Manuelle Medizin 2007 · 45:301–308
DOI 10.1007/s00337-007-0548-3
Online publiziert: 28. September 207

M. Jensen Stochkendahl[1] · H.W. Christensen[1] · J. Hartvigsen[1] · W. Vach[2] · M. Haas[3] · L. Hestbaek[4] · E. Adams[5] · G. Bronfort[6] · L. Beyer[7]

[1] Nordic Institute of Chiropractic and Clinical Biomechanics, Part of Clinical Locomotion Science, Odense
[2] The Department of Statistics, University of Southern Denmark, Odense
[3] Center for Outcomes Studies, Western States Chiropractic College, Portland
[4] The Back Research Center, Backcenter Funen; and Part of Clinical Locomotion Science, University of Southern Denmark, Odense
[5] Texas Chiropractic College, Pasadena
[6] Department of Research, Wolfe-Harris Center for Clinical Studies, Northwestern Health Sciences University, Bloomington
[7] Ärztehaus Mitte, Jena

Manuelle Untersuchung der Wirbelsäule

Ein systematischer, kritischer Review zur Reproduzierbarkeit

Originalpublikation

Stochkendahl MJ, Christensen HW, Hartvigsen J et al. (2006) Manual examination of the spine: a systematic critical literature review of reproducibility. J Manipulative Physiol Ther 29: 475–485
Mit freundlicher Genehmigung der National University of Health Sciences.

Metaanalysis

Es wird angenommen, dass biomechanische Dysfunktionen wesentlich zu Rückenschmerzen beitragen, und manuelle Palpation ist ein bei manuelle Medizin Praktizierenden weit verbreitetes Verfahren für die Diagnose solcher Dysfunktionen [1–3]. Anders als viele Kliniker annehmen, hat die Mehrzahl der bisher veröffentlichten Untersuchungen nicht akzeptable Reproduzierbarkeitsniveaus aufgezeigt, und in neueren Reviews wird der Nutzen manueller Untersuchungen der Wirbelsäule überhaupt infrage gestellt [4–7]. Besonders scharf sind die Studiendesigns der Originalveröffentlichungen kritisiert worden, beispielsweise die Aufnahme beschwerdefreier Individuen in Studienkollektive [4, 5], unerfahrene Untersucher [5], Paralleltests [4], unklare Definitionen für positive Befunde und Bewertungsskalen [4, 6] oder dürftige Beschreibungen von Studienergebnissen [4, 5, 7]. Insgesamt bestehe die Notwendigkeit umfassender Qualitätsverbesserungen [4, 7]. Auch die Abhängigkeit des κ-Koeffizienten nach Cohen (der in Studien zur Reproduzierbarkeit am meisten verwendete statistische Koeffizient) von der Prävalenz positiver Befunde und die Zusammensetzung der Studienkollektive waren Gegenstand kritischer Diskussionen [8, 9].

Leider unterliegen diese Reviews selbst ganz wichtigen Einschränkungen. Beispielsweise beschränken sich manche von ihnen auf eine Minderheit von Untersuchungsverfahren, so z. B. nur auf chiropraktische Verfahren [4], auf nur eine Region der Wirbelsäule oder nur auf Palpation in Bewegung [5]. Nur in drei Reviews war ein vorher festgelegtes System verwendet worden, um die Qualität der Studie zu beurteilen [4, 6, 7], und in keinem einzigen Review wurden – wie von van Tulder et al. [11–13] empfohlen – die Anzahl der Studien, ihre methodische Qualität und die Folgerichtigkeit ihrer Resultate berücksichtigt. Und schließlich war in keinem Review der Impact der vorher festgelegten Kriterien auf die Schlussfolgerungen untersucht worden. Insofern ist also der Wert der Palpation als diagnostisches Instrument gegenwärtig noch immer nicht geklärt und auch nicht die Fähigkeiten der Behandler, die manuelle Therapie anwenden, um vertebrale Dysfunktionen mithilfe der Palpation verlässlich zu diagnostizieren.

Daher beschlossen wir, dass ein weiterer systematischer Review gerechtfertigt ist, der die oben ausgeführten Probleme einbezieht. Auch eine Metaanalyse von miteinander vergleichbaren Studien mit ausreichenden methodischen Standards sowie eine Beurteilung der Folgerichtigkeit ihrer Resultate erachteten wir als außerordentlich hilfreich. Gegenstand dieser Veröffentlichung sind also ein systematischer Review und eine kritische Beurteilung des Designs und der statistischen Methoden der Literatur zur Reproduzierbarkeit der Wirbelsäulenpalpation. Dabei verwenden wir Kriterien, die zur Beurteilung von diagnostischen Instrumenten standardisiert sind. Eine Metaanalyse wurde durchgeführt, um die Folgerichtigkeit der Resultate der Studien zu evaluieren, und schließlich wurde der Evidenzgrad der Reproduzierbarkeit bestimmt.

Methoden

Definitionen

Palpation wurde gemäß Bergmann und Petersen definiert [1]. Die Ergebnisse der Primärstudien wurden anhand der Palpationsverfahren analysiert und mit Anmerkungen versehen, wie folgt: Bewegungspalpation („motion palpation", MP), statische Palpation („static palpation", SP, Palpa-

tion im Hinblick auf Ausrichtung und/oder Struktur), Knochenschmerzen („osteous pain", OP, durch die Palpation provozierte Schmerzen), Weichteilgewebeschmerzen („soft tissue pain", STP), Weichteilgewebeveränderungen („soft tissue changes", STC) und allgemeine Einschätzung („global assessment", GA). Letzteres war eingeführt worden, um zu beschreiben, dass 2 oder mehr der oben aufgeführten Verfahren für ein einzelnes Urteil hinsichtlich des Vorliegens mechanischer Dysfunktion angewendet worden waren. Jedes Palpationsverfahren konnte in 5 Positionen erfolgen: im Stehen, im Sitzen, in Pronation, Supination oder Seitenlage und auf unterschiedlichen segmentalen Ebenen. Folglich wird ein Palpationsverfahren, das in einer bestimmten Position auf einer segmentalen Ebene oder mehreren durchgeführt wird, als Test bezeichnet. Manche Arbeiten befassten sich mit einem einzelnen Test oder mehreren, mit nur einem Palpationsverfahren oder mit mehreren.

