0303-8408/96/050303-12 $ 1.50 + 0.20/0

Soz Präventivmed 1996; 41: 303–314

Ulrike Kliebsch[1], Hermann Brenner[2]

[1] *Münchner Forschungsverbund Public Health – Öffentliche Gesundheit, München*
[2] *Abteilung Epidemiologie der Universität Ulm*

Inter-Rater-Reliabilität von Instrumenten zur Beurteilung der Pflegebedürftigkeit: Ein Review der internationalen Literatur

Zusammenfassung

Ausgehend von der Notwendigkeit einer differenzierten, objektiven Erfassung der Pflegebedürftigkeit bei der Gewährung von Leistungen für Schwerpflegebedürftige im Rahmen des Sozialversicherungswesens wird mit diesem Literaturreview ein Überblick gegeben über international verbreitete Instrumente zur Beurteilung der Pflegebedürftigkeit unter besonderer Berücksichtigung der Inter-Rater-Reliabilität. Es wurden elf Instrumente identifiziert. Diese stammen bis auf zwei Ausnahmen aus den USA und wurden zwischen 1963 und 1988 entwickelt. Die Zahl und Qualität der Untersuchungen zur Inter-Rater-Reliabilität sind sehr heterogen. Für die Mehrzahl der Instrumente konnten gute Reliabilitätsergebnisse gefunden werden, während für einige Verfahren keine oder nur sehr lückenhafte Untersuchungen vorliegen. Auffällig ist, dass die Inter-Rater-Reliabilität für die Gesamtscores der Instrumente durchweg höher ausfällt, als die Reliabilität der Bewertung der einzelnen Merkmale. Aus den Ergebnissen zu den unterschiedlichen Beurteilungen für die einzelnen Merkmale können Anregungen für die Weiterentwicklung des im Bereich des Sozialversicherungswesens eingesetzten Begutachtungsverfahrens zum Vorliegen von Pflegebedürftigkeit gewonnen werden.

Im Rahmen des Gesundheitsreformgesetzes wurden in der Bundesrepublik Deutschland erstmals ab dem 01.01.1989 Ansprüche auf häusliche Pflegehilfe an die gesetzliche Krankenversicherung für zu Hause lebende Schwerpflegebedürftige wirksam. Mit Einführung der Pflegeversicherung ab dem 01.01.1995 wurden die bis dahin einheitlichen Sach- und Geldleistungen weiter ausgedehnt und nach Grad der Pflegebedürftigkeit differenziert. Das Antragsverfahren für diese Leistungen sieht eine standardisierte sozialmedizinische Begutachtung durch den Medizinischen Dienst der Krankenversicherung (MDK)[1] vor. Trotz Standardisierung bleibt jedoch für die Entscheidung über das Vorliegen von Schwerpflegebedürftigkeit bzw. den Schweregrad der Pflegebedürftigkeit ein subjektiver Ermessensspielraum des untersuchenden Arztes. In Anbetracht der enormen sozialrechtlichen Konsequenzen besteht die Forderung nach einer objektiven, reproduzierbaren und differenzierten Erfassung der Schwerpflegebedürftigkeit. Für eine Optimierung des Begutachtungsverfahrens sind wissenschaftliche Untersuchungen zur Reliabilität und Validität einschlägiger Begutachtungsverfahren vordringlich. Eine Reliabilitätsstudie zum Begutachtungsverfahren des MDK wurde letztes Jahr von uns in Bayern durchgeführt. Mit diesem Literaturreview soll den Fragen nachgegangen werden, welche Instrumente zur Beurteilung der Pflegebedürftigkeit international im Einsatz sind, welche Angaben zur Inter-Rater-Reliabilität dieser Instrumente vorliegen und welche Implikationen sich hieraus für eine Weiterentwicklung des Begutachtungsinstrumentariums in der Bundesrepublik ergeben könnten.

Begriffsdefinition

Zunächst gilt es, den Begriff Pflegebedürftigkeit zu definieren. Für diese Recherche wird Pflegebedürftigkeit, analog zur sozialmedizinischen Begutachtung des MDK, über den erforderlichen Hilfebedarf für die im deutschsprachigen Raum gebräuchlichen

„Aktivitäten im täglichen Leben (ATL)" festgelegt. In der internatinalen Literatur spricht man von den „Activities of daily living (ADL)" oder, so weit komplexere Funktionen betroffen sind, von den „Instrumental Activities of daily living (IADL)".
Hier geht es primär um Messinstrumente zur häuslichen oder ambulanten Pflege, so dass der Bereich der stationären Krankenpflege in medizinischen Versorgungseinrichtungen bewusst ausgespart wird. So weit wie möglich abzugrenzen sind ebenfalls Erhebungsinstrumente zur Lebensqualität, die mehr Wert legen auf Aspekte wie Lebenszufriedenheit („well-being") oder Schmerzempfinden, für die Defizite nicht zwangsläufig zu einem erhöhten Pflegebedarf führen. Eine strikte Trennung in diesem Bereich ist äusserst schwierig, so dass auch bei den hier recherchierten Instrumenten teilweise Überlappungen zu Lebensqualitäts-Skalen auftreten.

Methodik

Praktisches Vorgehen bei der Literaturrecherche

Die internationalen Datenbanken MEDLINE und SOMED wurden zur Recherche von Publikationen in medizinischen und sozialmedizinischen Zeitschriften verwendet. Als Einstieg wurde für den Zeitraum 1988–1994 in MEDLINE mit logischen Verknüpfungen (AND, OR) englischer Schlagworte des kontrollierten Vokabulars („Activities-of-daily-living", „Activities-of-daily-living-classification", „Disability-evaluation", „Health-status-indicators", „Patient-care-planning", „Occupational-therapy-methods", „Geriatric-Assessment", „Self-care", „Reliability", „Observer-Variation") sowie in SOMED mit logischen Verknüpfungen überwiegend deutscher Schlagworte („Pflegebedürftigkeit", „Sozialversicherungswesen", „Activities-of-daily-living", „Funktionstüchtigkeit", „Reliabilität", „Beobachtervariabilität") nach relevanten Publikationen gesucht. Desweiteren dienten einschlägige Handbücher neueren Datums als Informationsquellen[2,3]. Zur Komplettierung der Literaturrecherche, insbesondere bezüglich älterer Arbeiten, wurden die Literaturlisten der so gefundenen Publikationen systematisch nach weiteren relevanten Literaturstellen durchgesehen.

Einteilung der Instrumente

Als Einteilungskriterium der Instrumente dienten die erfassten Funktionen. Alle recherchierten Instrumente decken einen oder mehrere der folgenden drei Bereiche ab:

1. Elementare Aktivitäten des täglichen Lebens (ATL bzw. activities of daily living, ADL) wie „Essen", „Waschen", „WC-Benutzen", „An- und Ausziehen" etc.
2. Kognitive Funktionen, wie „Hören", „Sich Erinnern", „Ängstlich sein" etc. Einschränkungen in diesem Bereich können ebenfalls zu Pflegebedürftigkeit führen.
3. Komplexere Funktionen des täglichen Lebens (instrumental activities of daily living, IADL) wie „Einkaufen", „Mahlzeiten vorbereiten", „Wäsche waschen" etc.

