CHEST

Original Research

CRITICAL CARE

Comparison of APACHE III, APACHE IV, SAPS 3, and MPM_0III and Influence of Resuscitation Status on Model Performance

Mark T. Keegan, MB; Ognjen Gajic, MD, FCCP; Bekele Afessa, MD, FCCP

Background: **There are few comparisons among the most recent versions of the major adult ICU prognostic systems (APACHE [Acute Physiology and Chronic Health Evaluation] IV, Simplified Acute Physiology Score [SAPS] 3, Mortality Probability Model [$MPM]_0III$). Only MPM_0III includes resuscitation status as a predictor.**

Methods: **We assessed the discrimination, calibration, and overall performance of the models in 2,596 patients in three ICUs at our tertiary referral center in 2006. For APACHE and SAPS, the analyses were repeated with and without inclusion of resuscitation status as a predictor variable.**

Results: **Of the 2,596 patients studied, 283 (10.9%) died before hospital discharge. The areas under the curve (95% CI) of the models for prediction of hospital mortality were 0.868 (0.854-0.880), 0.861 (0.847-0.874), 0.801 (0.785-0.816), and 0.721 (0.704-0.738) for APACHE III, APACHE IV, SAPS 3, and MPM_0III, respectively. The Hosmer-Lemeshow statistics for the models were 33.7, 31.0, 36.6, and 21.8 for APACHE III, APACHE IV, SAPS 3, and MPM_0III, respectively. Each of the Hosmer-Lemeshow statistics generated *P* values $< .05$, indicating poor calibration. Brier scores for the models were 0.0771, 0.0749, 0.0890, and 0.0932, respectively. There were no significant differences between the discriminative ability or the calibration of APACHE or SAPS with and without "do not resuscitate" status.**

Conclusions: **APACHE III and IV had similar discriminatory capability and both were better than SAPS 3, which was better than MPM_0III. The calibrations of the models studied were poor. Overall, models with more predictor variables performed better than those with fewer. The addition of resuscitation status did not improve APACHE III or IV or SAPS 3 prediction.**

CHEST 2012; 142(4):851–858

Abbreviations: APACHE = Acute Physiology and Chronic Health Evaluation; AUC = area under the receiver operator characteristic curve; DNR = do not resuscitate; MPM = Mortality Probability Model; ROC = receiver operating characteristic; SAPS = Simplified Acute Physiology Score

The assessment of quality of care in the practice of medicine has become increasingly important.[1-3] The practice of critical care medicine has been especially scrutinized, at least in part because of the enormous costs of providing critical care services.[4,5] Aside from external pressures, monitoring and improvement of quality of care are important to clinicians.[6,7]

Misuse of quality measures may occur and "[risks] stigmatizing an entire institution,"[8] so it is imperative that any assessment of the quality of care delivered in the ICU involves a consideration of the severity of patient illness using a reliable measure. The Joint Commission[9] has proposed severity-adjusted mortality rate as a specific measure that should be recorded.

Prognostic scoring systems have been developed by the critical care community in an effort to quantify the severity of illness of a given patient or group of patients.[10-12] Adjustment for severity of illness enables monitoring of the performance of an ICU over time and for comparison of ICUs in the same or different hospitals. It is imperative that such severity adjustments be as accurate as possible.[13]

Many prognostic models exist, suggesting that the optimum model has not been established. The three most commonly used adult-ICU prognostic scoring systems are APACHE (Acute Physiology and Chronic Health Evaluation), the Simplified Acute Physiology Score (SAPS), and the Mortality Probability

Model (MPM). Any prognostic model will have a limited effective life span.[10,14] Changes in clinical practice over time and alterations in the provision of health care will alter the risk of mortality for a given clinical situation. Thus, prognostic models require updating. Major revisions of the prognostic models were published between 2005 and 2007, namely APACHE IV in 2006, SAPS 3 in 2005, and MPM_0III in 2007.[15-18]

A significant number of deaths in patients admitted to ICUs occur after a decision to forgo life-sustaining therapy.[19] Of the three major prognostic models, only MPM includes the presence of a do-not-resuscitate (DNR) order as a predictor variable. A patient's desire not to undergo CPR is not specifically accounted for in APACHE or SAPS. This is despite the fact that DNR status has been demonstrated to be an independent predictor of mortality in patients in the ICU.[20,21] It is possible that both SAPS and APACHE predict an inaccurately high likelihood of survival in certain circumstances because of the assumption that the full armamentarium of ICU resources will be used to support a patient for whom, in fact, limitations on the level of care have been placed.

We hypothesized that the performances of the three major ICU prognostic models for the prediction of mortality for patients in the ICU at Mayo Medical Center, Rochester, Minnesota, differ and that the performance of prognostic models with many variables is superior to the performance of models with fewer variables. Further, we hypothesized that the presence of a DNR order has a significant effect on mortality not accounted for in the APACHE and SAPS prognostic models. In this study, we compared the performance of APACHE III, APACHE IV, SAPS 3, and MPM_0III for the prediction of in-hospital mortality in a retrospective cohort of patients in the ICU at our institution and evaluated the impact of patients' DNR status on their first ICU day on the performance of the APACHE and SAPS prognostic models.

