Anaesthesist
1998 · 47:30–38 © Springer-Verlag 1998

Originalien

A. von Bierbrauer · S. Riedel · W. Cassel · P. von Wichert · Abteilung Medizinische Poliklinik – Intensivmedizin, Zentrum Innere Medizin, Philipps-Universität Marburg

Validierung des Acute Physiology and Chronic Health Evaluation (APACHE) III Scoringsystems

und Vergleich mit APACHE II auf einer deutschen Intensivstation

Zusammenfassung

Fragestellung: Ziel der Untersuchung war die systematische Validierung des APACHE III Scoringsystems hinsichtlich Klassifikation der Erkrankungsschwere und Prädiktion der Krankenhausletalität. Entsprechende Daten liegen bisher für ein größeres Patientenkollektiv einer deutschen Intensivstation nicht vor.
Methodik: Bei 531 konsekutiven Intensivpatienten (Liegezeit >4 h) wurde eine prospektive Erhebung der APACHE III Tag-1-Scores und die Berechnung der resultierenden Prädiktionen des Letalitätsrisikos durchgeführt und ein Vergleich mit den entsprechenden ebenfalls erhobenen Daten des etablierten APACHE II und der tatsächlich beobachteten Letalität angestellt.
Ergebnisse: Hinsichtlich der primären Validierungskriterien Diskriminationsfähigkeit (Fläche unter der ROC-Kurve für APACHE III 0,873, für APACHE II 0,859) und Kalibration (goodness-of-fit-test jew. $p>0{,}05$) mit dem Bezugspunkt Krankenhausletalität zeigten beide Systeme sehr befriedigende Resultate und unterschieden sich allenfalls marginal. Die tatsächliche Krankenhausletalität des Kollektivs (13,4%) stimmt mit der von APACHE III vorhergesagten (13,2%) nahezu exakt überein, die von APACHE II prognostizierte lag darüber (16,8%). Der durch den APACHE III vorgegebene Standard (Letalitätsindex nicht signifikant <oder>1,0) wurde erfüllt, der durch den APACHE II vorgegebene Standard wurde übertroffen. Die Scoremittelwerte und mittleren Letalitätsprognosen des Gesamtkollektivs unterschieden sich für Überlebende und Verstorbene jeweils hochsignifikant ($p<0{,}001$). Die individuellen Scorewerte beider Systeme zeigten eine enge Korrelation (r=0,922).
Schlußfolgerungen: Gruppenbezogen ermöglicht APACHE III (wie auch APACHE II) suffizient die Prädiktion der Krankenhausletalität und die Klassifikation der Erkrankungsschwere bei einem großen Patientenkollektiv einer deutschen internistischen Intensivstation und kann damit für vergleichbare deutsche Intensivpopulationen als validiert gelten. APACHE III scheint als Standard für die moderne Intensivmedizin besser als das etablierte, aber ältere APACHE II-System geeignet zu sein. Ob der Zuwachs an Information und Präzision den deutlichen Mehraufwand zur Scoreerhebung und Risikoberechnung bei Verwendung dieses Systems rechtfertigt, muß im Einzelfall in Abhängigkeit von der jeweiligen Fragestellung abgewogen werden.

Schlüsselwörter

Scoresysteme · APACHE III-Score · APACHE II-Score · Hospitalmortalität · Ergebnisforschung

Die moderne Intensivmedizin sieht sich mit dem Problem konfrontiert, daß dem ständig steigenden Bedarf an Intensivtherapieplätzen zunehmend begrenzte finanzielle und auch personelle Ressourcen gegenüberstehen. Aus diesem Grund sind Nutzungsoptimierung und Überprüfung von Therapieeffizienz als Elemente von Qualitätskontrolle und Qualitätsmanagement auch auf den Intensivstationen im Sinne einer vermehrten Transparenz der Leistungserbringung von essentieller Notwendigkeit. Hierfür finden insbesondere für den Bereich der Kontrolle der Ergebnisqualität Scoresysteme zunehmende Verwendung [4, 5, 7, 22]. Die meisten Systeme basieren dabei darauf, daß die Erkrankungsschwere durch die Quantifizierung des Grads der Abweichung physiologischer Parameter von der Norm quantitativ meßbar gemacht wird. Das zur Zeit gebräuchlichste und am besten validierte physiologisch basierte System ist das „Acute physiology and chronic health evaluation" (APACHE) II System [13], das zusätzlich über bestimmte Algorithmen die Berechnung der relativen Letalitätswahrscheinlichkeit einer Patientenpopulation ermöglicht. Als Weiterentwicklung dieses Systems wurde 1991 von Knaus et al. [14] das APACHE III System vorgestellt, das eine Verbesserung der Vorhersage der Letalitätswahrscheinlichkeit bzw. der Klassifikation der Erkrankungsschwere erbringen soll. Das APACHE III System ist bisher nicht systematisch an einem größeren Kollektiv für die deutsche Intensivmedizin validiert worden. Ziel der vorliegenden Arbeit war es daher, das APACHE III System hinsichtlich seiner Aussagekraft bezüglich Vorhersage des Letalitätsrisikos und Klassifikation der Erkrankungsschwere an einer repräsentativen Patientenpopulation einer deutschen Intensivstation zu validieren und mit dem weniger aufwendigen und eta-

Dr. A. von Bierbrauer
Abteilung Medizinische Poliklinik, Zentrum Innere Medizin, Philipps-Universität Marburg, Baldinger Strasse, D-35033 Marburg

Anaesthesist
1998 · 47:30–38 © Springer-Verlag 1998

A. von Bierbrauer · S. Riedel · W. Cassel
P. von Wichert

Scoring systems. Validation of APACHE III and comparison to APACHE II in a german intensive care unit

Abstract

Objectives: The aim of the study was to systematically validate the APACHE III scoring system concerning severity of illness classification and prediction of hospital mortality. Such data have not yet been determined in a large population of critically ill patients in germany.
Methods: 531 patients (ICU stay >4 hours) were prospectively and consecutively investigated. The day-1-scores and risk-of-death predictions of APACHE III and APACHE II were determined. A comparison was performed between both scoring systems, and the correlation with the observed hospital mortality was examined.
Results: For both main validation criteria, as were discrimination (areas under the ROC-curves: APACHE III 0.873; APACHE II 0.859) and calibration (goodness-of-fit testings; $p>0.05$), both scoring systems provided satisfying results concerning hospital mortality, no system showing a significantly superior performance. Compared to the observed hospital mortality (13.4%), the prediction of APACHE III (13.2%) was extremely accurate, whereas the prediction of APACHE II was higher (16.8%). The standard (mortality index not significantly <or>1.0) provided by APACHE III was fulfilled, while the standard given by APACHE II was surpassed. The mean scores and the mean risk-of-death predictions for non-survivors were significantly higher compared to survivors ($p<0.001$). The individual score values of both systems were found to have a strong correlation ($r=0.922$).
Conclusions: APACHE III (like APACHE II) provides a sufficient severity of disease classification and accurately predicts overall hospital mortality in a representatively large german population of a medical ICU. Therefore APACHE III can be regarded as validated for the use in comparable german ICUs. For use as a standard the more recently introduced APACHE III seems to be superior to the established but older APACHE II. However, each user will – depending on the particular questions to be addressed – carefully have to evaluate, if the improvement of prognostic accuracy really justifies the increased amount of workload necessary for calculating APACHE III score and risk prediction.

