

G. Stucki
S. Stucki
O. Sangha

Patienten-zentrierte Evaluation der Krankheitsauswirkungen bei muskuloskelettalen Erkrankungen: Auswahl und Testung von Outcome-Instrumenten

Patients-oriented outcome assessment in musculoskeletal disease: selection and testing of instruments

Zusammenfassung Die Untersuchung der Effektivität und Kosten-Wirksamkeit von Interventionen, die Evaluation von epidemiologischen Studien der Krankheitsauswirkungen sowie Versorgungsstudien und klinisches Qualitätsmanagement bedingen alle die standardisierte Erfassung von patienten-relevanten Parametern mit validierten, metrisch getesteten, zuverlässigen und verlauf-

empfindlichen Fragebögen. Dabei ist die sorgfältige Auswahl und Testung des Outcome-Instrumentariums in der Planungsphase aus wissenschaftlichen, ethischen und schließlich auch ökonomischen Gründen ausschlaggebend.

Die Auswahl von Instrumenten umfaßt die Recherche mit Hilfe von medizinischen Datenbanken (z. B. MEDLINE), die Prüfung der Face-Validität (mißt ein Instrument, was wir zu messen beabsichtigen?) und der Kompatibilität (wird das Instrument international verwendet?). Die Testung von Instrumenten umfaßt die Prüfung der Zuverlässigkeit (reliability), der internen Konsistenz (internal consistency) und der Empfindlichkeit (sensitivity). Wichtig ist die Prüfung der praktischen Eignung (Interpretation der Skalen und Scores, Akzeptanz in der Studienpopulation). Vor Einsatz eines Instrumentes empfiehlt sich die Rücksprache mit den Entwicklern (Copyright Fragen, Scoring, aktuelle Version).

and clinical quality management all rely on standardized assessment of disease consequences with psychometrically sound questionnaires. For scientific, ethical and economical reasons, careful selection and evaluation of instruments is critical.

Selection of instruments includes searches of medical databases (e.g. MEDLINE), testing of face-validity (does the instrument measure what we intend to measure?), and compatibility (is the instrument used internationally?). Evaluation of instruments includes the assessment of reliability, internal consistency and sensitivity. Most important is careful consideration of the practical usefulness (Interpretation of scores and scales, acceptance in the study population). Contact with instrument developers is advisable (Copyright issues, scoring, current version).

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

Schlüsselwörter Health Status – Outcome – Patienten Assessment – Instrumentenauswahl

Summary Effectiveness research, economic evaluation, epidemiologic studies of disease consequences

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

Hintergrund

Die Untersuchung der Effektivität und Kosten-Wirksamkeit von präventiven, therapeutischen oder rehabilitativen Interventionen und, auf einer globalen Ebene, die Evaluation von epidemiologischen Studien der Krank-

heitsauswirkungen („Out-come“) sowie Versorgungsstudien und klinisches Qualitätsmanagement bedingen alle die standardisierte Erfassung von patienten-relevanten Parametern mit validierten, metrisch getesteten, zuverlässigen und verlaufempfindlichen Fragebogen. Dabei ist die sorgfältige Auswahl und Testung des Outcome-

Instrumentariums in der Planungsphase aus wissenschaftlichen, ethischen und schlußendlich auch ökonomischen Gründen ausschlaggebend.

Eine exploratorische Analyse der kollektiven Erfahrung im Rahmen einer epidemiologischen Studie oder die Hypothesen testende Analyse im Rahmen einer Wirksamkeits-Studie sind nur dann aussagekräftig, wenn die Instrumente für die Fragestellung gültig sind (Validität), die Fallzahl ausreichend ist, um einen für Patienten relevanten Unterschied mitzuerfassen (Sensitivität), die Patienten das Instrumentarium akzeptieren, d. h. Fragebogen als relevant betrachten und z. B. die Studie nicht wegen einer fehlenden Akzeptanz vorzeitig abbrechen bzw. Fragebogen nur unvollständig ausfüllen (1).

Die Kompatibilität mit vergleichbaren Projekten und Studien ist nur dann gewährleistet, wenn standardisierte, international anerkannte Outcome-Instrumente zur Anwendung kommen. Dies ist vor allem dann wichtig, wenn z. B. bei Medikamentenstudien eine Zulassung der Substanz in verschiedenen internationalen Märkten vorgesehen ist. Letzteres hat in der Vergangenheit zur Definition von grundsätzlichen Parametern („core sets“) geführt (2, 3), die demzufolge in jede klinische Studie einbezogen werden sollten.

Das Gebiet der Outcome-Forschung befindet sich zur Zeit im Fluß. Die Auswahl eines für die jeweilige Fragestellung geeigneten Instrumentes läßt sich deswegen nur bedingt anhand von Büchern oder kochbuchartigen Listen durchführen. Ein Instrument, das heute geeignet erscheint, ist morgen eventuell bereits durch eine verbesserte Version oder ein gänzlich neues Instrument überholt. Zudem bedingt jede Fragestellung ein spezifisches Set von Outcome-Instrumenten.

Als Teil eines Qualitätssicherungsprogrammes, das jede klinische Studie begleiten sollte, ist es sinnvoll, die metrischen Eigenschaften der Instrumente, insbesondere deren Empfindlichkeit und Akzeptanz, in einer kleinen Pilot-Studie an einer repräsentativen Stichprobe von 10-20% der zu erwartenden Studiengröße zu erproben. Dies hat den zusätzlichen Vorteil, daß Fallzahlberechnungen basierend auf diesen Daten überprüft und relativiert werden können. Von entscheidender Bedeutung ist die Schätzung der Effektgröße („minimal important change“) (4, 5). Diese hängt einerseits von der Empfindlichkeit des Outcome-Instrumentes und andererseits von der Auswahl der Patienten ab und ist schlußendlich für die Berechnung der Stichprobengröße bestimmend. Die Literatur ist voll von Beispielen, bei denen Untersucher diese Vorsichtsmaßnahme nicht berücksichtigt haben und demzufolge wegen einer unzureichenden statistischen Aussagekraft („Power“) nicht in der Lage waren, klinisch relevante Unterschiede mit Sicherheit nachzuweisen (6).

