

E. Farin¹
C. Carl¹
S. Lichtenberg¹
W. H. Jäckel^{1,2}
B. Maier-Riehle³
E. Rütten-Köppel⁴

Die Bewertung des Rehabilitationsprozesses mittels des Peer-Review-Verfahrens: Methodische Prüfung und Ergebnisse der Erhebungsrunde 2000/2001 in den somatischen Indikationsbereichen

Evaluating the Rehabilitation Process by Means of Peer Review: Examination of the Methods Used and Findings of the 2000/2001 Data Collection in the Somatic Indications

Zusammenfassung

Berichtet wird von den Ergebnissen des Peer-Review-Verfahrens zur Bewertung des Rehabilitationsprozesses im Qualitätssicherungsprogramm der gesetzlichen Rentenversicherung. Die Daten beziehen sich auf die Erhebungsrunde 2000/2001 in den somatischen Indikationsbereichen. Die Gutachterübereinstimmung bezüglich des Urteils eines einzelnen Peers fällt nur in der Orthopädie zufrieden stellend aus. Da in dem Qualitätssicherungsprogramm Kliniken jedoch in der Regel auf der Basis des Mittelwerts über 20 Raterurteile bewertet werden, wird zusätzlich die Reliabilität dieses aggregierten Maßes abgeschätzt. Diesbezüglich ist für alle Indikationen von zumindest zufrieden stellenden Reliabilitätswerten auszugehen. Die Ergebnisse in den 561 am Programm beteiligten Kliniken zeigen, dass insbesondere diejenigen Prozessmerkmale von den Peers als verbesserungsbedürftig angesehen wurden, die eine Dokumentation der Einschätzungen und subjektiven Konzepte des Patienten erfordern (z. B. Erfassung des subjektiven Krankheitsverständnisses). Während sich zwischen den Erhebungsrunden 1998 und 1999 die Bewertungen der Berichte der Kliniken statistisch signifikant verbessert haben, ist zwischen 1999 und 2000/2001 über alle Kliniken hinweg keine weitere Verbesserung erkennbar. Allerdings haben die 1999 schlecht bewerteten Kliniken (unterstes Quartil der Verteilung) auch zwischen 1999 und 2000/2001 eine positive Entwicklung vollzogen. Im Diskussionsteil werden Ursachen für diesen Trend sowie Möglichkeiten zur Verbesserung der Interrater-Reliabilität des Peer-Review-Verfahrens erörtert.

Schlüsselwörter

Peer Review · Qualitätssicherung · Rehabilitation

Abstract

This paper reports the results of a peer review system that was implemented in the context of the quality assurance programme of the statutory German Pension Insurance scheme. The data reported refer to the 2000/2001 data collection period for medical rehabilitation in the somatic indications. Examination of interrater reliability for judgements of individual raters shows satisfactory results only in orthopaedics. In the quality assurance programme, rehabilitation centres are usually evaluated by the mean of 20 rater judgements. The reliability of this aggregated measure is satisfactory in all indications. The results of 561 rehabilitation centres show that those quality criteria are in particular need of improvement that refer to subjective concepts of patients (e. g., subjective theories of illness). Between peer review procedures in 1998 and 1999, the quality scores of rehabilitation centres had improved whereas between 1999 and 2000/2001, no further improvement can be shown. However, those rehabilitation centres with a low quality score in 1999 (lowest quartile of the distribution) underwent a positive development between 1999 and 2000/2001. Reasons for this trend and possibilities for improving interrater reliability of the peer review process as an element of the quality assurance programme of the German Pension Insurance scheme are discussed.

Key words

Peer review · quality management · rehabilitation

Institutsangaben

¹ Abteilung Qualitätsmanagement und Sozialmedizin, Universitätsklinikum Freiburg

² Hochrhein-Institut für Rehabilitationsforschung, Bad Säckingen

³ Verband Deutscher Rentenversicherungsträger, Frankfurt/Main

⁴ Bundesversicherungsanstalt für Angestellte, Berlin

Korrespondenzadresse

Dr. Erik Farin · Universitätsklinikum Freiburg, Abt. Qualitätsmanagement und Sozialmedizin · Breisacher Straße 62 · 79106 Freiburg · E-mail: farin@aq.s.ukl.uni-freiburg.de

Bibliografie

Rehabilitation 2003; 42: 323–334 · © Georg Thieme Verlag Stuttgart · New York · ISSN 0034-3536

Einleitung

Ein Peer-Review-Verfahren im Gesundheitswesen beinhaltet die Bewertung von Aspekten der Struktur-, Prozess- oder Ergebnisqualität durch „Peers“, also durch gleichgestellte Angehörige der jeweiligen Berufsgruppe (z. B. Ärzte, Pflegekräfte, Therapeuten). Die Bewertungsmaßstäbe sind dabei in der Regel in Form von Standards oder konsensuell abgestimmten Bewertungskriterien vorgegeben (vgl. z. B. [1–4]). Es kann unterschieden werden zwischen Beobachtungs- bzw. Interviewansätzen, die eine Begehung der zu bewertenden Einrichtung beinhalten (häufig auch als „Visitation“ bezeichnet; vgl. z. B. [5]), und Verfahren, die auf der Begutachtung von medizinischen Behandlungsunterlagen (Patientenakten) basieren. Beide Ansätze können auch kombiniert werden, wie z. B. in einem kanadischen Peer-Assessment-Programm für Allgemeinärzte [6] und im australischen Brain Injury Rehabilitation Program [1]. Peer Reviews gehören zu den am häufigsten eingesetzten Qualitätssicherungsverfahren in der Medizin und haben im akutmedizinischen Bereich wesentlich zum Wissen über medizinische Fehler und vermeidbare Behandlungskomplikationen beigetragen [7]. Im amerikanischen Gesundheitswesen fordern verschiedene Organisationen (z. B. Joint Commission on Accreditation of Healthcare Organizations JCAHO, Medicare, Medicaid) die Durchführung eines Peer-Review-Verfahrens bzw. integrieren ein solches in Akkreditierungsprozesse. Die Zielsetzungen von Peer-Review-Verfahren betreffen die Qualitätsmessung, aber auch die Förderung der Einhaltung der zur Bewertung herangezogenen Standards sowie die Stimulierung von Aktivitäten der Qualitätsverbesserung [2].

Im Bereich der gesetzlichen Rentenversicherung in Deutschland ist ein Peer-Review-Verfahren bezüglich anonymisierter Entlassungsberichte und individueller Therapiepläne seit 1999 routinemäßiger Bestandteil eines umfassenden Qualitätssicherungsprogramms [8, 9]. Es wird für alle stationären Rehabilitationseinrichtungen im Zuständigkeitsbereich der Rentenversicherungsträger durchgeführt. Die Grundlage der Prüfung wird durch eine „Checkliste qualitätsrelevanter Prozessmerkmale“ sowie ein dazugehöriges Manual mit indikationsspezifischen Bewertungskriterien gebildet. Für die somatischen Indikationsbereiche einerseits sowie für Einrichtungen der Entwöhnungsbehandlung und der psychosomatischen Rehabilitation andererseits liegen getrennte Versionen von Checkliste und Manual vor. Die Erarbeitung der Unterlagen erfolgte in zahlreichen Workshops unter der Mitwirkung von Experten aus Rehabilitationskliniken und von den Rentenversicherungsträgern [10, 11]. Für die Konsensus- und Entscheidungsfindung wurde eine modifizierte Delphi-Methodik angewandt (vgl. z. B. [12, 13]). Auf der Basis dieses Peer Reviews wurden auch für die Qualitätssicherungsprogramme der gesetzlichen Krankenkassen [14] und der Unfallversicherung [15] Peer-Review-Verfahren zur Bewertung des Rehabilitationsprozesses erarbeitet. Seit Ende 2002 liegt eine für den Bereich der Renten- und Krankenversicherung einheitliche Version des Peer Reviews vor, die zukünftig in den Qualitätssicherungsprogrammen beider Rehabilitationsträger eingesetzt werden soll.

Die Checkliste des Peer-Review-Verfahrens im Bereich der Rentenversicherung umfasst insgesamt 52 qualitätsrelevante Prozessmerkmale, die sich in die Bereiche „Anamnese“, „Diagnostik“, „Therapieziele und Therapie“, „sozialmedizinische Stellung-

nahme“, „Nachsorgekonzept“ und „Verlauf und Epikrise“ gliedern. Für die Bewertung der einzelnen Prozessmerkmale stehen die Antwortkategorien „keine Mängel“, „leichte Mängel“, „gravierende Mängel“ und „entfällt“ zur Verfügung. Jeder Bereich der Checkliste schließt mit einer zusammenfassenden Bewertung, wobei die Antwortkategorien „keine Mängel“, „leichte Mängel“, „deutliche Mängel“ und „gravierende Mängel“ (im Folgenden kurz „Mängelkategorien“ genannt) angekreuzt werden können. Ferner ist am Ende der Checkliste eine zusammenfassende Bewertung für die Qualität des gesamten Reha-Prozesses vorgesehen. Zusätzlich erfolgt bezüglich der zusammenfassenden Bewertungen (nicht bezüglich der Einzelmerkmale) eine Beurteilung nach „Qualitätspunkten“ (11-stufige Skala von 0–10 Punkten). Mit den Qualitätspunkten soll den Peers die Möglichkeit gegeben werden, in Bezug auf die übergeordneten Inhaltsbereiche eine differenziertere Qualitätsbeurteilung abzugeben. Darüber hinaus erschien eine zusätzliche, an Qualitätspunkten orientierte Bewertungsmetrik sinnvoll, da eine lediglich an „Mängeln“ orientierte Beurteilung, deren Optimum aus der Abwesenheit von Mängeln besteht, von den beteiligten Kliniken als wenig motivationsförderlich wahrgenommen wurde.

