Soz.- Präventivmed. 45 (2000) 258–266
0303-8408/00/060258-09 $ 1.50 + 0.20/0

Sebastian Schneeweiss[1,2], Oliver Sangha[1,3], Harald Siebert[4], Matthias Hübner[5], Jörg Fuhrmann[5], Manfred Wildner[1], Jens Witte[6]

[1] *Bayerischer Forschungsverbund Public Health und Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) der Ludwig-Maximilians-Universität, München*
[2] *Department of Medicine, Harvard Medical School, Boston*
[3] *Department of Health Policy & Management, Harvard University School of Public Health, Boston*
[4] *Medizinischer Dienst der Krankenversicherung in Bayern, Augsburg*
[5] *Medizinischer Dienst der Krankenversicherung in Hessen, Oberursel*
[6] *Klinik für Allgemein- und Viszeralchirurgie, Zentralklinikum Augsburg*

Reproduzierbarkeit eines deutschsprachigen Instruments zur Erfassung der Notwendigkeit von stationären Behandlungen in der Chirurgie

Zusammenfassung

Die Prüfung der Notwendigkeit von Krankenhausbehandlungen hat in den vergangenen Jahren in Deutschland zunehmend an Aktualität gewonnen. Die bisherigen Erfahrungen mit Prüfungen durch die Medizinischen Dienste der Krankenversicherung in den einzelnen Bundesländern verdeutlichen dabei den Bedarf an einem standardisierten, reproduzierbaren und validen Instrument zur Beurteilung der Notwendigkeit von Krankenhausaufenthalten. Ziel der vorliegenden Studie ist es die deutschsprachige Adaptierung eines solchen Instruments, basierend auf dem „Appropriateness Evaluation Protocol" (AEP), in Hinblick auf dessen Reproduzierbarkeit beim Einsatz in der chirurgischen Versorgung zu testen. Aus allen 2672 Aufnahmen eines Kalenderjahres aus der chirurgischen Abteilung eines Krankenhauses der Grundversorgung wurden 54 Patienten für die Untersuchung der Interrater- und 51 Patienten für die Intrarater-Reliabilität zufällig ausgewählt. Es wurden Gesamtübereinstimmungen, spezifische Übereinstimmungen und Kappa-Statistiken für die Beurteilung von Krankenhausaufnahmen und einzelnen Behandlungstage berechnet. Das deutsche AEP zeigte eine Interrater-Übereinstimmung von 74% (62–86%) bei den Krankenhausaufnahmen (Kappa = 0,44) und 84% (79–88%) für alle Behandlungstage (K = 0,55) sowie eine Intrarater-Übereinstimmung von 88% (79%–97%) für die Aufnahmen (K = 0,60) und 88% (85%–92%) für die Behandlungstage (K = 0,70). Die beobachteten Übereinstimmungen sind unabhängig von der Länge der Liegezeit und unabhängig vom Anteil angemessener oder nicht angemessener Krankenhaustage. Ein Standardinstrument mit hoher Reproduzierbarkeit ist für das interne Qualitätsmanagement von Krankenhäusern notwendig, um sich auf die Herausforderungen eines sich zunehmend konsolidierenden Gesundheitsmarktes vorzubereiten. Das deutsche AEP weist eine hinreichend hohe Reproduzierbarkeit auf, um Ineffizienzen im Bereich der stationären chirurgischen Versorgung zu identifizieren.

Nachdem ein Grossteil der Gesundheitsausgaben im Sektor der stationären Versorgung entstehen, werden zunehmend Anstrengungen unternommen, Wirtschaftlichkeitsreserven in diesem Bereich zu mobilisieren. Dabei haben die Prüfung der Notwendigkeit von Krankenhausbehandlungen, auch Fehlbelegungsprüfung genannt, in den vergangen Jahren in Deutschland zunehmend an Aktualität gewonnen. Die Verringerung unnötiger Krankenhausaufenthalte kann dabei nicht nur Kosten reduzieren, sondern gleichzeitig die Qualität der Versorgung verbessern[1]. Die Diskussion in Deutschland erlebte einen ersten Gipfel, nachdem der Medizinische Dienst der Spitzenverbände (MDS) im Frühjahr 1997 ein Gutachten veröffentlichte, welches die Prüfung von 63665 Patienten aus 11 Bundesländern zusammenfasste. Das Vorhaben hatte nach § 275a SGB V Modellcharakter und basierte auf einer freiwilligen Teilnahme von je einem Krankenhaus jeder Versorgungsstufe in 11 Bundesländern. Obschon das seinerzeit verwendete Verfahren erhebliche Probleme aufweist, haben die regionalen

medizinischen Dienste der Krankenversicherung (MDK) ihre Prüfung von Krankenhäusern fortgesetzt. Die vordringlichen Probleme bei den aktuellen Verfahren liegen in deren zweifelhaften und ungeprüften Reproduzierbarkeit (Reliabilität) und Gültigkeit (Validität). Entsprechend lassen Ergebnisse dieser Prüfungen erheblichen Spielraum für alternative Erklärungen unabhängig von dem tatsächlichen Ausmass der Fehlbelegung zu.
In verschiedenen Ländern wurden zum Teil unangemessene Krankenhausaufenthalte von erheblichem Umfang berichtet[2–11], aber nur wenigen Häusern ist es gelungen, diesen Anteil auf der Basis solcher Studien und Feedback-Mechanismen tatsächlich zu reduzieren[10,12,13].
Seit kurzem existiert eine deutsche Version des international verbreiteten Appropriateness Evaluation Protocol (AEP) für die Beurteilung von chirurgischen und internistischen Krankenhauspatienten[14].
Ziel der vorliegenden Studie ist es, die Reproduzierbarkeit dieses deutschen Instrumentes zur Beurteilung der Angemessenheit von Krankenhausaufnahmen und konsekutiven Behandlungstagen zu untersuchen. Insbesondere soll dabei die Reproduzierbarkeit des longitudinalen Verlaufs betrachtet werden.