Key words

Scoring systems · APACHE III score · APACHE II score · Hospital mortality · Outcome research

blierten APACHE II System zu vergleichen.

Patienten und Methodik

Patientengut

Untersucht wurden alle konsekutiven Intensivaufnahme im 12-Monats-Zeitraum von September 1993 bis September 1994 der allgemein-internistischen Intensivstation (7 Betten) der Abteilung Poliklinik des Zentrums Innere Medizin der Philipps-Universität Marburg. Auf dieser Station wird im ärztlichen und pflegerischen Bereich im Vollschichtdienst gearbeitet, ein Facharzt für Innere Medizin steht als Entscheidungsträger jederzeit zur Verfügung.

Datenerhebung

Bei allen Patienten, die im Untersuchungszeitraum auf die Station aufgenommen wurden, wurden prospektiv die zur Berechnung des Tag-1-Scores der APACHE III- und APACHE II-Systeme nötigen Daten gemäß den Anweisungen der Originalpublikationen [13, 14] erhoben und mittels einem von einem der Autoren (W.C.) entworfenen dBase-Programm auf einem PC gespeichert und der jeweilige Tag-1-Score berechnet. Basierend auf diesen Scorewerten wurden mittels der jeweiligen Algorithmen für jeden Patienten das individuelle Letalitätsrisiko berechnet. Die entsprechenden Koeffizienten sind für das APACHE II System von Knaus et al. [13] publiziert worden, für das APACHE III System sind sie uns von derselben Arbeitsgruppe für diese Forschungsarbeit zur Verfügung gestellt worden [11, 12].

Von der weiteren Analyse ausgeschlossen wurden lediglich Patienten mit einer Liegezeit von weniger als 4 h auf der Intensivstation. Bei den Patienten, die weniger als 24 h auf der Intensivstation lagen, wurden gemäß der Originalanweisung die schlechtesten gemessenen physiologischen Variablen dieser Liegezeit zur Scoreberechnung herangezogen, bei allen anderen Patienten wurden die schlechtesten Variablen der ersten 24 h zugrundegelegt. Fehlende Variablen wurden ebenfalls gemäß der Originalanweisung als normal gewertet. Die Bewertung der in den Score einfließenden „Glasgow Coma Scale" erfolgte bei spontanatmenden, nichtsedierten Patienten in üblicher Weise, der Bewußtseinszustand sedierter bzw. beatmeter Patienten wurden unter Hinzuziehen aller verfügbaren Informationen nach ihrem Zustand vor Sedation beurteilt bzw., falls nicht hinreichend einschätzbar, gemäß der Originalanweisung als unbeeinträchtigt gewertet.

Tabelle 1
Basisdaten des Gesamtkollektivs (*n*=531)

Parameter	(Einheit)	Wert
Geschlecht	Männer (*n*)	332 (62,5%)
	Frauen (*n*)	199 (37,5%)
Alter ∅	(Jahre)	58,5±17,7 (16–90)
Liegezeit ∅	Intensiv (Tage)	3,5±7,7 (<1–118)
	Krankenhaus (Tage)	18,1±20,5 (1–169)
Letalität	Intensiv (*n*)	50 (9,4%)
	Krankenhaus (*n*)	71 (13,4%)
Post-OP	(*n*)	14 (2,6%)
Beatmung	[invasiv] (*n*)	102 (19,2%)
	Dauer ∅ (Tage)	6,6±14,7 (1–118)

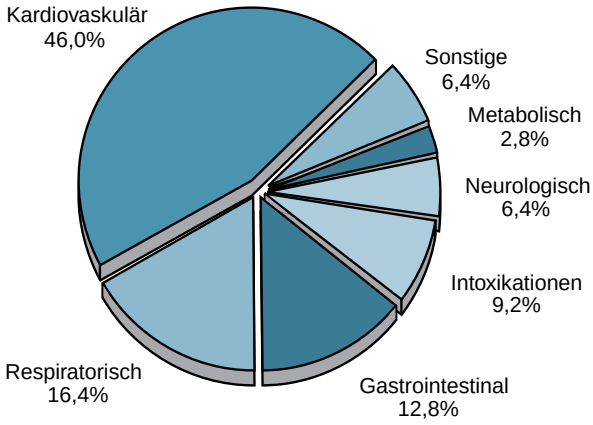

Abb. 1 ◀ **Zusammensetzung des Gesamtkollektivs (*n*=531) nach anteiliger Diagnose-subgruppenhäufigkeit**

Alle Patienten wurden bis zur Krankenhausentlassung bzw. bis zum Versterben nachverfolgt und dabei wurden neben dem tatsächlichen Intensiv- und Krankenhausoutcome (Überleben vs. Versterben) auch die Intensivliegedauer und die Krankenhausliegedauer registriert.

Statistik

Primäre Validierungskriterien waren die Diskriminationsfähigkeit und die Kalibration der Systeme. Die Diskriminationsfähigkeit der Systeme zwischen Versterben und Überleben wurde mittels der Erstellung von receiver-operating-characteristic (ROC)-Kurven anhand der Fläche unter der Kurve (AUC) überprüft [9]. Die Kalibration der Systeme, d.h. die homogene Übereinstimmung von tatsächlicher und prognostizierter Letalität in allen Bereichen der Letalitätsprädiktion (hohe – mittlere – niedrige Letalität), wurde mit einem goodness-of-fit-Test (nach Hosmer u. Lemeshow) geprüft [10]. Dabei weist ein niedriger sog. $\hat{H}$-Wert mit einem entsprechenden $p>0{,}05$ auf eine gute *und* homogene Übereinstimmung hin.

