

G. Stucki
S. Stucki
O. Sangha

Patienten-zentrierte Evaluation der Krankheitsauswirkungen bei muskuloskelettalen Erkrankungen: Adaptation und Neuentwicklung von Outcome-Instrumenten

Patients-oriented outcome assessment in musculoskeletal disease: cross-cultural adaptation and development of instruments

Zusammenfassung Die zuverlässige (reliable), gültige (valid) und verlaufsempfindliche (sensitive) Beurteilung von Krankheitsauswirkungen setzt die Verwendung von standardisierten Messmethoden voraus. Für die Erfassung der allgemeinen Gesundheit stehen dabei ausgezeichnete, hinreichend erprobte und für den deutschen Sprachraum adaptierte Instrumente wie der SF-36 zur Verfügung. Hingegen liegen krankheits- und problemspezifische Outcome-Instrumente oft nur in einer engli-

sehen Version vor und müssen für die deutsche Sprache adaptiert werden. Bei der Adaptation eines Instrumentes für den deutschen Sprachraum handelt es sich in der Regel nicht um eine simple Übersetzung. Damit Resultate des gleichen Instrumentes in verschiedenen Sprachräumen und Kulturen vergleichbar sind und ein Instrument z. B. in internationalen Multizenterstudien eingesetzt werden kann, ist einerseits eine *inhaltliche* und andererseits eine *metrische Äquivalenz* der Instrumente erforderlich. Das typische Vorgehen bei der Adaptation eines Instrumentes umfaßt die Übersetzung, Rückübersetzung, Review und Testung. Die Testung umfaßt einerseits die metrischen Eigenschaften (Faktoren, Zuverlässigkeit, interne Konsistenz) und andererseits die Gültigkeit des Instrumentes.

Die Neuentwicklung eines Instrumentes sollte man nur dann in Erwägung ziehen, wenn kein geeignetes Instrument zur Verfügung steht. Die Neuentwicklung auch von kurzen Instrumenten ist zeit- und kostenintensiv. Idealerweise weisen neuentwickelte Instrumente Intervalleigenschaften auf.

Summary Assessment of disease consequences (outcome) requires reliable, valid, and sensitive instru-

ments. Psychometrically sound generic health-status instruments such as the SF-36 have been validated for many languages and are available in German. Instead, most disease specific instruments have been developed in English and need to be adapted for the German language. To allow for cross-cultural comparisons and use of results across cultures, for instance, in international multicenter studies, instruments need to have both content and metric equivalence. Thus, adaptation of health-status instruments requires a standardized process including translation, backtranslation, review and assessment of metric properties (reliability, internal consistency, factors), and validity. Developments of new instruments from scratch are time and cost intensive and should only be considered if no current instrument is available. Ideally, newly developed instruments have interval-scale properties.

Schlüsselwörter Health Status – Outcome – Patienten Assessment – Instrumentenentwicklung

Key words Health status – outcome – patient self assessment – instrument development

Eingegangen: 25. August 1997
Akzeptiert: 26. September 1997

PD Dr. med. G. Stucki, MS (✉)
S. Stucki, MEd
Universitätsspital Zürich
Rheumaklinik und Institut
für physikalische Medizin
Gloriastrasse 25
CH-8091 Zürich

Dr. med. O. Sangha, MPH
Prigham and Womens Hospital
Harvard Medical School
75 Francis Street
Boston, MA

Hintergrund

Die zuverlässige (reliable), gültige (valide) und verlaufs-empfindliche (sensitive) Beurteilung von Krankheitsauswirkungen setzt die Verwendung von standardisierten Messmethoden voraus.

Für die Erfassung der allgemeinen Gesundheit stehen dabei ausgezeichnete, hinreichend erprobte und für den deutschen Sprachraum adaptierte Instrumente wie der SF-36 (1) oder das Nottingham Health Profile (2, 3) zur Verfügung. So wird der SF-36 weltweit professionell durch ein Netz von Psychometrikern betreut und weiterentwickelt. Als Nutzer und Anwender solcher Instrumente wird der Rheumatologe kaum direkt an der Entwicklung und kulturellen Adaptation solcher Instrumente beteiligt sein.

Hingegen werden krankheits- und problemspezifische Outcome-Instrumente wie der Health Assessment Questionnaire (HAQ) (4) und die Arthritis Impact Measurement Scales (AIMS) (5, 6) für die chronische Polyarthritis, der Western Ontario McMasters University (WOMAC) Osteoarthritis Index (7, 8) für die Arthrose durch die „Rheumatologische Gemeinschaft“ betreut.

Die zunehmende Anwendung und Verbreitung von Outcome-Instrumenten sowie deren Anwendung in multinationalen klinischen, epidemiologischen und ökonomischen Studien erfordert die kulturelle Adaptation der vorwiegend im angelsächsischen Sprachraum entwickelten Instrumente. Im deutschsprachigen Raum befassen sich verschiedene Forschergruppen mit diesen Adaptationen. Zur Zeit befindet sich beispielsweise der AIMS 2 für die chronische Polyarthritis in Validierung.

Jeder Forscher, insbesondere wenn er aus einer bestimmten (medizinischen) Kultur kommt, hat in der Regel eigene, starke Präferenzen, wie ein Instrument zur Erfassung eines bestimmten Problems auszusehen hat. Bevor man selbst mit der Entwicklung eines Outcome-Instrumentes beginnt, sollte man dennoch prüfen, ob nicht ein internationales, d. h. meist englischsprachiges Instrument den eigenen Vorstellungen einigermaßen entspricht. Sofern die Fragen aus der Sicht des Patienten (!) relevant sind und auch vom Format her gesehen akzeptabel scheinen, sollte man mit einer gewissen Toleranz über stilistische „Mängel“ hinwegsehen und auf die Entwicklung eines eigenen Instrumentes verzichten. Man muß sich bewußt sein, daß bereits eine sorgfältige kulturelle Adaptation einen beachtlichen Aufwand erfordert und oft Monate in Anspruch nimmt. Eine Eigenentwicklung dauert hingegen meist mehrere Jahre. Gerade die Erfahrungen mit dem SF-36, aber auch mit dem AIMS (AIMS 2) und dem HAQ (verbesserte Versionen sind in Entwicklung) (9) haben gezeigt, daß ein Instrument niemals auf Anhieb perfekt ist; so gesehen ist es vermessener anzunehmen, daß ein eigenes Instrument besser wäre als ein bereits vorhandenes mit zufriedenstellenden metrischen Eigenschaften. Zudem ist es

wenig wahrscheinlich, daß ein neues, in deutsch entwickeltes Instrument eine weite Verbreitung finden würde, wenn bereits ein ähnliches, international akzeptiertes und verbreitetes Instrument zur Verfügung steht.

