Clinical Research Article

Korean J Anesthesiol 2014 August 67(2): 115-122
http://dx.doi.org/10.4097/kjae.2014.67.2.115

Validation of the APACHE IV model and its comparison with the APACHE II, SAPS 3, and Korean SAPS 3 models for the prediction of hospital mortality in a Korean surgical intensive care unit

Hannah Lee, Yoon-Jung Shon, Hyerim Kim, Hyesun Paik, and Hee-Pyoung Park

Department of Anesthesiology and Pain Medicine, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Korea

Background: The Acute Physiology and Chronic Health Evaluation (APACHE) IV model has not yet been validated in Korea. The aim of this study was to compare the ability of the APACHE IV with those of APACHE II, Simplified Acute Physiology Score (SAPS) 3, and Korean SAPS 3 in predicting hospital mortality in a surgical intensive care unit (SICU) population.

Methods: We retrospectively reviewed electronic medical records for patients admitted to the SICU from March 2011 to February 2012 in a university hospital. Measurements of discrimination and calibration were performed using the area under the receiver operating characteristic curve (AUC) and the Hosmer–Lemeshow test, respectively. We calculated the standardized mortality ratio (SMR, actual mortality predicted mortality) for the four models.

Results: The study included 1,314 patients. The hospital mortality rate was 3.3%. The discriminative powers of all models were similar and very reliable. The AUCs were 0.80 for APACHE IV, 0.85 for APACHE II, 0.86 for SAPS 3, and 0.86 for Korean SAPS 3. Hosmer and Lemeshow C and H statistics showed poor calibration for all of the models ($P < 0.05$). The SMRs of APACHE IV, APACHE II, SAPS 3, and Korean SAPS 3 were 0.21, 0.11 0.23, 0.34, and 0.25, respectively.

Conclusions: The APACHE IV revealed good discrimination but poor calibration. The overall discrimination and calibration of APACHE IV were similar to those of APACHE II, SAPS 3, and Korean SAPS 3 in this study. A high level of customization is required to improve calibration in this study setting. (Korean J Anesthesiol 2014; 67: 115-122)

Key Words: Acute physiology and Chronic Health Evaluation II, Acute physiology and Chronic Health Evaluation IV, Intensive care unit, Simplified Acute Physiology Score 3, Validation.

Received: November 14, 2013. Revised: 1st, December 9, 2013; 2nd, January 24, 2014; 3rd, February 12, 2014. Accepted: February 12, 2014.
Corresponding author: Hee-Pyoung Park, M.D., Ph.D., Department of Anesthesiology and Pain Medicine, Seoul National University Hospital, Seoul National University College of Medicine, 101, Daehak-ro, Jongno-gu, Seoul 110-744, Korea. Tel: 82-2-2072-2466, Fax: 82-2-747-5639, E-mail: hppark@snu.ac.kr

Introduction

The scoring systems used widely in the field of intensive care are generic prognostic models that estimate the in-hospital mortality rate [1]. They are designed to express a patient's physical status numerically. Many clinicians utilize these systems to measure the severity of illness, predict patient prognosis, and gather information for clinical research. Since the development of the APACHE scoring system in 1981 [2], many scoring systems, such as the Simplified Acute Physiology Score (SAPS) in 1984 [3] and Mortality Probability Model in 1985 [4], have been introduced. Over time, these models were not only calibrated but also updated. The latest result of the effort is the Acute Physiology and Chronic Health Evaluation (APACHE) IV model, which was published in 2006 [5].

First presented two decades ago, APACHE II [6] and SAPS 2 [7] are old models. These are still in use because of their simplicity and easy accessibility. However, Zimmerman et al. [5] suggested that APACHE IV has better accuracy than the previous systems, and older models should not be used. The APACHE IV model showed good discrimination and calibration in the United States where the model was developed [5]. Outside the United States, recent studies have demonstrated that the discriminatory performance of APACHE IV was good [8-12]. However, its performance has not yet been validated in Korea.

The aim of the present study was 1) to validate APACHE IV and 2) to compare the ability of the APACHE IV with those of APACHE II, SAPS 3, and Korean SAPS 3 in terms of predicting hospital mortality in a Korean surgical intensive care unit (SICU) population.