Alle zwei Jahre werden je Rehabilitationseinrichtung bis zu 20 zufällig ausgewählte, anonymisierte Entlassungsberichte mit individuellen Therapieplänen dem Review durch Fachkollegen unterzogen. Um eine möglichst objektive Bewertung zu gewährleisten, werden alle Gutachter zuvor in dreitägigen Seminaren mit dem Peer-Review-Verfahren und der Anwendung der Beurteilungsmaßstäbe in Checkliste und Manual vertraut gemacht. Um Urteilstendenzen auszugleichen, werden die Behandlungsfälle einer Klinik per Zufall und auf viele verschiedene Peers verteilt.

Der vorliegende Beitrag stellt die Ergebnisse der Erhebungsrunde 2000/2001 im Peer-Review der Rentenversicherung in den somatischen Indikationsbereichen sowie Resultate methodischer Prüfungen zur Reliabilität des Verfahrens dar¹. Berichtet wird zunächst (dritter Abschnitt) von der Interrater-Reliabilität der Bewertung durch einen einzelnen Peer. Da Kliniken in den Qualitätssicherungsprogrammen der Rehabilitationsträger auf der Basis des Mittelwerts von bis zu 20 Peer-Urteilen bewertet werden, sind diese Resultate für die Praxis des Verfahrens nur begrenzt aussagekräftig. Es wird deshalb zusätzlich die Reliabilität des aggregierten Maßes analysiert. Abschließend wird der Frage nachgegangen, inwiefern sich mit einer rechnerischen Aggregation von Urteilsanteilen reliablere Gesamtbewertungen ergeben würden als mit einer dem einzelnen Peer überlassenen, den Einzelfall berücksichtigenden Zusammenfassung.

Im Anschluss an die methodische Prüfung werden im 4. Abschnitt die Ergebnisse des Peer-Reviews 2000/2001 referiert. Es wird dargestellt, wie die Bewertung der Behandlungsunterlagen in den einzelnen Indikationsbereichen im Durchschnitt erfolgte und in welchen Bereichen des Rehabilitationsprozesses die Peers besonders häufig Stärken bzw. Schwachstellen entdeckten. Der

¹ Das Projekt wurde vom Verband Deutscher Rentenversicherungsträger und der Bundesversicherungsanstalt für Angestellte finanziert.

folgende Abschnitt stellt dar, wie die zeitliche Entwicklung der Beurteilung bei denjenigen Kliniken aussieht, die 1999 und 2000/2001 mit der gleichen Fachabteilung am Peer-Review-Verfahren beteiligt waren. Vertiefende Auswertungen gehen der Frage nach, ob sich zeitliche Entwicklungen bei Kliniken in den Randbereichen der Verteilung (besonders gut bzw. eher schlecht bewertete Einrichtungen) anders darstellen als in der Gesamtgruppe aller Kliniken. Der letzte Abschnitt fasst die Ergebnisse zusammen und diskutiert die Konsequenzen im Hinblick auf mögliche Weiterentwicklungen des Verfahrens.

Datenerhebung und Methodik

Die hier dargestellten Analysen beziehen sich auf die Erhebungsrunde 2000/2001 des Peer Reviews der Rentenversicherung in den somatischen Indikationsbereichen. Der Zeitraum, aus dem per Zufall die zu bewertenden Behandlungsfälle gezogen wurden, fiel in das Jahr 2000, die Beurteilungen durch Peers und die statistischen Analysen erfolgten hauptsächlich im Jahr 2001. In die Auswertungen gehen 9627 Berichte (mit Therapieplan) aus 561 Kliniken ein. Tab. 1 verdeutlicht die Verteilung der Kliniken und der Berichte nach Indikation. Die mittlere Bearbeitungsdauer eines Falles durch die Peers betrug 30,5 Minuten (Standardabweichung 11,2). Bezüglich der zusammenfassenden Bewertungen liegt die Korrelation der Mängelkategorien mit den entsprechenden Bewertungen nach Qualitätspunkten zwischen 0,77 und 0,82 (Korrelation nach Spearman).

Tab. 1 Verteilung der Kliniken und der Berichte nach Indikation

Indikation	Anzahl der Kliniken	Anzahl der Berichte
Orthopädie	265 (47,3%)	4506 (46,8%)
Kardiologie	86 (15,3%)	1549 (16,1%)
Gastroenterologie	25 (4,4%)	472 (4,9%)
Onkologie	74 (13,2%)	1232 (12,8%)
Neurologie	61 (10,9%)	985 (10,2%)
Pneumologie	27 (4,8%)	499 (5,2%)
Dermatologie	9 (1,6%)	180 (1,9%)
Urologie	5 (0,9%)	81 (0,8%)
Gynäkologie	9 (1,6%)	123 (1,3%)
gesamt	561	9627

Für die Prüfung des Grads der Übereinstimmung zwischen den Peers (Interrater-Reliabilität) wurden pro Indikation 5 Behandlungsfälle ausgewählt, deren Unterlagen in identischer Form an alle Gutachter der jeweiligen Indikation versandt wurden. Für die Peers war kein Unterschied zwischen diesen sog. „Kontrollberichten“ und den üblichen Berichten erkennbar. Für einige onkologische Subindikationen (Bewegungsorgane, Atmungsorgane, ZNS, Haut und maligne Systemerkrankungen) wurden keine Kontrollberichte verteilt. Für Urologie und Gynäkologie liegen aufgrund der geringen Anzahl der Peers zu wenig Kontrollberichte vor, so dass keine aussagefähige Reliabilitätsanalyse erfolgen kann. Tab. 2 gibt einen Überblick über die in die Auswertung eingehenden Kontrollberichte.

Tab. 2 Verteilung der Peers und ihrer Beurteilungen der Kontrollberichte nach Indikation

Indikation	Anzahl der Peers	Anzahl der Beurteilungen (Kontrollberichte)
Orthopädie	230	1130
Kardiologie	83	401
Gastroenterologie	27	130
Onkologie/Verdauungsorgane	20	99
Onkologie/Urologie	18	88
Onkologie/Gynäkologie	21	101
Neurologie	54	257
Pneumologie	25	121
Dermatologie	9	42
gesamt	487	2369

Für die Berechnung der Interrater-Reliabilität bezüglich der *Mängelkategorien* wurden sowohl Kendalls Konkordanzkoeffizient W [16] als auch der Finn-Koeffizient [11,17] berechnet. Kendalls W ist eng verwandt mit einer durchschnittlichen Rangkorrelation und gibt den Anteil der Varianz zwischen den Ratingrängen der zu beurteilenden Objekte an der Gesamtvarianz der Ratingränge an. Der Finn-Koeffizient stellt einen Übereinstimmungskoeffizienten dar und wird gebildet, indem die beobachtete Varianz der Beurteilungen durch die bei zufälliger Kodierung zu erwartende Varianz geteilt wird. Nach Subtraktion dieses Quotienten von 1 erhält man den Anteil „nichtzufälliger“ Varianz [18]. Bei der Beurteilung des Finn-Koeffizienten kann der Anteil „nichtzufälliger“ Varianz von 0,5–0,7 als zufrieden stellend und von mehr als 0,7 als gut bezeichnet werden. Kendalls W stellt das wohl gebräuchlichste Verfahren bei der Bestimmung des Zusammenhangs ordinalskaliert Ratings bei einer Gruppe von Ratern dar. Der Finn-Koeffizient – der in der neueren Literatur nicht mehr unumstritten ist [19] – wird hier zusätzlich dargestellt, um einen Vergleich mit den Ergebnissen der Publikation früherer Daten des Peer-Review-Verfahrens [11] zu ermöglichen.

Bei den Berechnungen der Interrater-Reliabilität bezüglich der *Qualitätspunkte* wurden unadjustierte Intraklassenkorrelationen (ICC_{uni}) bestimmt, da bei den Qualitätspunkten von einem Intervallskalenniveau ausgegangen werden kann (zum Berechnungsverfahren siehe [17,19]). Die Intraklassenkorrelation setzt die Varianz zwischen den Beurteilungsobjekten (Entlassungsberichte) in Beziehung zur Varianz zwischen den Ratern. Es wurden unadjustierte Intraklassenkorrelationen berechnet, da die Raterstichprobe als eine Zufallsstichprobe aus der Population aller Rater aufzufassen ist und sich die zu gewinnende Reliabilitätsaussage nicht nur auf die Stichprobe der hier untersuchten Rater beziehen soll.

Die mithilfe der Kontrollberichte berechenbare Interrater-Reliabilität der Peer-Bewertung bezieht sich auf das Urteil eines einzelnen Raters. Um die Reliabilität der Bewertungen auf der Basis des Mittelwerts über 20 Raterurteile abschätzen zu können,

wurde die Spearman-Brownsche-Formel der Testverlängerung herangezogen [20].

Die Interrater-Reliabilität des Peer-Review-Verfahrens²

Die Interrater-Reliabilität der Bewertung durch einen einzelnen Peer

Die Ergebnisse der Reliabilitätsanalysen bezüglich der Mängelkategorien sind in Tab. 3 dargestellt, die der Qualitätspunkte in Tab. 4 (linke Spalten, ICC_{unj}). Es zeigt sich, dass die Interrater-Reliabilität *bezüglich des Urteils eines einzelnen Peers* nur in der Orthopädie in jeder Hinsicht zufrieden stellend ist. Der Finn-Koeffizient nimmt zwar auch in den anderen Indikationsbereichen in der Regel zufrieden stellende Werte an, doch sollten die teilweise recht deutlichen Abweichungen gegenüber den Werten von

Kendalls W Anlass sein, die Aussagekraft des Finn-Koeffizienten für die vorliegenden Daten vorsichtig zu bewerten: Die Verteilung der Bewertungen der Peers auf die Urteilkategorien unterscheidet sich zwar recht deutlich von einer Zufallsverteilung, die durchschnittliche Rangkorrelation der Urteile der Peers fällt jedoch nicht in allen Indikationsbereichen zufriedenstellend aus.