Steht wirklich kein geeignetes Instrument in der eigenen Sprache zur Verfügung oder läßt sich keines der vorhandenen Instrumente adaptieren, sollte die Möglichkeit einer Erweiterung/Ergänzung eines bestehenden Instrumentes in Betracht gezogen werden. So wurde z. B. der HAQ um 5 Fragen erweitert, um krankheitsspezifische Funktionseinschränkungen von Patienten mit Spondyloarthropathien zu erfassen (10). Eine solche Ergänzung ist methodologisch gesehen eine eigentliche Neuentwicklung und bedeutet einen dementsprechend großen Aufwand.

Entsprechend der Bedeutung kultureller Adaptationen wird anhand von Beispielen nachfolgend ein mögliches Vorgehen illustriert und nur eine kursorische Übersicht zur Erweiterung/Anpassung und Neuentwicklung von Instrumenten gegeben.

Kulturelle Adaptation

Bei der Adaptation eines Instrumentes für den deutschen Sprachraum handelt es sich in der Regel nicht um eine simple Übersetzung. Damit Resultate des gleichen Instrumentes in verschiedenen Sprachräumen und Kulturen vergleichbar sind ein Instrument z. B. in internationalen Multizenterstudien eingesetzt werden kann, braucht es einerseits eine *inhaltliche* und andererseits eine *metrische Äquivalenz* der Instrumente.

Das Erreichen der inhaltlichen Äquivalenz ist umso schwieriger, je verschiedener Zielsprache, Zielland und Zielkultur von der Quellenkultur sind. Hierbei ist zu berücksichtigen, daß Sprache, Kultur und Land je verschiedene Gegebenheiten sind. Es ist möglich, daß z. B. eine Adaptation eines Instrumentes für die Schweiz in Deutschland bei den Patienten auf Unverständnis stößt. So wurde z. B. eine von der Zürcher Gruppe vorgenommene Adaptation des HAQ von deutschen Kollegen kritisiert. Nachfragen ergaben, daß die in der Schweiz üblichen Milchpackungen („Tetrapack“) in Deutschland „Milchbeutel“ oder „Milchtüte“ genannt werden. Zudem können die Begriffe in der heutigen schnelllebigen Zeit auch wieder ändern. So werden offenbar in Deutschland in der Zwischenzeit auch „Tetrapackungen“ verwendet während in der Schweiz Milchbeutel aus Umweltschutzgründen wieder in Mode kommen. Auch wurde die in der Schweizer Version verwendete „Konfitüre“ (von „Konfi“ in der Umgangssprache) als nicht allgemein üblich taxiert. In Deutschland sei Konfitüre eine gehobene Ausdrucksweise und der Durchschnittsbürger verwende „Marmelade“. Eine umfassende

Tab. 1 Vorgehen bei der kulturellen Adaptation von Outcome-Instrumenten (11, 12)

Vorgehen	Kommentar
Übersetzung	Mindestens 3 Übersetzer mit Muttersprache = Zielsprache und ausgezeichneten Kenntnissen der Originalsprache
Rückübersetzung	Jede Übersetzung, Übersetzer mit Muttersprache = Originalsprache und ausgezeichneten Kenntnissen der Zielsprache
Review	Multidisziplinäres Komitee Strukturierte Techniken zur Lösung von Widersprüchen
Testung	Metrische Eigenschaften – Faktoren – Interne Konsistenz – Item-difficulty – Test-Retest-Zuverlässigkeit Gültigkeit
Anpassung	Überprüfung von Items Gewichtung von Items

Diskussion solcher Probleme findet sich in Übersicht-artikeln von Guillemin (11, 12) und Bullinger (13).

Das Erreichen der metrischen Äquivalenz ist ein zur Zeit nur ungenügend angegangenes Problem. Meist begnügt man sich damit, daß das Instrument die gleichen „Faktoren“ aufweist und die Fragelemente (Items) eines Konstruktes untereinander assoziiert sind, d. h. es wäre nicht notwendig, eine wortwörtliche Übersetzung vorzunehmen. Bei diesem Vorgehen wird die Übersetzung mit großer Wahrscheinlichkeit auch die metrischen Eigenschaften der Originalversion beibehalten. Ein Vergleich über verschiedene Sprachräume hinweg und konsequenterweise auch die Zusammenfassung von Daten verschiedener internationaler Studien (z. B. im Rahmen von internationalen Multizenterstudien) wäre demzufolge auch nicht in Frage gestellt. Auf der anderen Seite hat das Vorgehen einer „losen“, d. h. nur Konstrukt-bezogenen Übersetzung einige entscheidende Nachteile (14):

- Instrumente können auch in ihrer Originalversion Schwächen aufweisen, die sowohl inhaltlicher, sprachlicher oder metrischer Natur sein können.
- Es gibt Fragen, die sich nicht gut, auch nicht sinngemäß übersetzen lassen.
- Fragen, die für die Population, in der das Instrument ursprünglich entwickelt wurde, relevant waren, mögen für eine andere Population trivial sein. Sinngemäß können andere, für die Studienpopulation relevante Items in der Originalversion gänzlich fehlen.
- Damit Instrumentenscores in unterschiedlichen Sprachen und Kulturen vergleichbar sind, wäre es auch nötig, daß die Items der verschiedenen Instrumente den gleichen Schwierigkeitsgrad auf einer Skala aufweisen.