Reproduzierbarkeit bezieht sich auf die Fähigkeit eines einzelnen Untersuchers, mit dem gleichen diagnostischen Procedere zu 2 verschiedenen Zeitpunkten zum gleichen Ergebnis zu kommen (Intra-Untersucher-Übereinstimmung), und/oder auf die Fähigkeit zweier Untersucher, mit einem gegebenen diagnostischen Procedere bei demselben Patienten zu gleichen Ergebnissen zu kommen (Inter-Untersucher-Übereinstimmung; [14]).

Auswahl der Studien

Infrage kommende Untersuchungen wurden anhand einer umfassenden Recherche in den Datenbanken MANTIS (1966–2005), CINAHL (1982–2005) und MEDLINE (1965–2005) identifiziert. Indexbegriffe für die Abstract- und Volltextsuche waren Reproduzierbarkeit, Reliabilität, Untersucherabhängigkeit in Kombination mit Palpation, Bewegungspalpation, Techniken der körperlichen Untersuchung und Wirbelsäule. Auf die gefilterten Daten wurden nachfolgend folgende Ein- und Ausschlusskriterien angewendet:

Einschlusskriterien waren:

- publizierte Arbeit,
- englischsprachige Arbeit,
- Originalarbeit, prospektive Studie zur Reproduzierbarkeit,
- das verwendete manuelle Untersuchungsverfahren muss folgende Merkmale aufweisen: direkte Palpation relevanter Strukturen mit dem Ziel, mechanische Dysfunktion zu etablieren bzw. auszuschließen entweder direkt durch die Berührung oder indirekt durch Reaktionen des Patienten (Beschwerden, Schmerzen).

Ausschlusskriterien waren:

- Arbeiten, in denen indirekt angegeben ist, ob eine Dysfunktion besteht oder nicht,
- Fallbeschreibungen, Essays, in denen Spekulationen angestellt werden, Editorials,
- Arbeiten, die sich mit der Reproduzierbarkeit des Auffindens von Triggerpunkten befassen.

Exzerpierung der Daten

In Form von Checklisten wurden Daten aus den aufgenommenen Studien von 2 der Autoren unabhängig exzerpiert und aufgezeichnet. Die ausgefüllten Checklisten wurden verglichen, bei unterschiedlicher Auffassung wurde ein Konsens über Diskussionen erreicht. Konnte kein Konsens gefunden werden, so stand ein weiterer Autor (JH) zur Verfügung, um zu vermitteln.

Beurteilung der methodischen Qualität der Primärstudien

Zur Beurteilung der methodischen Qualität von Studien zur Reproduzierbarkeit gibt es keine standardisierte und validierte Methode. Daher entwickelten wir eine 6-Punkte-Skala, beruhend auf anerkannten Anforderungen an klinische Studien zur Reproduzierbarkeit und auf den üblichen Empfehlungen für systematische Bewertungen der Testgenauigkeit [12, 15, 16].

Die operativen Definitionen der Qualitätskriterien waren:

- Reihenfolge der Untersucher randomisiert [1],
- gemischtes Studienkollektiv („case mix"; [1]): symptomatische und nicht symptomatische Individuen. In Studien mit intendiert gemischtem Kollektiv sollte das Studienkollektiv einer tatsächlich vorkommenden klinischen Population entsprechen, d. h. unterschiedlichen Geschlechts und Alters sowie Beschwerden in verschiedenen Bereichen der Wirbelsäule. Zielt eine Studie auf Subpopulationen, z. B. nur symptomatische Patienten, so wird der Punkt ebenfalls vergeben,
- Verblindung aller Untersucher hinsichtlich der Befunde der anderen Untersucher [1],
- Verblindung der Studienteilnehmer hinsichtlich der Befunde (1 Punkt bei effektiver/kompletter Verblindung; 0,5 Punkte, wenn die Teilnehmer zwar nicht verblindet waren, die Ergebnisse aber nicht beeinflussen konnten),
- Kappa (κ, κ_o, κ_g, κ_w) oder ICC (1- bzw. 2-Wege-ICC oder Generalisierbarkeitskoeffizient für das Einzelrating jedes Untersuchers) verwendet für die Analyse [1].

Hohe Qualität einer Studie wurde angenommen, wenn der Score zur methodischen Qualität (angegeben als Prozentsatz des maximal erreichbaren Scores) bei mindestens 50% lag, niedrige Qualität bei einem Score <50%. Der Qualitätsscore spiegelt die Relevanz und die Angemessenheit dreier Dimensionen wider, die sich alle auf die Interpretation der Ergebnisse auswirken können: Studienkollektiv, Studiendesign und statistische Auswertung. Die Einordnung der Studien durch Anwendung des Scores wurde von 2 Autoren unabhängig voneinander vorgenommen; Differenzen wurden über Konsensbildung zwischen den beiden Autoren aufgehoben. Die Qualitätsscores der einzelnen Studien wurden im Rahmen der Bestimmung der Evidenzlage verwendet.

Metaanalyse

Um die Übereinstimmung von Ergebnissen der Studien, die in den systematischen Review aufgenommen waren, zu überprüfen, wurde eine Metaanalyse vorgenommen. Nicht infrage kamen

1. Studien geringer Qualität (>50%),
2. Studien ohne binär klassifizierte Ergebnisse,
3. Studien, in denen überhaupt keine Ergebnisse dokumentiert sind,

4. Studien mit zwar binär klassifizierten Ergebnissen, aber ohne Angabe der κ-Koeffizienten und
5. Studien ohne ausreichende Beschreibung der verwendeten Palpationsverfahren.

Einzelergebnisse [κ-Koeffizienten und Konfidenzintervalle (KI)] der aufgenommenen Studien wurden möglichst direkt den Originalpublikationen entnommen. Waren die KI in der Originalpublikation nicht angegeben, so wurden sie nach Altman [17] berechnet, wenn die dazu erforderlichen Angaben (Prävalenz und Stichprobenumfang) vorhanden waren. Nicht aufeinanderfolgende Ergebnisse für einzelne segmentale Level wurden gesondert in die Analyse aufgenommen. Bei multiplen Resultaten für Reproduzierbarkeit für mehrere Untersucherpaare oder mehrere Wirbelsäulensegmente nacheinander nahmen wir den Durchschnittswert der angegebenen κ-Koeffizienten und berechneten den KI, auch das mittels der Altman-Formel und dem ursprünglichen Stichprobenumfang. Dies ist ein konservativer Ansatz, bei dem dadurch, dass Durchschnittswerte verwendet werden, ein möglicher Zugewinn an Präzision vernachlässigt wird.