Methoden

Entwicklung des Instruments

Die Methodik, Angemessenheit von Krankenhausaufenthalten zu beurteilen, basiert auf dem AEP. Das AEP, ursprünglich von Gertman und Restuccia in den USA entwickelt[2], wurde später modifiziert, um sowohl Krankenhausaufnahmen als auch einzelne Krankenhaustage bezüglich ihrer Notwendigkeit zu beurteilen[15]. Das Instrument wurde weltweit in einer grossen Zahl von Krankenhäusern mit einer positiven Bewertung seines Nutzwertes angewandt[5,7,10,15–18]. Methodische Studien zum AEP konzentrierten sich auf die Beurteilung der Reliabilität und Validität[2,7,18–21], die Anpassung des ursprünglichen Instruments an Krankenhäuser mit unterschiedlichen Versorgungsaufträgen, Abteilungen oder Diagnosegruppen[17,22–25].
Das AEP verwendet 24 diagnoseunabhängige Kriterien der medizinischen Angemessenheit, um einzelne Tage eines Krankenhausaufenthalts zu beurteilen. Davon betreffen neun Kriterien medizinische Leistungen und Prozeduren, fünf Kriterien die pflegerischen Versorgung und 10 Kriterien den Gesundheitszustand des Patienten, die jeweils eine enge stationäre Überwachung notwendig machen[14]. Wurde ein Tag als unangemessen bewertet, d.h. es konnten in den Krankenhausakten keine Hinweise gefunden werden, die mindestens einem der 24 Kriterien zuordenbar sind, dann erlaubt das AEP eine Beschreibung der Gründe, die potentiell für die medizinische Unangemessenheit des stationären Aufenthalts verantwortlich waren. Diese werden einer komplementären Liste von Gründen und Verantwortlichkeiten entnommen. Das AEP misst zusätzlich die medizinische Notwendigkeit von Krankenhausaufnahmen anhand von 16 Kriterien, die sich auf den klinischen Zustand der Patienten, die Notwendigkeit von medizinischen Interventionen und geplanten chirurgischen Interventionen innerhalb von 24 Stunden beziehen. Eine Aufnahme wird für notwendig eingestuft, wenn mindestens eines dieser Kriterien zutrifft. Umgekehrt wird eine Aufnahme als nicht notwendig beurteilt, wenn keines der Kriterien zutrifft.
Das englische Originalinstrument wurde im Rahmen eines standardisierten Verfahrens übersetzt und rückübersetzt. Sämtliche Kriterien, deren Erläuterungen (Manual) und die Liste von Gründen und Verantwortlichkeiten wurden in mehrfachen Sitzungen eines Expertengremiums bestehend aus erfahrenen Medizinern, Vertretern von Fach- und Berufsverbänden (u.a. des Berufsverbandes der deutschen Chirurgen und der Deutschen Gesellschaft für Chirurgie), Prüfärzten und Gesundheitswissenschaftlern detailliert diskutiert und gegebenenfalls geändert. Änderungen an der englischen Originalversion wurden in Betracht gezogen, wenn sie Unterschiede in der deutschen Versorgungspraxis darstellten, wenn einzelne Leistungen in Deutschland nicht erbracht werden oder wenn neuere Entwicklungen der medizinischen Versorgung in dem ursprünglichen AEP noch nicht berücksichtigt waren. Das deutsche Instrument sowie das begleitende Manual sind an anderer Stelle veröffentlicht und im Internet verfügbar[14,26].

Studiendesign und Patientenstichprobe

Die Reliabilitätsstudie wurde in einem 400-Betten-Krankenhaus der Regelversorgung durchgeführt. Alle Patienten die im Laufe eines Kalenderjahres auf die Abteilung für Chirurgie aufgenommen wurden (2672), wurden für die retrospektive Analyse mit dem AEP in Betracht gezogen. Patienten die innerhalb dieses Jahres mehrfach aufgenommen wurden, hatten die Möglichkeit mehrfach in der Studie berücksichtigt zu werden. Patienten mit privater Krankenversicherung oder Sozialhilfe wurden von der Untersuchung ausgeschlossen (290).
Eine stratifizierte, zufällige Stichprobe von 292 Patienten, geschichtet nach Alter (< oder ≥ Median), Liegezeit (< oder ≥ Median) und Diagnose (14 häufigsten Diagnosen plus eine Gruppe „andere Diagnosen") wurden mit der Entlassungsstatistik des Krankenhauses identifiziert. Weitere 30 Patienten