In dieser Arbeit wird ein systematisches Vorgehen bei der Auswahl von Outcome-Parametern vorgeschlagen und die Pilottestung von Outcome-Instrumenten anhand von eigenen Beispielen illustriert (Tabelle 1).

Tab. 1 Systematisches Vorgehen bei der Auswahl von Outcome-Parametern

Vorgehen	Überlegungen
Definition Krankheitsmodell	Identifikation Krankheits/Problem relevante Outcome-Dimensionen Primäre Ergebnisgrößen? Sekundäre Ergebnisgrößen?
Recherche	Datenbanken (Medline, DIMDI) Durchsicht der verwendeten Instrumente in früheren Studien Expertenbefragung
Prüfung Face-Validität	Mißt ein Instrument, was wir zu messen beabsichtigen?
Kompatibilität	Instrumente international verwendet? Core sets von professionellen Organisationen abgedeckt?
Metrische Eigenschaften	Zuverlässigkeit Interne Konsistenz Empfindlichkeit Meßbereich Benötigte Stichprobengröße
Praktische Eignung	Interpretation Akzeptanz in der Studienpopulation Zeitliche Belastung
Rücksprache mit Instrumenten-Entwicklern	Instrument geeignet für Fragestellung? Scoring? Neueste Version? Copyright?

Definition des Krankheitsmodells

Zur Auswahl der Instrumente ist es entscheidend, ein *Modell der relevanten Outcome-Dimensionen* zu erstellen und die Ziel-Dimensionen zu definieren (7). Für die meisten Fragestellungen, sei es für ein Meßsystem im Rahmen einer Beobachtungs- oder einer Interventionsstudie, ist es in der Regel sinnvoll, alle patientenrelevanten Outcome-Dimensionen mitzuerfassen. Dies ermöglicht die Erstellung eines Profils der Veränderung während einer Interventions- oder einer Beobachtungsstudie. Allerdings ist es wichtig, primäre und sekundäre Zieldimensionen in der Planungsphase festzulegen. In der Regel können aus praktischen Gründen (Gesamtzahl der Fragen sollte möglichst gering sein) nur die primären Zieldimensionen mit differenzierten und entsprechend ausführlichen Outcome-Instrumenten erfaßt werden. Hingegen werden diejenigen Dimensionen, die in einer Beobachtungsstudie nur von untergeordneter Bedeutung sind oder bei denen im Rahmen einer Intervention keine Veränderung abzusehen ist, nur mittels Kurzversionen (8) oder nur mit einer globalen Frage („Wie stark ist Ihre funktionelle Beeinträchtigung?“ „Wie stark sind Ihre Schmerzen?“) erfaßt. Minimalstandards („core-sets“), z. B. für chronische Polyarthrit, basieren in der Regel auf diesem Konzept (2, 3).

Im Rahmen der Untersuchung von Rehabilitations-Interventionen mit dem Ziel der Lösung eines für jeden

Patienten individualisierten Problems ist der Einsatz von Instrumenten geeignet, die patientenspezifische Präferenzen erfassen (9). In der Rheumatologie ist dabei der McMaster Toronto Arthritis Patient Preference Disability Questionnaire (MACTAR) (9) am weitesten verbreitet. In einem teilstrukturierten Interview werden die Patienten gefragt, welche fünf Aktivitäten ihnen am meisten Mühe bereiten. Im Verlauf wird dann die Verbesserung in diesen fünf Schlüsselaktivitäten erfaßt.

Recherche

Die Durchsicht der Literatur nach geeigneten Outcome-Instrumenten ist immer lohnenswert (10–13). Medizinische Datenbanken (z. B. MEDLINE, DIMDI) haben in ihren aktuellen Versionen entsprechende Schlagwörter definiert (14), so daß eine gezielte Recherche möglich ist. Die Suche nach Textwörtern in Titeln und Abstracts erlaubt die Identifizierung von Studien, die ein bestimmtes Instrument eingesetzt haben. Entscheidend ist es, in den gefundenen Artikeln nach zitierten weiteren Instrumenten zu fahnden und über die Bibliographie zu suchen. Verschiedene Fachzeitschriften veröffentlichen in regelmäßigen Abständen Zusammenstellungen aller publizierten Instrumente (Medical Care 1990, Quality of Life Research 1995).

Neben der direkten Suche nach Outcome-Literatur lohnt sich auch der Weg über das Problem (15). Bei diesem Ansatz werden Studien, die eine ähnliche Fragestellung bearbeitet haben, identifiziert und nach den verwendeten Outcome-Instrumenten durchgesehen. Damit läßt sich erkennen, welche Instrumente häufig verwendet werden und in neueren und qualitativ hochstehenden Studien zum Einsatz kamen.

Vielleicht der wichtigste Ansatz überhaupt ist die Kontaktnahme mit auf dem entsprechendem Gebiet erfahrenen klinischen Forschern. Nur diese sind in der Regel über aktuelle Trends und noch nicht publizierte Richtlinien informiert. So befindet sich z. B. zur Zeit ein neues Instrumentarium der North American Spine Society (NASS) zur Erfassung der Krankheitsauswirkungen bei Rückenpatienten in Publikation (16). Das Instrument basiert auf den Erfahrungen verschiedener bisheriger Instrumente, wird durch eine standespolitisch wichtige Gesellschaft unterstützt und wird möglicherweise das dominante Instrument der Zukunft für Rückenpatienten.