Bei der Interpretation der teilweise sehr geringen Werte der Intraklassenkorrelation ist zu berücksichtigen, dass die ICC-Werte aufgrund der Berechnungsformel (Relation der Varianz zwischen den Beurteilungsobjekten – also hier die Kontrollberichte – zur Varianz zwischen den Ratern) entscheidend von der Streuung der Bewertungen der Kontrollberichte abhängen. So beträgt die Korrelation zwischen dem ICC_{unj}-Wert in Tab. 4 und der Streuung der mittleren Bewertungen der Kontrollberichte für die Qualität des Reha-Prozesses (Gesamtbewertung) über die 9 dar-

Tab. 3 Interrater-Reliabilität der Einzelratings bez. der Mängelkategorien (Finn-Koeffizienten und Kendalls W)

Bereich der Checkliste	Indikation																	
	Orthopädie		Kardiologie		Gastro- enterologie		Onkologie/ Verdauung		Onkologie/ Urologie		Onkologie/ Gynäkologie		Neurologie		Pneumologie		Dermatologie	
	Finn	W	Finn	W	Finn	W	Finn	W	Finn	W	Finn	W	Finn	W	Finn	W	Finn	W
Anamnese	0,69	0,57	0,67	0,47	0,64	0,21	0,61	0,39	0,69	0,30	0,58	0,54	0,65	0,53	0,72	0,21	0,64	0,36
Diagnostik	0,62	0,62	0,59	0,34	0,53	0,12	0,63	0,35	0,45	0,05	0,59	0,39	0,61	0,21	0,49	0,15	0,55	0,53
Therapieziele und Therapie	0,67	0,62	0,44	0,16	0,57	0,23	0,67	0,18	0,62	0,06	0,54	0,21	0,63	0,17	0,66	0,60	0,59	0,45
sozialmed. Stellungnahme	0,50	0,35	0,49	0,53	0,51	0,36	0,30	0,52	0,35	0,46	0,69	0,60	0,48	0,30	0,64	0,15	0,56	0,27
Nachsorgekonzept	0,69	0,33	0,47	0,24	0,45	0,14	0,65	0,39	0,36	0,06	0,50	0,38	0,43	0,26	0,47	0,18	0,37	0,14
Verlauf und Epikrise	0,58	0,54	0,58	0,48	0,64	0,17	0,47	0,21	0,63	0,37	0,56	0,81	0,52	0,20	0,59	0,44	0,72	0,21
Qualität des Reha-Prozesses	0,67	0,61	0,65	0,35	0,63	0,25	0,66	0,35	0,57	0,16	0,57	0,46	0,62	0,36	0,63	0,37	0,63	0,35

Tab. 4 Interrater-Reliabilität der Einzelratings bez. der Qualitätspunkte (ICC_{unj}) und die mit der Spearman-Brown-Formel berechnete Reliabilität des Mittelwerts über 20 Rater (ICC_M)^a

Bereich der Checkliste	Indikation																	
	Orthopädie		Kardiologie		Gastro- enterologie		Onkologie/ Verdauung		Onkologie/ Urologie		Onkologie/ Gynäkologie		Neurologie		Pneumologie		Dermatologie	
	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M	ICC _{unj}	ICC _M
Anamnese	0,51	0,95	0,35	0,91	0,14	0,77	0,19	0,82	0,24	0,83	0,25	0,87	0,29	0,89	0,12	0,72	0,17	0,62
Diagnostik	0,47	0,95	0,06	0,55	0,10	0,68	0,10	0,67	0,00	0,00	0,19	0,83	0,11	0,72	0,07	0,60	0,00	0,00
Therapieziele und Therapie	0,54	0,96	0,13	0,75	0,11	0,70	0,14	0,76	0,04	0,41	0,10	0,69	0,10	0,69	0,52	0,96	0,20	0,66
sozialmed. Stellungnahme	0,34	0,91	0,33	0,91	0,22	0,85	0,25	0,87	0,27	0,86	0,45	0,94	0,14	0,77	0,04	0,48	0,27	0,75
Nachsorgekonzept	0,26	0,88	0,12	0,73	0,10	0,69	0,20	0,83	0,00	0,00	0,28	0,89	0,11	0,72	0,14	0,77	0,00	0,00
Verlauf und Epikrise	0,52	0,96	0,31	0,90	0,12	0,73	0,13	0,74	0,16	0,75	0,34	0,91	0,11	0,72	0,23	0,86	0,12	0,51
Qualität des Reha-Prozesses	0,54	0,96	0,26	0,87	0,13	0,74	0,21	0,83	0,08	0,58	0,27	0,88	0,16	0,79	0,22	0,85	0,11	0,50

^a Bei Indikationen, für die weniger als 20 Peers zur Verfügung stehen (Dermatologie, Onkologie/Urologie und Onkologie/Verdauung), wurde zur Berechnung des Werts die tatsächliche Anzahl der zur Verfügung stehenden Peers verwendet.

² Für Diskussionen zur Analyse der Interrater-Reliabilität des Peer-Review-Verfahrens danken die Autoren/Innen Herrn Dr. M. Wirtz, Universität Freiburg, Abt. Rehabilitationspsychologie.

gestellten Indikationen $r = 0,96$. Der ICC_{unj} -Wert einer Indikation lässt sich also nahezu perfekt durch die Unterschiedlichkeit der Qualität der jeweils ausgewählten Kontrollberichte vorhersagen. Mithilfe einer Regressionsgleichung (Prädiktor Streuung und Kriterium ICC_{unj} -Wert) lässt sich berechnen, dass ein zufriedener ICC_{unj} -Wert von 0,70 ab einer Streuung der Qualitätspunkte der 5 Kontrollberichte von ca. 2,50 erreichbar ist (das entspricht z. B. einer Werteverteilung von 3, 4, 6, 8 und 9 Qualitätspunkten).

Zusammenfassend kann – auch unter Berücksichtigung der methodischen Einschränkungen der verschiedenen Reliabilitätsindizes – davon ausgegangen werden, dass die Interrater-Reliabilität *bezüglich des Urteils eines einzelnen Peers* nur teilweise zufriedenstellend ist. Kliniken sollten also nicht auf der Basis der Beurteilung eines einzelnen Peers bewertet werden. Dies geschieht allerdings auch nicht in den Qualitätssicherungsprogrammen der Rehabilitationsträger. Im Bereich der Rentenversicherung werden Kliniken üblicherweise bezüglich der mittleren Bewertung von bis zu 20 Fällen, die in der Regel durch 20 verschiedene Peers bewertet werden, verglichen³. Die Peers werden per Zufall den zu bewertenden Kliniken zugeordnet. Es stellt sich damit die Frage, wie die Reliabilität der Bewertung auf der Basis des Mittelwerts von ca. 20 Raterurteilen einzuschätzen ist.

Die Reliabilität der Bewertung auf der Basis von 20 Behandlungsfällen

Um die Reliabilität des aggregierten Maßes einer Mittelung über Raterurteile grob abzuschätzen, wurde in Tab. 4 (jeweils rechte Spalten, ICC_M) die Spearman-Brownsche-Formel der Testverlängerung auf die in der gleichen Tabelle dargestellten Interrater-Reliabilitätswerte für Einzelratings (ICC_{unj}) angewandt. Die ICC_M -Werte geben an, wie hoch die Reliabilität des Mittelwerts über 20 Rater bei Zugrundelegung der ICC_{unj} -Werte eines einzelnen Ratings ausfallen würde (vgl. auch [19], S. 192). Die Werte schwanken bezüglich der Gesamtbewertung (Qualität des Rehabilitationsprozesses) zwischen 0,50 und 0,96, bei Ausschluss der Indikationen mit weniger als 20 Peers (Dermatologie, Onkologie/Urologie und Onkologie/Verdauung) zwischen 0,74 und 0,96. Dies deutet auf eine zufriedenstellende bis gute Reliabilität hin.

Die Größe der ICC_M -Koeffizienten kann allerdings nur zu einer Abschätzung des Maximalwerts der Reliabilität des eingesetzten aggregierten Maßes herangezogen werden, da die 20 Peers nicht denselben, sondern 20 verschiedene Entlassungsberichte beurteilen (was aus Validitätsgesichtspunkten auch durchaus sinnvoll ist). Der tatsächliche Reliabilitätswert des in der Praxis des Peer-Review-Verfahrens eingesetzten aggregierten Klinikwerts wird zwischen den Werten von ICC_{unj} und ICC_M liegen.

Dass dies der Fall ist und dass der Wert unter Umständen auch relativ dicht an dem mit der Spearman-Brown-Formel berechneten Wert liegt, verdeutlicht beispielhaft ein anderer Berechnungsweg der Reliabilität: Fasst man Reliabilität als Reprodu-

zierbarkeit von Messungen auf, so lässt sich untersuchen, wie hoch in dem vorliegenden Datensatz die mittlere Bewertung von (zufällig ausgewählten) 10 Entlassungsberichten einer Klinik (die durch 10 verschiedene Peers bewertet werden) mit der mittleren Bewertung der restlichen 10 Berichte dieser Klinik zusammenhängt. Diese Korrelation beträgt z. B. über alle neurologischen Kliniken für die Qualitätspunkte hinsichtlich der zusammenfassenden Bewertung $r = 0,67$ ($p < 0,001$)⁴. Ähnliche Werte ergeben sich für die Kardiologie ($r = 0,62$, $p < 0,001$) und Orthopädie ($r = 0,59$, $p < 0,001$). Diese Werte würden noch höher ausfallen, wenn man – entsprechend der realen Auswertesituation – die Mittelwerte der Bewertungen von *jeweils 20 Berichten* korrelieren würde. Eine Abschätzung dieser Werte mittels der Spearman-Brownschen-Formel ergibt Reliabilitätswerte von 0,80, 0,77 und 0,74 (in dieser Reihenfolge für Neurologie, Kardiologie und Orthopädie).