Manuscript received August 27, 2011; revision accepted March 19, 2012.

Affiliations: From the Division of Critical Care, Department of Anesthesiology (Dr Keegan), the Division of Pulmonary and Critical Care, Department of Medicine (Drs Gajic and Afessa), and the Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC) group (Drs Keegan, Gajic, and Afessa), Mayo Clinic, Rochester, MN.

Funding/Support: This research was funded by grants to Dr Keegan from the Mayo Clinic Rochester Critical Care Research Subcommittee and the Mayo Clinic Rochester Quality Innovation Program. This project was supported by the US National Institutes of Health/National Center for Research Resources Clinical and Translational Science Awards [Grant UL1 RR024150].

Correspondence to: Mark T. Keegan, MB, Mayo Clinic Department of Anesthesiology, Charlton 1145, 200 First St SW, Rochester, MN 55905; e-mail: keegan.mark@mayo.edu

DOI: 10.1378/chest.11-2164

Materials and Methods

After institutional review board approval (number 2482-05) and waivers of informed consent were obtained, a retrospective cohort study was performed. Adult patients admitted to any of three ICUs at Mayo Medical Center, Rochester, between January 1, 2006, and December 31, 2006, were identified. The three ICUs included a 20-bed vascular/thoracic/orthopedic ICU, a 24-bed medical ICU, and a 20-bed mixed medical-surgical ICU. These ICUs were chosen because APACHE had been employed routinely in them for more than a decade. After power analysis, a cohort of 2,600 patients was identified using computerized randomization from approximately 5,000 admissions to the three ICUs studied. Patients who did not give consent for use of their medical record for research purposes and patients who remained in the ICU for fewer than 4 hours were excluded. Only first ICU admissions were included.

The medical records of each patient, the institutional APACHE database, and the ICU electronic "DataMart" were reviewed. The databases contained prospectively collected data, acquired as part of ongoing clinical care, quality improvement projects, and research activities. Our institutional experience with APACHE (including quality control measures), the nature of the ICU DataMart, and the staffing models in the ICUs have been described previously.[22,23] In addition to demographic variables, the data required for the calculation of APACHE III, APACHE IV, SAPS 3, and MPM_0III predictions of mortality were abstracted. There are a number of variables that are required for the calculation of SAPS 3 and MPM_0III that were not routinely collected as part of the APACHE dataset. These data were abstracted from individual patient records by trained abstractors. DNR status during the patient's first ICU day was also obtained from the databases and individual medical record review. DNR status was defined as the presence of an order in the medical record to not initiate basic or advanced life support measures in the event of a cardiac arrest. Patient vital status at hospital discharge (survivor or nonsurvivor) was obtained for all patients. APACHE IV, SAPS 3, and MPM_0III prediction of mortality were calculated using published formulae.[15-18] APACHE III predictions of mortality were obtained from the institutional APACHE III database using proprietary software provided by the Cerner Corporation.

Descriptive data were summarized as mean (SD), median (interquartile range), or percentage. χ^2 Tests were used to compare categorical variables and Student t test and rank sum tests were used to compare continuous variables. Evaluation of the predictive models was by assessment of calibration, discrimination, and overall performance. We determined the area under the receiver operating characteristic (ROC) curve (AUC) with its 95% CI for discrimination.[24] Discrimination was classified as perfect, excellent, very good, good, moderate, and poor if the AUCs, were 1.0, 0.9 to 0.99, 0.8 to 0.89, 0.7 to 0.79, 0.6 to 0.69, or <0.6, respectively.[25] Comparison of two AUCs (derived from the same set of patients) was performed by taking into account the correlation between the areas that is induced by the paired nature of the data.[26] Calibration for each model was assessed using the Hosmer-Lemeshow goodness-of-fit C statistic.[27] A nonsignificant P value was considered evidence of good calibration. The Brier score was assessed as a measure of overall model performance.[28,29] It measures the average squared deviation between predicted probabilities for a set of events and their outcomes. A lower score represents higher accuracy. The Brier score offers an

overall assessment of performance, involving elements of both discrimination and calibration.

In the cases of SAPS 3, APACHE III, and APACHE IV, the performances of the models with and without the inclusion of DNR status were assessed. APACHE III and IV and SAPS 3 generate an in-hospital mortality prediction on the first ICU day. Accordingly, the DNR status of the patient at the end of the first ICU day was considered the DNR status for the purpose of modeling. In addition to evaluation of the overall performances of the models, a comparison was made between the performances of the models for surgical and nonsurgical patients. Statistical analyses were performed using JMP, version 8.0 (SAS Institute, Inc); SPSS, version 11.5.0 (SPSS Inc); Medcalc, version 8.1.1.0 (MedCalc Software); and Confidence Interval Analysis, version 2.1.2 (University of Southampton).