Prüfung der Face – und inhaltlichen Validität

Nach Festlegung des Outcome-Modells und der Zieldimensionen werden verschiedene in Frage kommende Instrumente auf ihre Eignung hin inspiziert. Entscheidend

ist deren Gültigkeit: *Mißt das Instrument, was wir zu messen beabsichtigen?* Dies wird oft als die Face-Validität eines Instrumentes beschrieben und ist nichts anderes als ein subjektives Urteil darüber, ob eine Gruppe von Meßparametern das definierte Konstrukt (Zielinhalt) mißt. Ein wichtiger Aspekt dabei ist die sogenannte inhaltliche Gültigkeit („Content-Validity“): Werden mit dem Instrument alle relevanten Aspekte erfaßt? Wie die Face-Validität beruht auch die inhaltliche Validität auf einem subjektiven Urteil. Gemäß unserer klinischen Erfahrung wäre z. B. ein Patientenfragebogen zur physischen Funktionsfähigkeit bei Patienten mit chronischer Polyarthritis (cP), der nur die Funktionsfähigkeit der unteren Extremitäten messen würde, nicht umfassend. Obwohl ein solches Meßinstrument durchaus andere Validitätskriterien erfüllen würde und z. B. mit der Blutsenkung korreliert, wäre es zur Beurteilung der globalen physischen Funktionsfähigkeit bei Patienten mit chronischer Polyarthritis nicht gültig. Möchte man hingegen nur den Erfolg von Hüftoperationen bei cP-Patienten erfassen, hätte ein Instrument, das spezifisch die physische Funktionsfähigkeit der unteren Extremitäten erfaßt, durchaus Face-Validität. Je nach Fragestellung kann also das Instrument „valid“ oder „nicht valid“ sein. Entscheidend bei der Auswahl ist, daß Fragen und Inhalte eines Instrumentes im Detail studiert und bezüglich Face – und inhaltlicher Validität beurteilt werden.

Prüfung der Kompatibilität

Bei der Auswahl sollte der internationale Kontext vor Augen gehalten werden. In einer Ära verstärkter internationaler Zusammenarbeit auf dem Gebiet der klinischen Forschung und der Versorgungsforschung ist es wichtig, standardisierte Instrumente einzusetzen, die für verschiedene Kulturen übersetzt und adaptiert wurden (17). Ein ausgezeichnetes, aber nur in der deutschen Sprache vorliegendes Instrument mag durchaus für Studien zur Gesundheitsversorgung im eigenen Land geeignet sein. Ist die Studienfrage aber von internationaler Bedeutung oder ist die Publikation in einer englischsprachigen Zeitschrift vorgesehen, ist es sinnvoll, ein international anerkanntes Instrumentarium anzuwenden.

Mit der zunehmenden Bedeutung von Meta-Analysen (Integration der Resultate von verschiedenen Studien zur gleichen Fragestellung) ist es entscheidend, alle im Rahmen von Core-sets empfohlenen Parameter zu messen. Core sets stellen in der Regel den kleinsten gemeinsamen Nenner dar und beinhalten meist nur eine oder wenige Fragen pro Dimension. So beinhalten die vorläufigen Kriterien des American College of Rheumatology (ACR) zur Beurteilung von Basistherapeutikastudien bei der cP drei globale Fragen an den Patienten

(Schmerzen, Funktionsfähigkeit, Globaler Gesundheitszustand) (2). Selbstverständlich muß man darüber hinaus die Zieldimensionen mit differenzierteren Instrumenten genauer erfassen.

Metrische Eigenschaften

Zuverlässigkeit

Die Zuverlässigkeit eines Outcome-Instrumentes beinhaltet seine Reproduzierbarkeit. Mit einem zuverlässigen Instrument erreicht man bei wiederholten Messungen unter vergleichbaren Bedingungen Ergebnisse, die nur in einem sehr engen Bereich variieren. Bei der klinischen Untersuchung wird in diesem Zusammenhang die Intra – von der Inter-Untersucherzuverlässigkeit unterschieden. Die Intra-Untersucher-Zuverlässigkeit bestimmt dabei die Abweichung eines einzelnen Untersuchers bei wiederholten Messungen (wenn ein Patient in einem gewissen Abstand einen Fragebogen zweimal ausfüllt), während die Inter-Untersucherzuverlässigkeit die Meßvariation zwischen verschiedenen Untersuchern bestimmt (z. B. Patient und Physiotherapeut füllen unabhängig einen Fragebogen zur physischen Funktionsfähigkeit aus). Statistische und graphische Konzepte zur Bestimmung der Zuverlässigkeit werden im nächsten Teil beschrieben.

Das Fehlen von Angaben zur Zuverlässigkeit eines Instrumentes stellen dessen Eignung grundsätzlich in Frage. Da die Zuverlässigkeit nicht nur vom Instrument, sondern auch von der Population, in der ein Instrument zum Einsatz kommt, abhängt, ist es sinnvoll, eine Schätzung der Zuverlässigkeit im Rahmen einer Pilotstudie vorzunehmen.

Interne Konsistenz

Eine spezielle Form der Zuverlässigkeit ist die interne Konsistenz. Sie mißt die Beziehung einzelner Fragen eines Outcome-Instrumentes untereinander und wird meist mit dem Cronbach Koeffizienten Alpha (18,19) gemessen. Eine Diskussion dieses Konzeptes findet sich ebenfalls in der Arbeit zur Adaptation und Neuentwicklung von Outcome-Instrumenten.

Bei der Selektion eines Instrumentes sollten Angaben zur internen Konsistenz gezielt überprüft werden. Viele, z.T. weitverbreitete Instrumente lassen in der Literatur Angaben zur internen Konsistenz vermissen. So läßt sich anhand der Literatur nicht festlegen, inwieweit der Lequesne Index (20, 21) zur Symptomerfassung und Funktionseinschätzung bei Arthrosen der unteren Extremitäten jemals in Hinblick auf dessen interne Konsistenz untersucht wurde. In Untersuchungen zur Be-

stimmung der Leistungsfähigkeit bei Arthrose der unteren Extremitäten sollte man deswegen einem Instrument mit nachgewiesener interner Konsistenz, z. B. dem Western Ontario McMaster Universities (WOMAC) Osteoarthritis Index (22, 23), den Vorzug geben.

Ein Cronbach alpha von 0.65 wird im allgemeinen für Studienzwecke, d. h. wenn alle Patienten gemeinsam analysiert werden, als unterer Grenzwert erachtet. Hingegen wird ein Cronbach alpha von >0.95 für die Beurteilung eines individuellen Patienten verlangt (19).