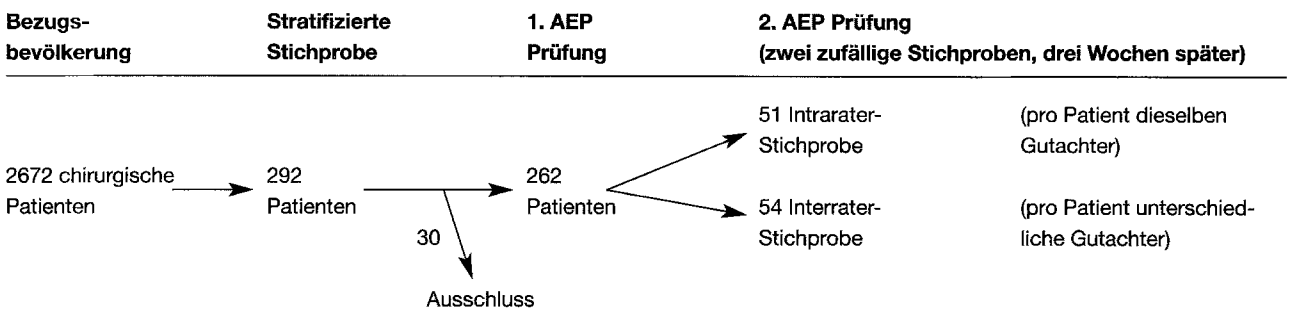

Abbildung 1. *Stichprobenplan der AEP Reliabilitätsstudie. Ausgehend von 2672 chirurgischen Patienten die im Jahre 1997 stationär aufgenommen worden waren, wurden zwei unabhängige Stichproben zur Bestimmung der Interrater- und Intrarater-Übereinstimmung gebildet.*

wurden ausgeschlossen, da deren Krankenunterlagen nicht zur Verfügung standen, unvollständig waren oder sich die Patienten nach dem Sampling als Sozialhilfeempfänger herausstellten. Von jedem der verbleibenden 262 Patienten wurde der gesamte Krankenhausaufenthalt einschliesslich Aufnahme- und Entlassungstag für die Beurteilung der Notwendigkeit der Krankenhausbehandlung berücksichtigt. Für die Untersuchungen zur Reproduzierbarkeit wurden von diesen Patienten zufällig je 54 Patienten für die Interrater-Reliabilität (Reproduzierbarkeit zwischen zwei Beurteilern) und 51 Patienten für die Intrarater-Reliabilität (Reproduzierbarkeit innerhalb eines Beurteilers) ausgewählt. Kein Patient war in beiden Stichproben gleichzeitig vertreten (siehe Abbildung 1).

Fachärzte für Chirurgie des MDK führten die Untersuchungen durch. Die Prüfer erhielten ein eintägiges Training mit Beispiel-Krankenakten einschliesslich Übungen zu der computer-basierten Version des deutschen AEP. Die auf portablen Notebooks implementierte Version verfügte ferner über ein ausführliches (online) Manual[26]. Die Fachärzte wurden explizit angeleitet, ihre Beurteilung über die Notwendigkeit einer Krankenhausaufnahme ausschliesslich an Hand des Aktenmaterials des Aufnahmetages und bis zu 24 Stunden nach Aufnahme abzugeben. Die Beurteilung eines einzelnen Behandlungstages wurde ausschliesslich auf den jeweiligen Tag basiert und nicht auf den gesamten Verlauf. Drei Wochen später führten zwei der Chirurgen (A+B) eine erneute Untersuchung der von ihnen bereits bewerteten Stichprobe von 51 Patient durch, um die Intrarater-Reliabilität zu bestimmen (Abbildung 1). Keiner der Beurteiler wusste während der ersten Prüfung, dass eine Untergruppe der Patienten ein zweites Mal untersucht werden sollte. Um die Interrater-Reliabilität zu bestimmen, führten zwei weitere Fachärzte für Chirurgie des MDK eine zweite Prüfung an der Stichprobe von 54 Patienten durch, die zuvor von zwei anderen Prüfern bewertet wurden.

Analyse

Reproduzierbarkeit zwischen Untersuchern (Interrater) und innerhalb eines Beurteilers (Intrarater) wurden mittels drei Methoden berechnet (siehe Abbildung 2): (1) Gesamtübereinstimmung, indem die Anzahl der Patienten mit übereinstimmender Beurteilung beider Ärzte (a+d) geteilt wurde durch die Anzahl aller Patienten an einem spezifischen Tag bzw. bei Aufnahme (a+b+c+d); (2) spezifische Übereinstimmung analog zur Gesamtübereinstimmung, jedoch getrennt für die Übereinstimmung bei angemessenen (a) und nicht

		Gutachter 1 (bzw. Zeitpunkt 1) angemessen	nicht angemessen
Gutachter 2 (bzw. Zeitpunkt 2)	angemessen	a	b
	nicht angemessen	c	d

Gesamtübereinstimmung	= (a+d) / (a+b+c+d)
Spezifische Übereinstimmung (angemessener Aufenthalt)	= a / (a+b+c+d)
Spezifische Übereinstimmung (nicht angemessener Aufenthalt)	= d / (a+b+c+d)

Abbildung 2. *Berechnung der Interrater- bzw. Intrarater-Übereinstimmung bei der Beurteilung der Angemessenheit von Krankenhausaufenthalten.*

angemessenen Tagen (d); (3) Übereinstimmung von Untersucherpaaren mittels Cohens-Kappa-Statistik, die für den Anteil an zufälliger Übereinstimmung adjustiert[27]. Wir weisen darauf hin, dass Kappa widersprüchlich niedrig ausfallen kann, wenn die Prävalenz angemessener bzw. nicht angemessener Tage sehr gering ist[28].