Alle Originalergebnisse stellten wir in einem Forest-Plot dar. Wir verzichteten auf eine formale Modellierung und auf eine Heterogenitätsanalyse, denn:

1. nicht in allen Studien lagen Informationen zur Präzision der Einzelergebnisse vor,
2. wir verwendeten zum Teil eine konservative Bewertung für die Einzelstudien, und
3. multiple Ergebnisse innerhalb einer Studie können nicht als unabhängige behandelt werden.

Übergreifende κ-Werte wurden berechnet, indem zunächst die Mittelwerte für κ innerhalb jeder Studie herangezogen wurden und dann ein Durchschnitt gebildet wurde. Konfidenzintervalle für die übergreifenden κ-Werte basieren auf der empirischen Variation der Mittelwerte der κ-Werte und wurden nur dann berechnet, wenn mindestens 4 Studien einen κ-Mittelwert darstellten.

In einer zusätzlichen Analyse wurde die Assoziation zwischen mehreren Ei-

Zusammenfassung · Abstract

Manuelle Medizin 2007 · 45:301–308 DOI 10.1007/s00337-007-0548-3

M. Jensen Stochkendahl · H.W. Christensen · J. Hartvigsen · W. Vach · M. Haas · L. Hestbaek · E. Adams · G. Bronfort · L. Beyer

Manuelle Untersuchung der Wirbelsäule. Ein systematischer, kritischer Review zur Reproduzierbarkeit

Zusammenfassung

Ziel. Immer wieder werden Berichte zur unzureichenden Reproduzierbarkeit der Wirbelsäulenpalpation publiziert, und Autoren kürzlich erschienener Reviews kritisieren die Qualität der Studien. Im vorliegenden Beitrag unterziehen wir die Literatur zur Inter- und Intra-Untersucher-Reproduzierbarkeit der Wirbelsäulenpalpation einer kritischen Analyse unter 2 Aspekten: Übereinstimmung von Untersuchungsergebnissen und Evidenzlage für die Reproduzierbarkeit.

Methoden. Durch Recherche in den elektronischen Datenbanken MEDLINE, MANTIS und CHINAL sowie durch Sichtung bibliographischer Zusammenstellungen wurde relevante, zwischen 1965 und 2005 veröffentlichte Literatur identifiziert. Diese wurde systematisch durchgesehen und einer Metaanalyse unterzogen. Deskriptive Ergebnisse der aufgenommenen Studien wurden von 2 Reviewern unabhängig exzerpiert. Eine 6-Punkte-Skala wurde entworfen, um die methodische Qualität der jeweiligen Originaluntersuchungen zu bestimmen. Die qualitativ hochwertigen Untersuchungen wurden einer Metaanalyse unterzogen im Hinblick auf die Übereinstimmung von Ergebnissen, jeweils getrennt nach Palpation in Bewegung und in Ruhe, im Hinblick auf Knochenschmerzen, Gewebeschmerzen, Weichteilgewebeveränderungen und allgemeine Einschätzung. Die Evidenzlage wurde mithilfe einer standardisierten Methode bestimmt.

Ergebnisse. Der Qualitätsscore der 48 eingeschlossenen Studien lag zwischen 0 und 100%. Es fand sich eine deutliche Evidenz dafür, dass die Inter-Untersucher-Reproduzierbarkeit hinsichtlich Knochen- und Weichteilgewebeschmerzen, die Intra-Untersucher-Reproduzierbarkeit hinsichtlich Weichteilgewebeschmerzen und allgemeiner Einschätzung klinisch akzeptabel ist. Für die übrigen untersuchten Verfahren gibt es widersprüchliche bzw. nur vorläufige Evidenz, andere sind nicht ausreichend reproduzierbar.

Schlüsselwörter

Reproduzierbarkeit von Ergebnissen · Palpation · Literaturreview · Diagnostik · Wirbelsäule · Metaanalyse

Manual examination of the spinal column. A systematic critical review of reproducibility

Abstract

Objective. Poor reproducibility of spinal palpation is continuously being reported and authors of recent reviews have criticized the quality of studies. This article critically analyzes the literature pertaining to the interobserver and intraobserver reproducibility of spinal palpation to investigate the consistency of study results and assess the level of evidence for reproducibility.

Methods. A systematic review and meta-analysis was performed on relevant literature published from 1965 to 2005, identified using the electronic databases MEDLINE, MANTIS, and CINAHL and checking of reference lists. Descriptive data from included articles were extracted independently by 2 reviewers. A 6-point scale was constructed to assess the methodological quality of original studies. A meta-analysis was conducted among the high-quality studies to separately examine the consistency of data on motion palpation, static palpation, osseous pain, soft tissue pain, soft tissue changes and global assessment. A standardized method was used to determine the level of evidence.

Results. The quality score of the 48 studies included in this analysis ranged from 0% to 100%. There was strong evidence to suggest that the interobserver reproducibility of osseous and soft tissue pain, the intraobserver reproducibility of soft tissue pain and global assessments are all clinically acceptable. Other spinal procedures are either not reproducible or the evidence is conflicting or preliminary.

Keywords

Reproducibility of results · Palpation · Literature review · Diagnostic tests · Spine · Meta-analysis

Tab. 1 Eckdaten der für den systematischen Review ausgewählten Beiträge

	Anzahl der Artikel	
Region	**Intra-Untersucher-Studien (n=48)**	**Inter-Untersucher-Studien (n=19)**
Zervikal	16	3
Thorax	5	2
Lumbal	19	8
Sakroiliakalgelenk	8	6
	Anzahl der berücksichtigten Tests	
Palpationstechniken	**Intra-Untersucher-Studien (n=58)**	**Inter-Untersucher-Studien (n=26)**
„Motion palpation" (MP)[a]	28	15
„Static palpation" (STP)[a]	2	0
Knochenschmerz (OP)	6	1
Weichteilgewebeschmerz (SP)	11	5
Weichteilgewebeveränderungen (STC)	3	0
Allgemeine Einschätzung (GA)	7	5

[a] „Static and motion palpation".
„Static palpation" beinhaltet die Palpation des die Wirbelsäule umgebenden Gewebes nach Verspannungen und Schmerzen, ebenso wie das Fühlen von fehlender Gewebecompliance oder Gewebespannung. Bei der „motion palpation" prüft der Untersuchende jedes komplexe Wirbelgelenk, um den Grad möglicher Bewegungseinschränkung zu bestimmen.

genschaften einer Studie (s. unten) und dem Mittelwert ihres κ-Koeffizienten anhand einer Kovarianzanalyse überprüft, auch die Art der Palpation, und zwar getrennt nach Intra-Untersucher- und Inter-Untersucher-Ergebnissen. Die überprüften Studieneigenschaften waren: Jahr der Veröffentlichung, Definition positiver Befunde, segmentale Region, Standardisierung (d. h. Einigung hinsichtlich des Verfahrens, schriftliche Unterweisung, praktische Schulung), Einsatzbedingung, Beruf sowie Erfahrung der Untersucher, Symptomstatus des Studienkollektivs und multiple Tests.