Entsprechend wurde die folgende Einteilung der Instrumente vorgenommen:

- Reine ADL-Skalen
- ADL-Skalen, erweitert um kognitive Funktionen
- Reine IADL-Skalen
- Umfassende Skalen aller drei Bereiche

Recherchierte Merkmale

Bei der Beurteilung von Instrumenten zur Feststellung der Pflegebedürftigkeit interessierten folgende Merkmale:

- Die von den Instrumenten erfassten Funktionen, eingeteilt in ADL, kognitive Funktionen und IADL
- Das Setting und Design der Studien zur Inter-Rater-Reliabilität
- Die berechneten Reliabilitätsmasse sowohl für den Gesamtscore der Instrumente, als auch für einzelne Merkmale.

Es sei an dieser Stelle erwähnt, dass sich die Übersetzung der englischsprachigen Instrumente an den deutschen Begriffen des MDK-Erhebungsinstrumentes orientiert, d.h. es wurde teilweise statt der wörtlichen eine sinngemässe Übersetzung vorgenommen, um direkte Vergleiche der Instrumente untereinander zu ermöglichen.

Beurteilung der Reliabilität

Die vorliegende Arbeit befasst sich mit der Inter-Rater-Reliabilität von Instrumenten, d.h. mit der Frage, inwieweit unterschiedliche Untersucher bei der Beurteilung desselben Probanden zu gleichen Ergebnissen kommen. Ausgeklammert werden Intra-Rater-Reliabilitäts- oder Test-Retest-Reliabilitäts-Aspekte, die die zeitliche Konstanz der Beurteilung bei identischen Untersuchern bzw. Erhebungsinstrumentarien betreffen.
Mit folgenden Kenngrössen bzw. Masszahlen werden Aussagen zur Reliabilität quantifiziert:
Für quantitative Merkmale kommt am häufigsten der Pearson-Korrelationskoeffizient r (Produkt-Moment-Korrelationskoeffizient)[4] zur Anwendung, der eine Normalverteilung des betreffenden Merkmals voraussetzt. Eine Alternative bildet der verteilungsfreie Spearman-Rang-Korrelationskoeffizient rs (Rho)[4]. Beide Korrelationskoeffizienten haben einen Wertebereich zwischen –1 und 1, wobei die Randwerte maximale Diskrepanz

bzw. maximale Übereinstimmung bedeuten. Sie können nur bei nicht mehr als zwei Messungen pro Patient verwendet werden. Ein weiterer Nachteil besteht darin, dass unterschiedliche Durchschnittswerte in beiden Messungen keine Berücksichtigung finden. Ein Mass, das diese Unterschiede berücksichtigt und zugleich allgemeiner für mehr als zwei Messungen pro Patient und für unterschiedliche Skalenniveaus einsetzbar ist, ist der aus der Varianzanalyse abgeleitete Intraclass-Korrelationskoeffizient ICC[5], dessen Maximalwert 1 beträgt. Ein Reliabilitätsmass für kategoriale Daten ist der Kappa-Koeffizient κ[5]. Er berücksichtigt die zufällige Übereinstimmung von verschiedenen Messungen und kann für ordinale Merkmale ungewichtet oder mit Gewichten für den Grad der Nichtübereinstimmung berechnet werden. Am häufigsten sind lineare (κ_w) oder quadratische (κ_w^2) Gewichte. Bei maximaler Übereinstimmung erreicht der Koeffizient den Wert 1. Bei nicht perfekter Übereinstimmung ist die Höhe des Kappa-Koeffizienten auch von der Prävalenz der Merkmalsausprägung abhängig. Grob orientierend werden Kappa-Koeffizienten zwischen 0,4 und 0,75 als mässig bis gut bezeichnet[6]. Entspricht die beobachtete gerade der zufällig erwartenden Übereinstimmung, so ist der Koeffizient 0. Ist sie sogar geringer als nach dem Zufallsprinzip zu erwarten, kann der Koeffizient negative Werte annehmen. Gelegentlich wird auch die prozentuale exakte Übereinstimmung[7] einzelner Items angegeben, die allerdings nur begrenzt aussagefähig ist, da sie die zufällige Übereinstimmung nicht mitberücksichtigt. Welche der vorgestellten Masszahlen zur Anwendung kommt, hängt also insbesondere vom jeweiligen Skalenniveau und von der Versuchsanordnung ab.

Skalentyp	Name bzw. Abkürzung	Autoren	Jahr
Reine ADL-Skalen	ADL (Katz)	Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW[9]	1963
	Barthel	Mahoney FI, Barthel DW[10]	1965
	Kenny SCE	Iversen IA, Silberberg NE, Stever RC, Schoening HA[11]	1973
	PSMS	Lawton MP, Brody E[12]	1969
ADL-Skalen, erweitert um kognitive Funktionen	FIM	Keith RA, Granger CV, Hamilton BB, Sherwin FS[13]	1987
	FSRS	Forer SK[14]	1981
	LTCMDS	National Center for Health Statistics[15]	1980
Reine IADL-Skalen	FAQ	Pfeffer RI, Kurosaki TT, Harrah CH, Chance JM, Filos S[16]	1982
	IADL-Skala (L & B)	Lawton MP, Brody EM[12]	1969
Umfassende Skalen	ICIDH	World Health Organization[17]	1980
	SMAF	Hebert R, Carrier R, Bilodeau A[18]	1988
	RDRS-2	Linn WM, Linn BS[19]	1982

Tabelle 1. *Recherchierte Instrumente.*

Ergebnisse

Recherchierte Instrumente

Tabelle 1 zeigt eine Auflistung der im Rahmen der Recherche identifizierten Erhebungsverfahren. Die Instrumente wurden zwischen 1963 und 1988 entwickelt. Bis auf das in Frankreich entwickelte „Functional Autonomy Measurement System" und die von der WHO in Genf entwickelte „International Classification of Impairments, Disabilities and Handicaps" stammen alle Messinstrumente aus den USA. Entsprechend der Fokussierung dieses Reviews auf die Inter-Rater-Reliabilität wurden Selbstbeurteilungsinstrumente (wie z.B. der im deutschsprachigen Raum verwendete Funktionsfragebogen Hannover[8]) nicht in die Übersicht aufgenommen.