Empfindlichkeit, Meßbereich und Stichprobengröße

Ein Meßparameter, der für eine bestimmte Fragestellung Gültigkeit hat und reproduzierbar ist, muß deswegen nicht zwangsläufig auch als Verlaufsparemeter geeignet sein. Ein Parameter ist empfindlich, wenn er bei einer klinischen Verschlechterung schlechtere Werte anzeigt, bei unverändertem klinischem Zustand gleich bleibt und sich bei einem klinisch positiven Effekt verbessert (5, 24). Für die Empfindlichkeit ausschlaggebend ist das sogenannte Signal-Rauschen Verhältnis („signal to noise ratio“) (24). Ein Parameter ist umso empfindlicher, je stärker das Signal und je geringer das (Hintergrund) Rauschen (Variation bei klinisch unverändertem Zustand) ist. Im Sinne einer weiteren Differenzierung des Konzeptes wird oft zwischen der Sensitivität („sensitivity“) als reine statistische Empfindlichkeit und des Ansprechverhaltens („responsiveness“) als Fähigkeit, einen minimalen, klinisch relevanten Unterschied (4, 25) zu erfassen, unterschieden (5).

Wenn verschiedene Instrumente grundsätzlich für eine bestimmte Studie in Frage kommen, kann durch einen direkten Vergleich der Empfindlichkeit der Instrumente in einer Pilotstudie das am besten geeignete Instrument identifiziert werden (26, 27) (Tabelle 2).

Zur Testung der Empfindlichkeit von Outcome-Instrumenten werden verschiedene statistische Techniken angewandt

- Der standardisierte Mittelwert der Empfindlichkeit („standardized response mean“) (26, 28, 29) als Mittelwert der Differenz von zwei Meßzeitpunkten dividiert durch die Standardabweichung der Differenzen (x/SD der Differenz),
- die Effektgröße („effect size“) (30) als Mittelwert der Differenz von zwei Meßzeitpunkten dividiert durch die Standardabweichung zum Zeitpunkt 1 (x/SD zum Zeitpunkt 1) und
- Guyatt's Empfindlichkeits-Statistik (31) („responsiveness statistic“) als Mittelwert der Differenzen dividiert durch die Standardabweichung bei stabilen Patienten (x/SD stabiler Zustand).

Bei einer Vergleichsuntersuchung zur Eignung der zwei am häufigsten verwendeten Instrumente zur Erfas-

Tab. 2 Beispiel zur Testung: Vergleich des Kurzfragebogens mit 36 Fragen (Short-Form 36 (SF-36)) (32) und des „Sickness Impact Profile“ (SIP) (33) zur Messung von Veränderungen im Krankheitsverlauf nach Hüftgelenksersatzoperation (27)

Konzept	Fragestellung	Methode	Globaler SIP	Körperlicher SIP	Globaler SF-36	Physischer SF-36
Relevanz	Wieviele Fragen der wichtigsten physischen Skalen sind relevant?	Anzahl der Fragen welche für >50% der Patienten ein Problem darstellen	na	Gehfähigkeit und Beweglichkeit: 1 of 22	na	physische Funktionsfähigkeit 9 of 10
	Wieviele Skalen sind relevant?	Anzahl der Skalen mit <40% der Patienten mit perfektem Score	5 von 12	3 von 3	7 von 8	2 von 2
Empfindlichkeit	Welches Instrument zeigt eine klinische Veränderung am stärksten an?	Standardized Response Mean (durchschnittliche Veränderung/durchschnittliche Standardabweichung der Veränderung)	0.96	0.88	1.06	1.26
	Welches Instrument korreliert am besten mit der Verbesserung der Lebensqualität?	Spearman Rangkorrelation	0.19	0.26*	0.45**	0.37**
Deckeneffekt	Häufung von Patienten am normalen Ende der Skala?	Verteilung: Anzahl der Patienten mit perfektem Score	Häufung nach 3 Monaten; 4 Patienten mit perfektem globalem SIP und 8 Patienten mit perfektem physischen SIP	Keine Häufung, keine Patienten mit perfekten Scores		
	Diskriminiert das Instrument den Gesundheitszustand bei Patienten mit nahezu perfekten Scores?	Korrelation mit einem klinisch relevanten äußeren Kriterium	Keine Korrelation mit der Fähigkeit zu arbeiten, Gartenarbeiten zu verrichten oder Sport zu betreiben	Korrelation mit der Fähigkeit zu arbeiten, Gartenarbeiten zu verrichten (r=0.79) oder Sport zu betreiben (r=0.88)		
Bodeneffekt	Häufen sich die Patienten am „schlechten“ Skalenende?	Verteilung: Anzahl Patienten mit schlechtest möglichem Score	Keine Häufung; keine Patienten mit schlechtest möglichem Score	Keine Häufung; keine Patienten mit schlechtest möglichem Score		

* Signifikant bei p<0.05
** Signifikant bei p<0.01
na=nicht anwendbar

sung des allgemeinen Gesundheitszustandes (Short-Form 36 (SF-36) (32) und dem Sickness Impact Profile (SIP) (33, 34)) haben wir zur Untersuchung den standardisierten Mittelwert der Empfindlichkeit untersucht (27). Dabei war die Skala zur physischen Funktionsfähigkeit des SF-36 mit 1.26 deutlich empfindlicher als die entsprechende Skala des SIP mit 0.88. In Anwendung eines von Deyo (35) vorgeschlagenen statistischen Verfahrens korrelierte die mit dem SF-36 gemessene Verbesserung auch besser mit der vom Patienten am Studienende angegebenen Verbesserung der allgemeinen Lebensqualität.