Materials and Methods

The present study was approved by the Institutional Review Board, and informed consent was obtained. The current retrospective study was conducted at the SICU of a 1200-bed university teaching hospital. The SICU, which comprised 32 beds, was managed using an open system.

Patient population

All patients admitted to the SICU from March 2011 to February 2012 were included in the present study. Patients who underwent vascular surgery, lung surgery, neurosurgery, orthopedic surgery, and general surgery were the main patient group. In addition, patients with serious medical or surgical postoperative complications admitted to the SICU were included. Pediatric patients (< 18 years of age), cardiac patients and medical department patients were excluded. We also excluded patients with an SICU stay < 24 h or a hospital stay > 365 days and those who were readmitted after initial ICU discharge. Patients who were cadaveric donors were also excluded from the main analysis.

Data collection

Two senior residents and one fellow retrospectively reviewed the electronic medical records. The electronic medical records provided all of the data required to predict the mortality rate using the APACHE IV, APACHE II, SAPS 3, and Korean SAPS 3 models. APACHE IV and APACHE II scores were derived from the worst laboratory findings obtained within 24 h after admission, and SAPS III scores were obtained from the worst laboratory findings 1 h after SICU admission. Predicted hospital mortalities were calculated using equations of each model, as follows: logit for APACHE II = −3.517 + (APACHE II) × 0.146; logit for SAPS 3 = −32.6659 + ln (SAPS 3 + 20.5958) × 7.3068; logit for Korean SAPS 3 = −35.1752 + ln (SAPS3 + 20.5958) × 7.7379; and the predicted mortality rate = $e^{Logit} / (1 + e^{Logit})$ [6,13]. The APACHE is a registered trademark of Cerner Corporation (Kansas City, MO, USA). The APACHE IV score and predicted mortality rate calculation on the website (http://www.mecriticalcare.net/icu_scores/apacheIV.php) was used in the present study. The performance of each model was evaluated both in total patients and in two subgroups of patients that were divided into the admission-type subgroup and admission diagnoses subgroup.

Definitions

To validate each prognostic model, discrimination and calibration were performed. Discrimination is defined as the ability of the model to separate survivors from non-survivors and is assessed using the area under the receiver operating characteristic curve (AUC) [14]. It is classified as excellent, very good, good, moderate, or poor according to the AUC values of 0.9 to 0.99, 0.8 to 0.89, 0.7 to 0.79, 0.6 to 0.69, and < 0.6, respectively [15,16]. A prognostic model with a high AUC suggests that the model can accurately predict the probability of death. Calibration is defined as the ability of a model to describe the mortality pattern in the data and is assessed using the Hosmer–Lemeshow goodness-of-fit test [17]. When the predicted mortality of the prognostic model differs significantly from the observed pattern, the calibration ability of this model is poor, and goodness-of-fit statistics are significant. The Hosmer–Lemeshow goodness-of-fit test evaluates the agreement between the observed and expected numbers of survivors and non-survivors across all of the strata with equal number of patients (C-statistics) or with 10 groups divided by expected mortality intervals (H statistics) [17]. The Brier score is assessed as a measure of overall model accuracy, involving elements of both discrimination and calibration.

It measures the average squared difference between predicted probabilities of outcomes [18,19]. A lower score represents higher accuracy. The standardized mortality ratio (SMR) is the ratio between the observed and predicted number of deaths. A SMR equal to 1.0 indicates that the number of observed mortality equals that of predicted mortality.

Statistical analysis

Statistical analysis was performed using SPSS version 19.0 for Windows (SPSS, Inc., Chicago, IL, USA). Data were reported as means ± standard deviation (SD) or medians with 25th and 75th quartiles for continuous variables and percentages for quantitative variables. Student's t-test, chi-squared test or Fisher's exact test were used, depending on whether the variables were continuous or categorical. P values less than 0.05 were deemed to indicate statistical significance. We used the AUC to measure the four models' discrimination for hospital mortality. Calibration was assessed using the Hosmer–Lemeshow goodness-of-fit C statistics, with a P value greater than 0.05 indicating good calibration [20]. The Brier score and SMR were also calculated.