Aus den individuellen Scorewerten und Letalitätsrisiken wurden die jeweiligen Mittelwerte für das Gesamtkollektiv und für die nach Diagnosegruppen getrennten Subkollektive berechnet. Die tatsächlichen Letalitäten wurden mit den prognostizierten Letalitäten verglichen und aus dem Quotienten der sog. „Letalitätsindex" inkl. 99%-Konfidenzintervall berechnet [8]. Dieser Letalitätsindex zeigt den durch das Scoresystem vorgegebenen Standard an, eine Abweichung des 99%-Konfidenzintervalls vom Wert 1 des Quotienten weist auf ein Übertreffen (<1) bzw. Nicht-Erfüllen (>1) des Standards hin.

Die mittleren Scorewerte und Letalitätsprädiktionen der Verstorbenen und der Überlebenden wurden für das Gesamtkollektiv getrennt berechnet und die Unterschiedlichkeit der Mittelwerte mit dem *t*-Test geprüft.

Die Überprüfung des Zusammenhangs zwischen der reinen Scorewerthöhe und der tatsächlichen Letalität erfolgte nach Kategorisierung der Scoresysteme (APACHE II=7 Kategorien; APACHE III=11 Kategorien) mittels χ^2-Test.

Die Korrelation zwischen den individuellen APACHE II und APACHE III Scorewerten wurde mittels linearer Regressionsanalyse geprüft. Als signifikant wurde, wenn nicht anders erwähnt, ein $p<0{,}05$ angenommen.

Ergebnisse

Während des Untersuchungszeitraums wurden 570 Patienten auf die Station aufgenommen. 39 Patienten (6,8%) hat-

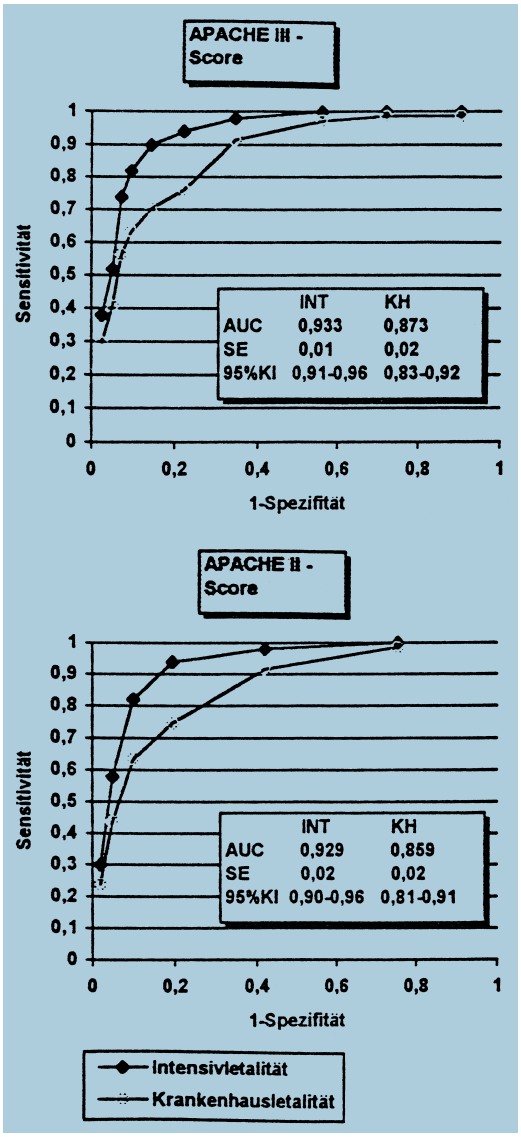

Abb. 2 ▶ **Überprüfung der gruppenbezogenen Diskriminationsfähigkeit der APACHE III- und APACHE II-Scoresysteme zwischen Überleben und Versterben mittels „receiver operating characteristic" (ROC)-Kurven jeweils für Intensiv (INT)- und Krankenhausletalität (KH). AUC=Fläche unter der Kurve, SE=Standardabweichung, 95%KI=95% Konfidenzintervall**

Tabelle 2
Mittlere Scorewerte und Letalitäts-Risikoprädiktionen (Standardabweichung) des Gesamtkollektivs und von Intensiv- bzw. Krankenhaus-Überlebenden vs.-Verstorbenen. Die Mittelwerte von Überlebenden und Verstorbenen unterscheiden sich sowohl für den Intensivaufenthalt als auch für den Krankenhausaufenthalt jeweils hochsignifikant voneinander (p<0,001; t-Test)

	Gesamtkollektiv (n=531)	Intensiv-Überlebende (n=481)	Intensiv-Verstorbene (n=50)	Krankenhaus-Überlebende (n=460)	Krankenhaus-Verstorbene (n=71)
APACHE III Score (Punkte)	41,2 (30,6)	35,5 (24,3)	96,3 (30,8)	34,7 (23,9)	83,7 (35,4)
APACHE II Score (Punkte)	12,1 (8,5)	10,6 (7,0)	27,0 (8,3)	10,4 (6,9)	23,3 (9,9)
APACHE III Letalitätsprädiktion (%)	13,2 (21,6)	8,6 (15,0)	57,0 (26,0)	8,1 (1,4)	46,1 (30,6)
APACHE II Letalitätsprädiktion (%)	16,3 (19,9)	12,6 (15,0)	52,2 (25,0)	12,1 (14,5)	43,4 (27,6)
Tatsächliche Letalität (%)	13,4				

ten eine Liegezeit von ≤4 h und wurden von der weiteren Betrachtung ausgeschlossen, so daß sich die Datenanalyse im weiteren auf 531 Patienten (=100%) bezieht. 14 Patienten (2,6%) wurden postoperativ betreut, bei den übrigen Patienten machte ein internistisches Krankheitsbild die Intensivaufnahme erforderlich. Die basalen demographischen und deskriptiven Daten der Patientenpopulation gehen aus Tabelle 1, die Diagnosesubgruppenhäufigkeit aus Abb. 1 hervor.

Die Flächen unter den ROC-Kurven waren bei beiden Systemen befriedigend groß (>0,7). Sie waren beim APACHE III System minimal größer als beim APACHE II, der Unterschied war allerdings nur marginal und statistisch nicht signifikant (Abb. 2). Die Flächen unter den ROC-Kurven waren jeweils für die Kurzzeitprognose (Intensivletalität) größer als bei Betrachtung der gesamten Krankenhausletalität.