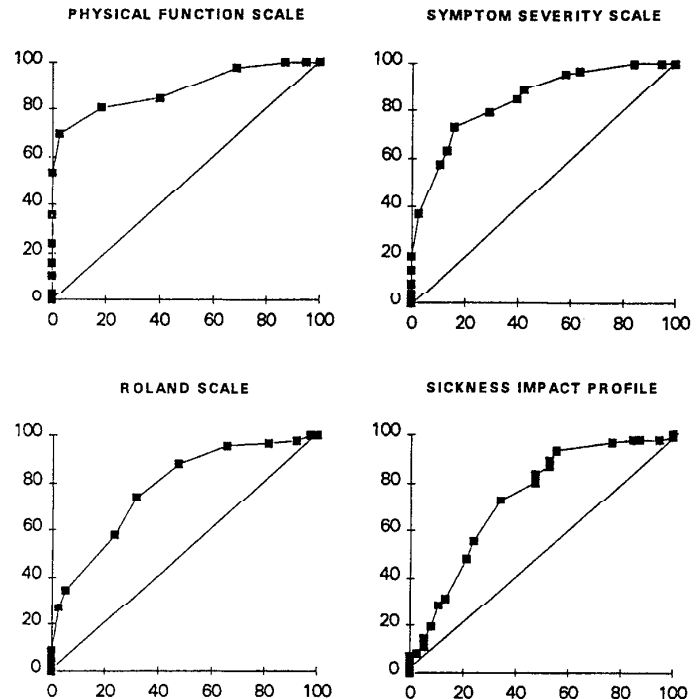
In einer Vergleichsuntersuchung zur Empfindlichkeit eines neu entwickelten Spinalstenose-Instrumentes haben wir neben den erwähnten statistischen Ansätzen zusätzlich ROC (Receiver Operator Characteristic)-Kurven (35) evaluiert (Abb.1) (26).

Dabei wurde die Zufriedenheit der Patienten mit dem Operationsergebnis als dichotomes (gut/schlecht) externes Kriterium verwendet. Für verschiedene Definitionen des als relevant betrachteten Unterschiedes zwischen Zeitpunkt 2 und Zeitpunkt 1 kann dann die echt-positive Rate (Sensitivität) auf der vertikalen (y) Achse und

die falsch-positive Rate (1-Spezifität) auf der horizontalen (x) Achse aufgetragen werden. Durch Verbinden der Punkte ergibt sich eine ROC-Kurve, die idealerweise möglichst weit nach „links oben“ verläuft (hohe Sensitivität und hohe Spezifität). Der Kurvenverlauf und die Fläche unter der Kurve (c-Werte aus einer logistischen Regression mit „Zufriedenheit“ als abhängige Variable) zeigten, daß die kurzen Symptom – und Funktionsskalen des Spinalstenose-Instrumentes besser zwischen zufriedenen und unzufriedenen Patienten differenzierten als der SIP oder nicht Spinalstenose-spezifische Roland-Index (26). Am Punkt, der am nächsten zur linken oberen Ecke des Quadranten liegt, beträgt die Sensitivität der Funktionsskala 80.4% und die Spezifität 81.6%. Hingegen betrug die Sensitivität des SIP im besten Fall nur 72.8% und die Spezifität nur 65.8%. Die gefundene höhere Empfindlichkeit des Spinalstenose-Instrumentes wurde durch die übereinstimmenden Resultate der Empfindlichkeitsstatistiken und der ROC-Kurven-Analyse bestätigt.

Unterschiede in der Empfindlichkeit von Outcome-Instrumenten haben wichtige Konsequenzen bei der Berechnung von Fallzahlen in klinischen Studien (36, 31,

Abb. 1 ROC (Receiver Operator Characteristic)-Kurven von 4 Outcome-Instrumenten für Patienten mit Spinalstenose (27)



26). Verwendet man z. B. den SIP als Outcome-Instrument zur Erfassung der physikalischen Funktionsfähigkeit bei Spinalstenose, beträgt die benötigte Stichprobengröße pro Behandlungsarm (zweiseitiger Alpha-Wert 0.05, und „power“ 0.80) 153 Patienten. Hingegen beträgt die benötigte Fallzahl bei Verwendung des Spinalstenose-Instrumentes nur 49 Patienten pro Behandlungsarm. Die Verwendung des auf dem SIP basierenden und speziell für Rückenprobleme adaptierten Roland-Fragebogen (26) würde 94 Patienten pro Behandlungsarm benötigen.

Unser Beispiel verdeutlicht, daß es sich durchaus lohnt, das empfindlichste Instrument für eine bestimmte Fragestellung in einer Pilotstudie zu identifizieren. Die Verwendung eines krankheitsspezifischen Instrumentes als primären Studienendpunkt ist dabei vor allem aus Effizienzgründen vorteilhaft.

Der Hauptgrund der besseren Effizienz der krankheitsspezifischen Instrumente im letzten Beispiel liegt im besseren „Signal-Rauschen“-Verhältnis. Da sich in der Regel alle krankheitsrelevanten Einzelfragen mit Veränderung des klinischen Zustandes verändern, erzeugen sie ein starkes Signal, und da sich für das Zustandsbild keine allgemeine Fragen im krankheitsspezifischen Instrument befinden, wird kaum „Rauschen“ erzeugt. Der Grund, weshalb der SF-36 in unserem ersten Beispiel (Patienten nach Hüftendoprothese) besser als der SIP zur Verlaufsbeurteilung geeignet ist, liegt hingegen

in den unterschiedlichen Meßbereichen beider Instrumente. Wiederum war es entscheidend, in einer Pilotstudie zu untersuchen, ob das Meßinstrument in der Lage ist, die Patienten in bezug auf das zu messende Konstrukt, insbesondere in den Extrembereichen, zu differenzieren (Abb.2).