Results

Characteristics of the study population

There were 2,952 admissions to our SICU during the study period. Of those, 1,314 patients comprised our final sample (Fig. 1). The basic patient characteristics and outcomes are shown in Table 1. General surgery included mainly liver transplantations, colorectal surgeries and stomach surgeries. Thoracic surgery included operations on the lung and esophagus. Neurosurgery included mainly brain tumor removals and intracranial or subarachnoid hemorrhage evacuations. Obstetric-gynecologic patients underwent mainly bleeding control after delivery and debulking operation for ovarian cancer. Orthopedic surgery included total knee replacement arthroplasties, total hip surgeries, and scoliosis correction operations.

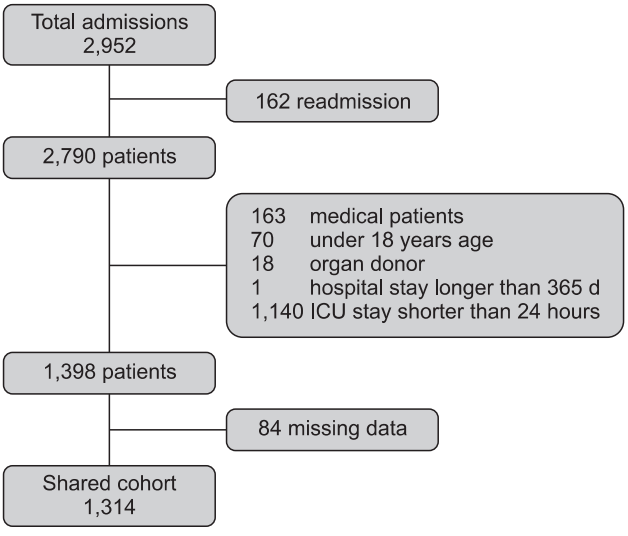

Fig. 1. Study population. ICU: intensive care unit.

Forty-three patients (3.3%) in the study population expired. They were in a poorer condition at the time of SICU admission than those who survived (P < 0.001). Non-survivors stayed longer in the ICU than survivors (P = 0.004; Table 1). Significant differences were found in the admission routes and surgery types between the survivor and non- survivor groups (P < 0.001, both).

Validation of the APACHE IV model

The APACHE IV model showed good discrimination and accuracy (AUC = 0.80; Brier score = 0.06) but poor calibration (C-statistics = 220.33; P < 0.001, Table 2). The model significantly overestimated the observed mortality (SMR = 0.21). The performance of the APACHE IV model varied among subgroups of admission types and admission diagnoses. Patients who received stomach cancer surgery showed good calibration (C-statistics = 11.51; H-statistics = 10.30; P > 0.05), whereas patients who had other surgeries showed poor calibration. All subgroups of admission types showed moderate discrimination and poor calibration.

Comparison of the performance of the APACHE IV and other prognostic models

Hosmer–Lemeshow statistics showed poor calibration for all four models: APACHE IV, APACHE II, SAPS 3, and Korean SAPS 3 (P < 0.001; Table 2; Figs. 2 and 3). Discrimination, as measured by the AUC, was generally very good for all models (AUCs > 0.8; Fig. 4). AUCs of all models showed no statistically significant difference (P > 0.05). In the subgroups of thoracotomy for lung cancer and stomach surgery, SAPS 3 and Korean SAPS 3 showed good calibration (P > 0.05) but poor discrimination (AUC < 0.6; Table 2). The APACHE II score showed poor calibration in all subgroups. Brier scores of the APACHE IV, SAPS 3, and Korean SAPS 3 models were significantly better than that of the APACHE II model. All prognostic models significantly overestimated the observed mortality (SMR < 1.0). The no-surgery subgroup showed less overestimation of the observed mortality rate in all of the models.

Table 1. Patient Characteristics and Scores and Predicted Mortality on each Prognostic Model