Results

Data were available for 2,596 patients, of whom 283 (10.9%) did not survive to hospital discharge. Table 1 shows the characteristics of the patients in the cohort and provides a comparison of survivors to hospital discharge vs nonsurvivors. The performance characteristics of each of the studied models are summarized in Table 2. The discriminatory performances of APACHE III and APACHE IV were similar ($P = .621$) and classed as very good. SAPS 3 also had very good discrimination, but was inferior to APACHE III and IV ($P < .001$). MPM_0III had good discrimination and was inferior to the other models ($P < .001$ for each comparison). A comparison of ROC curves is provided in Figure 1. Each of the Hosmer-Lemeshow statistics generated P values $< .05$, indicating poor calibration.

Influence of Surgical Status

Of the 2,596 patients in the cohort, 874 (33.7%) underwent surgical procedures before ICU admission. Fewer patients died in the surgical group (17 patients, 1.9% of the surgical patients) than in the nonsurgical group (266 patients, 15.4% of the nonsurgical group) ($P < .01$). These analyses for surgical and nonsurgical patients are shown in Table 3 and comparisons of the ROC curves are provided in Figure 2.

Influence of DNR Status

There was no significant difference between the discriminatory performance of APACHE III with or without including first-day DNR status ($P = 0.103$). Similarly, discriminatory capacities of both APACHE IV and SAPS 3 were not influenced by the presence or absence of DNR status in the model ($P = .145$ for APACHE IV and $P = .072$ for SAPS 3). The performance hierarchy of discrimination of the models in the patients with (n = 252) and without (n = 2344) DNR orders was similar to the pattern seen overall. Calibration was poor for all models, irrespective of the inclusion or noninclusion of DNR status. For each model, the Brier scores showed a small decrease (ie, slightly more accurate prediction) with the addition of DNR status.

Discussion

In our cohort of patients, APACHE III and IV had similar discrimination and both had better discrimination than SAPS 3, which, in turn, was better than MPM_0III. The calibration of each of the models was poor. Overall performance, as assessed by Brier scores, was best for APACHE IV and worst for MPM_0III. Our data demonstrate that the more complex prognostic scoring systems performed better than those with fewer variables. The addition of DNR status to the individual models did not alter the discriminatory capability or the calibration, although for each model the Brier scores showed a slight decrease with the addition of DNR. The relationship between the AUCs of the models was independent of inclusion of DNR status.

Previous Comparisons of Prognostic Models

There is a paucity of studies comparing the predictive accuracies of APACHE IV, SAPS 3, and MPM_0III. Costa e Silva et al[30] compared APACHE IV, SAPS 3, and MPM_0III in a narrow cohort of 366 patients with acute kidney injury. There were no differences in either discrimination or calibration between models. Kuzniewicz et al[31] compared the performance of APACHE IV, SAPS II (not 3), and MPM_0III in 11,300 patients in the ICU admitted to 35 California hospitals between 2001 and 2004. In findings similar to our study, they found that APACHE IV (AUC, 0.892) discriminated better than SAPS II (AUC, 0.873), which discriminated better than MPM_0III (AUC, 0.809) ($P < .001$). In contrast to the results of our investigation, the calibration of each model in the Kuzniewicz et al[31] study was satisfactory, perhaps reflecting differences in case mix, referral patterns, or other factors between the two populations studied. The investigators also assessed the burden of data collection. Abstraction time correlated with the number of variables required for each model. APACHE IV required 37.3 min (95% CI, 28.0-46.6) per patient; SAPS II, 19.6 min (95% CI, 17.0-22.2); and MPM_0III, 11.1 min (95% CI, 8.7-13.4).

Another study from the same group also documented a lower data collection burden associated with MPM_0III compared with the other models.[32] The automated collection of APACHE III and IV data in our cohort meant that overlapping SAPS 3 and MPM_0III data elements did not need to be

Table 1—*Characteristics of Patients in Cohort: Overall, Survivors to Hospital Discharge, and Nonsurvivors*