Die reinen ADL-Skalen können als Ursprung der wissenschaftlichen Aktivitäten für Fragestellungen der Hilfsbedürftigkeit in den 50er und 60er Jahren angesehen werden. Der ADL-Index von Katz[9] und der Barthel-Index[10] als älteste und etablierte Instrumente aus diesem Bereich werden häufig zum Vergleich bzw. Evaluation neuerer Instrumente herangezogen. Der „Kenny Self-Care-Evaluation"-Fragebogen (Kenny SCE)[11] zeichnet sich durch seine sehr detaillierte Erhebung aus und misst dadurch präziser als herkömmliche ADL-Skalen. Die „Physical Self-Maintenance-Scale" (PSMS)[12] wurde speziell für ältere Menschen über 60 Jahre entwickelt.

Die um kognitive Funktionen erweiterten ADL-Skalen bilden einen komplexeren Ansatz und berücksichtigen, dass die grund-

legenden Funktionen wie „Essen", „Waschen", „Gehen" etc. als Voraussetzungen zur selbständigen Lebensführung nicht ausreichen. Das in der Literatur häufig zitierte „Functional Independence Measure" (FIM)[13] ist Teil des „Uniform Data System for Medical Rehabilitation (UDSMR)". Es basiert auf dem Barthel-Index, ist jedoch um Kommunikations- und kognitive Funktionen erweitert. Dasselbe gilt auch für die „Functional Status Rating Scale" (FSRS)[14]. Beide enthalten, ergänzend zu den beurteilten Kommunikationsfähigkeiten, zusätzlich aus dem IADL-Bereich die Frage nach sozialen Kontakten. Ebenfalls in diesen Bereich gehört der „Long Term Care Minimum Data Set" (LTCMDS)[15] des „National Center for Health Statistics", der Verhaltensfunktionen genauer misst. Einen anderen Ansatz bilden die reinen IADL-Skalen, die zur Beurteilung der Selbstversorgungsfähigkeit eher eingehen auf soziale Funktionen wie „Einkaufen", „Finanzen regeln", „Haushalt führen" etc.. Dieser Gruppe sind der „Functional Activities Questionnaire" (FAQ)[16] sowie die zusammen mit der PSMS entwickelte IADL-Skala von Lawton und Brody[12] zuzuordnen.

Umfassende Skalen, die alle drei Bereiche abdecken, sind universaler einsetzbar. Die „International Classification of Impairments, Disabilities and Handicaps" (ICIDH)[17] ist ein umfassendes Klassifikationssystem der WHO, das die Funktionsfähigkeit auf den Ebenen „Impairments", „Disabilities" und „Handicaps" beurteilt. Dasselbe gilt auch für das aus Frankreich stammende „Functional Autonomy Measurement System" (SMAF)[18], welches konzeptuell der ICIDH sehr ähnlich ist. Die zweite Version der „Rapid Disability Rating Scale" (RDRS-2)[19] stellt ein sehr umfassendes und dennoch einfach anzuwendendes Forschungsinstrument dar.

Reliabilitätsuntersuchungen

Ergebnisse von Reliabilitätsstudien zum Gesamtscore der Instrumente sind in Tabelle 2 dargestellt. Unter den reinen ADL-Skalen liegen für den ADL-Index von Katz nur qualitative Aussagen vor. So berichten Asberg und Sonn über eine hohe Inter-Rater-Reliabilität, ohne diese Angabe näher zu quantifizieren[20]. Granger et al.[21] berichtet für den Barthel-Index Inter-Rater-Reliabilitätswerte von über 0,95, ohne jedoch nähere Angaben zur Studienpopulation, der Versuchsdurchführung und der Art der statistischen Masszahl zu machen. Erst 1991, also 26 Jahre nach Einführung des Barthel-Index, führte Wolfe et al.[22] eine Reliabilitätsuntersuchung zu diesem Instrument durch, in der drei Krankenschwestern insgesamt drei Untersuchungspaare bildeten, die jeweils zur Bestimmung der Inter-Rater-Reliabilität miteinander verglichen wurden. Der quadratisch gewichtete Kappa-Koeffizient erreichte hier Werte zwischen 0,88 und 0,98.

Zur Reliabilitätsprüfung des Kenny SCE dienten Videoaufnahmen der Patienten[23]. Dies ermöglichte sechs Beurteilungen am gleichen Probanden. Auffallend ist hier jedoch die ausserordentlich geringe Fallzahl von nur sechs Patienten insgesamt und der relativ niedrige Intraclass-Koeffizient von 0,67.

Lawton et al.[12] führte zur PSMS Reliabilitätstestungen mit Krankenschwestern als Untersucher und auch mit wissenschaftlichen Assistenten durch, wobei eine geringfügig höhere Reliabilität für wissenschaftliche Assistenten gemessen wurde ($r = 0.91$ vs. $r = 0.87$). Die parallel vom selben Autor durchgeführte Studie zur IADL-Skala ergab ein Reliabilitätsmass von $r = 0{,}85$. Green et al.[24] gaben sowohl für die PSMS als auch für die IADL-Skala einen Intraclass-Koeffizient von 1,0 an. Diese Studie zeichnet sich durch eine intensive Schulung der Untersucher aus, sowie die Einbeziehung der pflegeverantwortlichen Person bei der Begutachtung. Ferner wurden bei Anwendung der IADL-Skala nur solche Personen begutachtet, für die in einer Vorstudie die Beurteilung aller Items als relevant eingestuft wurden. Diese Vorgehensweise schränkte die Fallzahl auf 18 Probanden ein, die für die Begutachtung mit der IADL-Skala zur Verfügung standen.

Für die um kognitive Funktionen erweiterten ADL-Skalen liegen zum FIM drei Untersuchungen vor[25–27]. In der Studie von Hamilton et al.[25], in der eine Selektion besonders reliabler Untersucherpaare mittels einer Vorstudie erfolgte, wird ein Intraclass-Koeffizient von 0,97 für den Gesamtscore und zusätzlich werden Intraclass-Koeffizienten für die sechs Bereiche Selbstversorgung ($ICC = 0{,}96$), Sphinkter-Kontrolle ($ICC = 0{,}94$), Mobilität ($ICC = 0{,}96$), Fortbewegung ($ICC = 0{,}93$), Kommunikation ($ICC = 0{,}95$) und Sozialkognitive Fähigkeiten ($ICC = 0{,}94$) berechnet. In der Studie von Chau et al.[27] wurde die Übereinstimmung zwischen Krankengymnasten bzw. Beschäftigungstherapeuten und geschulten Erziehern bei der Beurteilung von 8–20jährigen Rehabilitationspatienten getestet. Aus beiden Untersuchergruppen wurde zur Beurteilung jeweils die am besten mit dem Patienten vertraute Person herangezogen. Für den Gesamtscore wird in dieser Studie ein gewichteter Kappa-Koeffizient von 0,91 berichtet, ohne jedoch die Gewichtung näher zu spezifizieren. Für den ebenfalls den erweiterten ADL-Skalen zugehörigen FSRS werden nur für die vorläufige Version Inter-Rater-Koeffizienten (zwischen 0,81 und 0,92, ohne exakte Angabe über die verwendete statistische Masszahl) berichtet[28], über die endgültige Version liegen keinerlei Angaben vor.

