter Riet G, Kleijnen J, Knipschild P. The effectiveness of acupuncture [in Dutch]. *Huisarts & Wetenschap* 1989;32:223–38.
52. Van der Heijden GJ, Beurskens AJ, Koes BW, et al. The efficacy of traction for back and neck pain: a systematic, blinded review of randomized clinical trial methods. *Phys Ther* 1995;75:93–104.
53. Vernon H. Spinal manipulation and headaches of cervical origin: A review of literature and presentation of cases. *J Manual Med* 1991;6:73–9.
54. Vernon H. Spinal manipulation and headaches: an update. *Topics in Clinical Chiropractic* 1995;2:34–47.
55. Winkel D, Aufdemkampe G, Matthijs O, et al. Diagnosis and treatment of the spine: Research on effectiveness of traction of the spinal column. Aspen Publishers; 1996:877–922.

■ Appendix A

Criteria Used to Assess the Scientific Quality of Selected Review Articles

Search methods (maximum score = 4)

1. Were the search methods used to find evidence (primary studies) on the primary question(s) stated?

2 points: Yes; includes description of databases searched, search strategy, and years reviewed. Described well enough to duplicate.

1 point: Partially; partial description of methods, but not sufficient to duplicate search

0 points: No; no description of search methods

2. Was the search for evidence reasonably comprehensive?

2 points: Yes; must include at least one computerized database search as well as a search of unpublished or nonindexed literature (for example: manual searches or letters to primary authors)

1 point: Cannot tell; search strategy partially comprehensive (for example: at least one of the strategies in the foregoing section were performed)

0 points: No; search not comprehensive or not described well enough to make a judgment

Selection methods (maximum score = 4)

3. Were the criteria used for deciding which studies to include in the review reported?

2 points: Yes; inclusion and exclusion criteria clearly defined

1 point: Partially; reference to inclusion and exclusion criteria can be found in the paper but are not defined clearly enough to duplicate

0 points: No; no criteria defined

4. Was bias in the selection of articles avoided?

2 points: Yes; key issues influencing selection bias were covered. Two of three of the following bias avoidance strategies were used: two or more assessors independently judged study relevance and selection using predetermined criteria, reviewers were blinded to identifying features of study (*i.e.*, journal title, author(s), funding source), and assessors were blinded to treatment outcome.

1 point: Cannot tell; if only one of the three strategies above were used

0 points: No; selection bias was not avoided or was not discussed

Validity assessment (maximum score = 4)

5. Were the criteria used for assessing the validity for the studies that were reviewed reported?

2 points: Yes; criteria defined explicitly

1 point: Partially; some discussion or reference to criteria but not sufficiently described to duplicate

0 points: No; validity or methodologic quality criteria not used or not described

6. Was the validity for each study cited assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?

2 points: Yes; the criteria used address the major factors influencing bias (for example: population, intervention, outcomes, follow-up)

1 point: Partially; some discussion of methodologic review strategy but not clearly described with predetermined criteria

0 points: No; criteria not used or not described

Synthesis (maximum score = 6)

7. Were the methods used to combine the findings for the relevant studies (to reach a conclusion) reported?

2 points: Yes; qualitative or quantitative methods are acceptable

1 point: Partially; partial description of methods to combine and tabulate; not sufficient to duplicate

0 points: Methods of combining studies not stated or described

8. Were findings of the relevant studies combined appropriately relative to the primary question the review addresses?

2 points: Yes; combining of studies appears acceptable

1 point: Cannot tell; should be marked if in doubt

0 points: No; no attempt was made to combine findings, and no statement was made regarding the inappropriateness of combining findings; should be marked if a summary (general) estimate was given anywhere in the abstract, the discussion, or the summary section of the paper, and the method of deriving

the estimate was not described, even if there is a statement regarding the limitations of combining the findings of the studies reviewed.

9. Were the conclusions made by author(s) supported by the data or analysis reported in the review?

2 points: Yes; data, not merely citations, were reported that support the main conclusions regarding the primary question(s) that the overview addresses

1 point: Partially

0 points: No; conclusions not supported or unclear

Scoring (maximum score = 18). How would you rate the scientific quality of this review? Add up the scores from questions 1–9. Maximum quality score is 18 points.

Adapted from Oxman DA, Guyatt GH, Singer J, et al.⁴⁴

Address reprint requests to

Jan Lucas Hoving
Institute for Research in Extramural Medicine
Vrije Universiteit
Van der Boechorststraat 7
1081 BT Amsterdam, The Netherlands
E-mail: jl.hoving.emgo@med.vu.nl