Parameter	Overall (N = 2,596)	Survivors (n = 2,313)	Nonsurvivors (n = 283)	*P* Value[a]
Mean age, y (SD)	63.2 (17.4)	62.7 (17.6)	66.6 (15.9)	< .01
Sex				.69
Female	1,173 (45.2)	1,042 (45.1)	131 (46.3)	
Male	1,423 (54.8)	1,271 (54.9)	152 (53.7)	
ICU				< .01
Mixed	760 (29.3)	680 (29.4)	80 (28.3)	
Medical	1,162 (44.8)	982 (42.5)	180 (63.6)	
Surgical	674 (26.0)	651 (28.2)	23 (8.1)	
Admission Source				< .01
Direct admission	227 (8.7)	206 (8.9)	21 (7.4)	
ED	624 (24.0)	556 (24.0)	68 (24.0)	
Floor	650 (25.0)	525 (22.7)	125 (44.2)	
Other ICU	22 (0.9)	15 (0.7)	7 (2.5)	
Operating room	339 (13.1)	331 (14.3)	8 (2.8)	
Other hospital	173 (6.7)	138 (6.0)	35 (12.4)	
Recovery room	535 (20.6)	526 (22.7)	9 (3.2)	
Respiratory care unit	26 (1.0)	16 (0.7)	10 (3.5)	
DNR status				
On ICU admission	211 (8.1)	156 (6.7)	55 (19.4)	< .01
At end of first day	252 (9.7)	160 (6.9)	92 (32.5)	< .01
CPR before admission	36 (1.4)	17 (0.7)	19 (6.7)	< .01
Mechanical ventilation within 1 h of ICU admission	630 (24.3)	531 (23.0)	99 (35.0)	< .01
Surgical procedure	874 (33.7)	857 (37.1)	17 (6.0)	< .01
Hospital discharge location				< .01
Death	283 (10.9)	0	283 (100)	
Home	1,542 (59.4)	1,542 (66.7)	0	
Rehabilitation center	80 (3.1)	80 (3.5)	0	
Skilled nursing facility	460 (17.7)	460 (19.9)	0	
Other hospital	122 (4.7)	122 (5.3)	0	
Other	109 (4.2)	109 (4.7)	0	
Acute physiology score, mean (SD)	42.5 (23.6)	38.8 (19.4)	72.3 (32.3)	< .01
APACHE III predicted hospital death, mean % (SD)	16.5 (21.4)	12.6 (16.7)	47.7 (28.4)	< .01
APACHE IV predicted hospital death, mean % (SD)	14.7 (19.4)	11.3 (14.8)	42.9 (27.9)	< .01
SAPS 3 score, mean (SD)	45.3 (13.2)	43.7 (12.3)	58.6 (13.0)	< .01
SAPS 3 predicted hospital death, mean % (SD)	16.4 (16.8)	14.2 (14.6)	34.3 (21.8)	< .01
MPM_0III predicted hospital death, mean % (SD)	13.9 (14.2)	12.5 (12.6)	25.4 (20.5)	< .01

Data given as No. (%) unless otherwise indicated. Percentages given are percentages of the overall group, survivor group, or nonsurvivor group. For example, there were 80 nonsurvivors in the Mixed ICU group, accounting for 28.3% (80 of 283) of the nonsurvivors overall. APACHE = Acute Physiology and Chronic Health Evaluation; DNR = do not resuscitate; MPM = Mortality Probability Model; SAPS = Simplified Acute Physiology Score.
[a]*P* value is for comparison between survivors and nonsurvivors.

collected separately. Approximately 12 to 15 min per medical record were required to abstract the required additional data for SAPS 3 and MPM_0III calculation.

Discrimination

In this cohort of patients, the discriminatory capabilities of APACHE III and APACHE IV were similar. The 95% CIs for the APACHE AUCs in our cohort overlapped each other and also overlapped the CIs of the AUCs in the original APACHE III and APACHE IV publications. The performances were classified as very good by the rating scheme defined at the outset of the study.

APACHE IV (142 variables) discriminated better than SAPS 3 (20 variables), which discriminated better than MPM_0III (16 variables), suggesting that the inclusion of more predictor variables in the prognostic model is associated with better discriminatory performance. However, with the use of additional variables comes the associated increase in data collection requirements. We did not specifically examine the resources required for data collection for each of the models. The improvement in discriminatory performance of complex prognostic scoring systems must be weighed against the financial and labor costs associated with maintaining mechanisms to collect data for such systems. The ongoing development of the electronic medical record and improvements to the interface between such records and resources to calculate prognostic scores may decrease the burden of calculation in the future.

Calibration

The calibration of each of the models was poor. There are many potential causes for suboptimal calibration.[33] It is influenced especially by case mix.

Table 2—*Performance Characteristics of the Prognostic Models Studied*

Prognostic Model	AUC (95% CI)	HLS	HLS *P* Value	Brier Score
APACHE III	0.868 (0.854-0.880)	33.66	< .05	0.0771
APACHE III DNR	0.876 (0.855-0.897)	29.25	< .05	0.0697
APACHE IV	0.861 (0.847-0.874)	31.00	< .05	0.0749
APACHE IV DNR	0.868 (0.846-0.891)	33.32	< .05	0.0700
SAPS 3	0.801 (0.785-0.816)	36.64	< .05	0.0890
SAPS 3 DNR	0.816 (0.791-0.841)	29.00	< .05	0.0780
MPM_0III[a]	0.721 (0.690-0.752)	21.80	< .05	0.0932

AUC = area under the receiver operating characteristic curve; HLS = Hosmer-Lemeshow statistic. See Table 1 legend for expansion of other abbreviations.

[a]MPM_0III includes DNR status in the model.

Although the cohort of patients used in our study represented a diverse group of critically ill patients, it differed in at least some respects from the case mix used to develop and validate the original models. Lemeshow and Hosmer,[27] Murphy-Filkins et al,[34] and others have highlighted the influence of sample size on calibration, and while our sample was large, it was relatively small when compared with the large cohorts (tens of thousands of patients) used in the development of the major scoring systems.[35,36] Smaller sample size tends to decrease the statistical power to detect lack of fit but the significance of the Hosmer-Lemeshow χ^2 statistics in our cohort might also be explained by ICU performance differences in low- vs high-risk patient groups.

