

Der Test der Tests: Wie sieht die Evidenz aus?

TESTS AUF DEM PRÜFSTAND: Das SIG ist eine mögliche Schmerzquelle bei Patienten mit unterem Rückenschmerz. Die Prävalenz liegt bei 10–53 %. Für sinnvolle Therapieentscheidungen müssen Therapeuten herausfinden, ob das SIG die Schmerzquelle ist. Am unteren Rücken ist das exakte Lokalisieren der Schmerzquelle jedoch schwierig, auch für erfahrene Therapeuten. Wer sich auskennt, kann mit über 50 Tests „experimentieren“. Arne Vielitz, der Autor dieses Artikels, wollte wissen, welche SIG-Tests für Therapeuten wirklich nützlich sind. Welche Tests können das Gelenk zuverlässig als Schmerzquelle identifizieren? Und welcher Weg führt zum Ziel?

Das SIG verursacht häufig Rückenschmerzen

Die Wahrscheinlichkeit im Laufe des Lebens Rückenschmerzen zu entwickeln, liegt bei ca. 70 % [27]. Das SIG verursacht zweifellos Rückenschmerzen mit oder ohne Ausstrahlungen in die Beine. Die Prävalenz für eine vom SIG ausgehende Symptomatik liegt zwischen 10 bis 53 % [9, 12, 17, 26]. Die Daten zur Prävalenz sind nach Meinung von Cohen [3] und auch Huijbregts [12] jedoch sehr speziell und rar.

Auf einer gemeinsamen Pressekonferenz 2008 stellten die Deutsche Gesellschaft für Orthopädie und Orthopädische Chirurgie (DGOOC) und die Deutsche Gesellschaft für Unfallchirurgie (DGU) eine Liste mit schmerzhaften Körperregionen vor. Rückenschmerzen stehen auf Platz 1: mit einer Häufigkeit von 47,8 %. Laut der Experten betragen die direkten medizinischen Kosten für die Behandlung von Rückenbeschwerden jährlich 10 Milliarden Euro [5]. Diagnostizieren Therapeuten möglichst genau, können sie die Patienten angemessen behandeln.

SIG-Block versus physiotherapeutische Untersuchung

Ein Goldstandard als Referenzstandard für die Diagnose einer SIG-Problematik existiert nicht. Deshalb empfehlen Kliniker zur Diagnostik einer SIG-Problematik zweimal Lokalanästhetika in das Gelenk zu injizieren (SIG-Block bzw. Block) [3, 9, 12]. Die Reliabilität dieser Methode ist jedoch nicht sicher bewiesen [1–3, 8, 9, 30]. Das Verfahren ist zudem schwierig und teuer [30], im Gegensatz zur physiotherapeutischen Untersuchung.

Viele Tests und viele Meinungen zur Evidenz

Physiotherapeuten können zwischen sehr vielen SIG-Tests wählen. Winkel [28] fand in seinem Review 54 verschiedene Tests der bekannten Kategorien:

- Positions-Palpationstests;
- Bewegungs-Palpationstests;
- Schmerz-Provokationstests.

Viele Autoren raten von Palpationstests der Position und der Bewegung ab [3, 7, 11, 22]. Zur diagnostischen Genauigkeit von

Schmerz-Provokationstests am SIG gehen die Meinungen in der Literatur hingegen auseinander. Hansen et al. [9] und Cohen [3] meinen, dass eine präzise Diagnosestellung aufgrund von Schmerz-Provokationstests nicht möglich ist. Laut Hancock et al. [8] und Robinson et al. [23] ist die Evidenz der Tests jedoch moderat bis gut. Stuber [26] meint, dass zwar die Evidenz einzelner Schmerz-Provokationstests mangelhaft ist, jedoch Testskombinationen die Aussagen verbessern.

Aufwendige Literaturrecherche

Wie gut eignen sich Schmerz-Provokationstests, um das SIG als Schmerzquelle bei Patienten mit unterem Rückenschmerz zu identifizieren? Wie gut ist ihre Evidenz? Auf der Suche nach Antworten durchforsteten Arne Vielitz und seine Kollegin Ulrike Honnens die internationalen medizinischen Datenbanken PubMed und PEDro sowie die Datenbank der Charité in Berlin

Sie benutzten die Suchworte: Sacroiliac Joint, Evidence-Based Medicine, Physical Therapy Modalities, Reproducibility of Results, Sensitivity and Specificity, Low Back Pain. Außerdem verwendeten sie die Funktion „Related Articles“ bei PubMed und nutzten die Literaturlisten der Studien und Reviews. Vielitz und Honnens wählten nur Studien oder Reviews der letzten 20 Jahre in englisch oder deutsch. Sie schlossen Artikel aus, die sich mit Schwangerschaft, speziellen Krankheitsbildern wie Inkontinenz, Tuberkulose oder Morbus Bechterew befassten, deren Teilnehmer Kinder, Sportler oder Personen ausschließlich über 60 Jahre waren, sowie Studien, die sich mit speziellen Therapieformen befassten.