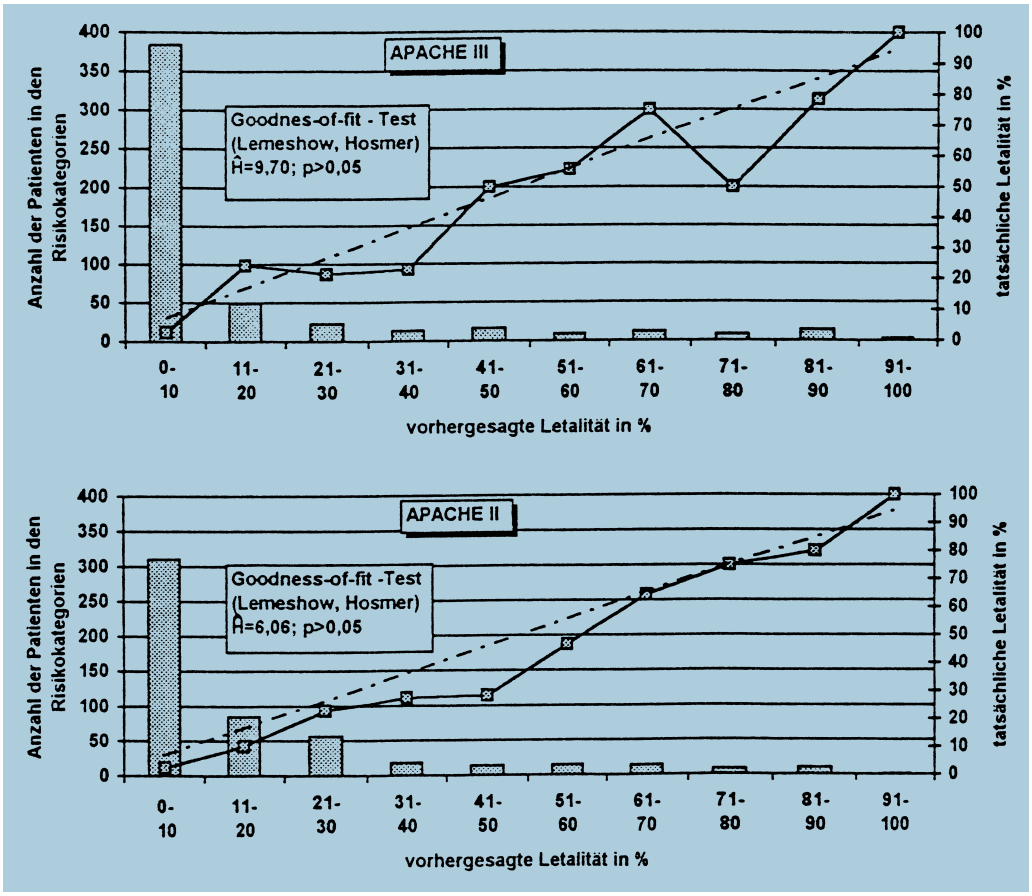

Abb. 3 ▶ **Überprüfung der Kalibration beider APACHE Systeme mittels Gegenüberstellung der vorhergesagten und der tatsächlich beobachteten Krankenhausletalitäten (Linien) für beide Scoresysteme mit Angabe der Patientenzahl in den jeweiligen Risikodezilen (Balken) und den Ergebnissen des goodness-of-fit Tests**

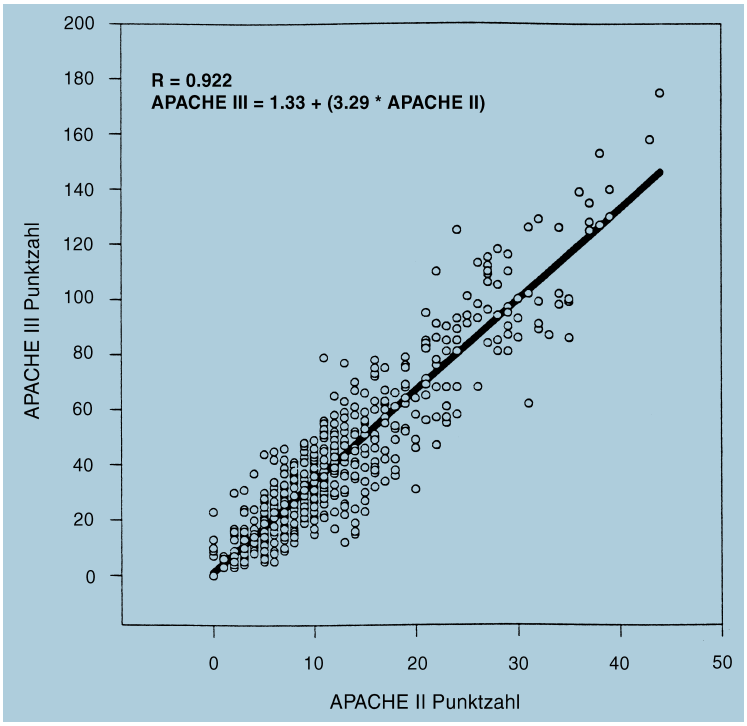

Abb. 4 ▲ **Lineare Regressionsanalyse der Korrelation zwischen den individuellen APACHE II und APACHE III-Scorewerten**

Die Überprüfung der Kalibration (Abb. 3) zeigte für beide Systeme eine homogen gute Beziehung zwischen prognostizierter und tatsächlicher Letalität in allen Prognosebereichen an (niedriger $\hat{H}$-Wert nach Hosmer u. Lemeshow [10], $p>0{,}05$), ebenfalls ohne relevante Unterschiede zwischen den Systemen. Das APACHE II System überschätzt die Letalität dabei offenbar systematisch, vor allem bei den niedrigen Risikoklassen (<60% Letalitätsprognose), während sich eine solche systematische Tendenz beim APACHE III System nicht ablesen läßt.

Die Ergebnisse der Berechnung der mittleren Scorewerte und der mittleren prognostizierten Letalitäten für das Gesamtkollektiv sowie getrennt betrachtet für Überlebende und Verstorbene sind in Tabelle 2 dargestellt. Es wird erkennbar, daß die Letalitätsprognose des APACHE III Systems (13,2%) überaus exakt mit der tatsächlichen Letalität des Gesamtkollektivs (13,4%) übereinstimmt, während die Prognose des APACHE II Systems (16,3%) deutlich mehr von dieser abweicht. Die Mittelwerte sowohl der Scores als auch der Letalitätsprädiktionen unterscheiden sich bei beiden Systemen für Überlebende und Verstorbene hochsignifikant voneinander ($p<0{,}001$).

Es findet sich eine sehr starke Korrelation der individuellen Scorewerte beider Systeme miteinander, der Korrelationskoeffizient betrug $r=0{,}922$ (Abb. 4).

Die Häufigkeitsverteilung und die mittlere tatsächliche Letalität der Patienten in den APACHE II- bzw. III-Kategorien geht aus Abb. 5 hervor. Es zeigt sich eine positive Beziehung im Sinne einer zunehmenden Letalitätsrate in Abhängigkeit von einer zunehmenden Scorewerthöhe sowohl für das APACHE II als auch das APACHE III System (jew. $p<0{,}001$).