Im Unterschied zu bisherigen internationalen Berichten zur Reproduzierbarkeit von Instrumenten zur Prüfung der Angemessenheit von Krankenhausbelegungen[2,7,15,20], berechneten wir die Reliabilität für jeden einzelnen aufeinanderfolgenden Behandlungstag im Laufe einer stationären Behandlung (d.h. -Aufnahme, 1. Tag, 2. Tag ... Entlassung). Übereinstimmung wurde nur für die Tage 1 bis 14 (Interrater) bzw. 13 (Intrarater) berechnet, da jeweils weniger als acht Patienten der Stichproben längere Krankenhausverweildauern aufwiesen. Gesamtübereinstimmung und spezifische Übereinstimmung wurde zusätzlich über die Summe aller Krankenhaustage berechnet. Da die Reproduzierbarkeit an einem Patiententag mit der des nächsten Tages korreliert sein kann, wurden zunächst die Übereinstimmungen für jeden Patienten über die Summe der jeweiligen Krankenhaustage errechnet und diese Übereinstimmungen dann über alle Patienten gemittelt. Aus den Standardabweichungen dieser Mittel können unverzerrte Schätzer der Konfidenzintervalle berechnet werden[29]. Gemittelte Kappa-Werte über den gesamten Krankenhausaufenthalt wurden analog zu Fleiss berechnet[30], nachdem die Gleichheit aller Kappa-Werte zwischen allen Tagen vom Krankenhaustag mittels Cochrans Q-Test geprüft wurde[31]. Für alle Schätzer werden 95%-Konfidenzintervalle berichtet. Alle Berechnungen wurden mit SAS Software durchgeführt[32].

	Fortlaufende Krankenhaustage seit Aufnahme													
	0[a]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Gesamt
Anzahl Patienten	54	50	47	40	36	35	29	26	20	14	13	11	9	8	8	[b]346
% Gesamtübereinstimmung	**0,74**	**0,9**	**0,81**	**0,85**	**0,86**	**0,86**	**0,76**	**0,77**	**0,75**	**0,86**	**0,85**	**0,82**	**0,89**	**0,88**	**0,88**	**0,84**
Untere Grenze[c]	*0,62*	*0,82*	*0,7*	*0,74*	*0,75*	*0,74*	*0,6*	*0,61*	*0,56*	*0,67*	*0,65*	*0,59*	*0,68*	*0,64*	*0,65*	*0,79*
Obere Grenze	*0,86*	*0,98*	*0,92*	*0,96*	*0,97*	*0,97*	*0,91*	*0,93*	*0,94*	*1*	*1*	*1*	*1*	*1*	*1*	*0,88*
% spezifische Übereinstimmung (angemessen)	0,54	0,62	0,62	0,75	0,72	0,71	0,69	0,62	0,55	0,71	0,69	0,64	0,78	0,75	0,75	0,67
% spezifische Übereinstimmung (nicht angemessen)	0,2	0,28	0,19	0,1	0,14	0,15	0,07	0,15	0,2	0,15	0,16	0,18	0,11	0,13	0,13	0,17
Kappa	**0,44**	**0,76**	**0,54**	**0,48**	**0,58**	**0,58**	**0,23**	**0,42**	**0,43**	**0,58**	**0,57**	**0,54**	**0,61**	**0,6**	**0,6**	**0,55**
Untere Grenze[c]	*0,21*	*0,59*	*0,27*	*0,12*	*0,25*	*0,25*	*0*	*0,03*	*0,01*	*0,05*	*0,03*	*0*	*0*	*0*	*0*	*0,45*
Obere Grenze	*0,67*	*0,96*	*0,8*	*0,84*	*0,91*	*0,9*	*0,63*	*0,8*	*0,85*	*1*	*1*	*1*	*1*	*1*	*1*	*0,66*

[a] Tag 0 steht für die Beurteilung der Notwendigkeit der Aufnahme.
[b] über die Tage 1–14 (inkl.), ohne Aufnahme.
[c] obere und untere Grenzen des 95% Konfidenzintervalls.

Tabelle 1. *Interrater-Reliabilität in einer Stichprobe von 54 chirurgischen Patienten im Verlauf von deren Liegezeit.*

Ergebnisse

Charakteristika der Patientenstichproben

Die Interrater-Stichprobe bestand zu 54,1% aus Frauen, im Schnitt waren die Patienten 54 Jahre alt (±19,5) und verbrachten im Mittel 9,5 Tage (±9,4) im Krankenhaus. Die Intrarater-Stichprobe bestand zu 53,9% aus Frauen, im Schnitt waren die Patienten 54 Jahre alt (±19,3) und verbrachten im Mittel 8,3 Tage (±8,9) im Krankenhaus.

Interrater-Reliabilität

Für die Interrater-Analyse wurden 54 chirurgische Patienten herangezogen. Tabelle 1 zeigt sowohl die Gesamtübereinstimmung und die spezifische Übereinstimmung als auch die Kappa Statistiken für Aufnahme und fortlaufende Liegetage bis zum 14. Tag einschliesslich der 95%-Konfidenzintervalle. Die Gesamtübereinstimmung zwischen zwei unterschiedlichen Beurteilern war 74% (95%-KI: 62–86%) bei der Aufnahme und 84% (79–88%) für die Summe aller fortlaufenden Behandlungstage. Kappa für die Aufnahme betrug 0,44 (0,21–0,67) und gemittelt über alle Behandlungstage 0,55 (0,45–0,66). Kappa-Werte unterschieden sich nicht signifikant von Tag zu Tag ($Q_K = 8,8$; $p = 0,79$).