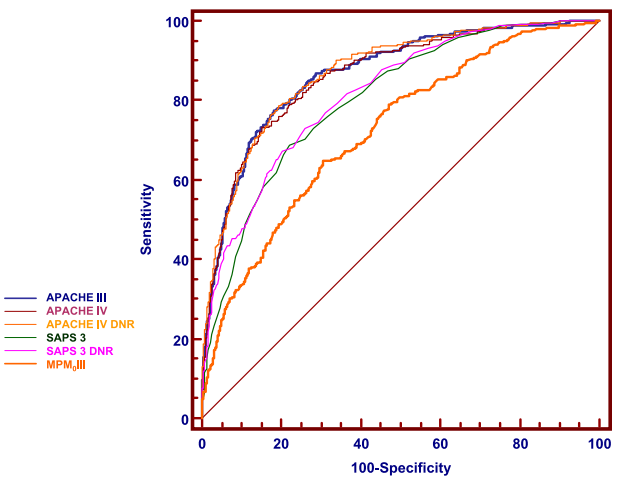

FIGURE 1. Comparison of receiver operating characteristic (ROC) curves for prediction of hospital death by APACHE III, APACHE IV (with and without DNR status), SAPS 3 (with and without DNR status), and MPM_0III. The discriminatory performances of APACHE III and APACHE IV were similar ($P = .621$). Both APACHE models discriminated better than SAPS 3 ($P < .001$). MPM_0III discrimination was inferior to all the other models ($P < .001$ for each comparison). Inclusion of DNR status in the APACHE or SAPS models did not significantly change the areas under the curve. APACHE III DNR is omitted as the curve is almost superimposed on APACHE III. APACHE = Acute Physiology and Chronic Health Evaluation; DNR = do not resuscitate; SAPS = Simplified Acute Physiology Score; MPM = Mortality Probability Model.

Overall Performance

Our data demonstrated a progressive improvement in the Brier score as the complexity of scoring system increased (Table 2). These results were consistent with the performances for discrimination.

Influence of Surgical Status on Model Performance

For each scoring system, the AUCs for surgical patients were similar to the AUCs for nonsurgical patients. The calibration of the models was better in the surgical patients. The calibration of APACHE III, APACHE IV, and MPM_0III in surgical patients was good. The number of events (ie, deaths) in the surgical group may have been too low to detect a significant *P* value for the χ^2 test for calibration. It is also conceivable, of course, that the models do, indeed, exhibit better calibration in the surgical group. The Brier scores in the surgical group (range: 0.0173-0.0297) were also lower than in the nonsurgical group (range: 0.0943-0.1259), perhaps indicating better overall performance.

Influence of Resuscitation Status on Model Performance

End-of-life issues are a feature of the practice of critical care medicine, and one in five Americans dies using ICU services.[37,38] In Olmsted County, Minnesota, in which Mayo Clinic, Rochester, is located, one in eight decedents received ICU care during a terminal hospital admission.[39] The principles of withholding and withdrawing life-sustaining treatments in the ICU have been established, and there is a growing acceptance by medical staff, patients, and patients' families of the appropriateness of such action.[38,40-46] The perceived probability of survival influences patient preferences regarding the desire to undergo CPR in the event of a cardiac arrest.[43] In an analogous fashion, the presence of a DNR order influences patient outcome.[47,48]

Based on previously published work and preliminary data, we postulated that the inclusion of a DNR order would alter the performances of APACHE and SAPS.[20,21] Our results refuted our hypothesis. No improvement in discrimination or calibration was seen and the minor improvement in the Brier scores is of uncertain practical significance. Our findings suggest that the inclusion of many other variables in the prognostic models compensates for the noninclusion of DNR status. Patients with many physiologic

Table 3—*Discrimination and Calibration of Prognostic Models (With and Without DNR Status) in Surgical and Nonsurgical Patients*

Prognostic Model	AUC (95% CI)	HLS	HLS *P* Value	Brier Score
Surgical patients (n = 874)				
APACHE III	0.848 (0.777, 0.920)	11.79	.161	0.0297
APACHE III DNR	0.858 (0.785, 0.931)	15.78	.046	0.0216
APACHE IV	0.823 (0.725, 0.920)	10.31	.244	0.0299
APACHE IV DNR	0.858 (0.774, 0.943)	8.94	.348	0.0219
SAPS 3	0.724 (0.616, 0.832)	16.57	.035	0.0234
SAPS 3 DNR	0.759 (0.648, 0.871)	11.18	.192	0.0173
MPM_0III	0.783 (0.658, 0.908)	8.91	.350	0.0287
Nonsurgical patients (n = 1,722)				
APACHE III	0.844 (0.818, 0.870)	22.75	.004	0.1022
APACHE III DNR	0.852 (0.828, 0.877)	59.30	< .001	0.0943
APACHE IV	0.841 (0.815, 08.68)	24.32	.002	0.0976
APACHE IV DNR	0.845 (0.820, 0.871)	58.72	< .001	0.0951
SAPS 3	0.746 (0.715, 0.777)	23.55	.003	0.1224
SAPS 3 DNR	0.765 (0.734, 0.795)	24.40	.002	0.1096
MPM_0III	0.681 (0.647, 0.716)	22.68	.004	0.1259