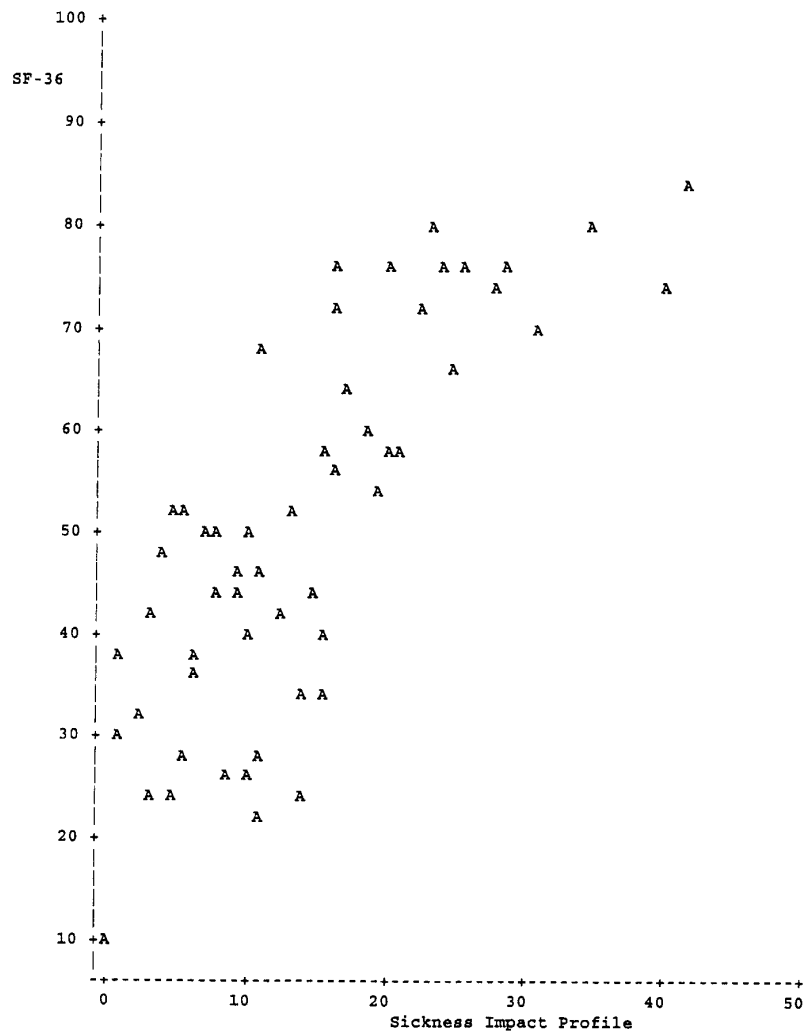
Falls die Patienten bereits zu Beginn (z. B. vor einer Therapie) einen perfekten (normalen) Wert aufweisen, hat das Meßinstrument einen „Deckeneffekt“ („ceiling-effect“) (27, 37). Umgekehrt, wenn eine größere Anzahl von Patienten den schlechtestmöglichen Wert aufweisen und damit eine mögliche Verschlechterung nicht mehr erfaßt werden kann, handelt es sich um einen „Boden“-Effekt („floor effect“) (38, 27).

Abb.3 zeigt den Deckeneffekt der physischen Skala des Sickness Impact Profile (P-SIP) (33, 34) für Patienten mit Hüftarthrose, welche sich einer Hüftgelenks-Operation unterzogen haben.

Die Abbildung illustriert, daß sich Patienten mit den besten SIP-Werten zu Studienbeginn nicht verbessern konnten. Entsprechend konnte für diese Patienten auch keine Korrelation mit einem externen Standard, z. B. der Arbeitsfähigkeit, Gartenarbeit oder der Ausübung von Sport, festgestellt werden (24).

Einen ausgeprägten „Deckeneffekt“ fanden wir auch in einer Vergleichsstudie (37) des Health Assessment Questionnaire (HAQ) und einer Kurzversion (modified HAQ „MHAQ“) dieses Instrumentes (39, 37) (Abb.4).

Abb. 2 Scatter-plot des SIP und SF-36 zu Studienbeginn. Der Plot illustriert den weiteren Wertebereich des SF-36 im Vergleich zum SIP. Zudem zeigt sich eine Anhäufung von Werten des SIP im tiefen Wertebereich (cave: die Werte des SIP waren nie höher als 40- die Skala wurde deshalb aus Platzgründen gekürzt) (27).



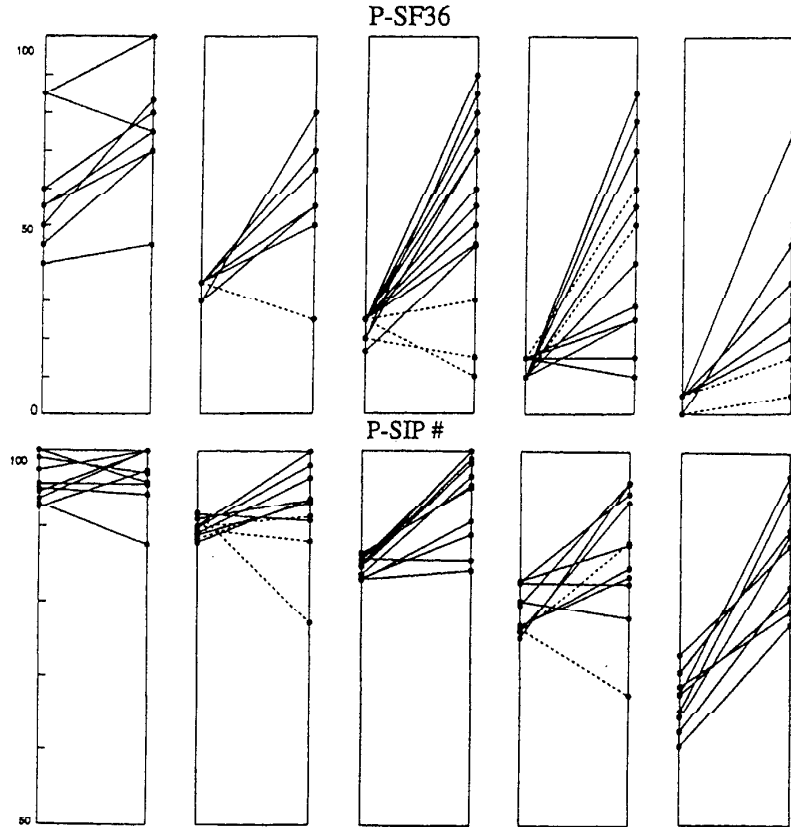
In den Stamm- und Blätterdiagrammen (40, 37) zeigt sich eine Anhäufung von Patienten mit normalen Werten (keine Einschränkung der physischen Funktionsfähigkeit entsprechend einem Score von 0), die vor allem für den MHAQ ausgeprägt war. Allerdings hat auch der HAQ eine deutlich zweigipflige Verteilung und eine Konzentration gegen 0. Ein „ideales“ Instrument würde die ganze Bandbreite der Skala ausnutzen, d. h. im Falle ausreichender Patientenzahlen zu keiner Konzentration an den Extremen führen. International werden zur Zeit erweiterte Versionen des HAQ mit schwierigeren ergänzenden Fragen geprüft, um die Skala zu spreizen. Basierend auf unserer Vergleichsuntersuchung des HAQ und des MHAQ verwenden wir in der klinischen Routine nur noch den HAQ zur Erfassung der physischen Funktions Einschränkung und verzichten auf die Verwendung des MHAQ.