Beurteilung des Evidenzgrades

Die Kriterien zur Bestimmung des Evidenzgrades für die Reproduzierbarkeit von Wirbelsäulenpalpation wurden aus den Leitlinien für akute Schmerzen im Bereich der unteren Wirbelsäule der „Agency for Health Care Policy and Research" übernommen [18]. Diese Methode ist auch verwendet worden, um mittels systematischer Reviews epidemiologischer Studien den Evidenzgrad von Risikofaktoren für Schmerzen im Bereich der unteren Wirbelsäule zu überprüfen [13, 19]. Mit dieser Methode werden alle zur Verfügung stehenden aufgenommenen Studien, die ein Palpationsverfahren beschreiben, Ergebnisse vorlegen und eine valide statistische Methode verwenden, d. h. κ bzw. κ_w oder den Intraklassenkorrelationskoeffizient („intraclass correlation coefficient", ICC; [8]), berücksichtigt.

Dieses System evaluiert die Evidenz unter Berücksichtigung

1. der Anzahl der Studien,
2. ihrer methodische Qualität, angegeben mittels Qualitätsscores, und
3. der Konsistenz der Ergebnisse.

Die Konsistenz wurde durch visuelle Prüfung der Forest-Plots überprüft. Das Rating-System wurde auf jedes Palpationsverfahren angewandt. Um Evidenzgrade zu beschreiben, verwendeten wir die folgenden 5 Kategorien:

- hohe Evidenz: allgemein konsistente Ergebnisse in mehreren (≥2) qualitativ hochwertigen Studien,
- mäßige Evidenz: allgemein konsistente Ergebnisse in einer hochwertigen und einer oder mehr Studien geringer Qualität oder in mehreren (≥2) Studien geringer Qualität,
- vorläufige Evidenz: nur eine Studie vorliegend,
- widersprüchliche Evidenz: nichteinheitliche Ergebnisse in mehreren (≥2) Studien,
- keine Evidenz: es waren keine Studien zu recherchieren.

Für Studien im Bereich der manuellen Medizin ist der Wert für akzeptable Reproduzierbarkeit traditionell – etwas willkürlich – bei κ>0,4 festgesetzt [8, 20–25]. Insofern wurde ein κ-Wert über 0,4 auch in diesem Review als Indikator für klinisch akzeptable Reproduzierbarkeit betrachtet, und für κ_w bzw. ICC wurden willkürlich 0,4 bzw. 0,8 angesetzt.

Analyse der Sensitivität

Um die Robustheit der den Gewichtungen der Evidenz zugrunde liegenden Annahmen zu überprüfen, unterzogen wir die vorher spezifizierten Cutoff-Werte für angemessene methodische Qualität (50%) und minimale klinisch akzeptable Reproduzierbarkeit (κ≥0,4) quantitativen Veränderungen: ±25% für den Qualitätsscore und ±0,1 für die Reproduzierbarkeit.

Ergebnisse

Ergebnisse der Literaturrecherche

Mehr als 900 Veröffentlichungen wurden aufgerufen, 48 zwischen 1980 und 2005 veröffentlichte Originalartikel wurden gemäß den Einschlusskriterien aufgenommen [20–67]. In allen 48 Arbeiten war die Inter-Untersucher-Reproduzierbarkeit angegeben, in 19 von ihnen ebenso die Intra-Untersucher-Reproduzierbarkeit (Appendices A und B, online abrufbar unter http://www.mosby.com/jmpt). Alle vorher definierten Kategorien, Palpation, vertebrale Segmente und Untersuchungsbedingungen, wurden evaluiert. In 25 Arbeiten war ein einzelner Test evaluiert worden, in 22 mehrere Tests (Paralleltests). Wegen unzureichender Beschreibung war bei 1 Studie [63] eine Zuordnung des Palpationsverfahrens nicht möglich. Insgesamt wurden 58 Tests für die Evaluierung der Inter-Untersucher-Reproduzierbarkeit und 26 für die der Intra-Untersucher-Reproduzierbarkeit in Betracht gezogen (**Tab. 1**). Nach der Palpation im Hinblick auf Schmerzreaktionen war die Bewegungspalpation das am meisten untersuchte Verfahren.

Methodische Qualität

Die methodische Qualität der Studien lag zwischen 0 und 100% (s. Appendi-

Tab. 2 Arbeiten, die in die Metaanalyse aufgenommen und zur Beurteilung des Evidenzgrades einzelner Palpationsverfahren herangezogen wurden

	Gesamtzahl der einbezogenen Artikel (n=HQ/NQ)		Artikel von hoher Qualität geeignet für Metaanalyse		Anzahl der einbezogenen Untersuchungsergebnisse in die Metaanalyse		Geeignete evidenzbasierte Artikel (n=HQ/NQ)		Widersprüchliche Resultate		Grad der Evidenz		Durchschnittlicher κ-Wert der Metaanalyse (95%-KI)[a]	
Technik	**Inter (30/18)**	**Intra (8/11)**	**Inter (n=22)**	**Intra (n=8)**	**Inter (n=57)**	**Intra (n=2)**	**Inter (25/6)**	**Intra (11/3)**	**Inter**	**Intra**	**Inter**	**Intra**	**Inter**	**Intra**
OP	8/2	1/1	5	1	5	1	8/1	1/0	Nein	–	Hoch	Vorläufig	0,53	0,91
STP	8/2	2/1	7	2	11	5	8/1	2/0	Nein	Nein	Hoch	Hoch	0,42	0,65
MP	22/14	7/8	16	6	27	15	20/3	6/2	Nein	Nein	Hoch	Hoch	0,17	0,35
STC	5/2	0/0	3	0	3	0	3/1	0	Nein	–	Hoch	Keine	0,03	–
SP	4/1	0/0	3	0	3	0	3/1	0	Ja	–	Widersprüchlich	Keine	–	–
GA	4/1	2/1	4	2	7	5	4/0	2/1	Ja	Nein	Widersprüchlich	Hoch	–	0,44