Bei einer kulturellen Adaptation eines Outcome-Instrumentes empfiehlt sich das in Tabelle 1 dargestellte Vorgehen (11, 12).

Übersetzung und Rückübersetzung

In einem ersten Schritt wird das Instrument durch bilinguale Übersetzer mit Muttersprache Deutsch in die deutsche Sprache übersetzt. Danach wird jede deutsche Version durch je einen bilingualen Übersetzer mit Muttersprache entsprechend der Quellsprache rückübersetzt.

Konsensus

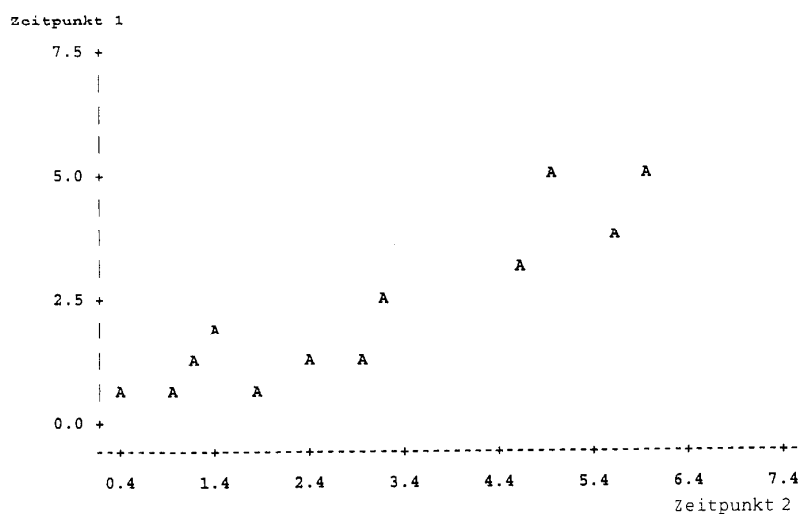
Ein Expertenteam mit möglichst breitem Wissens- und Erfahrungshintergrund untersucht die Resultate der 3 Übersetzungen/Rückübersetzungen und wählt diejenige Version aus, die der besten Rückübersetzung am nächsten kommt. In Situationen, bei denen keine exakte Rückübersetzung zustande kommt oder bei der verschiedene deutsche Übersetzungen zur exakten Rückübersetzung führen, wird ein Konsens über die Itemwahl unter den Experten gesucht. Das Ziel des Übersetzungs- Rückübersetzungsvorganges und der Konsensusfindung liegt nicht so sehr in der Identifikation der einzig korrekten Version (die es oft gar nicht gibt) als in der Aufdeckung von Problemen.

Bei der kulturellen Adaptation verschiedener Instrumente aus der nordamerikanischen Kultur in die deutsche Sprache (15, 8) zeigten sich nur geringe Probleme bei der Adaption von Fragen zu Symptomen und zur physischen Funktionsstörung. Hingegen mußten z. B. bei einer Adaptation des HAQ für Brasilien stärkere Anpassungen vorgenommen werden (16). Insbesondere mußte eine Frage zur Verwendung von Automobilen mit der Frage nach der Nutzung von öffentlichen Verkehrsmitteln (Bussen) ersetzt werden, da viele Brasilianer weder ein Automobil besitzen noch dieses regelmäßig benutzen. Die größten Schwierigkeiten ergeben sich erfahrungsgemäß für psychologische Konzepte.

Testung

Im nächsten Schritt wird das übersetzte Instrument an einer Gruppe von Patienten getestet. In der Regel werden dabei gleichzeitig eine Untersuchung der metrischen Eigenschaften, der Test-Retest-Zuverlässigkeit und verschiedene Validitätspriifungen vorgenommen. Dies garantiert, daß ein Instrument auch in der adaptierten Version in der Zielpopulation gültig und zuverlässig ist.

Abb. 1 Darstellung der Untersuchungswerte einer Test-Retest Prüfung im Koordinatensystem



Es ist kontrovers, ob die kulturelle Adaptation auch eine Prüfung der Sensitivität umfassen soll. Hier ist zu bemerken, daß aus rein metrischer Perspektive die Sensitivität mit der Zuverlässigkeit statistisch assoziiert ist. Anders als die Zuverlässigkeit ist die Sensitivität zudem nicht nur vom Instrument und der Population, sondern auch noch von einer Intervention abhängig. Genauso wie man die Zuverlässigkeit eines Instrumentes nur in bezug zu einer Population setzen kann, läßt sich auch die Sensitivität nur in bezug zur Untersuchungspopulation und der Intervention bestimmen. Damit wird deutlich, daß eine Prüfung der Sensitivität eher im Rahmen einer Pilotstudie zu einer spezifischen Studie und nicht vordringlich bei der kulturellen Adaptation sinnvoll ist. Wir verzichten deshalb an dieser Stelle auf eine Darstellung der Sensitivitätsprüfung und verweisen auf die Besprechung der Untersuchung der Sensitivität im letzten Teil der Arbeit zur Selektion von Outcome-Instrumenten.

Test-Retest-Zuverlässigkeit

Grundsätzlich ist die Test-Retest-Zuverlässigkeit kein statisches Charakteristikum eines Instrumentes, sondern eine Eigenschaft des Instrumentes in einer bestimmten Population. Entsprechend empfiehlt sich die Überprüfung der Zuverlässigkeit einerseits zur Testung des adaptierten Instrumentes und andererseits zur Abschätzung der Zuverlässigkeit in der Zielpopulation.