Um herauszufinden, ob länderspezifische Vorgaben existieren, suchten sie in Zeitschriften nach einheitlichen Empfehlungen für einen oder mehrere Schmerz-Provokationstests. Bei den Zeitschriften handelte es sich um: Australian Journal of Physiotherapy, manuelletherapie, Physical Therapy and Spine. Die Suche blieb ohne Ergebnis. Also nutzen sie auch Fachbücher.

Sowohl Vielitz als auch Honnens prüften die inkludierte Literatur. Insgesamt erfüllten 29 Quellen die Anforderungen, um sie genauer unter die Lupe zu nehmen.

Positions- und Bewegungs-Palpationstests sind ungeeignet

Mehrere Autorenteams [3, 7, 11, 22] meinen, dass Therapeuten Positions- und Bewegungs-Palpationstests nicht verwenden sollten, um eine vom SIG ausgehende Schmerzsymptomatik zu diagnostizieren. Sie halten Positions- und Bewegungs-Palpationstests für ungeeignet, weil das SIG individuell sehr unterschiedlich ist und seine Bewegungsamplitude sehr gering ist.

Stattdessen empfehlen sie **Schmerz-Provokationstests** zur Diagnostik. Aktuelle Studien untersuchten am gründlichsten den Thigh-Thrust-Test und den FABER-Test. Für den Kompressions- und Distractionstest interessierten sich weniger Wissenschaftler. Die wenigsten Studien untersuchten den Gaenslen- und Sacral-Thrust-Test.

Schmerz-Provokationstests im Evidenzcheck

Von den am häufigsten in der Literatur genannten Tests erreichen der Thigh-Thrust- und der Distractionstest die besten Werte. Der FABER- und der Kompressionstest erreichen ebenfalls gute Werte, sie sind jedoch weniger zuverlässig. Gaenslen- und Sacral-Thrust-Test erzielen die schlechtesten Ergebnisse, wobei der Sacral-Thrust-Test die geringste Zuverlässigkeit aufweist.

Thigh-Thrust-Test

Der Thigh-Thrust-Test (► **Abb. 1**) erreicht bei der Sensitivität Werte von 0,36 bis 0,88. Die Spezifität variiert von 0,5 bis 1,0. Der positive Vorhersagewert liegt bei 0,58, der negative bei 0,92. Die Intertester-Reliabilität ergibt Werte zwischen 82 bis 94,1%, der Kappa-Wert beträgt je nach Studie Werte von 0,64 bis 0,88.

Evidenzcheck: Von den sechs untersuchten Tests ist die Sensitivität des Thigh-Thrust-Tests eine der höchsten. Seine Spezifität ist ausgezeichnet. Broadhurst und Bond [1] ermittelten die besten Werte. Aufgrund der Injektionstechnik zweifelten Hansen et al. [9] die Resultate von Broadhurst und Bond [1] an. Dreyfuss et al. [4] liegen mit ihren Berechnungen deutlich unter dem Schnitt, sie ermittelten eine ungenügende Sensitivität und eine mäßige Spezifität. Der Thigh-Thrust-Test erreicht den höchsten Kappa-Wert aller Tests. Er hat gute bis ausgezeichnete Vorhersagewerte. Zu diesem Ergebnis kam jedoch nur eine Autorengruppe. Die gesamten Werte und die Quellenangaben zeigt ► **Tab. 1**.

Fazit: Der Thigh-Thrust-Test erzielt durchschnittlich gute bis ausgezeichnete Ergebnisse. Seine Evidenz ist gut.

Distractionstest

Die Sensitivität des Distractionstests (► **Abb. 2 u. 3**) liegt bei 0,55 bzw. 0,6. Die Spezifität berechneten die Forscher je nach Studie einmal mit 0,81 und einmal mit 1,0. Bei den Vorhersagewerten errechneten sie ebenfalls unterschiedliche Werte: für den positiven Vorhersagewert 0,6 bzw. 1,0 und für den negativen 0,67 bzw. 0,81. Die Intertester-Reliabilität ergab 82% und 88,2%, der Kappa-Wert lag bei 0,67 bzw. 0,69.

Evidenzcheck: Der Distractionstest hat nur eine mäßige Sensitivität, jedoch ebenfalls eine ausgezeichnete Spezifität. Sein Kappa-

TIPP

Wer sich für die genauen Testdurchführungen interessiert, findet gute Beschreibungen in Broadhurst und Bond [1], Cattley et al. [2], Hansen et al. [9], Laslett [17], Levin und Stenström [18].