Die tatsächlichen und die prognostizierten Letalitätsraten in den verschiedenen Diagnosesubgruppen sind

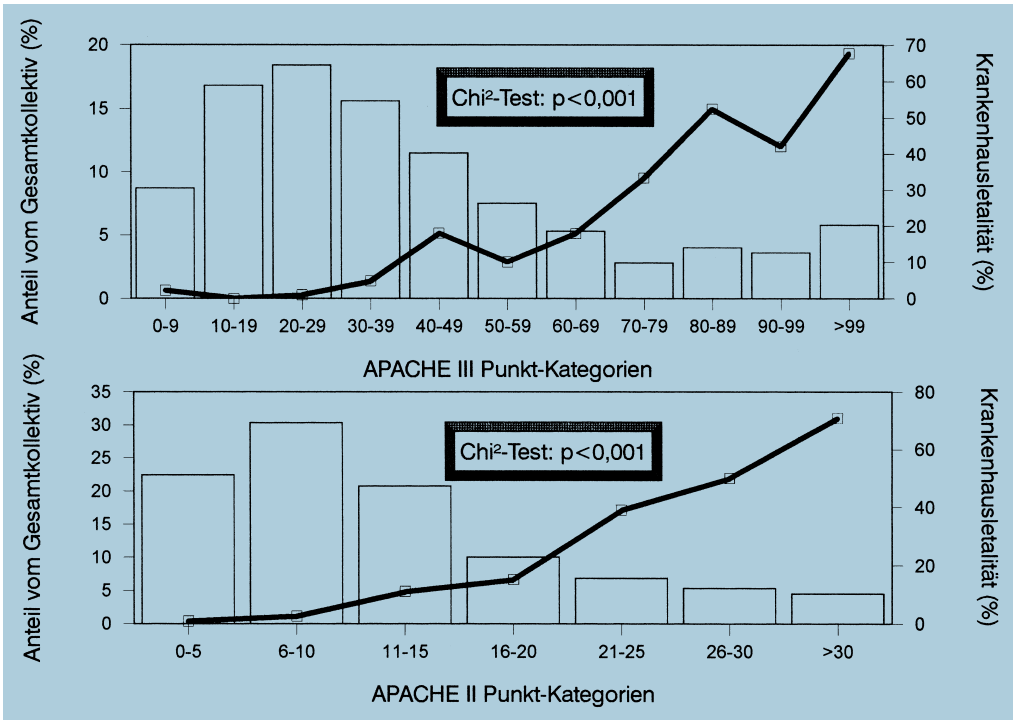

Abb. 5 ▶ **Darstellung der prozentualen Krankenhausletalität (Linien) in Abhängigkeit von der Scorewerthöhe der beiden APACHE Systeme und Ergebnisse der diesbezüglichen Zusammenhangsüberprüfung mit dem χ^2-Test. Die Häufigkeitsverteilung der Patienten (Balken) in den jeweiligen Scorewert-Kategorien ist ebenfalls dargestellt**

Tabelle 3
Gegenüberstellung der tatsächlichen und der durch die APACHE Systeme prognostizierten Letalitätsraten in den verschiedenen Diagnose-Subgruppen sowie der sich daraus ergebenden Letalitätsindizes (inkl. 99%-Konfidenzintervalle)

Subgruppe	Tatsächliche Letalität (%)	APACHE III Letalitätsprädiktion (%)	APACHE II Letalitätsprädiktion (%)	APACHE III Letalitätsindex (99%-Konfidenzintervall)	APACHE II Letalitätsindex (99%-Konfidenzintervall)
kardiovaskulär	13,9	12	14,1	1,17 (0,90–1,66)	0,99 (0,81–1,26)
respiratorisch	12,6	15,2	18,1	0,92 (0,60–1,36)	0,70 (0,53–1,02)
gastrointestinal	10,1	9,6	22,4	1,08 (0.79–1,68)	0,45 (0,37–0,58)
Intoxikation	0	0,7	0,6		
neurologisch	32,4	34,8	28,7	0,93 (0,65–1,66)	1,13 (0,80–1,94)
metabolisch	14,3	8,1	9,3	1,88 (0,94–4,77)	1,61 (0,95–5,11)
Sonstige	17,6	22,9	18,5	0,74 (0,37–1,71)	0,60 (0,42–1,05)
Beatmung (invasiv)	42,2	41,5	40,9	1,02 (0,86–1,25)	1,03 (0,89–1,24)
Gesamtkollektiv	13,4	13,2	16,3	1,02 (0,87–1,17)	0,82 (0,72–0,96)

in Tabelle 3 gegenübergestellt. Auffällig ist die gegenüber dem Gesamtkollektiv deutlich erhöhte Letalitätsrate der Patienten mit neurologischen Grunderkrankungen und das Überleben aller Patienten mit Intoxikationen. Die Abweichungen der prognostizierten Letalität von der tatsächlichen Letalität waren in den einzelnen Diagnosegruppen tendenziell beim APACHE III System zumeist geringer als beim APACHE II System. Dies dokumentiert sich in statistisch nachprüfbarer Weise durch die signifikante Abweichung des 99%-Konfidenzintervalls des Letalitätsindex vom Wert 1 für den APACHE II bzgl. des Gesamtkollektivs und der Subgruppe der Gastrointestinal-Erkrankungen. Auffällig ist ferner, daß die Krankenhausletalität der Gruppe der metabolischen Erkrankungen von beiden Systemen deutlich unterschätzt wird, die der respiratorischen Erkrankungen jedoch übereinstimmend überschätzt wird.

Diskussion

Scoresysteme haben in den letzten Jahren in der Intensivmedizin zunehmend an Bedeutung gewonnen [6, 24, 25, 27]. Einen großen Stellenwert haben hierbei die verschiedenen Versionen des APACHE-Systems erlangt. Die neueste Version, der APACHE III, ist dabei allerdings bisher, insbesondere hinsichtlich der Algorithmen zur prozentualen Vorhersage des Letalitätsrisikos, noch an keinem größeren Kollektiv für die deutsche Intensivmedizin systematisch validiert worden.

Die Resultate der vorliegenden Arbeit zeigen, daß das APACHE III-System gruppenbezogen eine suffiziente Klassifikation der Erkrankungsschwere und Vorhersage des Letalitätsrisikos für eine repräsentativ große Patientenpopulation einer deutschen internistischen Intensivstation gewährt. Das System kann damit für vergleichbare Populationen in der deutschen Intensivmedizin als validiert gelten und auf vergleichbaren Intensivstationen eingesetzt werden.