HQ hohe Qualität, NQ niedrige Qualität.
[a]Kalkuliert, falls mehr als 3 Resultate vorlagen.

ces C und D, online abrufbar unter http://www.mosby.com/jmpt). Von allen Studien waren zwar 30 (63%) von hoher Qualität, doch nur 8 von 19 (42%) Studien zur Intra-Untersucher-Reproduzierbarkeit. Der Anteil qualitativ hochwertiger Studien war höher in den Artikeln zu Hals- und Brustwirbelsäule als in denen zur Lumbalwirbelsäule und zu den Sakroiliakalgelenken (67 vs. 59%). Ein Trend hin zu höherer Qualität zeigte sich in neueren Beiträgen: Für vor 1988 veröffentlichte Artikel lag der durchschnittliche Qualitätsscore bei 27%, für zwischen 1988 und 1995 veröffentlichte bei 48% und für nach 1996 veröffentlichte bei 54%.

Metaanalyse

Von 48 Originalpublikationen, die sich mit der Inter-Untersucher-Reproduzierbarkeit befassten, wurden 22 anhand der vorher festgelegten Kriterien für hochwertig befunden und als geeignet für den Einschluss in die Metaanalyse, 26 wurden nicht eingeschlossen.

Acht Originalpublikationen, die sich mit der Intra-Untersucher-Reproduzierbarkeit befassten, wurden in unsere Metaanalyse eingeschlossen. Elf Originalpublikationen kamen nicht infrage, 10 wegen schlechter Qualität [34, 37, 48, 53, 60, 61, 63–66], eine, weil das Ergebnis nicht binär klassifiziert war [55]. Ergebnisse standen nur für 4 Untersuchungsverfahren (STP, OP, MP und GA) zur Verfügung. Mit Ausnahme der von Meijne et al. [39] schienen die Ergebnisse vergleichbar, sie weisen auf κ-Werte im mittleren bis hohen Bereich hin.

Im Hinblick auf Inter-Untersucher-Reproduzierbarkeit zeigen die meisten Ergebnisse für STP eine Reproduzierbarkeit im mittleren Bereich. Ausgenommen sind die Ergebnisse, zu denen Boline [58] kam, sie lagen im niedrigen Bereich, doch die κ-Schätzung war auch sehr wenig präzise (großes Konfidenzintervall). Für STC legen die Ergebnisse eine niedrige Reproduzierbarkeit nahe, für SP hingegen sind die Ergebnisse inkonsistent/nicht konsistent. Für OP deuten alle Ergebnisse κ-Werte im mittleren bis hohen Bereich an, die meisten Ergebnisse für MP deuten eine niedrige Reproduzierbarkeit an. Für GA waren die κ-Werte nicht konsistent, doch sie hatten weite, überlappende Konfidenzintervalle. Keinen signifikanten Effekt auf die κ-Werte hatten die folgenden Studieneigenschaften: Publikationsjahr, segmentaler Bereich, Standardisierung der Untersuchungen, Beruf sowie Erfahrung des Untersuchers, Symptomstatus des Studienkollektivs oder Anzahl der Tests (Daten nicht dargestellt). Unsere Nachprüfung ergab also, dass die meisten der oben angegebenen Eigenschaften nur wenig Einfluss auf die Ergebnisse der Studien hatten.

Eine bemerkenswerte Ausnahme zeigte sich beim Vergleich der Untersuchungsbedingungen: Palpation im Sitzen war assoziiert mit leicht geringeren κ-Werten und Palpation im Stehen mit deutlich geringeren κ-Werten. Diese Unterschiede waren signifikant (p=0,042) für die Studien zur Inter-Untersucher-Reproduzierbarkeit, aber auch in denen zur Intra-Untersucher-Reproduzierbarkeit zeigte sich eine Tendenz (nicht signifikant) in die gleiche Richtung. Wir möchten auch darauf hinweisen, dass wir in der Analyse zur Intra-Untersucher-Reproduzierbarkeit in den Studien ohne Paralleltests im Vergleich zu Studien mit Paralleltests (κ=0,61; nicht signifikant) eine Tendenz hin zu niedrigen κ-Durchschnittswerten (κ=0,23) beobachteten.

Evidenz für Reproduzierbarkeit

Zur Beurteilung des Evidenzgrades standen uns 31 Arbeiten zur Verfügung, darunter 6 ohne Ergebnismitteilung in binärer Form [20, 21, 25, 40, 47, 49]. Die in den 6 Studien in Form von gewichtetem κ (κ_w) bzw. ICC angegebenen Ergebnisse waren nicht direkt vergleichbar mit den Ergebnissen der Studien, die Ergebnisse als κ-Werte angegeben hatten, doch zeigten alle 6 ähnliche Trends: niedrige Inter-Untersucher-Übereinstimmung für MP und höhere für die Evaluierung von Schmerzen. Wir schlossen unter ähnlichen Aspekten 5 Studien niedriger methodischer Qualität ein, die ähnliche Trends aufwiesen [33, 36, 37, 56, 57].

Nahmen wir alle 31 Arbeiten zusammen, so zeigte sich eine hohe Evidenz für

klinisch akzeptable Intra-Untersucher-Reproduzierbarkeit ($\kappa \geq 0{,}4$] für STP und GA (**Tab. 2**). Nach den vorher definierten Kriterien für die Beurteilung der Evidenzlage fanden wir hohe Evidenz für die Inter-Untersucher-Reproduzierbarkeit von OP und STP, außerdem Evidenz dafür, dass die Intra-Untersucher-Reproduzierbarkeit von MP sowie die Inter-Untersucher-Reproduzierbarkeit von MP und STC klinisch nicht akzeptabel ist. Widersprüchliche Evidenz zeigte sich für die Inter-Untersucher-Reproduzierbarkeit von SP und GA. Vorläufige Evidenz für klinisch akzeptable Reproduzierbarkeit fand sich für die Intra-Untersucher-Reproduzierbarkeit von OP, keine Evidenz für die Intra-Untersucher-Reproduzierbarkeit von SP und STC.