Instrumente	Studienpopulation Art	Grösse	Untersucher	Anzahl Messungen am gleichen Probanden	Reliabilitätsmass[a]	Quelle
Barthel	keine Angabe	keine Angabe	keine Angabe	keine Angabe	≥ 0.95	Granger et al. 1979[21]
Barthel	Patienten mit einem höchstens drei Monate alten Apoplex	11 15 10	Krankenschwestern	2	$\kappa_w^2 = 0{,}88$ $\kappa_w^2 = 0{,}94$ $\kappa_w^2 = 0{,}98$	Wolfe et al. 1991[22]
Kenny SCE	Patienten mit funktionellen Einschränkungen verschiedenen Grades	6	Krankengymnastik-Schüler im letzten Ausbildungsjahr	6 Videoaufnahmen	ICC = 0,67	Kerner et al. 1981[23]
PSMS	Patienten mit unterschiedlichsten Behinderungen	36 14	Krankenschwestern wiss. Assistenten	2	$r = 0{,}87$ $r = 0{,}91$	Lawton et al. 1969[12]
PSMS	Zu Hause lebende Alzheimer-Patienten	27	geschulte wiss. Assistenten	2	ICC = 1,0	Green et al. 1993[24]
FIM	Hospitalisierte Rehabilitations-Patienten	263	Klinikärzte	2	ICC = 0,97	Hamilton et al. 1991[25]
FIM	Patienten mit Rückenmarksverletzungen	57	Krankenschwestern, Krankengymnasten, Beschäftigungstherapeuten	2	$r = 0{,}83$	Segal et al. 1993[26]
FIM	Rehabilitationspatienten	198	Krankengymnasten, Beschäftigungstherapeuten, Erzieher	2	$\kappa = 0{,}91$	Chau et al. 1994[27]
IADL-Skala (L & B)	Zu Hause versorgte Patienten	12	Sozialarbeiter	2	$r = 0{,}85$	Lawton et al. 1969[12]
IADL-Skala (L & B)	Zu Hause lebende Alzheimer-Patienten	18	geschulte wiss. Assistenten	2	ICC = 1,0	Green et al. 1993[24]
SMAF	Zu Hause versorgte Menschen	150	Gemeinde-, Krankenschwester, Sozialarbeiter	2	$\kappa_w = 0{,}75$	Hebert et al. 1988[18]

[a] r: Pearson-Produkt-Moment-Korrelationskoeffizient
κ_w: linear gewichteter Kappa-Koeffizient
κ_w^2: quadratisch gewichteter Kappa-Koeffizient
κ: gewichteter Kappa-Koeffizient, keine Spezifizierung der Gewichtung angegeben
ICC: Intraclass-Korrelationskoeffizient

Tabelle 2. *Untersuchungen zur Inter-Rater-Reliabilität des Gesamtscores.*

Aus dem Bereich der reinen IADL-Skalen konnten für die IADL-Skala von Lawton und Brody zwei Reliabilitätsstudien[12,24] gefunden werden, die jeweils in Zusammenhang mit der für die Beurteilung von ADL bestimmten PSMS durchgeführt wurden. Zum FAQ wurden dagegen keinerlei Reliabilitätsstudien identifiziert.

Für die Insrumente LTCMSD, ICIDH und RDRS-2 liegen Reliabilitätsuntersuchungen zu den einzelnen Merkmalen, nicht jedoch zu den Gesamtscores vor.

Zum SMAF, wie auch zum FIM, wurden nicht nur für den Gesamtscore, sondern auch für die Unterbereiche Reliabilitätsergebnisse berechnet. In der Untersuchung von Hebert et al.[18] liegen linear gewichtete Kappa-Werte für den Gesamtscore ($\kappa_w = 0{,}75$) und für die Bereiche ADL ($\kappa_w = 0{,}66$), Mobilität ($\kappa_w = 0{,}74$), Kommunikation ($\kappa_w = 0{,}53$), Mentale Funktionen ($\kappa_w = 0{,}58$) und IADL ($\kappa_w = 0{,}76$) vor.

Wie aus Tabelle 2 ersichtlich ist, bleibt, bis auf drei Ausnahmen[18,25,27], der Stichprobenumfang in den Studien weit unter 100 Probanden, was wohl in erster Linie auf die logistischen Probleme bei der Mehrfachuntersuchung von Probanden zurückzuführen ist. Auch die Anzahl der Beurteilungen am gleichen Probanden ist auf zwei beschränkt. Eine Ausnahme bildet die Untersuchung von Kerner et al.[23], in der Videoaufnahmen der Probanden als Grundlage der Begutachtung dienten. Als Untersucher werden vielfach Krankenschwestern, aber auch Ärzte, Sozialarbeiter und andere Personen, die häufig mit dem Beurteilten zu tun haben, genannt.

Für einige Studien wurden noch detailliertere Reliabilitätsergebnisse berichtet. Inter-Rater-Koeffizienten zu den einzelnen Erhebungsmerkmalen liegen für die reinen ADL-Instrumente Barthel-Index[29], Kenny SCE[23], PSMS[24] sowie für den um kognitive Funktionen erweiterten FIM[26,27] und den LTCMDS[30], die IADL-Skala (L & B)[24] und die alle Bereiche umfassenden ICIDH-D[7,31,32] und RDRS-2[19] vor. Die Ergebnisse sind in Tabelle 3 für die reinen ADL-Skalen und in Tabelle 4 für die übrigen Instrumente aufgelistet.

Die höchsten Werte aus dem Bereich der reinen ADL-Skalen erreichen die Untersuchungen von Green et al.[23] für die PSMS. Hier variiert der Intraclass-Koeffizient zwischen 0,98 („An-, Ausziehen", „Essen") und 1,0 („Gehen", „Baden", „Körperpflege"). Ähnlich hohe Reliabilität fanden dieselben Autoren für die IADL-Skala mit Intraclass-Koeffizienten zwischen 0,92 („Finanzen regeln") und 1,0 („Medikamenteneinnahme", „Einkaufen" und „Wäsche waschen"). Für den Kenny SCE und die RDRS-2, die ebenfalls den Intraclass-Koeffizienten berechnen, sind die Werte im Durchschnitt niedriger.