Zur Testung der Test-Retest-Zuverlässigkeit wird häufig der Pearson-Korrelationskoeffizient berechnet. Im Rahmen unserer Validierung des Western Ontario McMaster Universities (WOMAC) Osteoarthritis Index (8) fanden wir beim Vergleich der Schmerzskala einen hohen Pearson-Korrelationskoeffizienten von 0,92 zwischen den Werten der Ausgangsbestimmung und einer

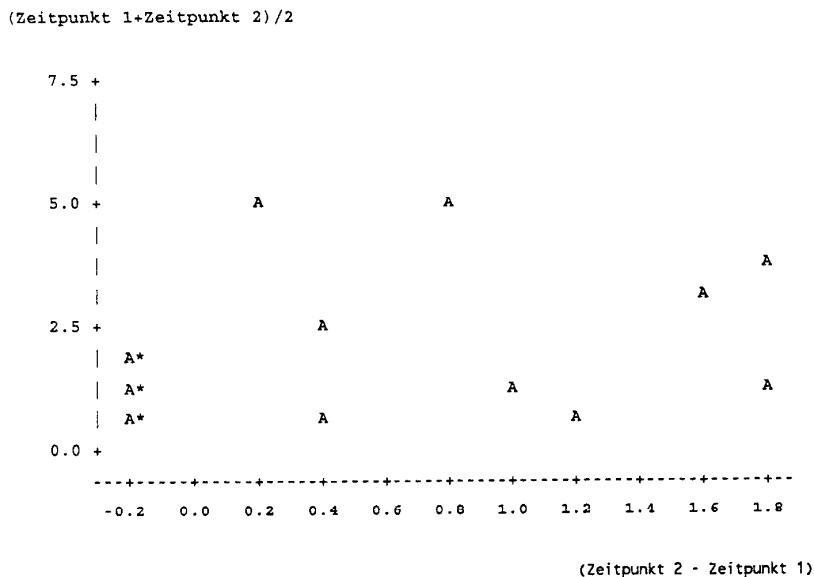
wiederholten Befragung nach 10 Tagen. Ein Wert von 1,0 stellt dabei eine perfekte, ein Wert von 0.0 eine fehlende Übereinstimmung dar. Eine etwas konservativere Einschätzung mit $r=0.87$ ergab sich bei Verwendung des Spearman Rang-Korrelationskoeffizienten, der für die ordinale WOMAC-Skala mit nicht-normaler Verteilung grundsätzlich vorzuziehen ist. Eine sicherere Bewertung der Test-Retest-Zuverlässigkeit setzt zudem eine größere Stichprobe voraus. Obwohl klare Richtlinien fehlen und für Validierungen wie in unserem Beispiel oft nur 10 Patienten getestet werden, sollten falls immer möglich ≥ 20 Patienten in die Prüfung mit einbezogen werden.

Das wesentliche Problem der Zuverlässigkeitsbeurteilung aufgrund von Korrelationskoeffizienten ist die Schwierigkeit des Erkennens systematischer Fehler. Bland und Altman (17) empfehlen deshalb neben der Angabe des Korrelationskoeffizienten zusätzlich die Darstellung der Untersuchungswerte in einem Achsen-Koordinatensystem und ergänzend die Gegenüberstellung der Mittelwerte aus den beiden Erhebungen und den Differenzen zwischen den Erhebungen. Die folgenden Abbildungen (Abb.1 und 2) zeigen die Resultate der Test-Retest-Prüfung der WOMAC-Schmerzskala (Pearson $r=0,92$, Spearman $r=0,87$).

In diesem Beispiel zeigt sich im XY-Koordinatensystem eine stark lineare Darstellung der Werte. Ein systematischer Unterschied zwischen den Zeitpunkten ist in der ersten Abbildung auf den ersten Blick nicht erkennbar. Hingegen genügt ein Blick auf die zweite Abbildung um festzustellen, daß nur drei Patienten einen geringeren Wert (markiert mit einem *), hingegen 9 Patienten einen höheren Wert bei der Zweituntersuchung aufwiesen.

Im Gegensatz zu den einfachen Korrelationskoeffizienten berücksichtigt und „bestraft“ der Intraklassen-Kor-

Abb. 2 Gegenüberstellung der Mittelwerte aus den beiden Erhebungen und den Differenzen zwischen den Erhebungen (Die 3 Patienten mit einem kleineren Wert bei der Zweituntersuchung sind mit einem * markiert)



relationskoeffizient („intraclass correlation coefficient“, ICC) (18) eine solche systematische Abweichung, da er nicht nur die Stärke der Assoziation (Korrelation), sondern auch eine Abweichung zwischen den Serien berücksichtigt. In unserem Beispiel war der ICC mit 0,70 zufriedenstellend, aber wie erwartet deutlich tiefer als der Korrelationskoeffizient.

Offen bleibt die Frage, ob diese Abweichung auf Zufall, einen Messfehler des Instrumentes oder eine wirkliche Zunahme der Schmerzen im Verlauf des 10-tägigen Intervalles zwischen der Beantwortung der Fragebogen zurückzuführen ist. Daß die Patienten zum Zeitpunkt 2 in der Tat stärkere Schmerzen hatten als zum Zeitpunkt 1 ist möglich, da mehr als die Hälfte in einer Übergangsfrage („transition-question“) angaben, stärkere Schmerzen zu verspüren. Zudem ist die systematische Abweichung zu relativieren: die maximal gefundene Differenz von 1,2 Punkten auf einer 1-10 Skala erscheint kaum relevant.

Interne Konsistenz: Eine spezielle Form der Zuverlässigkeit, die sogenannte interne Konsistenz, wird zur Beurteilung einer Gruppe von Fragen („items“) zur Erfassung eines definierten Konstrukts z. B. körperliche Funktionsfähigkeit) herangezogen. Entsprechend der Definition eines Konstruktes sind alle Parameter, die dieses reflektieren, miteinander assoziiert; demzufolge erwartet man eine signifikante Korrelation zwischen verschiedenen Parametern, die ein definiertes Konstrukt messen. Der Cronbach Koeffizient alpha ist eine statistische Messgröße zur Bestimmung des Zusammenhaltes oder der internen Konsistenz zwischen den Fragen (19). Ein Cronbach alpha von 0.65 wird im allgemeinen für

Tab. 2 Korrelation zwischen den Funktionsfragen (3A-3B) (21)

	3B	3C	3D
3A	0,76*	0,50*	0,66*
3B		0,49*	0,58*
3C			0,69*

* p<0.01

Tab. 3 Korrelation zwischen den Symptomfragen (21)

	1B	1C	1D	1E
1A	0,08	-0,07	0,20	0,41#
1B		0,41#	0,20	0,33
1C			0,15	-0,01
1D				0,22

p<0,05

* p<0,01

Studienzwecke, d. h. wenn alle Patienten gemeinsam analysiert werden, als unterer Grenzwert erachtet (18). Hingegen wird ein Cronbach alpha von >0,95 für die Beurteilung eines individuellen Patienten verlangt.