Ähnliche Ergebnisse erhält man für die Mängelkategorien, wenn man den mittleren Prozentsatz deutlicher und gravierender Mängel von (zufällig ausgewählten) 10 Entlassungsberichten einer Klinik mit dem mittleren Prozentsatz deutlicher und gravierender Mängel der restlichen 10 Berichte dieser Klinik korreliert. Die Korrelationen betragen für Neurologie, Kardiologie und Orthopädie (in dieser Reihenfolge) 0,62, 0,47 und 0,55. Die Abschätzung bezüglich des Werts bei 20 Berichten mithilfe der Spearman-Brownschen-Formel ergibt Reliabilitätswerte von 0,77, 0,64 und 0,71.

Dass die unterschiedliche „Bewertungsstrenge“ der Peers durch die Mittelung der Urteile von 20 verschiedenen Peers weitgehend kontrolliert werden kann und somit Voraussetzungen für eine zuverlässige Beurteilung gegeben sind, verdeutlicht auch folgende Analyse, welche beispielhaft für die Orthopädie durchgeführt wurde: Es wurde zunächst für jeden Peer ein „Härtefaktor“ berechnet, der angibt, wie groß die Differenz ist zwischen der Bewertung der Kontrollberichte (Qualitätspunkte bez. der Gesamtbewertung) durch den jeweiligen Peer und der mittleren Bewertung der Kontrollberichte durch alle anderen Peers der gleichen Indikation. Dieser „Härtefaktor“ (der auch als Ausdruck nicht-perfekter Interrater-Reliabilität verstanden werden kann) wurde anschließend als Bonus bzw. Malus zu jeder Bewertung des Peers addiert. Die so resultierende „strengeadjustierte“ Bewertung wurde mit den nicht-adjustierten Werten korreliert. Die Korrelation beträgt für alle orthopädischen Kliniken $r = 0,96$ ($p < 0,001$). Dies weist darauf hin, dass sich die unterschiedliche „Strenge“ der Peers – bedingt durch das gewählte Verteilungsverfahren von 20 zu bewertenden Berichten auf viele verschiedene Rater und die anschließende Mittelung – nur geringfügig auf die Bewertung von Kliniken auswirkt. Die beispielhaft für die Orthopädie vorgestellten Analysen führen in anderen Indikationen zu vergleichbaren Resultaten (die Korrelation zwischen adjustierten und nicht-adjustierten Werten beträgt z. B. für die gastroenterologischen Kliniken $r = 0,95$, $p < 0,001$). Einschränkend ist anzufügen, dass bei diesen Analysen zwar der „Härte-

³ Eine Abweichung von diesem Schema ergibt sich bei Indikationen, für die generell weniger als 20 Peers zur Verfügung stehen. Dies ist bezüglich der hier präsentierten Daten bei Dermatologie, Onkologie/Urologie und Onkologie/Verdauung der Fall.

⁴ In die diesbezügliche Analyse wurden alle Kliniken mit zumindest 16 bewerteten Berichten aufgenommen. Für Kliniken mit weniger als 20 bewerteten Berichten wurden entsprechend kleinere Hälften gebildet.

faktor“ der Peers, nicht aber die Interaktion zwischen Peers und Berichten kontrolliert wird.

Zusammenfassend ist zu sagen, dass die verschiedenen durchgeführten Analysen für eine zufrieden stellende Reliabilität des aggregierten Maßes einer Mittelung über 20 Raterurteile sprechen. Klinikvergleiche sollten – wie es bisher schon in den Qualitätssicherungsprogrammen geschieht – nur auf der Basis dieses aggregierten Maßes erfolgen.

Die Interrater-Reliabilität des Einzelratings bei einer regressionsanalytisch vorhergesagten Gesamtbewertung

Als zusammenfassendes Maß ist die Gesamtbewertung der „Qualität des Reha-Prozesses“ von besonderer Bedeutung. Es ist zu vermuten, dass die Peers bei der Bildung des Gesamturteils in individuell verschiedener Weise die Bewertungen der einzelnen Inhaltsbereiche des Peer Reviews (Anamnese, Diagnostik etc.) gewichten und es auch dadurch zu Unterschieden im Urteilsverhalten von Peers kommt. Es stellt sich die Frage, inwieweit die Interrater-Reliabilität der Gesamtbewertung besser ausfallen würde, wenn sie nicht aus einer individuellen Integration der Bereichsbewertungen resultieren würde, sondern aus einer adäquaten rechnerischen Summierung. Dies soll im Folgenden am Beispiel der Orthopädie und der Kardiologie für die Beurteilung nach Qualitätspunkten beispielhaft untersucht werden.

Als Funktion für die Integration der Bereichsbewertungen zu einem Gesamturteil bietet sich die lineare Regressionsgleichung an, die basierend auf den Bereichsbewertungen das Gesamturteil vorhersagt. Sie gibt – vereinfacht ausgedrückt – die mittlere Gewichtung der Bereichsbewertungen über alle Peers der jeweiligen Indikation wieder. Die Gleichung lautet im Datenpool (ohne Kontrollberichte) für Qualitätspunkte im Indikationsbereich Orthopädie:

$$QP_{ges_pred} = 0.158 \times \text{Anamnese} + 0.134 \times \text{Diagnostik} + 0.186 \times \text{Therapie} + 0.124 \times \text{Soz. med. Stellungnahme} + 0.060 \times \text{Nachsorgekonzept} + 0.352 \times \text{Epikrise} - 0.278$$

Alle Regressionskoeffizienten sind statistisch signifikant ($p < 0,001$). Die höchste Bedeutung besitzt der Bereich Epikrise, gefolgt von der Bewertung der Therapieziele und der Therapie.

Die entsprechende Gleichung für die Kardiologie lautet:

$$QP_{ges_pred} = 0.157 \times \text{Anamnese} + 0.137 \times \text{Diagnostik} + 0.180 \times \text{Therapie} + 0.118 \times \text{Soz. med. Stellungnahme} + 0.065 \times \text{Nachsorgekonzept} + 0.342 \times \text{Epikrise} - 0.150$$

Auch hier sind alle Regressionskoeffizienten statistisch signifikant ($p < 0,001$). Es zeigt sich, dass die Gewichtung der einzelnen Bereiche in der Kardiologie sehr ähnlich ausfällt. Berechnet man die Interrater-Reliabilität (ICC_{unj} für Einzelratings) bezüglich der mittels der Regressionsgleichung vorhergesagten Qualitätspunkte für die Gesamtbewertung (QP_{ges_pred}), so ergibt sich in der Orthopädie ein Reliabilitätswert für QP_{ges_pred} von 0,55. Die Interrater-Reliabilität fällt damit nur geringfügig besser aus als bei der üblichen Gesamtbewertung der Peers in diesem Bereich (s. Tab. 4). Für die Kardiologie ist eine leichte Verschlechterung der Interrater-Reliabilität erkennbar (Reliabilitätswert für

QP_{ges_pred} : 0,21), so dass festzuhalten bleibt, dass die Verwendung rechnerisch ermittelter Gesamtbewertungen zu keiner Verbesserung der Interrater-Reliabilität von Einzelratings führen würde.

Die Bewertung des Rehabilitationsprozesses in den am Programm beteiligten Kliniken

Nachdem im vorstehenden Abschnitt Daten dargestellt wurden, die die Reliabilität der Peer-Review-Bewertungen veranschaulichen, sollen nun die Ergebnisse des Abschneidens der Kliniken in der Erhebungsrunde 2000/2001 beschrieben werden.

Abb. 1 und 2 geben die Prozentverteilungen der zusammenfassenden Bewertungen des gesamten Reha-Prozesses in den einzelnen Indikationsbereichen wieder. Die Darstellungsweise, die alle Indikationsbereiche nebeneinander stellt, wurde gewählt, um die Ergebnisse kompakt zu veranschaulichen. Ein Vergleich der verschiedenen Indikationsbereiche ist dabei nicht intendiert und wäre auch nicht sinnvoll, da denkbar ist, dass die „Urteilsstrenge“ der Peers in den einzelnen Indikationen unterschiedlich ausgeprägt ist. Der Anteil „deutlicher“ und „gravierender Mängel“ bewegt sich über alle Indikationen zwischen 26% und 39%⁵. Der Anteil der Berichte mit „gravierenden Mängeln“ bewegt sich zwischen 3 und 11%, der Anteil der Behandlungsfälle, in denen der Peer insgesamt keine Mängel erkennen konnte, zwischen 7 und 24%.

Im Folgenden sollen nun gezielt diejenigen Bereiche herausgestellt werden, die von den Peers als problematisch angesehen wurden. Dazu wird für die 6 übergeordneten Bereiche der Checkliste jeweils der Prozentsatz der „gravierenden Mängel“ errechnet. Die Ergebnisse der Analyse für die einzelnen Indikationsbereiche sind Tab. 5 zu entnehmen. Die Prozentangaben beziehen sich dabei auf die in Tab. 1 dargestellte Anzahl der Berichte, wobei in die Auswertung aufgrund fehlender Angaben nicht alle Berichte einbezogen werden konnten.

Die häufigsten Problemstellen – gemessen am Prozentsatz „gravierender Mängel“ – liegen für fast alle Indikationen in der sozialmedizinischen Stellungnahme und im Bereich „Verlauf und Epikrise“. Der Anteil der Behandlungsunterlagen mit gravierenden Mängeln schwankt in diesen Bereichen in der Regel zwischen 3 und 10%. Der für die Gesamtbewertung des Reha-Prozesses zentrale Bereich „Therapieziele und Therapie“ wird zwischen den Indikationsbereichen recht unterschiedlich bewertet. Relativ wenige Berichte mit gravierenden Mängeln wurden in der Gynäkologie und Gastroenterologie gesehen (unter 4%), in manchen onkologischen Subindikationen liegt der Anteil gravierender Mängel hingegen zwischen 7 und 10%. Als vergleichsweise problemlos wurde der Bereich Diagnostik bewertet.