Abb. 1 Thigh-Thrust-Test (rechts).

Abb. 2 Distractionstest.

Abb. 3 Distractionstest gekreuzt.

Wert ist im Durchschnitt ebenfalls gut. Bei den Vorhersagewerten gehen die Berechnungen auseinander. Verschiedene Studie erzielten entweder für den positiven Vorhersagewert oder für den negativen gute bzw. ausgezeichnete Ergebnisse. Laslett et al. [16], deren Studie laut Stuber [26] eine hohe methodologische Qualität aufweist, berechneten den positiven als gut und den negativen als ausgezeichnet, Levin und Stenström [18] umgekehrt. Alle Werte und die Quellenangaben zeigt ► **Tab. 1**.

Fazit: Der Distractionstest erreichte durchschnittlich mäßige bis ausgezeichnete Ergebnisse. Seine Evidenz ist ebenfalls gut.

FABER-Test

Die Sensitivität des FABER-Tests (► **Abb. 4**) liegt zwischen 0,5 und 0,77. Seine Spezifität reicht von 0,16 bis 1,0. Die Forscher berech-

neten die Vorhersagewerte nicht. Seine Intertester-Reliabilität liegt zwischen 74% und 96%. Der Kappa-Wert zwischen 0,48 und 0,62.

Evidenzcheck: Sensitivität und Spezifität des FABER-Tests sind im Durchschnitt gut. Bei der Spezifität gehen die Berechnungen jedoch sehr stark auseinander. Dreyfuss et al. [4] berechneten für den Test eine sehr schlechte Spezifität, Broadhurst und Bond [1] eine ausgezeichnete. Der Kappa-Wert ist im Durchschnitt nur mäßig (alle Werte und Quellenangaben: ► **Tab. 1**).

Fazit: Der FABER-Test schneidet im Schnitt mäßig bis gut ab, auch seine Evidenz ist gut. Sie ist jedoch nicht so stark wie die des Distractionstests.

Tab. 1 Statistische Werte der Provokationstests in den einzelnen Studien.

Provokations-test	Se	Sp	PV	NV	ITR (%)	k
Thigh-Thrust-Test						
Robinson et al. [23]	-	-	-	-	84-87	0,74-0,76
Laslett et al. [16]/80%*	0,88	0,69	0,58	0,92	-	-
Broadhurst u. Bond [1]/70%*	0,80	1,00	-	-	-	-
Broadhurst u. Bond [1]/90%*	0,69	1,00	-	-	-	-
Dreyfuss et al. [4]/90%*	0,36	0,50	-	-	82	0,64
Laslett u. Williams [15]	-	-	-	-	94,10	0,88
Distractionstest						
Robinson et al. [23]	-	-	-	-	82	0,67
Laslett et al. [16]/80%*	0,60	0,81	0,60	0,81	-	-
Levin u. Stenström [18]	0,55	1,00	1,00	0,67-0,69	-	-
Laslett u. Williams [15]	-	-	-	-	88,20	0,69
FABER-Test						
Robinson et al. [23]	-	-	-	-	74-80	0,48-0,60
Broadhurst u. Bond [1]/70%*	0,77	1,00	-	-	-	-
Broadhurst u. Bond [1]/90%*	0,50	1,00	-	-	-	-

Tab. 1 Fortsetzung.

Provokations-test	Se	Sp	PV	NV	ITR (%)	k
Strenger et al. [25]	-	-	-	-	88-96	-
Dreyfuss et al. [4]/90%*	0,69	0,16	-	-	85	0,62
Kompressionstest						
Robinson et al. [23]	-	-	-	-	82-88	0,48-0,67
Laslett et al. [16]/80%*	0,69	0,69	0,52	0,82	-	-
Strenger et al. [25]	-	-	-	-	74-79	0,26
Laslett u. Williams [15]	-	-	-	-	88,2	0,73
Gaenslen-Test						
Laslett et al. [16]/80%*	0,5-0,53	0,71-0,77	0,47-0,50	0,76-0,77	-	-
Dreyfuss et al. [4]/90%*	0,71	0,26	-	-	82	0,61
Laslett u. Williams [15]	-	-	-	-	88,20	0,72-0,75
Sacral-Thrust-Test						
Laslett et al. [16]/80%*	0,63	0,75	0,56	0,80	-	-
Dreyfuss et al. [4]/90%*	0,53	0,29	-	-	66	0,30
Laslett u. Williams [15]	-	-	-	-	78	0,52

* = ≥ Schmerzlinderung bei SIG-Block; Se = Sensitivität; Sp = Spezifität; PV = Positiver Vorhersagewert; NV = Negativer Vorhersagewert; ITR = Intertester-Reliabilität; k = Kappa-Wert

Kompressionstest

Für den Kompressionstest (► **Abb. 5 u. 6**) berechneten die Wissenschaftler sowohl die Sensitivität als auch die Spezifität mit 0,69. Der positive Vorhersagewert liegt bei 0,52 und der negative bei 0,82. Bei der Intertester-Reliabilität variieren die Werte von 74% bis 88%. Für den Kappa-Wert errechneten je nach Studie Werte von 0,26 bis 0,73 (alle Werte und Quellenangaben: ► **Tab. 1**).