Das Beispiel des MHAQ verdeutlicht die Problematik von Kurzversionen. Sofern Einzelfragen, die das Ende eines Spektrums (hier z. B. besonders schwierige körperliche Tätigkeiten) ausgelassen werden, kommt es zu einem Decken- oder Bodeneffekt und damit auch zu einer Verminderung der Instrumenten-Empfindlichkeit.

Praktische Eignung

Neben den metrischen Eigenschaften der gewählten Meßparameter sind praktische Überlegungen von grosser Bedeutung. Sowohl bei klinischen Parametern, die durch den Arzt oder Therapeuten erhoben werden, als auch bei patientenzentrierten Outcomes gilt, sowenig wie möglich und soviel wie nötig. Aus Kosten- und Akzeptanz-

Abb. 3 Deckeneffekt der physischen Skala des Sickness Impact Profile (P-SIP) (33) für Patienten mit Hüftarthrose, welche sich einer Hüftgelenks-Operation unterzogen haben. Verlaufskurven einzelner Patienten sind stratifiziert nach Basiswert-Fünfteln dargestellt.



gründen sollte die Gesamtzahl möglichst klein gehalten, redundante Parameter vermieden und der Einsatz von für die Untersuchungspopulation irrelevanten Variablen vermieden werden. Entsprechend der Beurteilung metrischer Eigenschaften bedarf auch die Beantwortung der praktischen Eignung oft einer Pilotuntersuchung.

In unserer Untersuchung zur Evaluation verschiedener Instrumente zur Erfassung des allgemeinen Zustands bei Patienten nach Hüfttotalprothesenoperation verglichen wir die Relevanz der Einzelfragen für Patienten mit Gon- oder Coxarthrose (27). Die Patienten haben dabei systematisch die SIP-Skalen mit Fragen nach Schwierigkeiten beim Essen, beim Sprechen oder der Hygiene ausgelassen, da sie diese in keiner Hinsicht als für sich relevant betrachteten (Tabelle 2). Auch waren die Fragen zum Gehen und zur Beweglichkeit des SIP für die untersuchte Population weniger relevant als die Fragen zur physischen Funktionsfähigkeit des SF-36.

Neben der Relevanz der Fragen ist die benötigte Zeit zum Ausfüllen eines Fragebogens entscheidend. Sowohl in der klinischen Routine als auch im Rahmen von Studien sollte die benötigte Zeit 10-15 Minuten nicht überschreiten (1, 7). Verschiedene Instrumente erfordern den

Einsatz von geschulten Interviewern (z. B. Lequesne Index (20)) oder die Verwendung von standardisierten Hilfsmitteln (z. B. der Juvenile Arthritis Functional Assessment Scale, JAFAS (41)) zur Erfassung der Funktionsfähigkeit bei jugendlichen Patienten. Generell kann gesagt werden, daß Patientenfragebogen praktikabler sind als Interview-basierte Instrumente. Allerdings erlauben bei multiethnischen Populationen mit einem geringen Bildungsstand oft nur auf einem standardisierten Interview basierte Instrumente eine zuverlässige Informationserfassung.

Rücksprache mit den Entwicklern

Vor Durchführung einer größeren, prospektiven und aufwendigen Studie empfiehlt sich auf jeden Fall die Rücksprache mit den Entwicklern der Instrumente. „Serious“ Instrumenten-Entwickler „betreuen“ ihre Instrumente im Verlauf. So stellt z. B. die Entwicklergruppe des Medical Outcome Trust (u.a. SF-36 (32)) in regelmäßigen Abständen ein aktualisiertes Manual zur Verfügung, welches aktuelle Erfahrungen inkorporiert.

Abb. 4 „Deckeneffekt“ des Health Assessment Questionnaire (HAQ) und einer Kurzversion (modified HAQ „MHAQ“) dieses Instrumentes (37).

Baseline HAQ*	#
26 5	1
24	
22 588	3
20 0000222	7
18	
16 2555555	7
14 000	3
12 5588888	7
10 0000022	7
8	
6 25	2
4 0	1
2 555558888	9
0 000000022	9

Baseline MHAQ*	#
26	
24	
22	
20 5	1
18 8	1
16 5	1
14	
12 55	2
10 002	3
8 888	3
6 2225555555	10
4 0000	4
2 55555588	8
0 0000000000000000000002222222	23

Der Stamm reicht von 0 bis 26 (entsprechend einem Score von 0 bis 2.6).
 Jede Beobachtung wird mit einer Stamm - und Blätterkomponente dargestellt.
 # Gibt die Anzahl der Patienten mit gleichem Stammscore an.

Diese Entwicklergruppe unterhält in verschiedenen Ländern enge Kontakte zu Wissenschaftlern (im Falle des SF-36 zum Medizinisch Psychologischen Institut der Ludwig-Maximilians-Universität München), um länderspezifische Erfahrungen zu verarbeiten (17).

Die Rückfrage bei erfahrenen Outcome-Forschern und/oder Experten für die jeweilige Studienfrage erlaubt zudem die Einschätzung, welches der vorhandenen Instrumente grundsätzlich zum Einsatz kommen sollte. Wichtig ist auch, über die neueste Version und den aktuellen Auswertungsalgorithmus zu verfügen. Viele Instrumente werden kontinuierlich weiterentwickelt; so wurde die Ursprungsversion der Arthritis Impact Measurement Scales (AIMS) (42) um weitere Skalen und Fragen erweitert und grundsätzlich überarbeitet (AIMS 2) (43). Die Entwicklergruppe des SF-36 validiert zur Zeit eine verkürzte Version (SF-12), die für bestimmte Fragestellungen möglicherweise geeigneter ist als die Vollversion (44). Entwickler von Instrumenten klären auch über urheberrecht-

liche Bestimmungen auf. Die meisten in der Rheumatologie zur Anwendung kommenden Instrumente wurden durch nationale Forschungsprogramme finanziert und stehen daher akademisch tätigen Forschungsgruppen ohne Entschädigungspflicht zur Verfügung. Werden Instrumente aber in Industrie-finanzierten Medikamentenstudien verwendet, besteht meist eine Vergütungsverpflichtung, die individuell geklärt werden muß.