Analyse der Sensitivität

In die Metaanalyse wurden nur qualitativ hochwertige Untersuchungen eingeschlossen. Wären Untersuchungen unzureichender Qualität mit binär klassifizierten Ergebnissen und κ-Werten oder Untersuchungen hoher Qualität mit κ_w oder ICC eingeschlossen worden, wären die Resultate davon nicht beeinflusst worden (Daten nicht dargestellt).

Eine Erhöhung des Cutoff-Wertes für adäquate methodische Qualität von 50 auf 75% bzw. jegliche Verringerung des Cutoff-Wertes hatte keinen Einfluss auf das Gewicht der Evidenz oder zusammenfassende Schlussfolgerungen mit Ausnahme der bezüglich Intra-Untersucher-Reproduzierbarkeit von MP und GA: Hier würde eine Erhöhung auf 75% zu widersprüchlicher Evidenz für Intra-Untersucher-Reproduzierbarkeit von MP und zu mäßiger Evidenz für klinisch akzeptable Intra-Untersucher-Reproduzierbarkeit von GA führen. Eine Erhöhung des Cutoff-Wertes für die klinische Akzeptanz hat einen offensichtlichen Einfluss, wobei die Ergebnisse für Schmerzpalpation wegen der insgesamt hohen κ-Werte die stabilsten sind.

Diskussion

Zusammenfassung der Ergebnisse

Bei der Überprüfung von Studien zur Reproduzierbarkeit von manueller Palpation der gesamten Wirbelsäule einschließlich der Sakroiliakalgelenke fanden wir hohe Evidenz für klinisch akzeptable Intra- wie Inter-Untersucher-Reproduzierbarkeit für die Palpation im Hinblick auf knöcherne und Weichteilveränderungen sowie für klinisch akzeptable Intra-Untersucher-Reproduzierbarkeit von GA. Hohe Evidenz fanden wird für klinisch nicht akzeptable Intra-Untersucher- wie Inter-Untersucher-Reproduzierbarkeit von MP und STC. Die Intra-Untersucher-Reproduzierbarkeit war regelhaft höher als die Inter-Untersucher-Reproduzierbarkeit, und die Reproduzierbarkeit von Palpation im Hinblick auf Schmerzreaktionen war konsistent höher als die von Bewegungspalpation.

In der aktuellsten und umfassendsten Reviewarbeit zum Thema Reproduzierbarkeit der Palpation der Wirbelsäule [7] wurden andere allgemeine Review- und Einschlusskriterien angewendet, insofern wurden nur 27 von 44 Primärstudien eingeschlossen, und von 19 hochwertigen Studien wurden nur 9 in diesem Review evaluiert. Wir schlossen etliche aktuellere Veröffentlichungen sowie Arbeiten zu Sakroiliakalgelenken und GA ein, außerdem evaluierten wir Einzelergebnisse aus mehrteiligen Testregimes. Unsere Schlussfolgerungen beruhen auf einer bisher noch nicht verwendeten Methode: vorher festgelegte Kriterien und eine Konsistenzprüfung qualitativ hochwertiger Studien. Seffinger et al. [7] hingegen zogen ihre Schlussfolgerungen aus Studien mit hoher wie minderer Qualität und nahmen keine Konsistenzprüfung vor. Die Autoren schlossen, dass Palpationstests auf Schmerzreaktionen hin die verlässlichsten sind und dass paravertebrale Weichgewebspalpationen diagnostisch nicht verlässlich sind. Den 12 qualitativ hochwertigsten Studien zufolge war die Intra-Untersucher-Reproduzierbarkeit von Schmerzprovokation, Bewegungspalpation und Lokalisierung anatomisch-topographischer Orientierungspunkte reliabel, aber nicht immer die Inter-Untersucher-Reproduzierbarkeit unter ähnlichen Bedingungen. Die Reliabilität wurde durch die Disziplin oder die Erfahrung der Untersucher, Konsens hinsichtlich der verwendeten Verfahren, Schulung oder durch Untersuchung von symptomatischen Individuen nicht verbessert. Dies stimmt überein mit unseren Ergebnissen. Wir folgern überdies, dass Palpation im Hinblick auf Schmerz reproduzierbar ist sowohl von demselben Untersucher als auch von mehreren, MP hingegen möglicherweise von demselben Untersucher.

Methodische und klinische Überlegungen

Das experimentelle Design von Studien zur Reproduzierbarkeit ist schon in früheren Reviews kritisiert worden [4–7, 68–71]. Wir fanden, dass 26 von 48 Artikel von minderer methodischer Qualität waren, nicht valide statistische Methoden verwendet hatten oder Palpationsverfahren bzw. Ergebnisse nur unzureichend dargestellt hatten.

Die Vergleichbarkeit von Studien, die in ein Review aufgenommen werden, ist die wichtige Voraussetzung für valide Generalisierungen. Wir stellten zwar Vergleichbarkeit hinsichtlich der Palpationsverfahren sicher, doch bezüglich anderer Faktoren, z. B. Definition positiver Befunde, segmentale Region, Standardisierung, Beruf sowie Erfahrung der Untersucher, Symptomstatus des Studienkollektivs und Paralleltests, waren die Studien verhältnismäßig heterogen. Unsere Untersuchung zeigte dennoch, dass die meisten Faktoren wenig Einfluss auf die Ergebnisse hatten, eine Ausnahme bildeten die Untersuchungsbedingungen. Besonders die Palpation im Stehen war assoziiert mit sehr niedrigen κ-Werten. In allen Studien in unserem Review wurde Palpation am stehenden Patienten nur im Spine-Test (Gillet-Test) auf biomechanische SI-Dysfunktion angewendet, wir schlossen nur 2 Arbeiten zu dieser Dysfunktion ein [39, 59]. Beide trugen jedoch zur Evaluierung von Inter-Untersucher- und Intra-Untersucher-Übereinstimmung der MP bei. Lassen wir diese beiden Arbeiten außer Acht, so erhöht sich der durchschnittliche κ-Wert für die Inter-Untersucher-Übereinstimmung auf 0,19 (0,13–0,26) und der für Intra-Untersucher-Übereinstimmung auf 0,44 (0,14–0,73), damit kann die Intra-Untersucher-Übereinstimmung für MP als akzeptabel angesehen werden.