Evidenzcheck: Der Kompressionstest hat eine gute Sensitivität und eine gute Spezifität. Allerdings berechnete nur eine Forschergruppe beide Werte. Der durchschnittliche Kappa-Wert ist mäßig, die Berechnungen der Autoren differieren hier stark. Sie reichen von ungenügend [25] bis gut [15]. Sein positiver Vorhersagewert ist nur mäßig, der negative ist hingegen ausgezeichnet. Auch die beiden Vorhersagewerte berechnete nur eine Forschergruppe (alle Werte und Quellenangaben: ► **Tab. 1**).

Fazit: Der Kompressionstest erreicht durchschnittlich mäßige bis gute Ergebnisse. Seine Evidenz ist gut, liegt aber unter der des FABER-Tests.

Gaenslen-Test

Gaenslen-Test (► **Abb. 7**) erreicht bei der Sensitivität Werte zwischen 0,5 und 0,71. Bei der Spezifität variieren sie von 0,26 bis 0,77. Der positive Vorhersagewert liegt bei 0,47 bzw. 0,5 und der

negative bei 0,76 bzw. 0,77 (je nach getesteter Seite). Die Intertester-Reliabilität beträgt 82% und 88,2%, je nach Studie. Der Kappa-Wert liegt zwischen 0,61 und 0,75 (alle Werte und Quellenangaben: ► **Tab. 1**).

Evidenzcheck: Gaenslen-Test erreicht bei der Sensitivität und Spezifität durchschnittlich mäßige bzw. annähernd gute Ergebnisse. Laslett et al. [16] ermittelten eine mäßige Sensitivität, jedoch eine gute Spezifität. Dreyfuss et al. [4] erhielten entgegengesetzte Ergebnisse. Der Kappa-Wert des Tests ist gut, hier sind die Meinungen eindeutig. Der positive Vorhersagewert ist mäßig. Der negative Vorhersagewert ist hingegen gut. Allerdings berechnete nur eine Autorengruppe die Vorhersagewerte.

Fazit: Die durchschnittlichen Ergebnisse des Gaenslen-Tests sind mäßig bis gut, sie liegen jedoch unter denen der anderen Tests. Seine Evidenz ist gerade noch gut.

Sacral-Thrust-Test

Beim Sacral-Thrust-Test (► **Abb. 8**) reicht die Sensitivität von 0,53 bis 0,63 und die Spezifität von 0,29 bis 0,75. Sein positiver Vorhersagewert liegt bei 0,56 und der negative bei 0,8. Für die Intertester-Reliabilität errechnete man je nach Studie 66% bzw. 78%. Der Kappa-Wert liegt mal bei 0,3 und mal bei 0,52 (alle Werte und Quellenangaben: ► **Tab. 1**).

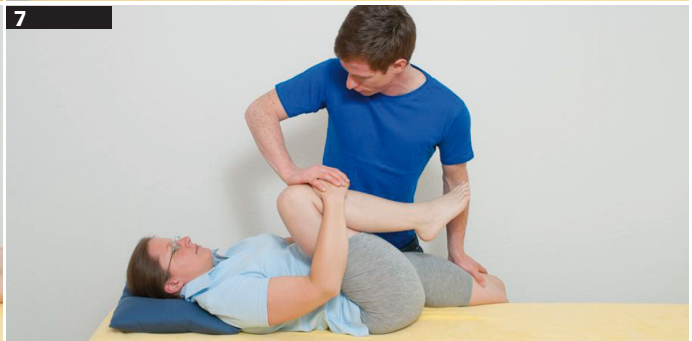


Abb. 4 FABER-Test (rechts).

Abb. 5 Kompressionstest aus Seitlage.

Abb. 6 Kompressionstest aus Rückenlage.

Abb. 7 Gaenslen-Test (rechtes Ilium in posteriorer Rotation, linkes Ilium in anteriorer Rotation).