Zusammenfassung

Die Auswahl geeigneter Meßinstrumente für die klinische Routine oder für Studienzwecke ist eine aufwendige, aber wichtige Aufgabe. Sie vermeidet unnötige Kosten und etwaige Frustrationen wenn erhoffte Ergebnisse aufgrund ungeeigneter Instrumente und möglicher Akzeptanzprobleme bei den Patienten nicht erfaßt werden können.

Referenzen

1. Sangha O, Stucki G, Liang MH (1996, im Druck) Outcome assessment in rheumatology. In: Oxford Textbook of Rheumatology
2. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, Katz LM, Lightfoot R, Paulus H, Strand V, Tugwell P, Weinblatt M, Williams HJ, Wolfe F, Kieszak S (1995) American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 38: 727-735
3. Lequesne M (1993) ILAR guidelines for testing slow acting drugs in osteoarthritis (SYSADOAs). *Rev Esp. Rheumatol* 20 (Suppl.1):220-221
4. Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Ascertaining the minimal clinically important difference. *Cont Clin Trials* 10:407-15
5. Fortin PR, Stucki G, Katz JN (1995) Measuring relevant change: an emerging challenge in rheumatologic clinical trials. *Arthritis Rheum* 38:1027-1030
6. Cohen J (1977) Statistical power analyses for the behavioral sciences. Academic Press, New York
7. Stucki G, Brühlmann P, Michel BA (1995) Verlaufsbeurteilung bei chronischer Polyarthritits: neue quantitative Ansätze für die Praxis. *Schweiz Med Wochenschr* 125:2003-2014
8. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH (1992) Comparative measurement sensitivity of short and longer health status instruments. *Med Care* 30:917-925
9. Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E (1987) The MACTAR Questionnaire-An individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *J Rheumatol* 14:446-451
10. Guyatt GH, Rennie D (1993) User's guides to the medical literature. *JAMA* 270:2096-2097
11. Oxman AD, Sackett DL, Guyatt GH (1993) User's guides to the medical literature. I. How to get started. *JAMA* 270:2093-2095
12. Guyatt GH (1993) The philosophy of health-related quality of life translation. *Qual Life Res* 2 (6):461-465
13. Jaeschke R, Guyatt G, Sackett DL (1994) User's guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA* 271:389-391
14. Lowe HJ, Barnett GO (1994) Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 271:1103-1108
15. Sackett DL, Haynes RB, Guyatt GH, Tugwell P (1991) *Clinical epidemiology. A basic science for clinical medicine.* Little Brown
16. Daltroy LH, Larson MG, Roberts WN, Liang MH (1988) A modification of the Health Assessment Questionnaire for the spondyloarthropathies. *J Rheumatol* 17:946-950
17. 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
18. Cronbach L (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297
19. Nunally JC (1978) *Assessment of Reliability.* In: *Psychometric Theory.* McGraw-Hill, New York
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 Fragebogenversion der Lequesne Cox- und Gonarthrose-Indices. *Z Rheumatol* 55: 50-7
22. Stucki G, Meier D, Stucki S, Michel BA, Tyndall AG, Dick W, Theiler R (1996) Evaluation einer deutschen Version des WOMAC (Western Ontario und McMaster Universities) Arthroseindex. *Z Rheumatol* 55:40-49
23. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW (1988) Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 15:1833-1840
24. Stucki G, Michel BA (1995) How to measure improvement: rules and fallacies. *Rheumatol Europe* 24 (Suppl. 2): 107-111
25. Redelmeier DA, Lorig K (1995) Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. *Arch Intern Med* 153: 1337-1342
26. Stucki G, Liang MH, Fossel AH, Katz JN (1995) Relative responsiveness of condition specific and generic health status measures. *J Clin Epidemiol* 48: 1369-1378
27. Stucki G, Liang MH, Philipps C, Katz JN (1995) The short form-36 is preferable to the SIP as a generic health status measure in patients undergoing elective total hip arthroplasty. *Arthritis Care Res* 8:174-181
28. Liang MH, Larson MG, Cullen KE, Schwartz JA (1983) Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 28:542
29. Liang MH, Fossel AH, Larson MG (1990) Comparison of five health status instruments for orthopedic evaluation. *Med Care* 28:632-642
30. Kazis LE, Anderson JJ, Meenan RF (1989) Effect sizes for interpreting changes in health status. *Med Care* 27:178-89
31. Guyatt G, Walter S, Norman G (1987) Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chron Dis* 40:171-178
32. 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
33. Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS (1976) The Sickness Impact Profile: Validation of a health status measure. *Med Care* 14:57-67
34. Bergner M, Bobbitt RA, Carter WB, Gilson BS (1981) The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 19:787-805
35. Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Contr Clin Trials* 12:142-158
36. Rosner B (1990) Estimation of sample size and power for comparing two means. In: *Fundamentals of Biostatistics.* PWS-Kent, Boston, pp 273-275
37. 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
38. Bindman AB, Keane D, Lurie N (1990) Measuring health changes among severely ill patients. The floor phenomenon. *Med Care* 28:1142-1152
39. 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
40. Tukey JW (1977) *Exploratory Data Analysis.* Addison-Wesley publishing company, Reading, Massachusetts
41. Lovell DH, Howe S, Shear E, Hartner S, McGirr G, Schulte M, Levinson J (1989) Development of a disability measurement tool for juvenile rheumatoid arthritis. *Arthritis Rheum* 32:1390-1395

42. Meenan RF, Gertman PM, Mason JH (1980) Measuring health status in arthritis: The Arthritis Impact Measurement Scales. *Arthritis Rheum* 23:146-152
43. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE (1992) AIMS 2. *Arthritis Rheum* 35:1-10
44. Ware JE, Kosinski M, Keller SD (1996) A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Med Care* 34:220-233