See Table 1 and 2 legends for expansion of abbreviations.

derangements and multiple comorbidities will generate high predictions of mortality. Similarly, such patients are more likely to have DNR orders on admission to the ICU or to have discussions regarding resuscitation status initiated early in their ICU stay. The significance of resuscitation status for prediction of in-hospital mortality in the original description of MPM_0III may perhaps be explained by the relatively small number of variables (n = 16) used in the model. In the case of SAPS 3 (20 variables), it may be that the statistical power of well-chosen variables cumulatively compensates for the noninclusion of DNR status in the same manner that the large number of variables used in APACHE IV (n = 142) obviates the need for inclusion of DNR status.

We did not explore the spectrum of limitations of care that occurs in our institution and throughout the United States, ranging from comfort measures only to full support up until the point of cardiac arrest.[38,40,41,46,49,50] For the purposes of our evaluation, we chose the presence or absence of a DNR order. Such orders are well defined in our institution and carefully documented in a specific area of the medical record. Evaluation of the influence of other degrees of care limitation on prognostic models might reveal different results. We must also consider the possibility that the frequency of occurrence of DNR was not high enough to influence the results and that evaluation of a larger cohort might lead to significance.

Strengths and Limitations

Our study compares APACHE III and IV, SAPS 3, and MPM_0III in a cohort of substantial size. It is one of the few independent validation studies of APACHE IV, SAPS 3, and MPM_0III. The single-center nature of the study limits its external validity, and our results may or may not be applicable to other ICUs where patient case mix, care models, and admission criteria, especially for those in whom care is limited, are different. We noted, however, that the overall pattern of discriminatory performance was maintained when the analyses were repeated for each ICU separately (data not shown), suggesting that the results were not markedly changed in units with different case mixes. Furthermore, evaluation of the models in a single institution avoided the potential confounding effects of institutionally related differences in management. The APACHE data were collected prospectively and stored in a database for review. It was necessary, however, to retrospectively abstract data from medical records for a number of the SAPS and MPM variables. Such retrospective data abstraction carries a risk of error, although steps were taken to ensure high-quality data abstraction. A further potential limitation is the dynamic nature of the prognostic scoring systems. The models lose calibration over time and the performance of the models in this cohort of patients may not reflect current performance.

Conclusions

Our data demonstrate that APACHE III and IV discriminate better than SAPS 3, which discriminates better than MPM_0III. Calibration of all the models was poor. Calibration was better for surgical patients. The discrimination and overall accuracy were related to the number of variables in the model, with more complex scoring systems performing better than simpler models. Neither the discrimination nor calibration