Evidenzcheck: Die durchschnittliche Sensitivität und Spezifität des Sacral-Thrust-Tests sind mäßig. Bei der Spezifität gehen die Meinungen stark auseinander, Laslett et al. [16] berechneten eine gute, Dreyfuss et al. [4] eine ungenügende Spezifität. Sein Kappa-Wert ist im Schnitt gerade noch mäßig. Die Ergebnisse sind mal mäßig [15], mal ungenügend [4]. Der positive Vorhersagewert ist ebenfalls nur mäßig. Der negative Vorhersagewert ist hingegen gut, wie auch die Wahrscheinlichkeitsverhältnisse. Auch beim Sacral-Thrust-Test berechnete nur eine Autorengruppe die Vorhersagewerte. Die durchschnittlichen Werte des Sacral-Thrust-Tests sind folglich mäßig. Sie variieren von gerade noch mäßig bis ausgezeichnet. Insgesamt liegen sie jedoch unter denen der anderen Tests.

Fazit: Die vorliegende Evidenz des Sacral-Thrust-Tests ist mäßig.

Reicht ein Schmerz-Provokationstest aus?

Nimmt man die durchschnittlichen Resultate aller Untersuchungsergebnisse pro Test, ist keiner sehr gut. Thigh-Thrust- und Distractionstest sind aber die besten von den sechs untersuchten Tests. FABER- und Kompressionstest erzielen insgesamt schlechtere Ergebnisse. Gaenslen- und Sacral-Thrust-Test liegen mit ihren Ergebnissen durchschnittlich unter denen des FABER-Tests und des Kompressionstests und deutlich unter denen des Thigh-Thrust-Tests und denen des Distractionstests.

Maigne et al. [19] fanden keine statistische Signifikanz für die von ihnen untersuchten Tests (Kompressions-, Distaktions-, FABER-, Gaenslen-, Sacral-Thrust-Test). Stuber [26] kritisiert in seinem Review die niedrige methodologische Qualität der Studie von Maigne. Angesichts der limitierten Qualität der Tests, meinen viele Autoren, dass Therapeuten am besten mehrere Schmerz-Provokationstests kombinieren.

Testkombinationen im Evidenzcheck

Einzelne Provokationstests am SIG sind uninformativ. Es gibt keinen Provokationstests mit herausragender Qualität [8, 26]. Deshalb empfehlen Experten mehrere Tests zu kombinieren [8, 31],

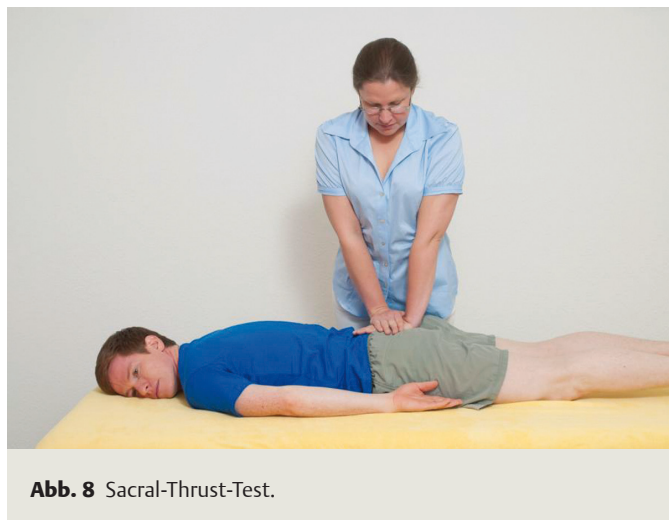


Abb. 8 Sacral-Thrust-Test.

da dies die Validität [8] und die Intertester-Reliabilität [12] erhöht.

Hinweis

Therapeuten verwenden am besten Tests mit hoher Sensitivität und Spezifität, sie erhalten so genauere Aussagen [26, 29, 31].

Kombinieren Therapeuten mehrere Tests, ist eine vorgegebene Mindestanzahl der positiven Tests wünschenswert, damit sie eine exaktere Diagnose stellen können. Beispielsweise: Zwei von vier Tests sollen positiv sein oder drei von fünf etc. Eine definitive Empfehlung gibt es zurzeit noch nicht [26].

Hinweis

Testkombinationen weisen jedoch deutlich bessere Ergebnisse auf als jeder einzelne Test.

Die Sensitivität reicht hier von 0,82 bis 0,94, die Spezifität liegt bei 0,57 bis 0,79. Der positive Vorhersagewert erreicht Werte von 0,47 bis 0,77, der negative 0,87 bis 0,96. Je nach Studie variiert das positive Wahrscheinlichkeitsverhältnis von 1,9 bis 4,29. Das negative Wahrscheinlichkeitsverhältnis variiert von 0,19 bis 0,8. Bei der Intertester-Reliabilität kombinierten die Forscher [13] folgende Tests:

- Kompressionstest;
- Distractionstest;
- Thigh-Thrust-Test;
- Gaenslen-Test;
- Sacral-Thrust-Test.

Bei der Intertester-Reliabilität wurden 93,59% errechnet, der Kappa-Wert beträgt 0,7. Alle Werte und Quellenangaben zeigen die ► **Tab. 2–5**.