Intrarater-Reliabilität

Intrarater-Reliabilität wurde bei 51 Patienten und anhand von zwei Beurteilern untersucht, die drei Wochen nach der ersten Beurteilung dieselben Patienten nochmals bewerteten. Tabelle 2 zeigt die Übereinstimmung innerhalb jeweils eines Beurteilers. Intrarater-Reproduzierbarkeit waren 88% Gesamtübereinstimmung für Aufnahmen (79%–97%) und 88% (85%–92%) für Krankenhaustage. Entsprechend waren Kappa 0,60 (0,31–0,89) und 0,70 (0,61–0,78). Beurteiler B hatte durchgehend eine deutlich niedrigere Reproduzierbarkeit als Beurteiler A, auch wenn dieser Unterschied nicht signifikant ausfällt ($Q_K = 20,5$; $p = 0,06$).
Tabelle 3 zeigt sowohl die Gesamtübereinstimmung und die spezifische Übereinstimmung als auch die Kappa-Statistiken für Aufnahme und fortlaufende Liegetage bis zum 13. Tag einschliesslich 95%-Konfidenzintervalle.

Diskussion

Das Appropriateness Evaluation Protocol (AEP) ist das international am häufigsten eingesetzte Instrument zur Beurteilung der Angemessenheit stationärer Versorgung. Die Vorzüge und Grenzen des Instruments sind in seinem Ursprungsland, den USA, in einer Reihe von Publikationen gut dokumentiert. Für den deutschen Sprachraum ist bisher kein standardisiertes Instrumentarium verfügbar, dessen Reproduzierbarkeit und Validität getestet wurde und das darauf basierend allgemein akzeptiert wurde.
Die Adaptierung des AEP an die Spezifika des deutschen Sprach- und Versorgungsraums wurde mit grosser Sorgfalt durchgeführt. Der Adaptierungsprozess in anderen europäischen Ländern im Rahmen eines EU-geförderten Projekts wurde detailliert studiert und entsprechende Expertise eingeholt[5,33,34]. Obwohl, wie in anderen Adaptierungen, einzelne Kriterien entfallen sind oder umformuliert wurden, wurden die Diagnoseunabhängigkeit und die expliziten Definitionen beibehalten.
Die Reliabiliät des deutschen Instruments ist bei Einsatz in

	Gutachter A			Gutachter B			Gesamt		
	% Übereinstimmung	Kappa	n	% Übereinstimmung	Kappa	n	% Übereinstimmung	Kappa	n
Aufnahme	0,96 (0,89–1)	0,87 (0,61–1)	27	0,79 (0,63–0,95)	0,32 (0–0,78)	24	0,88 (0,79–0,97)	0,60 (0,31–0,89)	51
Krankenhaustage[a]	0,92 (0,88–0,96)	0,80 (0,70–0,90)	173	0,84 (0,78–0,89)	0,59 (0,46–0,73)	168	0,88 (0,85–0,92)	0,70 (0,61–0,78)	341

[a] Tage 1 bis 13 (inkl.), ohne Aufnahme.

Tabelle 2. *Intrarater-Reliabilität von zwei Gutachtern in einer Stichprobe von 51 chirurgischen Patienten (Prozent Gesamtübereinstimmung, Kappa und 95%-Konfidenzintervalle).*

	Fortlaufende Krankenhaustage seit Aufnahme												
	0[a]	1	2	3	4	5	6	7	8	9	10	11	12	13	Gesamt
Anzahl Patienten	51	49	46	42	36	35	29	26	22	14	12	11	10	9	[b]341
% Gesamtüber-einstimmung	**0,88**	**0,96**	**0,78**	**0,88**	**1**	**0,94**	**0,86**	**0,88**	**0,86**	**0,86**	**0,75**	**0,82**	**0,9**	**0,67**	**0,88**
Untere Grenze[c]	*0,79*	*0,9*	*0,66*	*0,78*	*1*	*0,87*	*0,74*	*0,76*	*0,72*	*0,67*	*0,51*	*0,59*	*0,71*	*0,36*	*0,85*
Obere Grenze	*0,97*	*1*	*0,9*	*0,98*	*1*	*1*	*0,99*	*1*	*1*	*1*	*0,99*	*1*	*1*	*0,97*	*0,92*
% spezifische Übereinstimmung (angemessen)	0,76	0,71	0,59	0,64	0,78	0,71	0,66	0,65	0,64	0,79	0,67	0,73	0,8	0,56	0,68
% spezifische Übereinstimmung (nicht angemessen)	0,12	0,25	0,19	0,24	0,22	0,23	0,2	0,23	0,22	0,07	0,08	0,09	0,1	0,11	0,2
Kappa	**0,6**	**0,9**	**0,49**	**0,71**	**1**	**0,85**	**0,66**	**0,72**	**0,67**	**0,44**	**0,31**	**0,42**	**0,62**	**0,18**	**0,70**
Untere Grenze[c]	*0,31*	*0,75*	*0,21*	*0,48*	*1*	*0,65*	*0,37*	*0,42*	*0,34*	*0*	*0*	*0*	*0*	*0*	*0,61*
Obere Grenze	*0,89*	*1*	*0,76*	*0,95*	*1*	*1*	*0,95*	*1*	*1*	*1*	*0,78*	*1*	*1*	*0,84*	*0,78*

[a] Tag 0 steht für die Beurteilung der Notwendigkeit der Aufnahme.
[b] Tage 1–13 (inkl.), ohne Aufnahme.
[c] obere und untere Grenzen des 95 % Konfidenzintervalls.