	Total (n = 1,314)	Survivor (n = 1,271)	Non-survivor (n = 43)	P value
Age (yr)	57.8 ± 15.3	57.2 ± 15.3	61.2 ± 14.3	0.090
Male gender	759 (57.8)	731 (57.5)	28 (65.1)	0.350
Body mass index (kg/m^2)	23.3 ± 3.8	23.3 ± 3.8	22.8 ± 3.9	0.401
Main comorbidities				
Hypertension	460 (35.0)	445 (30.3)	15 (34.9)	0.611
Diabetes mellitus	245 (18.6)	233 (18.3)	12 (27.9)	0.327
Heart failure	21 (1.6)	20 (1.6)	1 (2.3)	0.563
Chronic liver disease	198 (15.1)	187 (14.7)	11 (25.6)	0.138
End stage kidney disease	53 (4.0)	53 (4.2)	0 (0.0)	0.255
Route of admission				< 0.001
OR/recovery room	1,193 (90.8)	1,167 (91.8)	26 (60.5)	
Emergency room	64 (4.9)	42 (3.3)	6 (1.4)	
Ward	48 (3.7)	55 (4.3)	9 (20.9)	
Other ICU	9 (0.7)	7 (0.5)	2 (4.7)	
Admission type				< 0.001
Elective surgery	928 (70.6)	916 (72.1)	12 (27.9)	
Emergency surgery	265 (20.2)	251 (19.7)	14 (32.6)	
No surgery	121 (9.2)	104 (8.2)	17 (39.5)	
Department				0.269
General surgery	448 (34.1)	429 (33.8)	19 (44.2)	
Neurosurgery	516 (39.3)	497 (39.1)	19 (44.2)	
Thoracic surgery (non-cardiac)	253 (19.3)	250 (19.7)	3 (7.0)	
Orthopedic surgery	45 (3.4)	44 (3.5)	1 (2.3)	
Obstetrics-gynecology	21 (1.6)	20 (1.6)	1 (2.3)	
Urology	14 (1.1)	14 (1.1)	0 (0.0)	
Otolaryngology	13 (1.0)	13 (1.0)	0 (0.0)	
Plastic surgery	4 (0.3)	4 (0.3)	0 (0.0)	
Admission diagnoses*				< 0.001
Craniotomy for brain neoplasm	194 (14.8)	192 (15.1)	2 (4.7)	
Intracranial hemorrhage surgery	64 (4.9)	53 (4.2)	11 (25.6)	
Thoracotomy for lung cancer	191 (14.5)	190 (15.0)	1 (2.3)	
Stomach cancer surgery	50 (3.8)	49 (3.9)	1 (2.3)	
Colorectal cancer surgery	45 (3.4)	44 (3.5)	1 (2.3)	
Liver transplantation (postoperative)	171 (13.0)	165 (13.0)	6 (14.0)	
DNR status	27 (2.1)	11 (0.9)	16 (37.2)	< 0.001
APACHE IV score	50.0 ± 22.9	49.0 ± 22.2	77.1 ± 26.2	< 0.001
APACHE IV predicted mortality	15.6 ± 17.6	15.0 ± 16.9	36.5 ± 24.6	< 0.001
APACHE II score	16.9 ± 6.8	16.6 ± 6.6	26.1 ± 6.9	< 0.001
APACHE II predicted mortality	28.7 ± 19.0	27.7 ± 18.2	55.8 ± 20.5	< 0.001
SAPS 3 score	42.5 ± 14.9	41.7 ± 14.0	67.4 ± 18.8	< 0.001
SAPS 3 predicted mortality	14.3 ± 19.3	13.1 ± 17.7	50.1 ± 27.8	< 0.001
Korean SAPS 3 predicted mortality	9.6 ± 15.9	8.6 ± 14.3	40.4 ± 26.9	< 0.001
ICU length of stay	2.1 (1.2–4.2)	2.1 (1.2–4.0)	7.0 (3.7–13.5)	0.004
Hospital length of stay after ICU discharge	14.0 (9.0–25.0)	14.0 (9.0–25.0)	15.0 (6.5–48.5)	0.139

Data are expressed as mean ± SD, median (interquartile range), or number (percent). OR: operating room, ICU: intensive care unit, DNR: do not resuscitate, APACHE: acute physiology and chronic health evaluation, SAPS: simplified acute physiology score. *Subpopulations based on admission diagnoses in the APACHE IV model.

Discussion

Our study demonstrated that the APACHE IV, APACHE II, SAPS 3, and Korean SAPS 3 models showed good discrimination but poor calibration. Additionally, these four models overestimated the observed mortality rate.