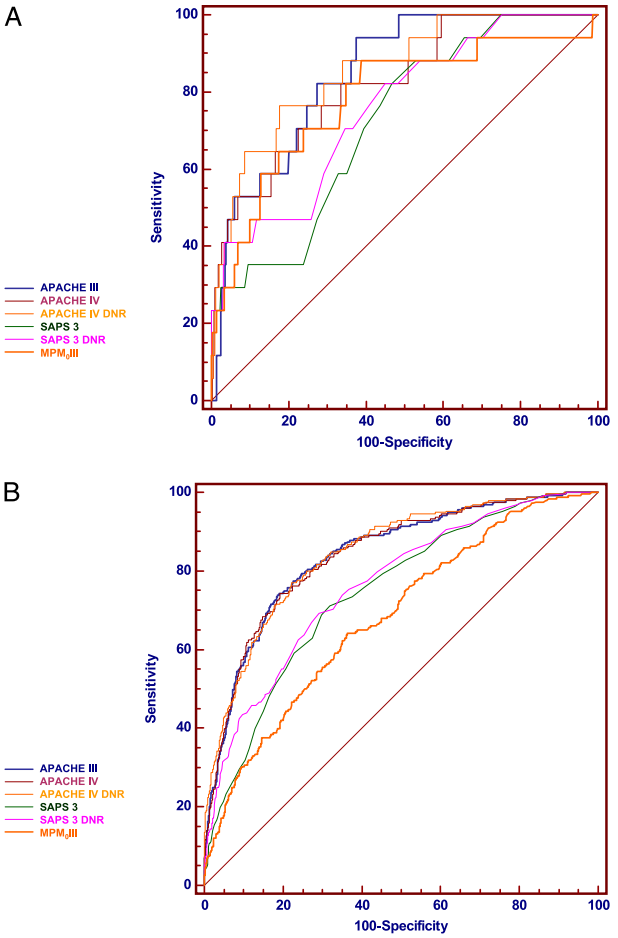

FIGURE 2. A, Comparison of ROC curves for prediction of hospital death by APACHE III, APACHE IV (with and without DNR status), SAPS 3 (with and without DNR status), and MPM_0III in surgical patients. The discriminatory performances of the models were similar. The 95% CIs for the areas under the curve (AUCs) in surgical patients were wide and overlapped. Inclusion of DNR status in the APACHE or SAPS models did not significantly change the AUCs. APACHE III DNR is omitted. B, Comparison of ROC curves for prediction of hospital death by APACHE III, APACHE IV (with and without DNR status), SAPS 3 (with and without DNR status), and MPM_0III in nonsurgical patients. The discriminatory performances of APACHE III and APACHE IV were similar ($P = .988$). Both APACHE models discriminated better than SAPS 3 ($P < .001$). MPM_0III discrimination was inferior to all of the other models ($P < .001$ for each comparison). Inclusion of DNR status in the APACHE or SAPS models did not significantly change the AUCs. APACHE III DNR is omitted. See Figure 1 legend for expansion of abbreviations.

of APACHE III, APACHE IV, or SAPS 3 was significantly improved by the inclusion of DNR status at the end of the first ICU day as a predictor variable.

ACKNOWLEDGMENTS

Author contributions: *Dr Keegan:* contributed to conception and design of the study; obtaining funding; data abstraction, analysis, and interpretation; writing and approval of the final manuscript; and served as principal author and guarantor of the manuscript.
Dr Gajic: contributed to study conception and design; data abstraction and interpretation; review and revising of the final manuscript for intellectual content; and approval of the final manuscript.
Dr Afessa: contributed to study conception and design, data interpretation, review and revising of the final manuscript for intellectual content, and approval of the final manuscript.
Financial/nonfinancial disclosures: The authors have reported to *CHEST* that no potential conflicts of interest exist with any companies/organizations whose products or services may be discussed in this article.
Role of sponsors: The sponsors had no role in the design of the study, the collection and analysis of the data, or in the preparation of the manuscript. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Blumenthal D. Part 1: quality of care—what is it? *N Engl J Med*. 1996;335(12):891-894.
2. Berwick DM, Calkins DR, McCannon CJ, Hackbarth AD. The 100,000 lives campaign: setting a goal and a deadline for improving health care quality. *JAMA*. 2006;295(3):324-327.
3. Angus DC, Black N. Improving care of the critically ill: institutional and health-care system approaches. *Lancet*. 2004; 363(9417):1314-1320.
4. Halpern NA, Pastores SM. Critical care medicine in the United States 2000-2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med*. 2010;38(1): 65-71.
5. Angus DC, Shorr AF, White A, Dremsizov TT, Schmitz RJ, Kelley MA; Committee on Manpower for Pulmonary and Critical Care Societies (COMPACCS). Critical care delivery in the United States: distribution of services and compliance with Leapfrog recommendations. *Crit Care Med*. 2006;34(4): 1016-1024.
6. Curtis JR, Cook DJ, Wall RJ, et al. Intensive care unit quality improvement: a "how-to" guide for the interdisciplinary team. *Crit Care Med*. 2006;34(1):211-218.
7. Pronovost PJ, Nolan T, Zeger S, Miller M, Rubin H. How can clinicians measure safety and quality in acute care? *Lancet*. 2004;363(9414):1061-1067.
8. Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet*. 2004;363(9415):1147-1154.
9. Joint Commission. The Joint Commission announces the 2006 National Patients Safety Goals and requirements. *Jt Comm Perspect*. 2005;25(7):1-10.
10. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. *Crit Care Med*. 2011;39(1): 163-169.
11. Rubenfeld GD, Angus DC, Pinsky MR, Curtis JR, Connors AF Jr, Bernard GR. Outcomes research in critical care: results of the American Thoracic Society Critical Care Assembly Workshop on Outcomes Research. The Members of the Outcomes Research Workshop. *Am J Respir Crit Care Med*. 1999;160(1):358-367.
12. Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. *Crit Care*. 2010;14(2):207.
13. Glance LG, Osler TM, Dick A. Rating the quality of intensive care units: is it a function of the intensive care unit scoring system? *Crit Care Med*. 2002;30(9):1976-1982.
14. Beck DH, Smith GB, Pappachan JV. The effects of two methods for customising the original SAPS II model for intensive care patients from South England. *Anaesthesia*. 2002;57(8):785-793.
15. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV:

hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34(5):1297-1310.
16. Metnitz PG, Moreno RP, Almeida E, et al; SAPS 3 Investigators. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. *Intensive Care Med*. 2005;31(10): 1336-1344.
17. Moreno RP, Metnitz PG, Almeida E, et al; SAPS 3 Investigators. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med*. 2005;31(10):1345-1355.
18. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med*. 2007;35(3):827-835.
19. Luce JM, Rubenfeld GD. Can health care costs be reduced by limiting intensive care at the end of life? *Am J Respir Crit Care Med*. 2002;165(6):750-754.
20. Azoulay E, Pochard F, Garrouste-Orgeas M, et al; Outcomerea Study Group. Decisions to forgo life-sustaining therapy in ICU patients independently predict hospital death. *Intensive Care Med*. 2003;29(11):1895-1901.
21. Afessa B, Gajic O, Keegan M, et al. End of life issues in the least sick ICU patients. *Intensive Care Med*. 2006; 32(suppl 1):S211.
22. Afessa B, Keegan MT, Hubmayr RD, et al. Evaluating the performance of an institution using an intensive care unit benchmark. *Mayo Clin Proc*. 2005;80(2):174-180.
23. Herasevich V, Pickering BW, Dong Y, Peters SG, Gajic O. Informatics infrastructure for syndrome surveillance, decision support, reporting, and modeling of critical illness. *Mayo Clin Proc*. 2010;85(3):247-254.
24. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
25. Afessa B, Gajic O, Keegan MT. Severity of illness and organ failure assessment in adult intensive care units. *Crit Care Clin*. 2007;23(3):639-658.
26. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843.
27. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. 1982;115(1):92-106.
28. Brier G. Verification of forecasts expressed in terms of probability. *Mon Wea Rev*. 1950;78(1):1-3.
29. Wagner DP. What accounts for the difference between observed and predicted? *Crit Care Med*. 2006;34(5):1552-1553.
30. Costa e Silva VT, de Castro I, Liano F, et al. Performance of the third generation models of severity scoring systems (APACHE IV, SAPS 3 and MPM-III) in acute kidney injury critically ill patients. *Nephrol Dial Transplant*. 2011;26(12): 3894-3901.
31. Kuzniewicz MW, Vasilevskis EE, Lane R, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. *Chest*. 2008;133(6):1319-1327.
32. Vasilevskis EE, Kuzniewicz MW, Cason BA, et al. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. *Chest*. 2009;136(1):89-101.
33. Metnitz PG, Valentin A, Vesely H, et al. Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Simplified Acute Physiology Score. *Intensive Care Med*. 1999;25(2):192-197.
34. Murphy-Filkins R, Teres D, Lemeshow S, Hosmer DW. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit. *Crit Care Med*. 1996;24(12):1968-1973.
35. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997;16(9):965-980.
36. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35(9):2052-2056.
37. Angus DC, Barnato AE, Linde-Zwirble WT, et al; Robert Wood Johnson Foundation ICU End-Of-Life Peer Group. Use of intensive care at the end of life in the United States: an epidemiologic study. *Crit Care Med*. 2004;32(3):638-643.
38. Siegel MD. End-of-life decision making in the ICU. *Clin Chest Med*. 2009;30(1):181-194.
39. Seferian EG, Afessa B. Demographic and clinical variation of adult intensive care unit utilization from a geographically defined population. *Crit Care Med*. 2006;34(8):2113-2119.
40. Rubenfeld GD, Curtis JR. Improving care for patients dying in the intensive care unit. *Clin Chest Med*. 2003;24(4): 763-773.
41. Prendergast TJ, Claessens MT, Luce JM. A national survey of end-of-life care for critically ill patients. *Am J Respir Crit Care Med*. 1998;158(4):1163-1167.
42. Curtis JR, Shannon SE. Transcending the silos: toward an interdisciplinary approach to end-of-life care in the ICU. *Intensive Care Med*. 2006;32(1):15-17.
43. Murphy DJ, Burrows D, Santilli S, et al. The influence of the probability of survival on patients' preferences regarding cardiopulmonary resuscitation. *N Engl J Med*. 1994;330(8): 545-549.
44. Levy MM. End-of-life care in the intensive care unit: can we do better? *Crit Care Med.* 2001;29(suppl 2):N56-N61.
45. Cook D, Rocker G, Giacomini M, Sinuff T, Heyland D. Understanding and changing attitudes toward withdrawal and withholding of life support in the intensive care unit. *Crit Care Med*. 2006;34(suppl 11):S317-S323.
46. Nelson JE, Angus DC, Weissfeld LA, et al; Critical Care Peer Workgroup of the Promoting Excellence in End-of-Life Care Project. End-of-life care for the critically ill: a national intensive care unit survey. *Crit Care Med*. 2006; 34(10):2547-2553.
47. Rocker G, Cook D, Sjokvist P, et al; Level of Care Study Investigators; Canadian Critical Care Trials Group. Clinician predictions of intensive care unit mortality. *Crit Care Med*. 2004;32(5):1149-1154.
48. Azoulay E, Adrie C, De Lassence A, et al. Determinants of postintensive care unit mortality: a prospective multicenter study. *Crit Care Med*. 2003;31(2):428-432.
49. Jayes RL, Zimmerman JE, Wagner DP, Knaus WA. Variations in the use of do-not-resuscitate orders in ICUS. Findings from a national study. *Chest*. 1996;110(5):1332-1339.
50. Prendergast TJ, Luce JM. Increasing incidence of withholding and withdrawal of life support from the critically ill. *Am J Respir Crit Care Med*. 1997;155(1):15-20.