Die Tabellen 2 und 3 illustrieren das Konzept der internen Konsistenz für die Korrelation zwischen den Symptomfragen und den Funktionsfragen des Lequesne-Arthrose-Index für das Knie (20, 21) (Abb.3). Die Funktionsfragen (3A-3B) waren alle miteinander korreliert (Spearman's Rangkorrelation) und wiesen entsprechend einen hohen Cronbach Koeffizienten alpha

Abb. 3 Lequesne-Arthrose-Index für das Knie (20, 21)

KNIEFRAGEBOGEN

Bitte beantworten Sie die folgenden Fragen (kreuzen Sie das zutreffende Kästchen an).

1. Schmerzen oder Beschwerden:

- A. Nachts im Bett?
 - keine
 - nur bei Bewegungen oder in einigen Positionen
 - sogar in Ruhelage
- B. Beim Aufstehen?
 - keine
 - während weniger als 1/4 Stunde
 - während mehr als 1/4 Stunde
- C. Beim Stehen?
 - Nein
 - Ja
- D. Beim Gehen?
 - keine
 - erst nach einer gewissen Distanz
 - rasch und mit zunehmender Tendenz
- E. Beim Aufstehen aus einem Sessel ohne Mithilfe der Arme?
 - Nein
 - Ja

2. Wie weit können Sie maximal gehen?

- unbegrenzt
- begrenzt, jedoch über 1 km
- ungefähr 1 km (15 Min.)
- 500 - 900 m (8-15 Min.)
- 300 - 500 m
- 100 - 300 m
- weniger als 100 m

Welche Gehhilfen brauchen Sie dabei?

- keine
- 1 Stock oder 1 Krücke
- 2 Stöcke oder 2 Krücken

3. Schwierigkeiten im Alltag?

		ohne Schwierigkeiten	mit leichten Schwierigkeiten	mit mittleren Schwierigkeiten	mit grossen Schwierigkeiten	unmöglich
A.	Ein Stockwerk hinaufsteigen					
B.	Ein Stockwerk hinuntersteigen					
C.	In die Hocke gehen					
D.	Auf unebenem Boden gehen					

von 0,86 auf (Tabelle 2). Die Schmerzfragen (1A-1D) zeigten hingegen keinen, bzw. nur einen geringen Zusammenhang (Tabelle 3).

Entsprechend fand sich ein unbefriedigend tiefer Cronbach Koeffizient alpha von 0,55 (19, 18). Zwischen verschiedenen Fragen wie „Schmerzen nachts im Bett“ und „Schmerzen beim Stehen“ ($r = -0,07$) ergaben sich gar negative Rangkorrelationen. Eine mögliche Erklärung für die fehlende interne Konsistenz der Le-

quesne-Symptomfragen ist die Verwendung unterschiedlicher Gradierungen bei einzelnen Schmerzfragen. So wird einerseits nach „Dauer der Schmerzen“, „Auftreten von Schmerzen nach einer gewissen Distanz“ und andererseits nach „Schmerzen bei Bewegung/Ruhigstellung“ gefragt und deren Antwortprofile in einer Skala subsummiert. Entsprechend allgemeingültiger klinischer Erfahrungen und der Ergebnisse unserer Analyse kann nicht zwangsläufig davon ausgegangen werden, daß Pa-

Abb. 4 ACR-Funktionskategorien für Patienten mit cP (24)

ACR-Funktionsklasse	Beurteilung
I	Der Patient ist in keinem Bereich des täglichen Lebens (Selbstversorgung, berufliche Tätigkeit, Freizeit) eingeschränkt
II	Der Patient ist fähig, sich selbst zu versorgen und seiner beruflichen Tätigkeit nachzugehen; er ist jedoch in seiner Freizeit eingeschränkt
III	Der Patient ist fähig, sich selbst zu versorgen; er ist jedoch in seiner beruflichen Tätigkeit und in seiner Freizeit eingeschränkt
IV	Der Patient ist in allen Bereichen des täglichen Lebens (Selbstversorgung, berufliche Tätigkeit, Freizeit) eingeschränkt

tienten mit nächtlichen Schmerzen auch Schmerzen beim Gehen haben müssen. Interessanterweise waren im Gegensatz zu den Lequesne-Symptomfragen die entsprechenden Fragen des WOMAC, die ein einheitliches Gradierungsschema aufweisen (keine bis extreme Schmerzen), intern konsistent. Diese Beobachtungen unterstützten unsere Vermutung, daß nicht der Inhalt, sondern eher die Gradierung der Schmerzfragen für die unbefriedigende, interne Konsistenz der Symptomskala und daraus resultierend auch des Lequesne-Globalscores verantwortlich ist.

Übereinstimmungs-Validität (concurrent validity)

Zur Testung der *Übereinstimmungs-Validität* von Outcome-Instrumenten wird in der Regel die Patientenbeurteilung mit der auf einer externen Beobachtung beruhenden Beurteilung (z. B. Arzt, Therapeut) verglichen. Im Rahmen der Validierung der Originalversion des Health Assessment Questionnaire wurden z. B. die Patientenangaben zur Funktionsfähigkeit mit den Beobachtungen eines Forschungsassistenten verglichen, der jeden der Patienten zu Hause besuchte (4). Vergleichbare Studien zum AIMS (22) und zum HAQ für Patienten mit Sklerodermie (23) fanden insgesamt eine gute Übereinstimmung zwischen der Patienten- und der externen Expertenbeurteilung.