Tabelle 3. *Intrarater-Reliabilität in einer Stichprobe von 51 chirurgischen Patienten im Verlauf von deren Liegezeit.*

chirurgischen Abteilungen hoch. Mit einer Interrater-Übereinstimmung von 74% der Aufnahmen und 86% aller Behandlungstage und einer Intrarater-Übereinstimmung von 88% der Aufnahmen und 88% der Behandlungstage liegt es im Bereich der internationalen Erfahrungen[21]. Das Instrument behält seine hohe Übereinstimmung unabhängig vom Anteil angemessener oder nicht angemessener Krankenhaustage und unabhängig von der Krankenhausverweildauer.

Für internistische Patienten konnten noch höhere Interrater-Übereinstimmungen von 92% und Intrarater-Übereinstimmungen von 96% bei der Bewertung der Notwendigkeit von stationären Aufnahmen durch das AEP ermittelt werden. Die Übereinstimmungen von einzelnen Behandlungstagen für internistische Patienten lagen bei 76% bzw. 93% [35].

Mit diesen Reproduzierbarkeiten entspricht das deutsche AEP den Grundvoraussetzungen für ein standardisiertes Instrument zur Beurteilung der Notwendigkeit von Krankenhausbehandlungen auch wenn eine Bestätigung unserer Ergebnisse an grösseren Patientenkollektiven noch aussteht[36].

Kappa-Werte fallen insbesondere an Tagen, die fast ausnahmslos als „notwendig" bewertet wurden, aufgrund der algebraischen Eigenschaften von Kappa niedrig aus[28]. Obwohl dieses häufig beschriebene Phänomen die Interpretation erschwert, liegen alle Kappa-Werte für Aufnahmen und der Summe aller Behandlungstage über 0,4. Landis und Koch beschrieben diese Werte als gute Übereinstimmung[37]. Dennoch deutet die niedrige Intrarater-Reliabilität eines der beiden Gutachter (B) darauf hin, dass ein eindeutiges und ausführliches Training essentiell ist, um eine hohe Reproduzierbarkeit zu erreichen.

Das von uns vorgeschlagene longitudinale Design untersucht einzelne

Patientenverläufe. Der wesentliche Vorteil einer longitudinalen Erfassung besteht darin, dass die Beurteiler nicht nur einen isolierten Tag im Verlauf einer Krankenhausbehandlung betrachten, sondern sich mit dem gesamten Verlauf vertraut machen können. Ein praktischer Vorteil besteht darin, dass weniger Krankenakten angefordert werden müssen. Die Bewertungen einzelner Patiententage können jedoch untereinander hoch korreliert sein. Wird beispielsweise der Aufenthalt eines Patienten gegen Ende des Aufenthaltes als unangemessen bewertet, so werden mit grosser Wahrscheinlichkeit die folgenden Tage ebenfalls als unangemessen bewertet. Damit müssen komplexere Methoden zur Analyse von „Fehlbelegungsraten" herangezogen werden und die notwendige Stichprobengrösse an der Gesamtzahl an Tagen, nicht Patienten, kann sich leicht erhöhen.

Der AEP ist ein aktenbasiertes Verfahren, d.h. Gutachter beurteilen anhand der Patientenakten, ob eines der 24 diagnoseunabhängigen Kriterien für einen gerechtfertigten stationären Aufenthalt zutrifft. Ärzte und Klinikmanager waren besorgt, dass eine rückwirkende (retrospektive) Beurteilung auf der Basis von Krankenakten das Ausmass der Fehlbelegung zu hoch einschätzen könnte[38].

Dies wird damit begründet, dass einerseits in der konkreten klinischen Situation mehr Informationen zur Verfügung stehen als dokumentiert werden und andererseits in der Krankenakte auch Ergebnisse von zeitlich später stattgefundenen Untersuchungen zu finden sind, die zum Zeitpunkt der klinischen Entscheidung nicht vorhanden waren[38]. In einer Studie an 67 chirurgischen Patienten wurde eine Übereinstimmung zwischen zeitgleicher und retrospektiver Untersuchung innerhalb derselben Prüfer von 86% für Aufnahmen und 96% für einzelne Behandlungstage mit dem deutschen AEP ermittelt. Ähnlich hohe Werte ergaben sich bei dem Vergleich zwischen zwei unterschiedlichen Gutachtern, einer prüfte zeitgleich, ein anderer retrospektiv (86% bzw. 90%)[39].

Zusammenfassend kann das AEP in der derzeitigen Situation des deutschen Gesundheitssystems, das erneut mit erheblichen Einsparmassnahmen konfrontiert wird, hilfreich sein, diese Diskussion auf mehr Empirie zu stützen. Kostenträger und Versorger müssen jedoch möglichst schnell den Rahmen für den Einsatz des AEP festlegen, um ansonsten unausweichlichen Konflikten aktiv entgegenzutreten.