Instrumente:	**Barthel-Index**	**Kenny SCE**	**PSMS**
Reliabilitätsmass:	**ex. Üb.[a]**	**ICC**	**ICC**
Anzahl Patienten:	**n = 25**	**n = 6**	**n = 27**
Quelle:	**Collin et al. 1988[29]**	**Kerner et al. 1981[23]**	**Green et al. 1993[24]**
Elementare Funktionen „ADL-functions"			
An-, Ausziehen	0.72	0.89	0.98
Transfer	0.6	0.81	
Gehen	0.88	0.46	1.0
Treppensteigen	0.92		
Mobilität			
Duschen/Baden	1.0		1.0
Körperpflege	0.72	0.72	1.0
WC-Benutzung	0.68		
Harninkontinenz	0.84		0.99[b]
Stuhlinkontinenz	0.76		
Essen	0.76		0.98

[a] Anteil der exakten Übereinstimmung
[b] nicht unterschieden zwischen Harn- und Stuhlinkontinenz

Tabelle 3. *Inter-Rater-Reliabilität der Beurteilung spezifischer Funktionen: Untersuchungen für reine ADL-Skalen.*

In der Untersuchung zum ICIDH-D von Behrens et al.[7] an 18 sehgeschädigten Personen wird als statistisches Mass der Kappa-Koeffizient verwendet. In dieser Studie treten zweimal negative Koeffizienten auf, was darauf hindeutet, dass die beurteilten Items („Hören", „Orientierung") für diese Studienpopulation nicht geeignet sind. In den Studien von van Triet et al.[31] und van den Berg et al.[32] ebenfalls zum „Disability"-Code der ICIDH fallen die berechneten Kappa-Koeffizienten wesentlich höher aus und erreichen sogar dreimal den Maximalwert 1,0 für die Merkmale „Hören", „Sprechen" und „Hausarbeit"[32].

Die Pearson-Korrelationskoeffizienten in der Untersuchung zum FIM an Patienten mit Rückenmarksverletzungen[21] zeigen recht niedrige Werte. Für das Item „Sich ausdrücken, Sprechen" wird nur ein Korrelationskoeffizient von 0,02 gefunden. Demgegenüber sind die Kappa-Koeffizienten in der Studie von Chau et al.[27] zum FIM als sehr hoch einzuschätzen. Sie variieren zwischen 0,63 („Gedächtnis") und 0,92 („An-, Ausziehen" für die obere Körperhälfte).

Zu beachten sind die sehr unterschiedlichen Stichprobenumfänge, die zwischen sechs Probanden für den Kenny SCE (hier liegen allerdings sechs Messungen für jeden Probanden vor) und 290 Probanden für den LTCMDS liegen. Erwähnenswert ist weiterhin, dass die Ergebnisse des Gesamtscores aus Tabelle 2 durchschnittlich höher ausfallen als die in Tabelle 3 und 4 dargestellten Reliabilitätskoeffizienten der Einzelmerkmale.

Diskussion

Mit dieser Literaturstudie wurde ein Überblick über die Inter-Rater-Reliabilität international verbreiteter Messinstrumente zur Pflegebedürftigkeit gegeben. Die Messverfahren stammen bis auf zwei Ausnahmen (SMAF, ICIDH) aus den USA, was die längere Tradition wissenschaftlicher Aktivitäten in den USA auf diesem Gebiet widerspiegelt.

Die Zahl und Qualität der Studien bezüglich Reliabilität der einzelnen Instrumente sind sehr heterogen. So konnten für einige Instrumente zwei (für den FIM und die ICIDH drei) Reliabilitätsstudien gefunden werden, während für andere Instrumente keine oder nur sehr unvollständige Untersuchungen vorliegen. Gerade die fehlenden Reliabilitätsuntersuchungen für den ADL-Index von Katz als eines der ältesten und etablierten Instrumente auf diesem Gebiet sind überraschend, ebenso sind die fehlenden Reliabilitätsuntersuchungen des gut strukturierten und detaillierten Erhebungsinstrumentes FSRS bedauerlich. Insgesamt muss der Wissensstand bezüglich der Inter-Rater-Reliabilität als sehr begrenzt betrachtet werden, was angesichts der langjährigen Anwendung der Instrumente (teilweise seit über 20 Jahren!) überraschend ist. Diese Diskrepanz zwischen Einsatz und Evaluation der Verfahren unterstreicht den Forschungsbedarf für intensivere anwendungsbegleitende Untersuchungen zu Reliabilität und Validität der Instrumente. Eine erhebliche Einschränkung der Aussagekraft der meisten Studien liegt in den teilweise eher niedrigen Fallzahlen begründet. Bis auf eine Ausnahme wurden zur Inter-Rater-Reliabilitätsbestimmung ferner nur zwei Untersuchungen an der gleichen Probandengruppe durchgeführt. Die minimale Zahl von zwei Messwiederholungen mag vielfach durch logistische und ethische Gründe bedingt sein, um eine zu starke Belastung der Probanden durch Mehrfachuntersuchungen zu vermeiden. Auch würden bei mehrfach wiederholten Untersuchungen zunehmend „Lerneffekte" eine Rolle spielen. Dennoch wären Studien zur Inter-Rater-Reliabilität mit mehr als zwei Untersuchern vielfach aussagekräftiger. Eine Lösung dieses Konfliktes wurde in der Studie von Kerner et al.[23] versucht. Hier dienten Videoaufnahmen der Patienten als Grundlage von sechs Begutachtungen. Die Vor- und Nachteile dieses Verfahrens liegen auf der Hand. Einerseits können auf diese Weise beliebig viele Beurteilungen unter völlig identischen Versuchsbedingungen erfolgen, ohne die Probanden wiederholt belästigen zu müssen. Andererseits wird das Verhalten der Probanden möglicherweise durch die ungewohnten Umstände der Filmaufnahme beeinflusst.

Die Ergebnisse zur Inter-Rater-Reliabilität sind für den Gesamtscore der verschiedenen Instrumente zumeist zufriedenstellend, wobei jedoch eine erhebliche Variation zwischen den Instrumenten zu verzeichnen ist. So wird für den Kenny SCE in der Untersuchung von Kerner et al.[23] ein relativ niedriges Ergebnis (ICC = 0,67) erreicht im Vergleich zur Untersuchung von Green et al.[24] zur PSMS (ICC = 1,0), die ebenfalls den reinen ADL-Skalen zuzuordnen ist. Ein Grund mag in der Studiendurchführung liegen. Für die Beurteilung mit dem Kenny SCE wurden, wie oben erwähnt, Videoaufnahmen der Patienten zugrunde gelegt. Wenngleich diese Vorgehensweise beliebig viele, völlig identische Beurteilungen ermöglicht, besteht dennoch für den Untersucher keine Möglichkeit der Interaktion, beispielsweise um Unklarheiten mit den Patienten selbst oder Angehörigen zu klären. In der Untersuchung von Green et al.[24] wurde dagegen neben einer intensiven Untersucher-Schulung auf das Gespräch mit der pflegeverantwortlichen Person grossen Wert gelegt. Ein weiterer Grund mag in der Struktur der verschiedenen Instrumente selbst liegen. So ist der Kenny SCE ein sehr detail-