Grundsätzlich ist vor Einführung eines Scoresystems (z.B. im Rahmen qualitätssichernder Maßnahmen) auf einer Intensiveinheit dessen Validität hinsichtlich der spezifischen Fragestellung (z.B. Letalitätsprädiktion) zu überprüfen [16]. Dies kann entweder durch eine Testphase in der entsprechenden Einheit geschehen, oder es müssen Validitätsprüfungen an *vergleichbaren* Kollektiven vorliegen, wobei neben Faktoren wie der Erkrankungsgruppen-Zusammensetzung und Altersverteilung besonders zu berücksichtigen ist, daß durch Scoresysteme erhobene Daten (insbesondere Standards) nicht ohne erneute Validitätsüberprüfungen auf andere Gesundheitsysteme (Länder) übertragen werden dürfen.

Validierungskriterien

Internationaler Goldstandard [16] für die Validierung von Scoresystemen für bestimmte Populationen sind zum einen die Bestimmung der Diskriminationsfähigkeit (Überleben vs. Versterben) eines Systems und zum anderen seine Kalibration (homogen gute Aussagefähigkeit über alle Risikoklassen). Diese Validierungskriterien sind der Überprüfung von Mittelwertsunterschieden von Überlebenden und Verstorbenen und der Bestimmung von Sensitivitäten und Spezifitäten hinsichtlich der Unterscheidung von Überleben und Versterben an willkürlich gewählten Schwellwerten deutlich überlegen [16].

Die Diskriminationsfähigkeit gemessen mit der Fläche unter der ROC-Kurve war in der vorliegenden Untersuchung sowohl für das APACHE III als auch für das APACHE II System sehr befriedigend, der Unterschied zwischen beiden Systemen war marginal. Die Flächen unter der Kurve lagen dabei in der Größenordnung der Werte der Originalpublikation [14] bzw. anderer vergleichbarer internationaler Einzeluntersuchungen der Systemversionen [19, 21, 28]. Die Fläche unter der Kurve war dabei für die Einschätzung der Kurzzeitprognose (Intensivletalität) jeweils größer als für die Beurteilung der Krankenhausletalität, was darauf hinweist, daß die Krankenhausletalität noch zusätzlich durch andere Faktoren beeinflußt wird, die durch die Scoresysteme in den ersten 24 h des Krankenhausaufenthalts nicht erfaßt werden.

Die Kalibration war ebenfalls bei beiden Systemen sehr befriedigend, was bedeutet, daß die Letalitätsprädik-

tion und die tatsächliche beobachtete Letalität sowohl bei den Patientengruppen mit niedrigem Risiko als auch bei solchen mit mittlerem und hohem Risiko eine homogen gute Übereinstimmung zeigen. Sicher relevante Unterschiede zwischen beiden Systemen haben sich nicht gezeigt, tendenziell scheint das APACHE II System die Letalität systematisch in allen Risikoklassen etwas zu überschätzen, während eine solche Tendenz beim APACHE III nicht ablesbar ist.

Gemäß den beiden primären Validierungskriterien Diskriminationsfähigkeit und Kalibration ergaben sich also keine sicheren Vorteile für das APACHE III System. Eine – wie auch in unserer Untersuchung – diskret größere Fläche unter der Kurve für das APACHE III System bei der ROC-Analyse fand sich auch bei der einzigen vorliegenden vergleichbaren Untersuchung an einer deutschen kleineren (n=150) Patientenpopulation einer anästhesiologischen Intensivstation [2], wobei allerdings lediglich die reinen Scorewerte und nicht die Algorithmen zur prozentualen Vorhersage des Letalitätsrisikos eingesetzt wurden. Die AUC-Differenz war dort zwar größer, aber ebenfalls nicht statistisch signifikant, und die dort geäußerte Vermutung, daß sich möglicherweise bei größeren Kollektiven signifikante Unterschiede zwischen den beiden Systemen ergeben würden, kann aufgrund der Ergebnisse der vorliegenden Untersuchung nicht bestätigt werden.

Letalitätsprädiktion

Die Gesamtletalität der Population lag mit 13,4% extrem nahe an der von APACHE III vorhergesagten (13,2%), was einem Letalitätsindex von 1,02 entspricht. Die von APACHE II vorhergesagte Letalität lag hingegen mit 16,8% doch deutlich über der tatsächlichen Letalität (Letalitätsindex 0,80). Dies bedeutet, daß der durch den APACHE III vorgegebene Standard erfüllt wird. Der durch den APACHE II vorgegebene Standard wird statistisch signifikant übertroffen, da das 99%-Konfidenzintervall unterhalb des Werts 1 liegt [8]. Eine mögliche Erklärung hierfür ist, daß sich der intensivmedizinische Standard (und damit die Prognose der Patienten) in den Jahren zwischen der Entwicklung des APACHE II und des APACHE III verbessert hat, und somit die Verwendung der APACHE II-Letalitätsprädiktion ein systematisches Überschätzen der gegenwärtig zu erwartenden Letalität beinhaltet!

Die Beobachtung der systematischen Überschätzung der tatsächlichen Letalität durch das APACHE II System wurde auch in einer größeren (n=844) Vergleichsuntersuchung an einem rein chirurgischen Kollektiv gemacht [1]. In dieser Untersuchung überschätzte bei einer sehr niedrigen Gesamtletalität (9,1%) allerdings auch das APACHE III System die tatsächliche Letalität systematisch, und darüber hinaus waren die Letalitätsprognosen des APACHE III Systems für die Gesamtpopulation und in allen Untergruppen jeweils signifikant höher als die des APACHE II Systems. Diese Beobachtung der höheren Letalitätsprognosen des APACHE III konnte in unserer Untesuchung eines überwiegend internistischen Patientenguts überhaupt nicht bestätigt werden und bedarf der gezielten Überprüfung an anderen großen chirurgischen Patientenkollektiven.

Neurostatus analgosedierter Patienten

Ein grundsätzliches Problem ist die Bewertung des in die Scoreberechnung ganz entscheidend einfließenden Neurostatus bei analgosedierten Patienten. Hierzu fehlen zwar klar formulierte Empfehlungen, es ist jedoch unzweifelhaft, daß die neurologische Einschätzung solcher Patienten gemäß ihrem (medikamentös induzierten) aktuellen Status irreführend ist und zu einem „Überscoring“ führen kann. Diesbezüglich findet das bei [2] vorgeschlagene und auch bei uns hinsichtlich analgosedierter Patienten angewandte Verfahren der Einschätzung der wahrscheinlichen Vigilanz unter Berücksichtigung aller zur Verfügung stehenden anamnestischen, klinischen und apparativen Informationen zunehmend Akzeptanz. Für die Vergleichbarkeit von scorebasierten Studienergebnissen ist eine einheitliche Bewertung des Neurostatus überaus wichtig; wir halten in diesem Zusammenhang die auch in dieser Studie angewandte Vorgehensweise für empfehlenswert.