Literaturverzeichnis

1 *Brennan TA, Leape LL, Laird NM*, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med 1991; *324*: 370–6.

2 *Gertman PM, Restuccia JD.* The appropriateness evaluation protocol: a technique for assessing unnecessary days of hospital care. Med Care 1981; *19*: 855–71.

3 *Winickoff RN, Restuccia JD, Fincke BJ.* Concurrent application of the Appropriateness Evaluation Protocol to acute admissions in Department of Veteran Affairs Medical Centers. Med Care 1991; *29*: AS64–AS75.

Summary

Reliability of a German instrument to assess the appropriateness of hospital utilization in surgery

During the past years, the assessment of the appropriateness of hospital utilization has become increasingly important in the German health care system. Previous evaluations by regional review organizations in several states demonstrated the need for a standardized, reliable, and valid instrument to evaluate the appropriateness of inpatient care. Objective of the study is to test the reliability of a German adaptation of the "Appropriateness Evaluation Protocol" (AEP). Among all 2672 admissions from the department of surgery of a regional medical center during one calendar year, 54 patients were randomly selected to evaluate the inter-rater reliability and 51 patients to test intra-rater reliability. Overall agreement, specific agreement and Kappa statistics were estimated for every hospital admissions and all consecutive hospital days. The German AEP showed an inter-rater agreement of 74% (62–86%) for hospital admissions (Kappa = 0.44) and 84% (79%–88%) for all hospital days (K = 0.55). Intra-rater reliability was 88% (79%–97%) for hospital admissions (K = 0.60) and 88% (85%–92%) for all hospital days (K = 0.70). The observed agreement is independent of length of hospital stay and proportion of appropriate days. A standardized instrument with known metric properties is essential for quality management in hospitals to prepare for an increasingly consolidating health care market in Germany. The German AEP is a reliable instrument, which will allow to identify inefficiencies in the management of surgical inpatients.

Résumé

Reproducement d'une version allemande pour l'évaluation de la nécessité d'un traitement à l'hôpital

L'examen de la nécessité d'un traitement hospitalier est devenu de plus en plus actuel dans les dernières années en Allemagne. Jusqu'à maintenant, ceci fut réalisé par un service spécial des assurances médicaux des différentes régions allemandes. Les résultats ont montré le besoin d'un instrument standardisé, valide et reliable pour estimer la nécessité d'un traitement à l'hôpital. Le but de l'étude suivante ítait de tester la reproductibilité d'une version allemande adaptée du «appropriateness Evaluation protocol» (AEP). Parmis les 2672 admissions d'une année d'un services de chirurgie, 54 patients ont été élus pour l'investigation sur la reproductibilité entre les différents experts et 51 patients pour juger la reproductibilité parmis chaque expert. La concordance générale ainsi que différentes concordances spécifiques et la statistique de Kappa furent calculées pour le jugement des admissions à l'hôpital de même que pour le jugement de jours particuliers de traitement. La version allemande de l'AEP montre une concordance de 74% (62%–86%) entre les experts pour le jugement des admissions hospitalier et de 84% (79%–88%) pour tous les jours de traitement hospitalier et une concordance intra-individuelle parmis les experts de 88% (79%–97%) pour les admissions et de 88% (85%–92%) pur les jours de traitement. Les concordances observées sont indépendantes au nombre de jours de traitement ainsi qu'à la proportion des jours jugés comme convenable ou non-convenable. Un instrument standardisé avec des propriétés métriques connues est nécessaire pour le management interne de la qualité dans les hôpitaux, pour pouvoir se préparer aux défis d'un future marché consolidé. Avec la version allemande de l'AEP, il existe maintenant un instrument testé metriquement, qui permet d'identifier les secteurs inefficients dans les services de chirurgie.

4 *Siu AL, Sonnenberg FA, Manning WG*, et al. Inappropriate use of hospitals in a randomized trial of health insurance plans. N Engl J Med 1986; *315*: 1259–66.

5 *Apolone G, Alfieri V, Braga A*, et al. A survey of the necessity of the hospitalization day in an Italian teaching hospital. Qual Ass Health Care 1991; *3*: 1–9.

6 *Bare ML, Prat A, Lledo L, Asenjo MA, Salleras L.* Appropriateness of admissions and hospitalization days in an acute-care teaching hospital. Rev Epidemiol Sante Publique 1995; *43*: 328–36.

7 *Rishpon S, Lubacsh S, Epstein LM.* Reliability of a method of determining the necessity for hospitalization days in Israel. Med Care 1986; *24*: 279–82.

8 *Alonso J, Munoz A, Anto JM.* Using length of stay and inactive days in the hospital to assess appropriateness of utilisation in Barcelona, Spain. J Epidemiol Community Health 1996; *50*: 196–201.

9 *Hynes M, O'Herlihy BP, Laffoy M, Hayes C.* Patients 21 days or more in an acute hospital bed: appropriateness of care. Ir J Med Sci 1991; *160*: 389–92.

10 *Payne SMC, Ash A, Restuccia JD.* The role of feedback in reducing medically unnecessary hospital use. Med Care 1991: AS91–AS106.

11 *Chopard P, Perneger TV, Gaspoz JM*, et al. Predictors of inappropriate hospital days in a department of internal medicine. Int J Epidemiol 1998; *27*: 513–9.

12 *Restuccia JD.* The effect of concurrent feedback in reducing inappropriate hospital utilization. Med Care 1982; *20*: 46–62.