Möglicherweise spiegelt die schlechte Reproduzierbarkeit von MP eher das Design der Reproduzierbarkeitsstudien wider als die Qualität des Untersuchungsverfahrens [29, 30, 72]. Bessere Reproduzierbarkeit könnte erreicht werden, wenn positive Befunde in benachbarten Wirbelsäulensegmenten bei der Bewertung der Übereinstimmung berücksichtigt würden [29]. Dafür müsste jedoch ein neuer diagnostischer Test definiert werden, der dann auch jenseits der Erhöhung der κ-Werte einer klinischen Rationale für seine Bedeutung bedürfte [8]. Darüber hinaus scheinen Paralleltests (Untersuchungsregimes) die klinische Entscheidungsfindung zu fördern und so die Reproduzierbarkeit zu erhöhen [30, 42], dies ist eine Tendenz, die wir auch beobachteten. Auch die akzeptable Intra-Untersucher-Reproduzierbarkeit für GA entspricht dieser Beobachtung. Bei der Evaluierung einer Kombination von Tests stehen jedoch nur Informationen zu Reproduzierbarkeit des einzelnen Tests als Teil genau dieser Kombination zur Verfügung [14, 73]. Wir müssen uns außerdem klar darüber sein, dass Schlussfolgerungen, die einen Test betreffen, aber in einer Studie mit mehreren Tests entstanden sind, möglicherweise nur valide sind, wenn dieser Test auch wieder als Teil genau dieser Kombination angewendet wird. Von einer klinischen Perspektive aus betrachtet, weist die höhere Reproduzierbarkeit bei Paralleltests darauf hin, dass Kliniker ihre Diagnose nicht auf einem einzelnen klinischen Befund basieren lassen sollten, sondern vielmehr eine ganze Reihe von Tests durchführen sollten. Es wäre allerdings verfrüht, jetzt schon klinische Empfehlungen für die Anwendung von Palpation zu geben, denn viele Aspekte, so beispielsweise die Validität, müssen noch erforscht werden.

Die Reproduzierbarkeit von Palpation auf Schmerzreaktionen hin ist konsistent höher als die von Palpation hinsichtlich Bewegung, und ihre Reproduzierbarkeit ist innerhalb eines Untersuchers wesentlich höher, als wenn mehrere Untersucher das Verfahren anwenden. Dennoch gehört sowohl zu Studien mit Fragestellungen zu Palpation und Schmerz als auch zu Intra-Untersucher-Studien ganz allgemein das Problem der Verblindung der Untersucher. Bei Intra-Untersucher-Studien kann die Verblindung durch bewusste wie unbewusste Signale ganz unmöglich werden, die Unabhängigkeit kann nicht mehr gewährleistet werden. In Studien mit Fragestellungen zu Palpation und Schmerz ist die Verblindung der Subjekte nicht möglich. Das Risiko, die Reproduzierbarkeit zu überschätzen, ist bei beiden Ausgangssituationen gegeben. Es sollte auch bedacht werden, dass die Intra-Untersucher-Reproduzierbarkeit definitionsbedingt etwas höher ist als die Inter-Untersucher-Reproduzierbarkeit (abhängig vom Ausmaß der Interaktion zwischen Untersucher und Untersuchtem; [74]). Beim Design von Untersuchungen zur Reproduzierbarkeit ergibt sich ein Dilemma zwischen hoher interner Validität und klinischer Anwendbarkeit. In Trainingsstudien beispielsweise wird eine maximale (ideale) Reproduzierbarkeit in Kontrast gesetzt mit der in der Praxis tatsächlich zu erzielenden Reproduzierbarkeit. Um die interne Validität zu erhöhen, sollten strenge Testbedingungen festgelegt werden, die auch Verblindung, Randomisierung, Standardisierung, Schulung und Paralleltests berücksichtigen. Doch die strikte Durchsetzung der Testbedingungen führt oft zu Abweichungen von der Situation in der Klinik, was die externe Validität verringert. In der klinischen Situation stehen die manuelle Medizin Praktizierenden am ehesten einem Mix aus symptomatischen und beschwerdefreien Patienten gegenüber. Das Studienkollektiv sollte also sowohl aus symptomatischen als auch nichtsymptomatischen Individuen bestehen, sodass die Reproduzierbarkeit der diagnostischen Methode sich auf die Charakteristika des Studienkollektivs bezieht [14]. Schließlich ist zu bedenken, dass diagnostische Verfahren nicht unbedingt die klinische Entität evaluieren, die sie evaluieren sollen, auch wenn sie in der täglichen Routine verwendet werden. Deshalb ist es wichtig, diagnostische Verfahren auch inhaltlich zu diskutieren [14, 75].

Statistische Überlegungen

Der κ-Koeffizient nach Cohen ist allgemein anerkannt als statistische Methode der Wahl um auszudrücken, inwieweit Beobachtungen zu einem binär klassifizierten Merkmal übereinstimmen [8]. Es ist allerdings nicht unproblematisch, κ als einziges Maß für Untersucher-Übereinstimmung zu verwenden, denn bei der Zusammenfassung einer Vierfeldertafel in eine Zahl gehen Informationen verloren. Wir wissen also nicht, ob ein moderater κ-Wert in einer Studie zur Reproduzierbarkeit Folge einer Differenz in Prävalenzschätzungen mehrerer Untersucher ist oder Folge einer mangelnden Übereinstimmung trotz ähnlicher Prävalenz.

Es ist kritisiert worden, dass der κ-Wert von der Prävalenz positiver Befunde abhängig ist. Das limitiert seine Relevanz in Metaanalysen, in denen ja in der Regel Primärstudien mit schwankender Prävalenz miteinander verglichen werden. Doch die Zusammensetzung des Studienkollektivs kann κ stärker beeinflussen als die Prävalenz positiver Befunde [9]. Um Teil unserer Metaanalyse zu werden, mussten die Primärstudien sowohl ein binär klassifiziertes Ergebnis als auch einen κ-Wert haben. Aber binäre Outcomes können je nach der Definition für positive Befunde variieren, d. h., die Prävalenz ist direkt abhängig von der Definition für positive Befunde. Um ein Beispiel zu geben: Wird der Untersucher aufgefordert, jedes hypomobile Segment einer vertebralen Region zu ermitteln, so kann die Prävalenz je nach Studienkollektiv zwischen 0 und 100% liegen. Soll der Untersucher das am deutlichsten hypomobile Segment identifizieren, wird die Gesamtprävalenz positiver Befunde 100% betragen, für jedes einzelne untersuchte Segment kann die Prävalenz zwischen 0 und 100% liegen. Wir fanden jedoch keine Assoziation zwischen der Prävalenz positiver Befunde und κ-Werten. Das unterstützt Vachs Hypothese [9], dass die Zusammensetzung des Studienkollektivs wahrscheinlich von größerer Bedeutung als die Prävalenz positiver Befunde ist.