Instrumente:	FIM	FIM	FIM	LTCMDS	LTCMDS	IADL	ICIDH-D	ICIDH-D	ICIDH-D	ICIDH-D	RDRS-2
Reliabilitätsmass:	r	ex. Üb.[a]	κ[b]	κ[c]	ex. Üb.[a]	(L & B)	κ[d]	ex. Üb.[a]	κ[c]	κ[c]	ICC
Anzahl Patienten:	n = 57	n = 57	n = 198	n = 290	n = 290	ICC	n = 18	n = 18	n = 50	n = 39	n = 100
Quelle:	Segal et al. 1993[26]	Segal et al. 1993[26]	Chau et al. 1994[27]	Hogan et al. 1986[30]	Hogan et al. 1986[30]	18 < n < 27 Green et al. 1993[24]	Behrens et al. 1987[7]	Behrens et al. 1987[7]	van Triet et al. 1990[31]	van den Berg et al. 1990[32]	Linn et al. 1982[19]
Elementare Funktionen „ADL-functions"											
An-, Ausziehen	0.67; 0.46[e]	0.32; 0.42[e]	0.92; 0.82[e]	0.46; 0.29[f]	0.65; 0.85[f]		0.25	0.93	0.57	0.82	0.73
Transfer	0.72; 0.65; 0.6[g]	0.47; 0.6; 0.63[g]	0.88; 0.85; 0.81[g]	0.69	0.81		0.13	0.88		0.87; 0.77[h]	
Gehen	0.62	0.44	0.7	0.74	0.86		0.56; 0.44[i]	0.17; 0.2[i]	0.91	0.89	0.98
Treppensteigen	0.32	0.95	0.79				0.41	0.61	0.74	0.81	
Mobilität											0.82
Waschen									0.83		
Duschen/Baden	0.62	0.32	0.8	0.39	0.58					0.82	0.81
Körperpflege	0.63	0.42	0.88				0.21	0.81			0.77
WC-Benutzung	0.35	0.46	0.85	0.68	0.84						0.88
Harninkontinenz	0.10	0.65	0.84	0.51	0.77		0.24[j]	1.0[j]			0.89[j]
Stuhlinkontinenz	0.16	0.63	0.88	0.52	0.8						
Essen	0.77	0.61	0.85	0.28	0.41		0.43	0.94		0.93	0.93
Kognitive Funktionen											
Sehen				0.42	0.86		0.44	0.67		0.88	0.98
Hören				0.18	0.76		−0.03	0.94		1.0	0.96
Sich ausdrücken, Sprechen	0.02	0.81	0.79	0.53	0.76		0.63	0.89		1.0	0.92
Verwirrtheit				0.13	0.56						0.68

[a] Anteil der exakten Übereinstimmung
[b] gewichteter Kappa-Koeffizient, keine Spezifizierung der Gewichtung angegeben
[c] Kappa-Koeffizient, keine Gewichtung angegeben
[d] Kappa-Koeffizient nach Fleiss für unterschiedliche Beurteilerpaare
[e] für obere (1. Wert) und untere (2. Wert) Körperhälfte
[f] Anziehen (1. Wert), Ausziehen (2. Wert)
[g] Transfer aus dem Bett (1. Wert), Transfer auf Toilette (2. Wert) und Transfer zum Baden bzw. Duschen (3. Wert)
[h] Transfer vom Liegen zum Sitzen (1. Wert), vom Sitzen zum Stehen (2. Wert)
[i] auf ebenem Gelände (1. Wert) und unebenem Gelände (2. Wert)
[j] nicht unterschieden zwischen Harn- und Stuhlinkontinenz
[k] Gedächtnis für die Zeit (1. Wert), für Personen (2. Wert) und für Ereignisse (3. Wert)

Tabelle 4. *Inter-Rater-Reliabilität der Beurteilung spezifischer Funktionen.*

Instrumente:	FIM	FIM	FIM	LTCMDS	LTCMDS	IADL	ICIDH-D	ICIDH-D	ICIDH-D	ICIDH-D	RDRS-2
Reliabilitätsmass:	r	ex. Üb.[a]	κ[b]	κ[c]	ex. Üb.[a]	(L & B)	κ[d]	ex. Üb.[a]	κ[c]	κ[c]	ICC
Anzahl Patienten:	n = 57	n = 57	n = 198	n = 290	n = 290	ICC	n = 18	n = 18	n = 50	n = 39	n = 100
Quelle:	Segal et al. 1993[26]	Segal et al. 1993[26]	Chau et al. 1994[27]	Hogan et al. 1986[30]	Hogan et al. 1986[30]	18 < n < 27 Green et al. 1993[24]	Behrens et al. 1987[7]	Behrens et al. 1987[7]	van Triet et al. 1990[31]	van den Berg et al. 1990[32]	Linn et al. 1982[19]
Kognitive Funktionen											
Depression				0.12	0.61						0.62
Unruhe				0.12	0.56						
Unentschlossenheit				0.10	0.39						
Unkooperativität											0.7
Orientierung							–0.1	0.22			
Gedächtnis	0.39	0.32	0.63	0.15; 0.29; 0.13[k]	0.56; 0.77; 0.56[k]					0.86	
Probleme lösen	0.3	0.19	0.81								
Schreiben							0.31	0.44		0.76	
Verstehen	0.08	0.81	0.73	0.46	0.8						
Soziale Funktionen „IADL-functions"											
Hausarbeit						0.99	0.35	0.5	0.69	1.0	
Zubereitung von Mahlzeiten						0.94	0.29	0.46	0.56		
Medikamenten-einnahme						1.0					
Einkaufen						1.0			0.71		
Finanzen regeln						0.92					
Wäsche waschen						1.0					
Telefonieren						0.99					
Soziale Kontakte	0.25	0.23	0.69								
Benutzung öffentl. Verkehrsmittel						0.97	0.3	0.48			

Tabelle 4. *(Fortsetzung)*

liertes Erhebungsinstrumentarium, welches auf der Beurteilung von 85 (!) vom Probanden auszuführenden Bewegungsabläufen beruht. Im Gegensatz dazu beurteilt die PSMS sechs Bereiche des täglichen Lebens jeweils als Einheit auf einer 5-Punkte-Skala.

Im Vergleich der einzelnen Merkmale fällt auf, dass zum Teil recht grosse Unterschiede bei der Begutachtung desselben Merkmals mit verschiedenen Instrumenten auftreten. So wird beispielsweise für das Merkmal „Gehen" bei Anwendung des Kenny SCE ein ICC von 0,46 angegeben. Bei der Beurteilung durch die PSMS oder die RDRS-2 wird ein ICC≥0,98 ereicht. Diese Diskrepanz ist teilweise in der notwendigen Übersetzung der Instrumente aus dem Englischen bzw. Amerikanischen begründet, ergänzt durch unterschiedliche Auslegungen des Begriffs innerhalb der Instrumente. So wird in der RDRS-2 das Item „Walking" als Gehfähigkeit an sich bewertet, während es im Kenny SCE unter dem Begriff „Locomotion" alle Aspekte der Fortbewegung mitberücksichtigt. Weitere Ursachen für Abweichungen bei der Begutachtung desselben Merkmals könnten in der Qualifikation der Untersucher und in der Art der Studienpopulation zu suchen sein. Angesichts der teilweise sehr geringen Fallzahlen sind darüber hinaus Zufallsschwankungen zu berücksichtigen.