⁵ Die in den Abbildungen nicht dargestellte Rate der fehlenden Werte liegt üblicherweise um 2% und ist damit akzeptabel. Eine Ausnahme stellt die Gynäkologie mit 22% fehlenden Werten dar. Detailanalysen zeigen, dass dieses Problem auf einen einzelnen Peer zurückzuführen ist, der vermutlich die Instruktion falsch verstanden hat.

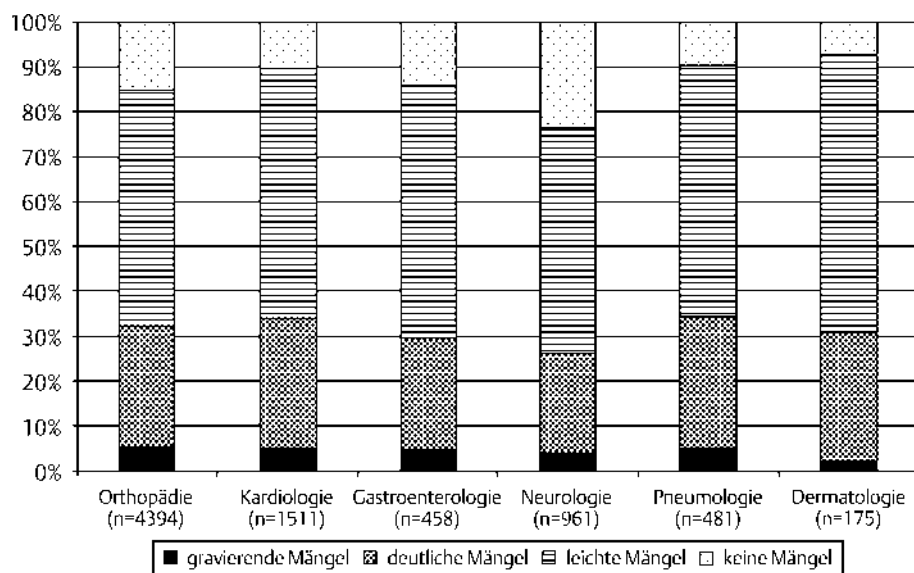


Abb. 1 Verteilung der zusammenfassenden Bewertung des gesamten Reha-Prozesses in verschiedenen Indikationen (I).

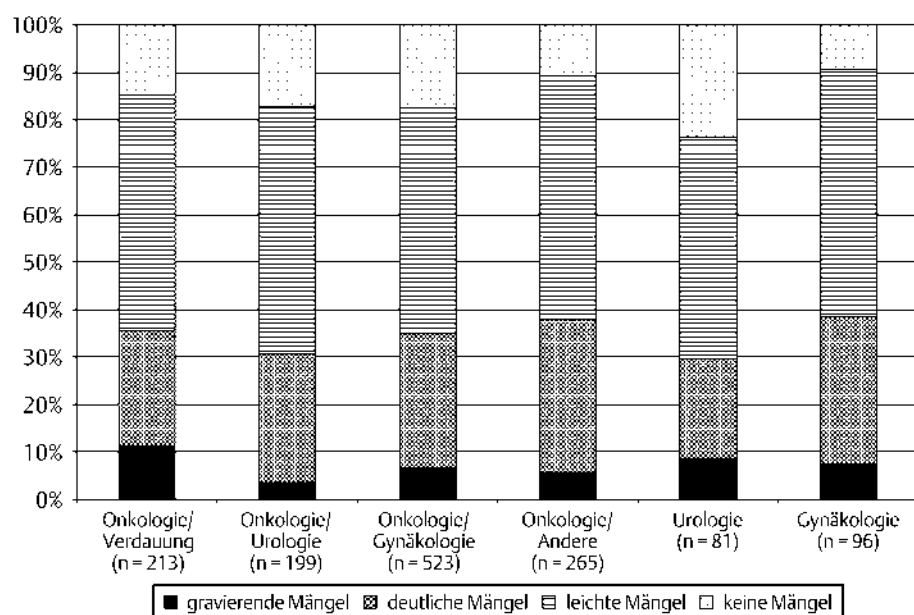


Abb. 2 Verteilung der zusammenfassenden Bewertung des gesamten Reha-Prozesses in verschiedenen Indikationen (II).

Tab. 5 Häufigkeit „gravierender Mängel“ für die zusammenfassenden Bewertungen der Checkliste, bezogen auf die einzelnen Indikationsbereiche (für Indikationen mit mindestens 100 beurteilten Behandlungsfällen; für jede Indikation sind die drei Bereiche mit der größten Häufigkeit „gravierender Mängel“ durch Kursivschrift hervorgehoben)

Bereiche der Checkliste	Indikationen (alle Angaben in %)										
	<i>Orthopädie</i>	<i>Kardiologie</i>	<i>Gastroenterologie</i>	<i>Neurologie</i>	<i>Pneumologie</i>	<i>Dermatologie</i>	<i>Gynäkologie</i>	<i>Onkologie/Verdauung</i>	<i>Onkologie/Urologie</i>	<i>Onkologie/Gynäkologie</i>	<i>Onkologie/andere</i>
Anamnese	5,4	6,0	3,2	4,3	5,8	5,1	4,9	5,0	4,5	4,1	8,0
Diagnostik	4,9	3,9	1,9	2,9	4,4	2,3	4,9	6,5	6,0	7,0	7,5
Therapieziele und Therapie	4,3	4,9	3,3	4,0	4,9	5,1	2,0	10,7	6,6	7,1	8,5
sozialmed. Stellungnahme	6,3	6,5	5,0	6,8	7,2	9,4	9,4	11,4	6,1	10,5	14,3
Nachsorge	4,1	5,8	3,3	4,6	8,1	10,7	2,0	10,2	7,1	7,5	6,2
Verlauf und Epikrise	6,7	7,6	4,8	5,1	7,7	2,8	5,9	8,4	6,5	10,0	9,4
Qualität des Reha-Prozesses	5,4	5,1	4,6	4,0	5,2	2,3	7,3	11,3	3,5	6,7	5,7

Abschließend soll auf der Ebene der einzelnen qualitätsrelevanten Prozessmerkmale untersucht werden, bei welchen Aspekten die Peers häufig Schwachstellen des Rehabilitationsprozesses erkannten. Dazu werden in Tab. 6 für jede Indikation die vier Prozessmerkmale mit dem größten Anteil gravierender Mängel dargestellt. Es zeigt sich, dass insbesondere diejenigen Prozessmerkmale als kritisch beurteilt wurden, die eine Erhebung der Sichtweise und der Selbsteinschätzung des Patienten erfordern (das Krankheitsverständnis in der Anamnese, die Selbsteinschätzung der beruflichen Leistungsfähigkeit in der sozialmedizinischen Stellungnahme sowie die Patienteneinschätzung des Abschlussbefunds in der Epikrise). Dies gilt mehr oder weniger für alle Indikationen. Ein weiterer häufig bemängelter Punkt betrifft die Darstellung der Initiative zur Reha-Antragstellung. In einigen Indikationsbereichen stellen darüber hinaus Aussagen zu Funktionseinschränkungen, zur Motivation und Kooperation des Patienten sowie die Darstellung des Untersuchungsbefunds in der Epikrise Bereiche mit einem hohen Anteil gravierender Mängel dar.

Die zeitliche Entwicklung der Qualitätswerte im Peer-Review-Verfahren

Nach 1999 stellt die Peer-Review-Erhebungsrunde 2000/2001 die zweite dar, an der alle Rehabilitationskliniken im Zuständigkeitsbereich der Rentenversicherungsträger beteiligt waren. Es

ist damit möglich, zu untersuchen, ob sich die Bewertungen des Rehabilitationsprozesses in diesem Zeitraum verändert haben. Insgesamt haben 459 Kliniken, die bereits 1999 in das Peer-Review-Verfahren einbezogen waren, am Erhebungsverfahren 2000/2001 teilgenommen und gleichzeitig ihre Indikationszuordnung nicht verändert. Die 459 Kliniken verteilen sich auf die Indikationsbereiche wie folgt:

– Orthopädie	231 Kliniken
– Kardiologie	77 Kliniken
– Gastroenterologie	22 Kliniken
– Onkologie	39 Kliniken
– Neurologie	53 Kliniken
– Pneumologie	20 Kliniken
– Dermatologie	7 Kliniken
– Urologie	5 Kliniken
– Gynäkologie	5 Kliniken

In die folgenden Auswertungen gehen insgesamt 8167 Berichte aus dem Peer Review 1999 und 7941 Berichte aus dem Peer Review 2000/2001 ein. In Tab. 7 werden die Ergebnisse bez. der zusammenfassenden Bereiche der Checkliste dargestellt. Tab. 8 stellt ergänzend für die sechs am stärksten vertretenen Indikationsbereiche die Veränderung in der Bewertung der Qualität des Reha-Prozesses insgesamt dar.