Verschiedene Begriffe und Konzepte sind zur Evaluierung der Stärke der Reproduzierbarkeit verwendet worden, aber für die Interpretation guter Übereinstimmung gibt es keine klar umrissenen Richtlinien [8, 76]. Darüber hinaus ist bisher wenig zur minimalen klinische Reproduzierbarkeit geforscht worden, und vielleicht ist es sogar noch wichtiger, die Indizes für quantitative Reproduzierbarkeit im

Hinblick auf ihre klinische Anwendung zu evaluieren, als die Stärke der Übereinstimmung zu qualifizieren [8].

Beschränkungen unseres Reviews

Mehrere Methodologien sind für systematische Reviews von Studien zu therapeutischer Wirksamkeit empfohlen worden [12], aber zur Bewertung der Qualität von Studien zur Reproduzierbarkeit gibt es kaum Konsens. Wir haben uns entschieden, die Evidenzstärke mit einer Methode der „best-evidence synthesis" zu evaluieren. Das ist einer der wichtigsten Unterschiede dieses Reviews zu den bisher veröffentlichten Reviews zum gleichen Thema. Der konzeptuelle Ansatz der „best-evidence synthesis" kann studienübergreifende Heterogenitäten (Untersuchungsverfahren, Einschlusskriterien, Studiendesign und Ergebnispräsentation/Präsentation der Ergebnisse) maskieren. Bei den Studien, die wir eingeschlossen haben, war eine erhebliche Heterogenität zu verzeichnen. Trotzdem hat die Metaanalyse insgesamt sehr konsistente Befunde ergeben und nur moderate Auswirkungen von spezifischen Designeigenschaften auf die Ergebnisse der jeweiligen Studie.

Studien ohne binär klassifizierte Ergebnismitteilungen von der Metaanalyse auszuschließen stellt einen weiteren wesentlichen Unterschied zwischen unserem Review und früher veröffentlichten dar. Voraussetzungen für die Vergleichbarkeit von Studien zur Reproduzierbarkeit sind gleicher Ergebnistyp und gleiche statistische Methode. Deswegen mussten wir 5 methodisch hochwertige Studien von der Metaanalyse ausschließen. Die Ergebnisse dieser Studien sind zwar nicht direkt vergleichbar mit denen der aufgenommenen, doch zeigen alle 5 Studien ähnliche Trends: geringe Intra-Untersucher-Übereinstimmung für MP und höhere Inter-Untersucher-Übereinstimmung bei der Evaluierung von Schmerzen. In die Beurteilung des Evidenzgrades wurden die Ergebnisse dieser Studien aufgenommen. Für 3 Kategorien gibt es wegen der limitierten Anzahl von Studien nur eine vorläufige bzw. keine Aussage zur Evidenzlage. Dafür war die Stärke der Schlussfolgerungen zu Palpation für die Untersuchung von Schmerzen und Bewegung umso überwältigender. Doch für manche Kategorien beruhten die Resultate auf einer relativ kleinen Anzahl von Primärstudien, sodass die Schlussfolgerungen sehr anfällig sind – einige wenige qualitativ hochwertige Studien mit divergierenden Resultaten würden ausreichen. In allen qualitativ hochwertigen Studien mit binärer Klassifikation wurde ein κ-Wert angegeben, er musste also nicht mittels einer Vierfeldertafel errechnet werden. Wir machten nicht den Versuch, ursprüngliche Ergebnisse oder Materialien von den Autoren der Primärstudien zu erhalten.

Auch wenn wir nichts unversucht gelassen haben, alle veröffentlichten Reproduzierbarkeitsstudien zu finden, ist ein Selektionsbias nicht auszuschließen, weil wir uns auf englischsprachige Artikel beschränkt haben. Ein Publikationsbias mag zu einer Überschätzung der Reproduzierbarkeit von Untersuchungen geführt haben, denn Arbeiten mit positiven Schlussfolgerungen werden eher publiziert [77, 78]. Darüber hinaus ist auch das Reviewer-Bias eine mögliche Beschränkung dieses Reviews: Bei der Bewertung der methodischen Qualität der einzelnen Studien konnte keine Verblindung bezüglich der Autoren bzw. der Studienergebnisse erfolgen, denn die Literatur war uns zu gut bekannt.

Auch wenn sie unseren Kriterien zufolge hinsichtlich der Qualität akzeptabel waren, so wiesen viele Arbeiten trotzdem methodische Mängel auf bzw. gaben – bestenfalls – keine hinreichende Beschreibung ihrer Methoden. Dennoch ist die Reproduzierbarkeit manueller Palpation der Wirbelsäule sehr gründlich erforscht worden, und mehr als 40 Primärstudien sind in diesem Review bewertet worden. Um Aufschlüsse zur klinischen Relevanz der Palpation zu geben, muss ihre Validität überprüft werden, und innovative Forschung muss die damit einhergehenden Probleme lösen, einen Goldstandard für Bewegungsuntersuchungen auszuwählen. Künftige Forschung sollte sich auch der Frage zuwenden, welchen Stellenwert Palpation für die Gesamteinschätzung von Patienten mit Wirbelsäulenbeschwerden hat und welche Relevanz der Palpation als Teil der vollständigen klinisch-körperlichen Diagnostik aller Patienten zukommt.

Fazit für die Praxis

Die Palpation im Hinblick auf Schmerz ist auf einem akzeptablen Niveau reproduzierbar, und zwar sowohl vom gleichen Untersucher als auch von mehreren. Palpation für GA ist nur vom gleichen Untersucher reproduzierbar, nicht von mehreren. Der Evidenzgrad für diese Schlussfolgerungen ist hoch. Die Reproduzierbarkeit von MP, STC und SP ist klinisch nicht akzeptabel. Der Evidenzgrad für die Inter-Untersucher-Reproduzierbarkeit von MP und STC ist hoch, für die von SP und für die Intra-Untersucher-Reproduzierbarkeit von STC hingegen gibt es keine oder nur widersprüchliche Evidenz. Die Ergebnisse sind insgesamt robust im Hinblick auf die vorher definierten Schwellenwerte für akzeptable Qualität. Sie sind allerdings empfindlich gegenüber Veränderungen des vorgegebenen Niveaus für klinisch akzeptable Reproduzierbarkeit und gegenüber der Anzahl der eingeschlossenen Primärstudien.