Ein wesentliches Hindernis für die Verbreitung des APACHE III Systems ist die Tatsache, daß die entsprechenden Algorithmen und Koeffizienten zur Berechnung der Letalitätsprädiktion nur käuflich erworben werden können und nicht publiziert worden sind. Kostenfrei werden sie von den Autoren lediglich für Forschungsvorhaben – wie für die vorliegende Untersuchung – zur Verfügung gestellt. Aus diesem Grund haben wir für beide Systeme auch die Beziehung der reinen Scorewerthöhe, unabhängig von der Letalitätsprädiktion, zur Krankenhausletalität gruppenbezogen überprüft. Hierbei zeigte sich eine signifikante Beziehung im Sinne einer zunehmenden Krankenhausletalität bei zunehmender Scorewerthöhe. Dies bedeutet den Nachweis, daß mit dem APACHE III (ebenso wie mit dem APACHE II) gruppenbezogen eine suffiziente Klassifikation der Erkrankungsschwere möglich ist, wenn man voraussetzt, daß eine erhöhte Letalitätsrate in einer Patientenkategorie eine erhöhte Erkrankungsschwere bedeutet. Dieses Axiom ist essentieller Bestandteil unseres täglichen praktischen ärztlichen Handelns [23]. Damit ist das APACHE III System mittels der reinen Scorewerterhebung, auch wenn die Algorithmen zur Letalitätsprädiktion nicht zur Verfügung stehen, für die Klassifikation der Erkrankungsschwere von Patienten, und damit auch für die Vergleichbarmachung von Kollektiven, z.B. im Rahmen von Multicenterstudien geeignet.

Die vorliegenden Ergebnisse zeigen, daß die für das Gesamtkollektiv gemachten Aussagen hinsichtlich der gruppenbezogenen Letalitätsprädiktion im wesentlichen auch für die einzelnen Diagnosesubgruppen gelten. Die tatsächlichen Letalitätsraten in den einzelnen Gruppen lagen dabei tendenziell etwas näher an den Letalitätsprädiktionen des APACHE III als an denen des APACHE II. Auffällig ist dabei eine geringere tatsächliche Letalität als von beiden Systemen prädiziert in der Gruppe der respiratorischen Erkrankungen und eine höhere tatsächliche Letalität als übereinstimmend von beiden Systemen vorhergesagt in der Gruppe der metabolischen Erkrankungen.

Mag bezüglich der metabolischen Erkrankungen hierfür die Heterogenität der Erkrankungsgruppe und die geringe Fallzahl ausschlaggebend sein, so vermuten wir, daß im Hinblick auf die

respiratorischen Erkrankungen der hohe Anteil von nicht-invasiver Beatmung auf unserer Station die Prognose gegenüber dem APACHE III-Originalkollektiv verbessert hat. Denn zur Zeit der Erstellung des APACHE III (Ende der 80er Jahre) hatte das Verfahren der nicht-invasiven Beatmung mit seinen gegenüber der herkömmlichen invasiven Beatmung deutlich niedrigeren Letalitätsraten noch keinen Eingang in die intensivmedizinische Praxis gefunden, und dementsprechend noch nicht zu Veränderungen der Bewertungskriterien und -koeffizienten schwerer pulmonaler Erkrankungen geführt. Dies macht deutlich, daß alle Prognosesysteme zum einen in wiederholten Abständen erneuten Validierungsuntersuchungen unterworfen werden müssen, und daß zum anderen auch beispielsweise einmal gefundene Scorehöhen-Grenzwerte (oberhalb derer bisher kein Patient bestimmter Populationen überlebt hat), nie als alleiniges Entscheidungskriterium für die Verweigerung von Intensivtherapie herangezogen werden können [15, 20]. Letzteres würde die entsprechenden Patienten ungerechtfertigterweise von möglicherweise die Prognose verbessernden Neuerungen der Intensivmedizin ausschließen [23].

Die Überschätzung der Letalität in der Subgruppe der metabolischen und respiratorischen Erkrankungen besteht zwar übereinstimmend bei beiden Systemen, eine statistisch signifikante Abweichung vom vorgegebenen Standard (99%-Konfidenzintervall des Letalitätsindex <oder>1) findet sich jedoch nicht. Dieser Standard wird hinsichtlich der Diagnosesubgruppen lediglich für den APACHE II bei den Gastrointestinal-Erkrankungen übertroffen, eine Unterschreitung des Standards findet sich in keinem Fall. Ursächlich für die gegenüber dem APACHE II-Standard verbesserte Prognose dieser Patienten ist am ehesten das im letzten Jahrzehnt verbesserte Management schwerer gastrointestinaler Blutungen.

Die Eignung der Systeme für die *Individual*prognose auf der Basis der Tag-1-Scores bleibt im Gegensatz zur *Gruppen*prognose hingegen weiterhin unbefriedigend, da eine für die Basierung von Individualentscheidungen wie z.B. Therapielimitierung ausschließlich auf Scorewerten zu fordernde 100-prozentige Prognosesicherheit nur für einen sehr geringen Prozentsatz der Patienten möglich ist. Die Systeme eignen sich daher gegenwärtig primär zum Gruppenvergleich und können für Individualentscheidungen allenfalls als ein Co-Faktor unter vielen anderen Faktoren beigezogen werden, niemals jedoch die alleinige Basis von Therapieentscheidungen darstellen [17]. Eine Verbesserung ihrer individualprognostischen Aussagekraft ist möglicherweise in Ansätzen zu finden, die die Verlaufsentwicklung der Scorewerte bei täglich wiederholter Erhebung in der Longitudinalbetrachtung berücksichtigen [3, 18, 26, 29].

Schlußfolgerungen

Synoptisch erbrachten sowohl das APACHE III- als auch das APACHE II-System auf die Gesamtpopulation bezogen valide und vergleichbare Resultate hinsichtlich der Vorhersage der Letalitätswahrscheinlichkeit (Diskriminationsfähigkeit und Kalibration) und der Klassifikation der Erkrankungsschwere (reine Scorewerte). Damit kann das APACHE III System für die internistische Intensivmedizin in Deutschland als validiert gelten und auf Stationen mit vergleichbarem Krankengut zur Anwendung gebracht werden.