Testkombination 1

Kokmeyer et al. [13] kombinierten den Kompressions-, Distaktions-, Thigh-Thrust-, Gaenslen-Test von rechts und von links sowie den Sacral-Thrust-Test. Sie ermittelten bei ihrer Kombination gute Ergebnisse für den Kappa-Wert, wenn **drei oder mehr Tests von fünf positiv** waren (► **Tab. 2**).

Testkombination 2

Laslett et al. [16] untersuchten die gleiche Testkombination. Sie erhielten gute bis ausgezeichnete Ergebnisse bei drei oder mehr positiven Tests von insgesamt fünf (Berechnung 2a).

Das Team empfiehlt Therapeuten, den Gaenslen-Test nicht zu verwenden, da er in ihrer Studie am schlechtesten abschnitt. Laslett et al. [16] berechneten für die Kombination der restlichen vier Tests ebenfalls gute bis ausgezeichnete Ergebnisse, **wenn zwei Tests positiv waren** (Berechnung 2b). ► **Tab. 3** zeigt alle Werte.

Testkombination 3

Van der Wurff et al. [31] kombinierten den Kompressions-, Distaktions-, Thigh-Thrust-, FABER- und Gaenslen-Test. Sie sind der Meinung, dass **drei oder mehr positive Tests von fünf** auf das SIG als Schmerzsache hinweisen und ermittelten ausgezeichnete bis annähernd ausgezeichnete Ergebnisse (► **Tab. 4**).

■ **Tab. 2** Statistische Werte der Testkombination 1.

Kokmeyer et al. [13]	ITR (%)	k	Bbk	PBbk
1 oder mehr von 5	83,33	0,63	0,07	0,65
2 oder mehr von 5	92,31	0,74	0,74	0,85
3 oder mehr von 5	93,59	0,7	0,7	0,87
4 oder mehr von 5	96,15	0,71	0,71	0,92
5 oder mehr von 5	98,72	0,66	0,66	0,97

Verwendete Tests: Kompressions-, Distractions-, Thigh-Thrust-, Gaenslen-, Sacral-Thrust-Test. ITR = Intertester-Reliabilität; k = Kappa-Wert; Bbk = Biasbereinigter Kappa; PBbk = Prävalenz und biasbereinigter Kappa

■ **Tab. 4** Statistische Werte der Testkombination 3.

van der Wurff et al. [31]/50%*	Se	Sp	PV	NV	+LR	-LR
1 oder mehr von 5	1,00	0,42	0,59	1,00	1,74	0,00
2 oder mehr von 5	0,93	0,58	0,64	0,90	2,18	0,13
3 oder mehr von 5	0,85	0,79	0,77	0,87	4,01	0,19
4 oder mehr von 5	0,26	0,81	0,54	0,58	0,43	0,91
5 oder mehr von 5	0,00	1,00	0,00	0,55	0,00	1,00

Verwendete Tests: Kompressions-, Distractions-, Thigh-Thrust-, Gaenslen-, FABER-Test. * = \geq Schmerzlinderung bei SIG-Block; Se = Sensitivität; Sp = Spezifität; PV = Positiver Vorhersagewert; NV = Negativer Vorhersagewert; +LR = positives Wahrscheinlichkeitsverhältnis; -LR = negatives Wahrscheinlichkeitsverhältnis

Testkombination 4

Stanford und Burnham [24] wählten eine andere Testzusammensetzung. Sie ließen den Distractionstest weg, da viele Patienten den Druck auf die Spinae iliacae anteriores superiores (SIAS) als unangenehm empfinden. Die Patienten können sich dann schlechter auf den Schmerz am SIG konzentrieren. Ihre Beurteilung ist erschwert. Stanford und Burnham [24] kombinierten daher den Kompressions-, Thigh-Thrust-, Gaenslen-Test ipsilateral und kontralateral, den FABER- und den Sacral-Thrust-Test. Diese Testkombination erreichte annähernd gute bis ausgezeichnete Ergebnisse (► **Tab. 5**).

Auch Testkombinationen stoßen an Grenzen

Die Ergebnisse der Kombinationen 1–4 sind etwa gleichwertig. Schmerz-Provokationstests zu kombinieren, ist im Schnitt besser als jeder untersuchte Einzeltest. Aber: Belasten die Schmerz-Provokationstests nur das SIG? An der Validität ihrer Aussagekraft muss man zweifeln. Mögliche Schmerzverursacher sind auch andere das SIG umgebende und beeinflussende Strukturen, wie Muskeln und Bänder, Lendenwirbelsäule oder Hüftgelenke.

Die Validierung der Schmerz-Provokationstests mittels intraartikulärer Injektion eines Lokalanästhetikums ist umstritten [2, 3, 8]. Stanford und Burnham [24] entschieden sich für ein Studiendesign, in dem sie die SIG der Probanden vor und nach Lokalanästhesie tes-

■ **Tab. 3** Statistische Werte der Testkombination 2.