Kriterium-Validität (criterion validity)

Bei der Entwicklung eines neuen Instrumentes oder der kulturellen Adaptation kann ein Vergleich mit einem bereits etablierten Instrument durchgeführt werden. Während die Neuentwicklung eines Instrumentes oft eine bessere Differenzierung des gemessenen Konstruktes anstrebt, kann umgekehrt auch ein differenziertes Instru-

Tab. 4 Gegenüberstellung der ACR-Funktionsklassen und der physischen Funktionseinschränkung gemessen mit dem HAQ (24)

	Durchschnittliche HAQ-Scores gemäß ACR 1991 Funktionskriterien			
	I n=9	II n=19	III n=32	IV n=2
Essen	0,11	0,68	1,53	2,50
Reichen	0,11	0,79	1,81	3,00
Greifen	0,00	0,79	1,75	2,50
Anziehen	0,22	0,58	1,56	3,00
Hygiene	0,00	0,63	1,81	3,00
Aufstehen	0,00	0,21	0,84	1,50
Gehen	0,11	0,63	1,19	2,00
Aktivität	0,11	0,89	1,88	3,00
Global-Score	0,08	0,65	1,57	2,57

$p < 0,001$ für alle Dimensionen (Kruskal-Wallis-Test)

ment als externer Standard zur Beurteilung eines neuen vereinfachten und globaleren Instrumentes dienen.

So haben wir zur Validierung der modifizierten ACR-Funktionskategorien bei Patienten mit cP in einer Schweizer Population einen Vergleich mit dem HAQ, einem umfassenden Instrument zur Erfassung der physischen Funktionsfähigkeit, vorgenommen (24). Dabei fanden wir, daß die 4 ACR-Funktionsklassen (Abb.4), die primär zur Charakterisierung von Patienten in klinischen und epidemiologischen Studien und nicht zur Verlaufsbeurteilung gebraucht werden, das Kontinuum einer physischen Funktionseinschränkung wie sie der HAQ erfaßt, ausgezeichnet reflektierten (Tabelle 4).

Konstrukt-Validität (construct validity)

Die Konstrukt-Validierung basiert im Prinzip auf der Testung einer Hypothese. Dabei wird angenommen, daß

ein Instrument, welches ein Konstrukt wiedergibt, auch mit einem externen Kriterium statistisch korreliert, von dem man weiß, daß es mit dem Konstrukt assoziiert ist. Entsprechend wird analysiert, ob die Assoziation eines Instrumentes mit dem externen Standard in die erwartete Richtung zeigt.

Bei der Prüfung der Konstrukt-Validität einer deutschen Version des HAQ (15) wurden die Hypothesen getestet, daß diejenigen Patienten mit den besten HAQ-Werten eine kürzere Krankheitsdauer, eine geringere Krankheitsaktivität und eine weniger ausgeprägte radiologische Destruktion haben als die Patienten mit den schlechtesten HAQ-Werten.

In einer Vergleichsuntersuchung des modifizierten HAQ (MHAQ) (25) und des HAQ testeten wir auch die Hypothese, daß die physische Funktionsfähigkeit mit entsprechenden Veränderungen in klinischen und radiologischen Parametern assoziiert ist (9). Aufgrund der „Deckeneffekte“ des MHAQ war jedoch nur der HAQ mit der Veränderung in klinischen und radiologischen Parametern assoziiert. Gemäß unserer Hypothese ist damit der HAQ, nicht aber der MHAQ ein gültiges Instrument zur Verlaufserfassung der physischen Funktionsfähigkeit bei cP in einer ambulanten Population.

Aufgrund dieser Ausführungen wird deutlich, daß ein Messparameter oder eine Gruppe von Messparametern nicht per se valide ist, sondern daß die Validität schlußendlich nur im Hinblick auf eine gezielte Fragestellung beurteilt werden kann. Ein Outcome-Instrument mag durchaus valide sein für einen bestimmten, aber nicht für einen anderen Zweck.

Adaptation

Bevor man ein Outcome-Instrument vollständig neu entwickelt, sollte man unbedingt versuchen, ein bestehendes Instrument durch problemspezifische Fragen zu ergänzen. Man muß sich bewußt sein, daß die meisten Fragen bestehender Instrumente aus einem großen Fundus von immer wieder verwendeten Fragen kommen und bei einer Neuentwicklung die Gefahr besteht, daß auch ein „neues“ Instrument schlußendlich redundant ist. So decken z. B. der HAQ und der AIMS ein weites Spektrum von physischen Funktionseinschränkungen bei entzündlich-rheumatischen Krankheiten ab. Allerdings werden z. B. die spezifischen Probleme von Patienten mit Spondylarthropathien von keinem der beiden Instrumente berücksichtigt. Bei der Entwicklung von Instrumenten zur Erfassung der Funktionsfähigkeit von anderen entzündlichen Krankheiten wie der Spondylarthropathien ist es daher sinnvoll, auf den HAQ oder den AIMS zurückzugreifen und mit problemspezifischen Fragen zu ergänzen. Diesen Ansatz hat Daltroy bei der Entwicklung des HAQ-S für Spondylarthropathien gewählt (10). Es ist deshalb der

Ansatz einer sorgfältigen Ergänzung und Weiterentwicklung von bestehenden Instrumenten im Gegensatz zu einer Neuentwicklung zu befürworten. Dies läßt sich am Beispiel des HAQ-S illustrieren.

Item Auswahl

Für die Item-Auswahl kommt grundsätzlich ein Experten- oder ein empirischer Ansatz in Frage (26). Bei einem Expertenansatz werden mögliche Fragen durch Experten, oft nach einer eingehenden Literaturstudie und informellen Patientenbefragungen zusammengestellt. Dieser Ansatz hat den Nachteil, daß möglicherweise nicht alle von Patienten als relevant erachteten Funktionseinschränkungen berücksichtigt werden. Der Hauptnachteil dürfte aber darin liegen, daß mit diesem Ansatz der von den Patienten verwendete und damit für andere Patienten verständliche Wortlaut nicht gewährleistet ist.