Für Orthopädie, Kardiologie und Onkologie/Gynäkologie haben sich in den Kategorien „leichte Mängel“ und „deutliche Mängel“

Tab. 6 Prozessmerkmale mit einem hohen Anteil „gravierender Mängel“ (für Indikationen mit mindestens 100 beurteilten Behandlungsfällen; für jede Indikation werden die 4 Merkmale mit dem höchsten Anteil wiedergegeben)

Bereiche der Checkliste	Indikationen (alle Angaben in %)										
	Orthopädie	Kardiologie	Gastroenterologie	Neurologie	Pneumologie	Dermatologie	Gynäkologie	Onkologie/Verdauung	Onkologie/Urologie	Onkologie/Gynäkologie	Onkologie/andere
Anamnese											
Funktionseinschränkungen im Alltag							38,2		20,1		
Funktionseinschränkungen im Beruf	24,5										
Krankheitsverständnis/ Krankheitsverarbeitung	32,5	39,0	32,8	30,8	46,7	43,3	26,8	23,6		30,8	34,6
Initiative zur Reha-Antragstellung			37,8	23,5	37,3	45,6	31,7	30,0	23,5	23,5	34,6
Diagnostik											
klinische Untersuchung: Fähigkeits- und Funktionsstörungen					43,5			25,2			
sozialmedizinische Stellungnahme											
Selbsteinschätzung des Patienten zur beruflichen Leistungsfähigkeit	33,2	38,4	43,6	34,0	47,4	58,3		21,4	25,3	34,0	33,5
Verlauf und Epikrise											
Motivation und Kooperation des Patienten											24,5
Untersuchungsbefund		28,0						24,0	22,4		
Patientenselbsteinschätzung		38,7	35,8	24,3		38,3				24,3	
Übereinstimmung von Beschwerden und Befund	26,5										

Tab. 7 Vergleich der Ergebnisse aus den Jahren 1999 und 2001 (Angaben in %), alle Indikationsbereiche (Anzahl Berichte: 1999: n = 8167, 2001: n = 7941)

	keine Mängel		leichte Mängel		deutliche Mängel		gravierende Mängel	
	1999	2001	1999	2001	1999	2001	1999	2001
Anamnese	24,7	22,0	47,3	49,3	22,6	23,6	5,4	5,1
Diagnostik	33,1	28,2	44,7	47,7	17,9	19,8	4,3	4,2
Therapieziele/ Therapie	33,0	29,9	45,4	47,6	16,9	18,2	4,7	4,2
sozialmed. Stellungnahme	35,4	33,5	41,0	42,6	16,8	17,4	6,9	6,5
Nachsorge- konzept	52,8	47,5	32,4	35,8	10,1	11,5	4,7	5,1
Verlauf und Epikrise	24,2	20,7	46,1	48,5	22,8	24,6	6,9	6,3
Qualität des Reha-Prozesses	18,3	15,1	50,9	53,3	25,1	26,6	5,7	4,9

Tab. 8 Vergleich der Ergebnisse aus den Jahren 1999 und 2001 (Angaben in %). Zusammenfassende Bewertung der Qualität des Reha-Prozesses in den 6 am stärksten vertretenen Indikationsbereichen

	keine Mängel		leichte Mängel		deutliche Mängel		gravierende Mängel	
	1999	2001	1999	2001	1999	2001	1999	2001
Orthopädie	19,4	15,6	50,2	53,2	25,1	26,3	5,2	4,9
Kardiologie	15,6	10,7	51,7	55,4	26,1	29,1	6,6	4,7
Neurologie	22,6	23,7	52,1	50,1	20,2	22,4	5,1	3,8
Gastro- enterologie	18,7	14,4	54,3	54,7	23,6	26,3	3,4	4,6
Pneumologie	13,8	11,0	53,0	56,3	29,2	27,9	4,1	4,8
Onkologie/ Gynäkologie	20,6	17,7	45,4	49,4	27,0	27,1	7,1	5,8

die Häufigkeiten bei der Befragung 2001 gegenüber 1999 in der Regel leicht erhöht, während sie bei den Kategorien „keine Mängel“ und „gravierende Mängel“ zurückgegangen sind. Dies kann als „Trend zur Mitte“ interpretiert werden. In der Gastroenterologie ist ein geringfügig negativer Trend erkennbar. In den Indikationsbereichen Neurologie und Pneumologie sind nur geringe Bewertungsveränderungen erkennbar, die darüber hinaus keinem bestimmten Muster folgen. Alles in allem ist im Jahr 2000/2001 gegenüber 1999 keine deutlich unterschiedliche Bewertung erkennbar.

Berücksichtigt man zusätzlich die Daten aus der Pilotphase des Peer Reviews von 1998 (s. Tab. 9), so sinkt zwar die Fallzahl erheblich, da 1998 nur ca. 100 Kliniken – auf freiwilliger Basis – am Verfahren beteiligt waren, das Bild der zeitlichen Veränderungen vervollständigt sich jedoch: Zwischen 1998 und 1999 ist eine deutliche und statistisch signifikante positive Entwicklung erkennbar. Zwischen 1999 und 2000/2001 hingegen ist keine weitere Verbesserung festzustellen.

Tab. 9 Vergleich der Ergebnisse aus den Jahren 1998, 1999 und 2001 (Prüfung der Signifikanz), alle Indikationsbereiche (Anzahl Berichte: 1998: n = 1118, 1999: n = 1060, 2001: n = 1024)

	1998 vs. 1999	1998 vs. 2000/2001	1999 vs. 2000/2001
	Anamnese	↑	↑
Diagnostik	↑	↑	–
Therapieziele/ Therapie	↑	↑	–
sozialmedizinische Stellungnahme	↑	↑	–
Nachsorgekonzept	↑	↑	–
Verlauf und Epikrise	↑	↑	–
Qualität des Reha-Prozesses	↑	↑	–

Mann-Whitney-U-Test

↑ positiver und statistisch signifikanter ($p < 0,01$) zeitlicher Trend

– kein signifikanter Trend

Analysiert man lediglich die Gruppe der Kliniken, bei denen 1999 ein besonderer Optimierungsbedarf bestand, weil sie vergleichsweise schlecht bewertet wurden (das unterste Quartil der Verteilung bezüglich der Gesamtbewertung des Reha-Prozesses), so ergibt sich folgendes Bild (s. Tab. 10): Die 25% der Einrichtungen, die 1999 am schlechtesten bewertet wurden, verbesserten sich auch zwischen 1999 und 2000/2001 deutlich: Der Anteil der gravierenden Mängel hat sich bei der Gesamtbewertung des Reha-Prozesses um 40% reduziert, der Anteil der Berichte, die mit „keine Mängel“ beurteilt wurden, hat sich um 70% erhöht.

Einschränkend ist zu sagen, dass die Möglichkeit besteht, dass die hier aufgezeigten positiven Veränderungen zumindest teilweise Ausdruck eines statistischen Artefakts, der sog. „Regression zur Mitte“, sind. Gemeint ist damit das Phänomen, dass sich Werte, die bei einer ersten Messung eine starke positive oder negative Ausprägung aufweisen, bei einer Wiederholungsmessung in den mittleren Bereich verschieben. Um die Möglichkeit dieser Erklärung zu prüfen, wurde ein sog. „residualized change score“ (kurz RCS) berechnet [21]. Dieser stellt in dem hier untersuchten Fall den Residualwert (Differenz des vorhergesagten zum tatsächlichem Wert) bei der Vorhersage der Veränderung 1999 vs. 2000/2001 durch den Wert der Klinik im Jahr 1999 dar. Das heißt, der RCS-Wert kontrolliert den Ausgangswert 1999 und gibt somit an, welche Veränderung zu erwarten gewesen wäre, wenn alle Kliniken auf gleichem Niveau begonnen hätten. Aufgrund dieses Umstandes korrigiert der RCS-Wert den Einfluss von Regressionseffekten⁶.

⁶ Um den Einfluss der Regressionseffekte nach dem hier beschriebenen Verfahren abschätzen zu können, wurde das arithmetische Mittel der Bewertungen nach Mängelkategorien innerhalb einer Klinik bestimmt. Diese Operation ist streng genommen bei den nur ordinalskalierten Mängelkategorien nicht erlaubt. Da die Ergebnisse jedoch nur zur Orientierung dienen, scheint uns die Analyse bei vorsichtiger Interpretation vertretbar. Mögliche Alternativen (z. B. die Verwendung des Medians) wären mit anderweitigen Nachteilen verbunden. Die Verwendung der intervallskalierten Qualitätspunkte war nicht möglich, da diese 1999 noch nicht routinemäßig eingesetzt wurden.

Tab. 10 Zusammenfassende Bewertungen der einzelnen Bereiche der Checkliste und des gesamten Reha-Prozesses (Angaben in %) für die 25% im Jahr 1999 am schlechtesten beurteilten Kliniken (alle Indikationsbereiche, Anzahl Berichte 1999: n = 1989; 2001: n = 1938)

	keine Mängel		leichte Mängel		deutliche Mängel		gravierende Mängel	
	1999	2001	1999	2001	1999	2001	1999	2001
Anamnese	7,5	12,0	38,7	42,7	39,6	34,1	14,2	11,2
Diagnostik	18,0	19,8	43,6	45,9	28,7	26,3	9,7	8,0
Ziele und Therapie	14,4	20,8	46,4	47,4	29,0	25,2	10,2	6,6
sozialmed. Stellungnahme	19,5	25,6	41,7	42,5	24,6	22,2	14,2	9,7
Nachsorgekonzept	38,4	40,3	36,0	38,3	15,9	14,0	9,7	7,4
Verlauf und Epikrise	8,3	13,0	39,9	44,1	36,2	32,1	15,6	10,8
Qualität des Reha-Prozesses	5,0	8,5	37,5	45,9	43,2	37,0	14,3	8,6

Sollte sich zeigen, dass auch der RCS-Wert in der Gruppe der im Jahr 1999 am schlechtesten bewerteten Einrichtungen besser ausfällt als in der Restgruppe, so spräche dies dafür, dass der Regressionseffekt den obigen Befund nicht gänzlich erklären kann, dass also tatsächlich die 1999 nicht so gut bewerteten Kliniken eine positive Entwicklung durchlaufen haben. Für die Analyse wurden drei Gruppen gebildet:

Gruppe 1: das oberste Viertel in der Verteilung der Bewertung des Reha-Prozesses insgesamt (1999)

Gruppe 2: die mittleren 50% in der Verteilung der Bewertung des Reha-Prozesses insgesamt (1999)

Gruppe 3: das unterste Viertel in der Verteilung der Bewertung des Reha-Prozesses insgesamt (1999)

Insgesamt gehen die Daten von 459 Kliniken ein. Für die drei Gruppen wurde das mittlere standardisierte Residuum bei der Vorhersage der Veränderung 1999 vs. 2000/2001 durch den Wert 1999 bestimmt (jeweils anhand der Mittelwerte über alle vorliegenden Bewertungen). Das standardisierte Residuum ist negativ, wenn die Veränderung zwischen 1999 und 2000/2001 besser ausfällt als aufgrund des Ausgangswerts zu erwarten gewesen wäre, und es ist positiv, wenn die Veränderung zwischen 1999 und 2000/2001 schlechter ausfällt, als aufgrund des Ausgangswerts zu erwarten gewesen wäre.