Laslett et al. [16]/80%*	Se	Sp	PV	NV	+LR	-LR
1 oder mehr von 5	1,00	0,44	0,47	1,00	1,78	0,00
2 oder mehr von 5	0,93	0,66	0,58	0,96	2,73	0,10
3 oder mehr von 5 (2a)	0,94	0,78	0,68	0,96	4,29	0,80
4 oder mehr von 5	0,60	0,81	0,60	0,81	3,20	0,49
5 oder mehr von 5	0,27	0,88	0,50	0,72	2,13	0,84
2 oder mehr von 4 (2b)	0,88	0,78	0,67	0,93	4,00	0,16

Verwendete Tests Berechnung 2a: Kompressions-, Distractions-, Thigh-Thrust-, Gaenslen-, Sacral-Thrust-Test; 2b: ohne Gaenslen-Test.

* = \geq Schmerzlinderung bei SIG-Block; Se = Sensitivität; Sp = Spezifität; PV = Positiver Vorhersagewert; NV = Negativer Vorhersagewert; +LR = positives Wahrscheinlichkeitsverhältnis; -LR = negatives Wahrscheinlichkeitsverhältnis

■ **Tab. 5** Statistische Werte der Testkombination 4.

Stanford u. Burnham [24]/80%*	Se	Sp	PV	NV	+LR
3 oder mehr von 6	0,82	0,57	0,47	0,87	1,9

Verwendete Tests: Kompressions-, Thigh-Thrust-, Gaenslen-Test ipsilateral und kontralateral, Sacral-Thrust-, FABER-Test.

* = \geq Schmerzlinderung bei SIG-Block; Se = Sensitivität; Sp = Spezifität; PV = Positiver Vorhersagewert; NV = Negativer Vorhersagewert; +LR = positives Wahrscheinlichkeitsverhältnis

Anmerkung: Für die Untersuchung von Stanford liegen keine weiteren Ergebnisse vor, da bisher lediglich ein Forschungsbericht und kein kompletter Artikel veröffentlicht wurde.

teten. Die nach der Injektion durchgeführten Schmerz-Provokationstests hatten keinen diagnostischen Wert. Als Ursache vermuten die Autoren, dass die Tests unterschiedliche Strukturen reizen.

Kann man falsch-positive Ergebnisse verhindern?

Laslett [17] merkt in seinem Übersichtsartikel an, dass SIG-Tests bei Patienten mit Rückenschmerz häufiger positiv sind, als es die Prävalenz für das SIG vermuten lässt. Er geht davon aus, dass einzelne Tests häufig falsch positiv sind, z. B. bei Patienten mit Nerven-Wurzel-Schmerzen oder einem Diskus-Prolaps.

Untersuchen Therapeuten ihre Patienten nach McKenzie und findet eine Zentralisation der Schmerzen statt, sind falsch-positive Ergebnisse seltener. Die Zentralisation ist eine bekannte klinische Beobachtung bei der standardisierten McKenzie-Untersuchung mit wiederholten und gehaltenen Bewegungen. Sie wurde wiederholt als reliabel und valide beschrieben und ist sehr spezifisch für diskogenen Schmerz.

Bei Patienten mit **schmerzhaftem SIG** oder **Facettengelenk** ist die **Zentralisation nicht zu finden**.

- Patienten mit diskogenen und gleichzeitig durch das SIG verursachten Schmerzen sind sehr selten.
- Nimmt man an, dass bei 30% der Patienten mit unterem Rückenschmerz das SIG die Ursache ist und findet man bei der

Untersuchung eines Patienten mit Rückenschmerzen drei oder mehr positive SIG Schmerz-Provokationstests, gibt es eine 59%-ige Chance, dass tatsächlich eine SIG Problematik vorliegt.

- Ist allerdings eine McKenzie-Untersuchung vorangegangenen und Therapeut beobachtete keine Zentralisation des Schmerzes, steigt die Wahrscheinlichkeit auf eine vom SIG ausgehende Problematik auf 77%.
- Würde ein Therapeut den gleichen Clinical-Reasoning-Prozess bei einer Gruppe Schwangeren anwenden, wäre die Zahl noch viel höher [17]. Bei Schwangeren mit Beckengürtelschmerzen umfasst die Diagnostik jedoch weitere Tests, wie etwa den Active-Straight-Leg-Raise-Test (ASLR) [20].

Welche Tests sind praxistauglich?

Aufgrund der Literatur ist keine einheitliche Aussage zur Evidenz von Schmerz-Provokationstests am SIG bei Patienten mit unterem Rückenschmerz möglich. In den Studien fallen die Ergebnisse sehr unterschiedlich aus, sodass die Frage nach ihrer Evidenz nicht eindeutig zu klären ist.