13 *Vardi A, Modan B, Blumstein Z, Lusky A, Schiff E, Barzilay Z.* A controlled intervention in reduction of redundant hospital days. Int J Epidemiol 1996; *25*: 604–8.

14 *Sangha O, Wildner M, Schneeweiss S*, et al. Fehlbelegung im Krankenhaus. Entwicklung eines standardisierten Verfahrens zur Beurteilung der Notwendigkeit von vollstationären Krankenhausbehandlungen. Chirurg 1999; 38: 201–10.

15 *Restuccia JD, Kreger BE, Payne SM, Gertman PM, Dayno SJ, Lenhart GM.* Factors affecting appropriateness of hospital use in Massachusetts. Health Care Finance Rev 1986; *8*: 47–54.

16 *Payne SM, Campbell D, Penzias BG, Socholitzky E.* New methods for evaluating utilization management programs. Qrt Qual Rev Bull 1992; *18*: 340–7.

17 *Payne SM.* Identifying and managing inappropriate hospital utilization: a policy synthesis. Health Serv Res 1987; *22*: 709–69.

18 *Kaza S, Erdem Z, Dogrusoy S, Halici N.* Reliability of a hospital utilization review method in Turkey. Int J Qual Health Care 1999; *10*: 53–8.

19 *Siu AL, Leibowitz A, Brook RH, Goldman NS, Lurie N, Newhouse JP.* Use of the hospital in a randomized trial of prepaid care. JAMA 1988; *259*: 1343–6.

20 *Strumwasser I, Paranjpe NV, Ronis DL, Share D, Sell LJ.* Reliability and validity of utilization review criteria. Appropriateness Evaluation Protocol, Standardized Medreview Instrument, and Intensity-Severity-Discharge criteria. Med Care 1990; *28*: 95–111.

21 *Inglis AL, Coast J, Gray SF, Peters J, Frankel SJ.* Appropriateness of hospital utilization. Med Care 1995; *33*: 952–7.

22 *Davido A, Nicoulet I, Levy A, Lang T.* Appropriateness of admission in an emergency department: reliability of assessment and causes of failure. Qual Ass Health Care 1991; *3*: 227–34.
23 *Gloor JE, Kissoon N, Joubert GI.* Appropriateness of hospitalization in a Canadian pediatric hospital. Pediatrics 1993; *91*: 70–4.
24 *Kemper KJ, Fink HD, McCarthy PL.* The reliability and validity of the pediatric appropriateness evaluation protocol. Qrt Qual Rev Bull 1989; *15*: 77–80.
25 *Kemper KJ.* Medically inappropriate hospital use in a pediatric population. N Engl J Med 1988; *318*: 1033–7.
26 *Sangha O, Wildner M, Schneeweiss S,* et al. Die deutsche Version des Appropriateness Evaluation Protocol (AEP). URL: http://www.BFV-online.de/BFV/AEP.
27 *Cohen J.* A coefficient of agreement for nominal scales. Educ Psychol Measure 1960; *20*: 37–46.
28 *Feinstein AR, Cichetti DV.* High agreement but low kappa: 1. The problem of two paradoxes. J Clin Epidemiol 1990; *43*: 543.
29 *Diggle PJ, Liang KY, Zeger SL.* Analysis of longitudinal data. Oxford: Clarendon Press, 1994: 147–53.
30 *Fleiss JL.* Statistical methods for rates and proportions. 2nd ed. New York: Wiley, 1981: 217–22.
31 *Cochran WG.* The comparison of percentages in matched samples. Biometrika 1950; *37*: 256–66.
32 SAS/STAT Software: changes and enhancements through release 6.12. Cary, NC: SAS Institute Inc., 1996.
33 *Bentes M, Gonsalves ML, Santos M, Pina E.* Design and development of a utilization review program in Portugal. Int J Qual Health Care 1995; *7*: 201–12.
34 *Santos-Eggimann B, Paccaud F, Blanc T.* Medical appropriateness of hospital utilization: an overview of the Swiss experience. Int J Qual Health Care 1995; *7*: 227–32.
35 *Schneeweiss S, Sangha O, Siebert H* et al. Erfassung der Notwendigkeit von stationären Behandlungen in der Inneren Medizin. Reproduzierbarkeit eines für Deutschland entwickelten Verfahrens. Dtsch Med Wochenschr, 2000; *125*: 894–9.
36 *Streiner DL, Norman GR.* Health Measurement Scales. 2nd ed. Oxford: Oxford University Press, 1994: 104–27.
37 *Landis JR, Koch GG.* The measurement of observer agreement for categorical data. Biometrics 1977; *33*: 159–74.
38 *Santos-Eggimann B, Sidler M, Schopfer D, Blanc T.* Comparing results of concurrent and retrospective designs in a hospital utilization review. Int J Qual Health Care 1997; *2*: 115–20.
39 *Schneeweiss S, Sangha O, Siebert H,* et al. Übereinstimmung einer zeitgleichen Bewertung von Fehlbelegung mit deren retrospektiver Beurteilung anhand von Patientenakten. Gesundheitswesen 2000; *62*: 207–10.

Korrespondenzadresse

Dr. med. Oliver Sangha
IBE der LMU München
Marchioninistrasse 15
D-81377 München
Tel.: ++89 69349-100
Fax: ++89 69349-104
e-mail: san@ibe.med.uni-muenchen.de