External validation is necessary before implementing prediction models in countries other than that in which the prognostic model was first developed [21,22]. To our knowledge, this is the first study to validate the APACHE IV model and compare it with other prognostic models in a Korean ICU. Since the APACHE IV system was developed in the United States in 2006, it has been implemented worldwide and applied to general ICUs and specific patient groups [5,8-10]. A major advantage of the APACHE IV model is its ability to select 116 detailed admitting diagnostic options, which promote outcome analysis in

Table 2. Performance of APACHE IV, APACHE II, SAPS 3, and Korean SAPS 3 Models on Prediction of Hospital Mortality

Model	AUC (95% CI)	Hosmer-lemeshow goodness-of fit test				Brier score	SMR (95% CI)
		C-test	P value	H-test	P value		
APACHE IV	0.80 (0.74–0.86)	220.33	< 0.001	252.77	< 0.001	0.06	0.21 (0.15–0.28)
Admission diagnosis*							
Craniotomy for brain neoplasm	0.64 (0.17–1.12)	18.48	0.018	18.75	0.016	0.02	0.12 (0.01–0.42)
Intracranial hemorrhage surgery	0.78 (0.65–0.90)	7.39	0.495	21.41	0.006	0.14	0.82 (0.41–1.47)
Thoracotomy for lung cancer	0.84 (0.79–0.89)	21.87	0.005	23.98	0.002	0.02	0.05 (0.00–0.29)
Stomach cancer surgery	0.29 (0.16–0.41)	11.51	0.175	10.30	0.244	0.05	0.15 (0.00–0.86)
Colorectal cancer surgery	0.71 (0.57–0.84)	28.94	< 0.001	35.53	< 0.001	0.14	0.08 (0.00–0.45)
Liver transplantation (postoperative)	0.81 (0.68–0.94)	74.37	< 0.001	77.54	< 0.001	0.12	0.13 (0.05–0.28)
Admission type							
Elective surgery	0.77 (0.63–0.9)	108.83	< 0.001	113.85	< 0.001	0.03	0.12 (0.06–0.20)
Emergency surgery	0.73 (0.62–0.83)	105.21	< 0.001	117.12	< 0.001	0.12	0.22 (0.12–0.36)
No surgery	0.62 (0.49–0.76)	47.58	< 0.001	41.71	< 0.001	0.18	0.46 (0.27–0.73)
APACHE II	0.85 (0.80–0.90)	560.94	< 0.001	621.38	< 0.001	0.11	0.11 (0.08–0.15)
Admission diagnosis*							
Craniotomy for brain neoplasm	0.89 (0.84–0.93)	58.70	< 0.001	64.01	< 0.001	0.07	0.05 (0.01–0.17)
Intracranial hemorrhage surgery	0.89 (0.82–0.97)	18.95	0.015	19.45	0.013	0.16	0.43 (0.21–0.77)
Thoracotomy for lung cancer	0.87 (0.82–0.91)	58.80	<0.001	75.85	< 0.001	0.07	0.02 (0.00–0.13)
Stomach cancer surgery	0.51 (0.37–0.65)	19.20	0.014	19.01	0.015	0.10	0.07 (0.00–0.41)
Colorectal cancer surgery	0.46 (0.31–0.60)	56.73	< 0.001	56.99	< 0.001	0.21	0.06 (0.00–0.32)
Liver transplantation (postoperative)	0.92 (0.85–0.98)	104.49	< 0.001	110.34	< 0.001	0.15	0.10 (0.04–0.21)
Admission type							
Elective surgery	0.81 (0.69–0.94)	304.74	< 0.001	334.63	< 0.001	0.08	0.05 (0.03–0.10)
Emergency surgery	0.83 (0.74–0.91)	202.96	< 0.001	200.72	< 0.001	0.20	0.13 (0.07–0.22)
No surgery	0.71 (0.61–0.82)	87.66	< 0.001	98.89	< 0.001	0.22	0.34 (0.20–0.54)
SAPS 3	0.86 (0.79–0.92)	202.85	< 0.001	226.64	< 0.001	0.06	0.23 (0.17–0.31)
Admission diagnosis*							
Craniotomy for brain neoplasm	0.92 (0.89–0.96)	15.09	0.035	22.64	0.004	0.03	0.13 (0.02–0.46)
Intracranial hemorrhage surgery	0.77 (0.60–0.94)	19.25	0.014	21.87	0.005	0.15	0.56 (0.28–1.01)
Thoracotomy for lung cancer	0.53 (0.47–0.60)	10.77	0.096	10.32	0.243	0.01	0.10 (0.00–0.57)
Stomach cancer surgery	0.41 (0.28–0.54)	6.81	0.558	4.41	0.818	0.03	0.24 (0.01–1.34)
Colorectal cancer surgery	0.36 (0.22–0.51)	28.98	< 0.001	26.97	< 0.001	0.13	0.10 (0.00–0.54)
Liver transplantation (postoperative)	0.93 (0.88–0.98)	59.11	< 0.001	65.32	< 0.001	0.09	0.15 (0.05–0.32)
Admission type			< 0.001		< 0.001		
Elective surgery	0.80 (0.64–0.96)	62.88	< 0.001	67.37	< 0.001	0.02	0.17 (0.09–0.29)
Emergency surgery	0.76 (0.62–0.91)	109.92	< 0.001	117.09	< 0.001	0.13	0.19 (0.11–0.32)
No surgery	0.79 (0.69–0.89)	56.15	< 0.001	61.14	< 0.001	0.18	0.38 (0.22–0.61)
Korean SAPS 3	0.86 (0.79–0.92)	89.67	< 0.001	100.26	< 0.001	0.04	0.34 (0.25–0.46)
Admission diagnosis*							
Craniotomy for brain neoplasm	0.92 (0.89–0.96)	6.83	0.556	11.86	0.158	0.02	0.22 (0.03–0.78)
Intracranial hemorrhage surgery	0.77 (0.60–0.94)	26.83	0.001	21.60	0.006	0.13	0.77 (0.38–1.37)
Thoracotomy for lung cancer	0.54 (0.48–0.61)	5.89	0.659	4.48	0.811	0.01	0.19 (0.00–1.06)
Stomach cancer surgery	0.42 (0.29–0.55)	8.41	0.394	2.87	0.942	0.03	0.43 (0.01–2.41)
Colorectal cancer surgery	0.36 (0.22–0.51)	24.46	0.002	15.57	0.049	0.09	0.14 (0.00–0.77)
Liver transplantation (postoperative)	0.93 (0.88–0.98)	30.12	< 0.001	34.10	< 0.001	0.06	0.21 (0.08–0.47)
Admission type							
Elective surgery	0.81 (0.66–0.96)	26.98	< 0.001	29.41	< 0.001	0.01	0.29 (0.15–0.51)
Emergency surgery	0.76 (0.62–0.91)	56.79	< 0.001	57.89	< 0.001	0.09	0.27 (0.15–0.46)
No surgery	0.79 (0.69–0.89)	26.88	< 0.001	30.66	< 0.001	0.14	0.50 (0.29–0.80)