Die tatsächliche Letalität der Gesamtpopulation stimmte überaus exakt mit der durch das APACHE III System vorhergesagten Letalität überein, während das APACHE II System teilweise signifikant das Letalitätsrisiko (vgl. auch Letalitätsindices) in unserer Population überschätzte. Dies legt nahe, daß das neuere APACHE III System als Standard, z.B. für Maßnahmen der Ergebnisqualitätskontrolle, in der Intensivmedizin besser geeignet ist als das ältere APACHE II System.

Bei Verwendung der APACHE-Systeme muß jeder Anwender in Abhängigkeit von der jeweiligen Fragestellung entscheiden, welche Systemversion auf seiner Intensivstation zur Anwendung kommen soll. Es muß dabei im Einzelfall abgewogen werden, ob der Zuwachs an Information und Präzision durch das APACHE III System den Mehraufwand zur Scoreerhebung und Risikoberechnung bei Verwendung dieses Systems rechtfertigt.

Fazit für die Praxis

Scoresysteme haben in den letzten Jahren in der Intensivmedizin zunehmend an Bedeutung gewonnen. Der zur Zeit gebräuchlichste und am besten validierte physiologisch basierte Score ist der „Acute physiology and chronic health evaluation" (APACHE) II Score. Als Weiterentwicklung dieses Scores wurde 1991 von Knaus et al. der APACHE III Score vorgestellt. Ziel der vorliegenden Untersuchung war die systematische Validierung des APACHE III Scores auf der Basis von 531 konsekutiven internistischen Intensivpatienten.

Hinsichtlich der gruppenbezogenen Prädiktion der Krankenhausletalität zeigte sowohl der APACHE III als auch der APACHE II Score befriedigende Ergebnisse, wobei die individuellen Werte beider Score-Systeme eine enge Korrelation aufwiesen. Der APACHE III Score scheint dabei insgesamt eine höhere Vorhersagepräzision als der ältere APACHE II zu haben.

Literatur

1. Barie PS, Hydo LJ, Fischer E (1995) **Comparison of APACHE II and III scoring systems for mortality prediction in critical surgical illness.** Arch Surg 130:77–82
2. Bein T, Fröhlich D, Frey A, Meth Ch, Taeger K (1995) **Vergleich von APACHE II und APACHE III zur Einschätzung der Erkrankungsschwere von Intensivpatienten.** Anaesthesist 44:37–42
3. Chang RWS, Bihari DJ (1994) **Outcome prediction for the individual patient in the ICU.** Unfallchirurg 97:199–204
4. Civetta JM, Hudson-Civetta JA, Nelson LD (1990) **Evaluation of APACHE II for cost containment and quality assurance.** Ann Surg 212:266–276
5. Cullen DJ (1982) **The importance of comparative data in critical care analyses.** Crit Care Med 10:618–619
6. Elsasser S, Zuber M, Weber W, Planta Mv, Ritz R (1989) **Wertigkeit der prognostischen Scores Apache 1,2 und TISS in der Intensivmedizin.** Intensivmed 26:80–84
7. Farmer JC (1989) **Intensive care: How do we measure outcome?** Problems Crit Care 3:511–513
8. Feinstein AR (1985) **The architecture of clinical research.** In: Feinstein AR (ed) Clinical Epidemiology. Saunders, Philadelphia, pp 114–115
9. Hanley JA, Mc Neil BJ (1982) **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** Radiology 143–29–36
10. Hosmer DW, Lemeshow S (1989) **Applied logistic regression.** Wiley, New York

11. Knaus WA, Draper EA (1991) **APACHE III Data Collection Manual, Version 1.3.** APACHE Medical Systems, McLean, USA
12. Knaus WA, Draper EA (1991) **APACHE III hospital mortality predictive equation.** APACHE Medical Systems, McLean, USA
13. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) **APACHE II: A severity of disease classification system.** Crit Care Med 13:818–829
14. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M (1991) **The APACHE 3 prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults.** Chest 100:1619–1636
15. Lemeshow S, Klar J, Teres D (1994) **Outcome prediction for individual intensive care patients: useful, misused, or abused?** Intensive Care Med 21:770–776
16. Lemeshow S, Le Gall JR (1994) **Modeling the severity of illness of ICU patients.** JAMA 272:1049–1055
17. Luce JM, Wachter RM (1994) **The ethical appropriateness of using prognostic scoring systems in clinical management.** Crit Care Clin 10:229–241
18. Moser KH, Bouillon B, Troidl H, Köppen L (1989) **Validation of the continous APACHE score (CAPS) for a better prediction of outcome in surgical ICU patients.** Theor Surg 3:192–197
19. Oh TE, Hutchinson R, Short S, Buckley T, Lin E, Leung D (1993) **Verification of the acute physiology and chronic health evaluation scoring system in a Hong Kong intensive care unit.** Crit Care Med 21:698–705
20. Rogers J, Fuller HD (1994) **Use of daily Acute Physiology and Chronic Health Evaluation (APACHE) II scores to predict individual patient survival rate.** Crit Care Med 22:1402–1405
21. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP (1993) **Intensive care society's APACHE II study in Britain and Ireland II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method.** Br Med J 307:977–981
22. Schuster HP (1988) **Hypothese: Score-Systeme optimieren die Intensivmedizin.** Med Klin 83:68–70
23 Schuster HP (1996) **Die Bedeutung von Scoresystemen für die Voraussage von Behandlungsergebnissen in der Intensivmedizin.** Internist 37:1237–1243
24. Schuster HP, Hesse M, Tröster S (1992) **Prognoseeinschätzung durch Ärzte in der Intensivmedizin.** Intensivmed 29:55–60
25. Schuster HP, Wilts S, Ritschel P (1996) **Analyse einer Ergebnisqualitätskontrolle in der Intensivmedizin mittels des SAPS II.** Med Klin 91:343–348
26. Wagner DP, Knaus WA, Harrell FE, Zimmerman JE, Watts CM (1994) **Daily prognostic estimates for critical ill adults in intensive care units: Results from a prospective multicenter inception cohort analysis.** Crit Care Med 22:1359–1372
27. Watts CM, Knaus WA (1994) **The case for using objective scoring – systems to predict intensive care unit outcome.** Crit Care Clin 10:73–89
28. Wong DT, Crofts SL, Gomez M, McGuire GP (1995) **Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients.** Crit Care Med 23:1177–1183
29. Zimmerman JE, Wagner DP, Draper EA, Knaus WA (1994) **Improving intensive care unit discharge decisions: Supplementing physician judgement with predictions of next day risk for life support.** Crit Care Med 22:1373–1384