Daltroy wählte deshalb zusätzlich zum Experten- den empirischen Ansatz. Dabei wurden 300 britische Patienten mit Spondylarthropathie in bezug auf ihre funktionellen Einschränkungen befragt. Die Patientenaussagen wurden dabei wortwörtlich erfaßt und dann entsprechend der Bedeutung einzelner Problembereiche in einer Prioritätenliste geordnet.

Testung

Im nächsten Schritt wurden die am häufigsten genannten Symptome und Funktionsstörungen in ihrem typischen Wortlaut in einem Fragebogen zusammengestellt und einer neuen Patientengruppe vorgelegt. In einer statistischen Analyse wurde dann der Zusammenhalt der verschiedenen Fragen untereinander untersucht und redundante Fragen identifiziert. Das Ziel war es, einerseits so wenig Fragen wie möglich zu verwenden und andererseits eine ausreichende interne Konsistenz zu gewährleisten. Erfahrungsgemäß nimmt diese bei weniger als 5 Fragen deutlich ab. Daltroy fand, daß sich die spezifische Rückenproblematik von Patienten mit Spondylarthropathien mit 5 Fragen umfassend und zuverlässig erfassen läßt.

Validierung

Bei der Entwicklung wird das Instrument mit statistischen Analysen für die Patienten der Testpopulation maßgeschneidert. In einer anderen Population werden die metrischen Eigenschaften in der Regel deshalb etwas weniger gut sein. Es war deshalb wichtig, daß die metrischen Eigenschaften des HAQ-S in einer neuen Population getestet wurden. Der HAQ-S hat sich auch in anderen Populationen bewährt und es wurde kürzlich

in einer Vergleichsstudie gezeigt, daß der IIAQ-S im Hinblick auf die Erfassung der physischen Funktionsfähigkeit anderen Spondyloarthropathie-spezifischen Instrumenten überlegen war (27).

Neuentwicklung

Die Neuentwicklung eines Instrumentes sollte man dann vornehmen, wenn kein geeignetes Instrument zur Verfügung steht oder wie am Beispiel des HAQ-S ergänzt werden kann. Zwei Einschränkungen sollten an dieser Stelle jedoch nicht unerwähnt bleiben: Erstens ist die Neuentwicklung auch von Kurzinstrumenten in der Regel zeit- und kostenintensiv. Die Beschreibung des Prozesses zur Erweiterung des HAQ für Patienten mit Spondyloarthropathien um nur wenige Fragen verdeutlicht diesen Punkt. Zweitens wird in den letzten Jahren zunehmend die „item response theory“ anstatt der „generalizability theory“ für die Entwicklung von Outcome-Instrumenten verwendet (28, 29). Dabei wird die Unidimensionalität eines Konstruktes nicht mehr ausschließlich über das Konzept der internen Konsistenz und Faktorenanalyse studiert, sondern zunehmend durch die in der Psychologie und zum Teil in der Rehabilitationsmedizin schon seit längerem angewandte Rasch-Methodologie (30). Die Theorie der Item-Empfindlichkeit hat dabei offensichtliche theoretische Vorteile. Sie ermöglicht die Entwicklung von Skalen mit Intervalleigenschaften, was einen unschätzbaren Vorteil sowohl in der Interpretation als auch der Analyse darstellt. In einer Rasch-Analyse werden die Fragen nach ihrem Schwierigkeitsgrad geordnet. Nur Fragen mit einem gut definierten Schwierigkeitsgrad werden in die Konstruktskala integriert. Zudem wird angestrebt, über den ganzen Schwierigkeitsbereich in gleichen Abständen entsprechende Fragen einzuschließen. Bei Mehrfachrepräsentation eines Schwierigkeitsgrades wird die Frage mit der geringeren Zuverlässigkeit im Rahmen der Rasch-Rangordnung ausgeschlossen oder eine Gewichtung vorgenommen.

Intervall-Daten (z. B. die Körpertemperatur in Grad Celsius) haben gegenüber ordinalen Daten (Antwortkategorien mit ungleichen, respektive nicht definierten Abständen, z. B. kein, leicht, mittel, schwer) den großen Vorteil, daß sie mit parametrischen Methoden ana-

lysiert werden können. Praktisch alle in der Rheumatologie verwendeten Instrumente haben hingegen ordinale Eigenschaften. Ordinale Skalen haben den Nachteil, daß z. B. der Abstand zwischen 1 (z. B. leicht) und 2 (mittelstark) nicht gleich groß sein muß wie zwischen 2 (mittelstark) und 3 (stark). 2 ist dementsprechend auch nicht unbedingt doppelt so groß wie 1. Bei der statistischen Analyse müssten deshalb Methoden verwendet werden, die den ordinalen Eigenschaften dieser Instrumente Rechnung tragen. Generell kommen in Querschnitts-Untersuchungen nicht-parametrische statistische Methoden, z. B. der Spearman Rangkorrelationskoeffizient, zur Anwendung. Im Falle von Längsuntersuchungen muß die Richtungsänderung und nicht zwingend die Stärke einer Assoziation berechnet werden. Entsprechend haben wir z. B. bei der Untersuchung der Assoziation von HAQ-Werten und einer Veränderung der Muskelkraft im Langzeitverlauf über ein Jahr die Tendenz der Veränderung (besser, unverändert, schlechter) und nicht die Stärke der Änderung in IIAQ-Werten evaluiert (9).