Ein interessanter Aspekt ist in der Studie von Segal et al.[26] zum FIM zu berücksichtigen: hier wurden die Beurteilungen in zwei örtlich getrennten Einrichtungen vorgenommen. Zunächst erfolgte die Begutachtung in einem Akutkrankenhaus und ein zweites Mal, bei Verlegung innerhalb drei Tagen, in einem Rehabilitationszentrum. Das gute Reliabilitätsergebnis (r = 0,83 für den Gesamtscore) zeigt, dass sich ein eventuell zu erwartender negativer Effekt unterschiedlicher Begutachtungsorte nicht bestätigt hat. In dieser Untersuchung wird ein weiteres Phänomen deutlich, welches auch in anderen Studien auftritt: für den Gesamtscore des FIM wird ein Pearson-Korrelationskoeffizient von 0,83 berechnet, während für die Übereinstimmung der einzelnen Items der Koeffizient nur Werte zwischen 0,02 und 0,77 erreicht. Solche Diskrepanzen können sich daraus ergeben, dass gleiche Gesamtscores nicht unbedingt dieselben Behinderungsmuster beim Patienten widerspiegeln. Gerade bei numerischen Scores, die keine Gewichtung von Einzelitems vorsehen, können völlig unterschiedliche Beurteilungen einzelner Funktionseinschränkungen beim Patienten zum gleichen Gesamtscore führen. Durch Vergleiche von einzelnen Items kommen solche Unterschiede im Detail zum Vorschein, während sie beim Vergleich der Gesamtscores unentdeckt bleiben. Daneben ist natürlich auch das durch die Codierung bedingte unterschiedliche Spektrum möglicher Merkmalsausprägungen zu berücksichtigen.

Durch die lückenhafte Erhebung der Reliabilität für Einzelmerkmale und die Verwendung unterschiedlicher Skalen und Koeffizienten ist die Identifikation besonders „reliabler" bzw. objektiver Merkmale nur schwer möglich. Tendenziell lässt sich dennoch folgendes zusammenfassen: die meisten Informationen zur Inter-Rater-Reliabilität liegen für die elementaren, sogenannten „ADL"-Funktionen vor, wobei die Merkmale „Gehen" und „Essen" im Durchschnitt die höchsten Reliabilitätsergebnisse aufweisen, dagegen ist für die Beurteilung der „IADL"-Funktionen das gesamte Umfeld des Betroffenen mitzuberücksichtigen und daher eine Erhebung recht aufwendig. Kognitive Funktionen sind erwartungsgemäss am schwierigsten „reproduzierbar" zu erfassen und zu beurteilen.

Aus sozialmedizinischer Sicht ist die Tatsache, dass die Instrumente trotz grösserer Diskrepanzen bei Einzelmerkmalen bezüglich der Gesamtbeurteilung eine hohe Inter-Rater-Reliabilität aufweisen, durchaus beruhigend.

Auch die insgesamt als zufriedenstellend zu bezeichnenden Reliabilitätsergebnisse der „ADL"-Funktionen sind für die sozialmedizinische Begutachtung der Pflegebedürftigkeit, bei der die Beurteilung dieser Grundfunktionen eine zentrale Rolle spielt, als sehr positiv zu bewerten.

Generell sei bemerkt, dass die in der vorliegenden Arbeit besprochenen Instrumente ihre Hauptanwendungen in wissenschaftlichen (z.B. epidemiologischen) Untersuchungen haben. Eine unmittelbare Anwendung dieser Instrumente für die Praxis der sozialmedizinischen Untersuchung ist nicht ohne weiteres möglich. Beispielsweise fehlen in sämtlichen Verfahren Angaben zu Anamnese, Therapie und Prognose der Untersuchten, die für die sozialmedizinische Begutachtung von grosser Bedeutung sind.

Dennoch können aus den recherchierten Untersuchungen wertvolle Hinweise für die Weiterentwicklung und Verbesserung des sozialmedizinischen Begutachtungsverfahrens zur Pflegebedürftigkeit gewonnen werden. So könnte die Einführung numerischer Scores für einzelne Merkmale und deren Addition zu einem Gesamtscore zur weiteren Objektivierung des Verfahrens des Medizinischen Dienstes der Krankenversicherung beitragen. In der Vergangenheit erfolgte die Begutachtung der Pflegebedürftigkeit ausschliesslich durch Ärzte des MDK. Die hier vorgestellten Untersuchungen lassen durchaus auch andere Berufsgruppen wie z.B. Pflegekräfte als Untersucher geeignet erscheinen. Dieses Ergebnis wurde bereits im Rahmen der Pflegeversicherung durch den verstärkten Einsatz

von Pflegekräften bei der Begutachtung berücksichtigt. Schliesslich legen die vergleichsweise niedrigen Reliabilitätswerte für den Bereich der kognitiven und IADL-Funktionen nahe, dass hier kontinuierliche begleitende qualitätssichernde Massnahmen z.B. eine sorgfältigste Schulung der Untersucher besonders wichtig sind.

Summary

Inter-rater-reliability of instruments for assessing disability and functional dependence: A review of the international literature

Objective and reliable rating of disability and functional dependence is of utmost importance for both medical and rehabilitation research and the practice of social medicine. The present paper provides an overview on internationally used disability scales and on studies that were carried out to determine their inter-rater reliability. Most of the scales were developed in the United States. The number and quality of inter-rater reliability studies strongly vary for various scales. In general, reliability was found to be high for summary scores of disability, whereas reliability strongly varied from extremely poor to excellent for single items of disability. This variation provides valuable suggestions for improving rating of disability.

Résumé

Fiabilité entre différents observateurs des méthodes d'évaluation des déficiences et handicaps: Une revue de la littérature

Une évaluation objective et détaillée des déficiences et handicaps est nécessaire dans le cadre de la recherche médicale, de la réhabilitation et de la médecine sociale. Cet article donne une vue d'ensemble des méthodes mondialement connues d'évaluation des déficiences en regard de la fiabilité entre des évaluateurs différents. La majorité des méthodes citées sont originaires des États-Unis et ont été développées entre 1963 et 1988. Le nombre et la qualité des études de fiabilité diffèrent beaucoup pour chaque méthode. En général, on remarque que la fiabilité entre des observateurs différents pour le score global sont habituellement élevés alors que ceux des divers paramètres varient grandement. On pourra tirer de cet article des suggestions importantes pouvant améliorer les méthodes d'évaluation des déficiences et handicaps.