Die Residuen fallen in den drei Gruppen wie folgt aus:

Gruppe 1: -0,029
 Gruppe 2: 0,041
 Gruppe 3: -0,046

Es zeigt sich also, dass die Verbesserung der Werte von Kliniken im untersten Quartil nicht hinreichend durch den Regressionseffekt erklärt werden kann. Auch nach der Kontrolle von Regres-

sionseffekten machen die Kliniken im untersten Quartil die positivste Entwicklung durch.

Zusammenfassung und Diskussion

Zusammenfassend ist zur Reliabilität des Peer Reviews zu sagen, dass die Gutachterübereinstimmung bezüglich des Urteils eines einzelnen Peers nur in der Orthopädie zufrieden stellend ausfällt. Bezüglich der für die Praxis des Verfahrens viel relevanteren Bewertung auf der Basis des Mittelwerts über die Beurteilung von 20 Berichten durch 20 Peers sieht die Situation anders aus: Auch wenn hier ein exakter Reliabilitätswert aufgrund der vorliegenden Daten nicht berechenbar ist, deuten die dargestellten Analysen darauf hin, dass die Reliabilität des aggregierten Werts in der Regel zufrieden stellend ist. Auch Hofer et al. [21] konnten zeigen, dass solche aggregierten Maße eine deutlich höhere Reliabilität aufweisen. Die Autoren empfehlen eine Mittelung über zumindest 20 Gutachter, um trotz einer nicht optimalen Interrater-Reliabilität aussagekräftige Klinikvergleiche durchführen zu können. Auch im kanadischen „Peer Assessment Program“ des College of Physicians and Surgeons of Ontario wird eine Zahl von ca. 20–30 Behandlungsfällen als ausreichend für die Sicherung der Reliabilität der Peer-Bewertung angesehen. Man kann somit davon ausgehen, dass das im Bereich der Renten- und Krankenversicherung gewählte Vorgehen einen sinnvollen Kompromiss zwischen ökonomischen und methodischen Anforderungen darstellt. Erforderlich wäre allerdings eine exakte empirische Analyse der Reliabilität einer aggregierten Bewertung, die mit der hier vorgelegten Studie noch nicht geleistet werden konnte. Hierzu müsste eine komplette Mehrfachbewertung von Einrichtungen (d. h. beispielsweise die Bewertung von 2-mal 20, also 40 Berichten pro Klinik) erfolgen.

Die auf der Basis der Einzelratings nur teilweise zufrieden stellende Interrater-Reliabilität entspricht den Ergebnissen verschiedener vorliegender Arbeiten und Literaturübersichten, die aufzeigen, dass selbst bei strukturierten Verfahren und geschulten Gutachtern die Interrater-Reliabilität eines Peer-Reviews im Gesundheitswesen selten sehr gut ausfällt [2,22–25]. Für ein Verfahren, welches eine kontinuierliche Weiterentwicklung anstrebt, stellt sich trotz der auf der aggregierten Ebene zufrieden stellenden Reliabilität die Frage, welche Maßnahmen zur Verbesserung der Gutachterübereinstimmung ergriffen werden können. In diesem Zusammenhang sind folgende Maßnahmen denkbar:

• Rückmeldung der „Härtefaktoren“ an die Peers

Es kann vermutet werden, dass die Rückmeldung der „Härtefaktoren“ (Abweichung vom Mittelwert der anderen Peers) ähnlich wie in Modellen formaler Konsensusverfahren (vgl. z. B. [13,26]) zu einer Annäherung der extrem Urteilenden an den Gruppenmittelwert führt. Dieses würde bez. des Finn-Koeffizienten unmittelbar zu einer Erhöhung der Interrater-Reliabilität führen, bez. der Intraklassenkorrelationskoeffizienten nur dann, wenn damit nicht gleichzeitig auch eine Reduktion der Varianz der Bewertung unterschiedlicher Berichte verbunden wäre. Damit wird deutlich, dass die Auswirkungen einer Rückmeldung der „Härtefaktoren“ auf die Interrater-Reliabilität vorab nicht eindeutig vorhergesagt werden können.

Von Vorteil wäre die Rückmeldung, wenn sie die Peers dazu anleiten würde, unter Beachtung der „wahren“ Variabilität der Qualität der Behandlungsfälle homogener zu urteilen. Sollte die Rückmeldung jedoch dazu führen, dass Gutachter unabhängig von der Qualität der Berichte dazu tendieren, Bewertungen im Mittelbereich zu vergeben, würde die Diskriminationsfähigkeit des Verfahrens leiden. Zu beachten ist, dass die Rückmeldung des individuellen „Härtefaktors“ eine Intervention darstellt, die die Urteilsstrenge des Gutachters auf eine nicht eindeutig vorhersehbare Weise verändern wird. Das heißt, dass nach jedem „Härtefaktor“-Feedback eine erneute Erhebung des Härtefaktors des jeweiligen Peers erfolgen muss. Es wird nicht möglich sein, über mehrere Erhebungen hinweg eine auf vielen Beurteilungen basierende und damit reliablere Bestimmung des Härtefaktors von Peers vorzunehmen. Einer differenzierten, bereichsspezifischen Rückmeldung (für Anamnese, Diagnostik etc.) wäre der Vorzug zu geben, da sie dem Peer auch eine differenzierte Anpassung seiner Urteilsstrenge nahelegt. Bei der Rückmeldung eines einzigen globalen „Härtefaktors“ wäre die Wahrscheinlichkeit vermutlich größer, dass Gutachter die oben beschriebene, zu geringerer Diskriminationsfähigkeit führende unspezifische Tendenz zur Mitte zeigen würden.

- **Intensivierung von Schulungs- und Informationsmaßnahmen**

Bekannt ist, dass fallbezogene Gruppendiskussionen die Interrater-Reliabilität erhöhen (vgl. z. B. [27]). Die von der Rentenversicherung durchgeführten Peer-Schulungen besitzen somit neben dem Schulfungseffekt vermutlich auch einen positiven Einfluss auf die Gutachterübereinstimmung. Neben Schulungs- und Diskussionsforen als Maßnahmen zur Verbesserung der Reliabilität wäre auch an die Möglichkeit der Versendung schriftlicher Informationsmaterialien an die Peers zu denken (z. B. exemplarische Berichtsbewertungen zur Verdeutlichung einer manualgemäßen Bearbeitung).

- **Begrenzung des Kreises der Peers**

Wenn eine intensivere Auseinandersetzung der Peers mit den Bewertungsmaßstäben des Verfahrens gewünscht wird, wäre es ferner denkbar, diese indirekt durch eine Begrenzung des Kreises der Peers und damit durch eine erhöhte Frequenz bewerteter Berichte pro Person zu erreichen. Durch eine entsprechende Selektion wäre es u. U. auch möglich, eher motivierte Peers anzusprechen, die bereit sind, sich intensiver mit dem Verfahren zu befassen. Ferner bestünde die Möglichkeit, Peers mit extremen Ausprägungen des „Härtefaktors“ vom Bewertungsverfahren auszuschließen.

- **„Qualitätssicherung der Peers“**

Mangelnde Gutachterübereinstimmung kann auf das Instrument, auf den Gutachter oder auf beide zurückzuführen sein. Fokussiert man den Gutachter (die Fokussierung des Instruments erfolgt im nächsten Abschnitt), so stellt sich die Frage, ob der Peer die ihm vorgegebenen „Standards“ der Beurteilung auch einhält. Unter diesem Gesichtspunkt wäre es denkbar, eine Form der „Qualitätskontrolle der Peers“ einzuführen: Für einige geeignete, prototypische Berichte werden von einem Expertenteam unter strenger und systematischer Beachtung aller Beurteilungsmaßstäbe des Peer-Review-Manuals Bewertungen erstellt, die den Anspruch erheben können, musterhaft zu sein. Die auf diese Weise beurteilten Berichte werden als Kontrollberichte eingesetzt, so dass nicht nur die

Variabilität der Peer-Bewertungen bezüglich identischer Berichte analysiert werden kann, sondern auch die individuelle Abweichung von der Bewertung des Musterberichts. Bei Peers mit extremen Abweichungen von der Musterbewertung wären Maßnahmen wie Nachschulung oder Ausschluss aus dem Bewertungsverfahren denkbar.

- **Präzisierung des Manuals**

Im vorherigen Abschnitt wurde davon ausgegangen, dass es primär die Gutachter sind, bei denen im Rahmen der Verbesserung der Interrater-Reliabilität anzusetzen ist. Eine solche Sichtweise wäre sicherlich einseitig. Unter Umständen sind auch Teile des Peer-Review-Manuals uneindeutig, so dass Peers selbst bei bestem Willen und Vermögen nicht zu homogenen Urteilen kommen können. Empfehlenswert wäre deshalb – zusätzlich zu den bisher schon vorliegenden einzelfallbezogenen Hinweisen – eine systematische Prüfung der Items und Formulierungen, mit denen Peers Bewertungsprobleme haben, da ihnen das zu Bewertende und die dabei anzuwendenden Maßstäbe nicht eindeutig klar sind.

Entsprechende Maßnahmen sind parallel zur Durchführung der hier referierten Auswertungen im Frühjahr und Sommer 2002 vorgenommen wurden. Im Rahmen einer generellen Überarbeitung der Peer-Review-Checkliste, die zur Erstellung einer für Rentenversicherung und gesetzliche Krankenkassen einheitlichen Version geführt hat, wurden u. a. auch die Manualtexte geprüft und ggf. präzisiert (entsprechende Publikation ist in Vorbereitung). Ob diese Maßnahmen zu einer Verbesserung der Interrater-Reliabilität führen werden, wird sich im Rahmen der kommenden Peer-Review-Erhebungen in den Qualitätssicherungsprogrammen von Renten- und Krankenversicherung zeigen.