Zur Zeit sind verschiedene Entwickler von Outcome-Instrumenten, z. B. des SF-36 (31), damit beschäftigt, ihre Instrumente bezüglich ihrer Intervalleigenschaften zu untersuchen und durch eine neue Gewichtung der Fragen im Rahmen der Auswertung Skalen mit Intervalleigenschaften zu produzieren. Sinnvollerweise sollten auch bei jeder Neuentwicklung eines Instrumentes die Möglichkeiten der Rasch-Methodologie geprüft werden. Dies bedingt die intensive Zusammenarbeit mit einem Biostatistiker oder Psychometriker mit Erfahrung in der Rasch-Methodologie. Wegen der hohen methodologischen Ansprüche und der Notwendigkeit, ausreichende Fallzahlen einzuschließen, ist dies in der Regel mit einem sehr hohen Aufwand verknüpft.

Zusammenfassung

Die kulturelle Adaptation bzw. die Erweiterung vorhandener Outcome-Instrumente ist einer Neuentwicklung unbedingt vorzuziehen. Gleichzeitig müssen jedoch strikte Kriterien erfüllt werden, um die Gültigkeit, Zuverlässigkeit und Empfindlichkeit eines Instrumentes in einer anderen Sprache oder in einer anderen Kultur zu gewährleisten.

Referenzen

1. Ware JE, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 30:473-483
2. McDowell IM, Martini CJM, Waugh W (1978) A method for self-assessment of disability before and after hip replacement operations. *BMJ* 2:857-859
3. Hunt SM, McKenna S, McEwen J, Williams J, Papp E (1981) The Nottingham health profile: subjective health status and medical consultations. *Soc Sci Med* 15A:221-229
4. Fries JF, Spitz PW, Kraines RG, Holman HR (1980) Measurement of patient outcome in arthritis. *Arthritis Rheum* 23:137-145
5. Meenan RF, Gertman PM, Mason JH (1980) Measuring health status in arthritis: the Arthritis Impact Measurement Scales. *Arthritis Rheum* 23:146-152
6. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE (1992) AIMS 2. *Arthritis Rheum* 35:1-10
7. Bellamy N (1995) WOMAC Osteoarthritis Index. A user's guide. University of Western Ontario, London, Ontario, Canada
8. Stucki G, Meier D, Stucki S, Michel BA, Tyndall AG, Dick W, Theiler R (1996) Evaluation einer deutschen Version des WOMAC (Western Ontario and McMaster Universities) Arthritisindex. *Z Rheumatol* 55:40-49
9. Stucki G, Stucki S, Brühlmann P, Michel BA (1995) Ceiling effects of the health assessment questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. *Ann Rheum Dis* 54:461-465
10. Daltroy LH, Larson MG, Roberts WN, Liang MH (1988) A modification of the Health Assessment Questionnaire for the spondyloarthropathies. *J Rheumatol* 17:946-950
11. Guillemin F, Bombardier C, Beaton D (1993) Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol* 12:1417-1432
12. Guillemin F (1995) Measuring health status across cultures. *Rheumatol Europe (EULAR)* 24 (Suppl. 2):102
13. Bullinger M, Anderson R, Cella D, Aaronson N (1993) Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Qual Life Res* 2 (6):451-459
14. Guyatt GH, Sackett DL, Cook DJ (1993) User's guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 270:2598-2601
15. Brühlmann P, Stucki G, Michel BA (1994) Evaluation of a German version of the physical dimensions of the health assessment questionnaire in patients with rheumatoid arthritis. *J Rheumatol* 21:1245-1249
16. Ferraz MB, Oliveira LM, Araujo PM, Atra E, Tugwell P (1990) Crosscultural reliability of the physical ability dimension of the health assessment questionnaire. *J Rheumatol* 17 (6):813-817
17. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 307-310
18. Nunally JC (1978) Assessment of Reliability. In: *Psychometric Theory*. McGraw-Hill, New York
19. Cronbach L (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297
20. Lequesne M, Méry C, Samson M, Gérard P (1987) Indexes of severity for osteoarthritis of the hip and knee. *Scand J Rheumatol* 65 (Suppl):85-89
21. Stucki G, Meier D, Stucki S, Michel BA, Tyndall AG, Elke R, Theiler R (1996) Evaluation einer deutschen Fragebogen-version der Lequesne Cox- und Gonarthrose-Indices. *Z Rheumatol* 55:50-57
22. Jacobs JWG, Oosterveld FGJ, Deurbouts N, Rasker JJ, Taal E, Dequeker J, Uytendhoeven K (1992) Opinions of patients with rheumatoid arthritis about their own functional capacity: how valid is it. *Ann Rheum Dis* 51:765-768
23. Poole JL, Williams CA, Bloch DA, Holak B, Spitz P (1995) Concurrent Validity of the Health Assessment Questionnaire Disability Index in Scleroderma. *Arthritis Care Res* 8 (3):189-193
24. Stucki G, Stoll T, Brühlmann P, Michel BA (1995) Construct validation of the ACR 1991 revised criteria for global functional status in rheumatoid arthritis. *Clin Exp Rheumatol* 13:349-352
25. Pincus T, Summey JA, Soraci SA, Wallston KA, Hummon NP (1983) Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. *Arthritis Rheum* 26:1346-1353
26. Bombardier C, Tugwell P (1982) A methodological framework to develop and select indices for clinical trials: Statistical and judgmental approaches. *J Rheumatol* 9:753-757
27. Kuzis K, Ward MM (1994) Measurement of functional disability in ankylosing spondylitis: a comparison of four self-report questionnaires. *Arthritis Rheum* 9 (Suppl.):226
28. Merbitz C, Morris J, Grip JC (1989) Ordinal scales and foundations of misinference. *Arch Phys Med Rehab* 70:308-312
29. Silverstein B, Fisher WP, Kilgore KM, Harley JP, Harvey RF (1992) Applying psychometric criteria to functional assessment in medical rehabilitation: II. Defining interval measures. *Arch Phys Med Rehab* 73:507-518
30. Wright BD (1977) Solving measurement problems with the Rasch model. *J Educ Res* 14:97-116
31. Haley SM, McHorney CA, Ware JE (1994) Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): I. Unidimensionality and reproducibility of the Rasch Item Scale. *J Clin Epidemiol* 47:671-684