- Keiner der Einzeltests besitzt eine herausragende Qualität.
- Der Thigh-Thrust- und der Distractionstest erreichten die besten Werte. Auch wenn diese Tests einzeln verwendet positiv sind, ist die Wahrscheinlichkeit gering, dass man durch sie eine vom SIG ausgehende Schmerzsymptomatik erkennt. Ein negatives Ergebnis ist ein starker Hinweis, dass das SIG die Schmerzen nicht verursacht.
- Für den FABER- und den Kompressionstest gilt diese Aussage ebenfalls. Beide Tests sind jedoch weniger zuverlässig.
- Gaenslen- und Sacral-Thrust-Test kann man nach den vorliegenden Ergebnissen nicht empfehlen.

Testkombinationen erhöhen die Validität und Reliabilität

Verwenden Therapeuten Testkombinationen, erhöht dies die Validität und Reliabilität. Am besten kombiniert man:

- Thigh-Thrust-Test;
- Distractionstest;
- FABER-Test;
- Kompressionstest.

Sind zwei oder mehr der Tests positiv, weist dies deutlich auf eine Störung des SIG hin. Handelt es sich dabei um den Thigh-Thrust- und Distractionstest, ist die Wahrscheinlichkeit einer Störung sehr hoch. Sind beide negativ, kann man davon ausgehen, dass das SIG nicht die Schmerzursache ist.

McKenzie vor der Testkombination erhöht die Validität

Positive SIG Schmerz-Provokationstests allein ermöglichen jedoch keine sichere Aussage zur Schmerzquelle. Sie reichen für eine Diagnose nicht aus. Führt man sie sehr standardisiert aus, haben sie zwar eine gewisse Validität. Kann man aber sicher sein, dass die Patienten keine andere, z. B. diskogene Schmerzursache haben, verbessert sich die Validität deutlich. Für die Untersuchung empfiehlt es sich also, die Tests in einen Clinical-Reasoning-Prozess einzubetten.

Hinweis

Ergibt z. B. die vorherige McKenzie-Untersuchung keine Zentralisation der Symptome, steigt die diagnostische Stärke der Testkombinationen, denn die falsch-positive Rate der Schmerz-Provokationstests sinkt.

Kritische Anmerkungen

Die vorgestellten Studien sind schlecht zu vereinheitlichen, da die Studiendesigns zum Teil sehr unterschiedlich sind. Nicht alle Forscherteams verblindeten, setzten Kontrollgruppen oder SIG-Blockaden mittels Injektionen ein. Einige verwendeten einen Block, andere zwei. Setzten die Forscher SIG-Blockaden ein, gaben sie unterschiedliche Grenzwerte der Schmerzbefreiung an, ab denen sie den Block positiv werteten. Auch die Testdurchführungen definierten die Forscherteams unterschiedlich. Die aufzuwendenden Kräfte und die jeweilige Richtung, in der sie wirken sollten, waren nicht genau definiert.

Fehlender Goldstandard

Für zukünftige Studien, die die Validität und Reliabilität der empfohlenen Testkombination untersuchen möchten, sind einheitliche Kriterien für die Durchführung der Tests unbedingt notwendig. Diese Kriterien sollten die Größe der Kraft und die Krafrichtung beinhalten. Sie sollten außerdem die maximale Haltezeit für den Test definieren. Entwickeln Forscher den Gold-Standard, können wir die Gültigkeit von Schmerz-Provokationstests genauer überprüfen.

Danksagung

Dieser Artikel basiert auf meiner Bachelor-Thesis, die ich zusammen mit Ulrike Honnens geschrieben habe. Für die gemeinsame Arbeit und das Einverständnis zur Veröffentlichung der Ergebnisse möchte ich ihr herzlich danken.

Die Literatur finden Sie online unter [thieme connect: www.thieme-connect.de/ejournals/toc/manuelletherapie](http://thieme-connect.de/ejournals/toc/manuelletherapie)

AUTOR

Arne Vielitz B.A.

Arne Vielitz ist seit 2005 Physiotherapeut. Er studierte von 2006 bis 2009 Physiotherapie an der Hogeschool Fysiotherapie Thim van der Laan in den Niederlanden und arbeitet in einer physiotherapeutischen Praxis in Eutin. Als Grundlage für diesen Artikel diente seine Bachelor-Thesis und die fachliche Neugier eines Manual- und Cred. MDT-Therapeuten.
Hundestraße 73
23552 Lübeck
vielitz@pvtvd.de



BIBLIOGRAFIE

DOI 10.1055/s-0032-1304755
manuelletherapie 2012; 16: 18-24
© Georg Thieme Verlag KG
Stuttgart · New York · ISSN 1433-2671