Literaturverzeichnis

1 Medizinischer Dienst der Spitzenverbände der Krankenversicherung: Begutachtungsanleitung Schwerpflegebedürftigkeit (§§ 53 ff SGB V). Essen: Reimar Hobbing Verlag, 1990.

2 *Westhoff G.* Handbuch psychosozialer Messinstrumente. Göttingen: Hogrefe-Verlag, 1993.

3 *McDowell I, Newell C.* Measuring Health: A Guide to Rating Scales and Questionnaires. New York: Oxford University Press, 1987.

4 *Hartung J.* Statistik: Lehr- und Handbuch der angewandten Statistik. München: Oldenbourg Verlag, 1987: 546 ff.

5 *Dunn G.* Design and Analysis of Reliability Studies. London: Edward Arnold-Verlag, 1989: 32 ff.

6 *Fleiss JL.* Statistical methods for rates and proportions. 2nd ed. New York: Wiley and Sons, 1981: 218.

7 *Behrens E, Brambring M.* Beurteilerübereinstimmung einer deutschen Version der „International Classification of Impariments, Disabilities, and Handicaps (ICIDH)" der Weltgesundheitsbehörde. Int J Rehabil Res 1987; *10*: 391–404.

8 *Raspe HH, Hagedorn U, Kohlmann T, Matussek S.* Der Funktionsfragebogen Hannover (FFbH): Ein Instrument zur Funktionsdiagnostik bei polyartikulären Gelenkerkrankungen. In: Siegrist J. Wohnortnahe Betreuung Rheumakranker. Ergebnisse sozialwissenschaftlicher Evaluation eines Modellversuchs, Stuttgart: Schattauer Verlag, 1990: 164–182.

9 *Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW.* Studies of Illness in the Aged. The Index of ADL: A standardized measure of biological and psychological function. JAMA 1963; *185*: 914–919.

10 *Mahoney FI, Barthel DW.* Functional evaluation. The Barthel Index. Md State Med J 1965; *14*: 61–65.

11 *Iversen IA, Silberberg NE, Stever RC, Schoening HA.* The revised Kenny Self-Care Evaluation: A numerical measure of independence in activities of daily living. Minneapolis, Minnesota: Sister Kenny Institute 1973.

12 *Lawton MP, Brody EM.* Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living. Gerontologist 1969; *9*: 179–186.

13 *Keith RA, Granger CV, Hamilton BB, Sherwin FS.* The Functional Independence Measure: a new tool for rehabilitation. In: Eisenberg MG, Grzesiak RC (Hrsg.): Advances in clinical rehabilitation. New York: Springer, 1987; 16–18.

14 *Forer SK.* Revised functional status rating instrument, Glendale, California. Rehabilitation Institute, Glendale Adventist Medical Center, December 1981.
15 National Center for Health Statistics. Long Term Health Care Minimum Data Set. U.S. Department of Health and Human Services Publication PHS 80-1158, Washington, DC 1980.
16 *Pfeffer RI, Kurosaki TT, Harrah CH, Chance JM, Filos S.* Measurement of functional activities in older adults in the community. J Gerontol 1982; *37*:323–329.
17 World Health Organization. International classification of impairments, disabilities, and handicaps. Genf: WHO, 1980.
18 *Hebert R, Carrier R, Bilodeau A.* The Functional Autonomy Measurement System (SMAF): Description and Validation of an Instrument for the Measurement of Handicaps. Age Ageing 1988; *17*: 293–302.
19 *Linn WM, Linn BS.* The Rapid Disability Rating Scale-2. J Am Geriatr Soc 1982; *30*:378–383.
20 *Asberg KH, Sonn U.* The cumulative structure of personal and instrumental ADL. Scand J Rehabil Med 1988; *21*:171–177.
21 *Granger CV, Albrecht GL, Hamilton BB.* Outcome of Comprehensive Medical Rehabilitation: Measurement by PULSES-Profile and the Barthel Index. Arch Phys Med Rehabil 1979; *60*: 145–154.
22 *Wolfe CDA, Taub NA, Woodrow EJ, Burney PGJ.* Assessment of Scales of Disability and Handicap for Stroke Patients. Stroke 1991; *22*:1242–1244.
23 *Kerner JF, Alexander J.* Activities of Daily Living: Reliability and Validity of Gross vs Specific Ratings: Arch Phys Med Rehabil 1981; *62*:161–166.
24 *Green CR, Mohs RC, Schmeidler J, Aryan M, Davis KL.* Functional Decline in Alzheimer's Disease: A Longitudinal Study. J Am Geriatr Soc 1993; *41*:654–661.
25 *Hamilton BB, Laughlin JA, Granger CV, Kayton RM.* Interrater Agreement of the Seven Level Functional Independence Measure (FIM). Arch Phys Med Rehabil 1991; *72*:790.
26 *Segal ME, Ditunno JF, Staas WE.* Interinstitutional agreement of individual functional independence measure (FIM) items measured at two sites on one sample of SCI patients. Paraplegia 1993; *31*: 622–631.
27 *Chau N, Daler S, Andre JM, Patris A.* Inter-rater agreement of two functional independence scales: the Functional Independence Measure (FIM) and a subjective uniform continuous scale. Disabil Rehabil 1994; *16*:63–71.
28 *Forer SK, Miller LS.* Rehabilitation Outcome: Comparative Analysis of Different Patient Types. Arch Phys Med Rehabil 1980; *61*:359–365.
29 *Collin C, Wade DT, Davies S, Horne V.* The Barthel ADL-Index: a reliability study. Int Disabil Stud 1988; *10*:61–63.
30 *Hogan AJ, Smith DW, Jameson J.* Functional Assessment of nursing home patients: Reliability and Relevance. Evaluation and the Health Professions 1986; *9*:339–360.
31 *van Triet EF, Dekker J, Kerssens JJ, Curfs EC.* Reliability of the assessment of impairments and disabilities in survey research in the field of physical therapy. Int Disabil Stud 1990; *12*:61–65.
32 *van den Berg JP, Lankhorst GJ.* Inter-rater and intra-rater reliability of disability ratings based on the modified D Code of the ICIDH. Int Disabil Stud 1990; *12*:20–21.

Danksagung

Diese Arbeit wurde im Rahmen des durch das Bundesministerium für Forschung und Technologie geförderten Forschungsvorhabens „Epidemiologische Untersuchungen zur Schwerpflegebedürftigkeit auf der Basis der Gutachten des Medizinischen Dienstes der Krankenversicherung" (Projekt A2-1 des Münchner Forschungsverbunds Public Health) erstellt.

Korrespondenzadresse

Ulrike Kliebsch
Institut für Med. Informationsverarbeitung
Biometrie u. Epidemiologie
Marchioninistr. 15
D-81377 München