Bezüglich der Bewertung des Rehabilitationsprozesses in den über 550 am Verfahren beteiligten Einrichtungen ist zu sagen, dass über alle Indikationsbereiche hinweg in den Bereichen Anamnese, sozialmedizinische Stellungnahme und Verlauf/Epi-krise von den Peers die meisten Mängel gesehen wurden. Dies betraf insbesondere Prozessmerkmale, die eine Erhebung der Sichtweise und Selbsteinschätzung des Patienten erfordern (z. B. das Krankheitsverständnis oder die Selbsteinschätzung der beruflichen Leistungsfähigkeit). Die Bedeutung der Wahrnehmung und der subjektiven Konzepte des Patienten für die Kooperation im Rahmen der Rehabilitation, aber auch für die berufliche Reintegration und die poststationäre Inanspruchnahme von Gesundheitsdienstleistungen scheint noch nicht in allen Kliniken hinreichende Berücksichtigung in den Behandlungsprinzipien und Therapiekonzepten gefunden zu haben.

Erfreulich ist, dass im Jahr 2001 trotz der in manchen Kliniken nicht einfachen Personalsituation das Niveau der Bewertungen von 1999 und damit der positive Trend gegenüber 1998 gehalten werden konnte. Die in 1999 schlecht bewerteten Kliniken konnten sich im Jahr 2001 sogar leicht verbessern. Es konnte gezeigt werden, dass dieses Phänomen nicht vollständig auf einen statistischen Artefakt (Regressionseffekt) zurückzuführen ist. Ähnliche Tendenzen (Verbesserung der anfangs schlechter bewerteten Einrichtungen) fanden auch Norton et al. [6] bei dem schon erwähnten Peer-Review-Programm unter kanadischen Allgemeinärzten.

Wenn man versucht, diese Daten zu erklären, so könnte man die Vermutung aufstellen, dass das Peer-Review-Verfahren zu einer Qualitätsentwicklung insbesondere bei den Einrichtungen beiträgt, deren Qualitätsniveau im Vergleich zu anderen Einrichtungen nicht sehr gut ist (z. B. durch die „aufrüttelnde“ Wirkung einer schlechten Rückmeldung). Es scheinen aber auch die schon sehr gut bewerteten Kliniken davon zu profitieren (u. U. durch den motivierenden Effekt eines positiven Feedbacks). Kliniken, die sich hingegen im Mittelfeld bewegen, haben sich zwischen 1999 und 2001 eher leicht verschlechtert.

Die Ursachen des Fehlens eines anhaltend positiven Trends über alle Kliniken und Erhebungszeitpunkte hinweg sind beim jetzigen Wissensstand nicht eindeutig identifizierbar. Es lassen sich nur verschiedene alternative Erklärungsansätze aufzeigen und ihre Wahrscheinlichkeit abwägen. Denkbar wäre, dass sich die Kliniken hinsichtlich der mit dem Peer Review erfassten Qualitätsindikatoren nicht weiter verbessert haben, weil weniger günstige Rahmenbedingungen – die unabhängig vom Peer-Review-Verfahren bestehen – einer positiven Qualitätsentwicklung entgegenstanden. Eine weitere Erklärungsmöglichkeit bestünde darin, dass sich die Kliniken hinsichtlich der mit dem Peer-Review erfassten Qualitätsindikatoren zwar im Grunde verbessert haben, dass sich dies aber nicht abbildet, da sich gleichzeitig die Bewertungsmaßstäbe der Peers verändert (verschärft) haben. Während in der Anfangsphase der Einführung des Peer-Reviews leichte Abweichungen von den Anforderungen unter Umständen toleriert wurden, da sich die Kliniken auf die neuen Anforderungen erst einzustellen hatten, könnte es sein, dass diese Toleranz mit zunehmender Routinisierung des Verfahrens abnimmt. Vermutet werden könnte auch, dass mit den massiven Verbesserungen zwischen 1998 und 1999 das unter den gegebenen Rahmenbedingungen an Verbesserungen Mögliche erreicht wurde und weitere Optimierungen Ressourcen erfordern würden, die zurzeit nicht von allen Kliniken bereitgestellt werden. Zur Klärung der Hintergründe müssten die aufgezeigten Erklärungsmodelle empirisch untersucht werden.

Literatur

- 1 Moseley A, Rotem W. Establishing standards for the provision of brain injury services. *Brain Injury* 1995; 9: 355–364
- 2 Sheahan SL, Simpson C, Rayens MK. Nurse practitioner peer review: process and evaluation. *Journal of the American Academy of Nurse Practitioners* 2001; 13: 140–145
- 3 Saturno PJ, Palmer RH, Gascon JJ. Physician attitudes, self-estimated performance and actual compliance with locally peer-defined quality evaluation criteria. *International Journal for Quality in Health Care* 1999; 11: 487–496
- 4 Wakefield DS, Helms CM. The role of peer review in a health care organization driven by TQM/CQI. *The Joint Commission Journal of Quality Improvement* 1995; 21: 227–231
- 5 Shaw CD. External quality mechanisms for health care: Summary of the ExPeRT project on visitation, accreditation, EFQM and ISO assessment in European Union countries. *International Journal for Quality in Health Care* 2000; 12: 169–175
- 6 Norton PG, Dunn EV, Beckett R, Faulkner D. Long-term follow-up in the peer assessment program for nonspecialist physicians in Ontario, Canada. *The Joint Commission Journal of Quality Improvement* 1998; 24: 334–341
- 7 Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, Newhouse JP, Weiler PC, Hiatt HH. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *New England Journal of Medicine* 1991; 324: 370–376
- 8 Verband Deutscher Rentenversicherungsträger. Das Qualitätssicherungsprogramm der gesetzlichen Rentenversicherung in der medizinischen Rehabilitation. Instrumente und Verfahren. DRV-Schriften, 2000; (18)
- 9 Egner U, Gerwin H, Schliehe F. Das bundesweite Reha-Qualitätssicherungsprogramm der gesetzlichen Rentenversicherung. *Zeitschrift für Ärztliche Fortbildung und Qualitätssicherung* 2002; 96: 4–9
- 10 Jäckel WH, Maier-Riehle B, Protz W, Gerdes N. Peer-Review: Ein Verfahren zur Analyse der Prozessqualität stationärer Rehabilitationsmaßnahmen. *Die Rehabilitation* 1997; 36: 224–232
- 11 Maier-Riehle B, Gerdes N, Protz W, Jäckel WH. Übereinstimmung und Unterschiede zwischen Beurteilern bei einem Peer-Review-Verfahren. *Gesundheitswesen* 1998; 60: 290–296
- 12 Normand S-LT, McNeil BJ, Peterson LE, Palmer RH. Eliciting expert opinion using the Delphi technique: identifying performance indicators for cardiovascular disease. *International Journal for Quality in Health Care* 1998; 10: 247–260
- 13 Buetow SA, Coster GD. New Zealand and United Kingdom experiences with the RAND modified Delphi approach to producing angina and heart failure criteria for quality assessment in general practice. *Quality in Health Care* 2000; 9: 222–231
- 14 Farin E, Jäckel WH. Qualitätssicherung in der medizinischen Rehabilitation. *Die Betriebskrankenkasse* 2001; (8): 376–381
- 15 Vogel H, Neuderth S, Schieweck R, Gerlich C, Weber-Falkensammer H, Mehrhoff F. Qualitätssicherung in den Kliniken zur Berufsgenossenschaftlichen stationären Weiterbehandlung der gesetzlichen Unfallversicherung [Abstract]. In: *Verband Deutscher Rentenversicherungsträger (Hrsg). Tagungsband, „Wissenstransfer zwischen Forschung und Praxis“*, 10. Rehabilitationswissenschaftliches Kolloquium, 12. bis 14. März 2001 in Halle/Saale. DRV-Schriften 2001; (26): 48–49
- 16 Kendall MG, Smith BB. The problem of m rankings. *Annals of Mathematical Statistics* 1939; 10: 275–287
- 17 Asendorpf J, Wallbott HG. Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie* 1979; 10: 243–252
- 18 Finn RH. A note on estimating the reliability of categorical data. *Educational and Psychological Measurement* 1970; 30: 71–76
- 19 Wirtz M, Caspar F. Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen: Hogrefe, 2002
- 20 Lienert GA, Raatz U. Testaufbau und Testanalyse. Weinheim: Psychologie Verlags Union, 1998
- 21 Campbell JL, Kenny DA. A primer on regression artifacts. New York: Guilford Press, 1999
- 22 Goldman RL. The reliability of peer assessments: A meta-analysis. *Evaluation & The Health Professions* 1994; 17: 3–21
- 23 Hofer TP, Bernstein SJ, De Monner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Medical Care* 2000; 38: 152–161
- 24 Weingart SN, Mukamal K, Davis RB, Davies DT, Palmer RH, Cahalane M, Hamel MB, Phillips RS, Iezzoni LI. Physician-reviewers' perceptions and judgements about quality of care. *International Journal for Quality in Health Care* 2001; 13: 357–365
- 25 Neale G, Woloshynowych M. Retrospective case record review: a blunt instrument that needs sharpening. *Quality and Safety in Health Care* 2002; 12: 2–3
- 26 Brook RH, Chassin MR, Fink A, Solomon DH, Koscoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *International Journal of Technology Assessment in Health Care* 1986; 2: 53–63
- 27 Levine RD, Sugarman M, Schiller W, Weinshel S, Lehning EJ, Lagasse RS. The effect of group discussion on interrater reliability of structured peer review. *Anesthesiology* 1998; 89: 507–515