APACHE: acute physiology and chronic health evaluation, SAPS: simplified acute physiology score, AUC: the area under the receiver operating characteristic curve, SMR: standardized mortality ratio. *Subpopulations based on admission diagnoses in the APACHE IV model.

specific subgroups [5]. Conversely, the large number of variables requires a relatively longer time for data abstraction [23]. In the current study, the APACHE IV model had very good discrimination and accuracy (AUC = 0.82; Brier score = 0.05) but poor calibration (C-statistics = 309.27). Such findings are supported by the results of a previous study that performed external validation of APACHE IV, in which APACHE IV showed very good discrimination and accuracy (AUC = 0.87; Brier score = 0.10) but poor calibration (C-statistic = 822.67) in Dutch ICUs [8]. A previous study conducted in American ICUs showed very good

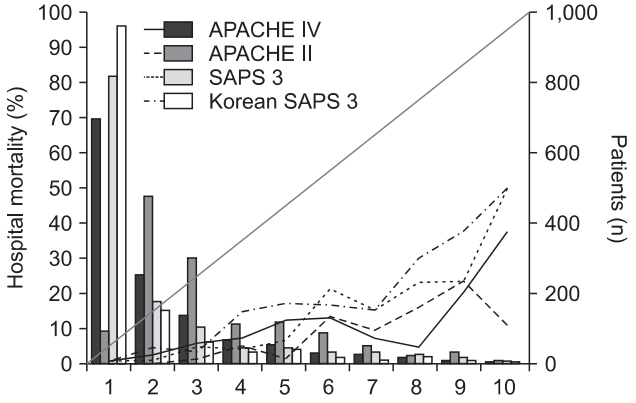

Fig. 2. Calibration plots of four different models (H-statistics). X-axis represents a decile predicted mortality of 10 groups on each model. The diagonal line indicates ideal prediction for hospital mortality. The bars represent the number of patients. APACHE: acute physiology and chronic health evaluation, SAPS: simplified acute physiology score.

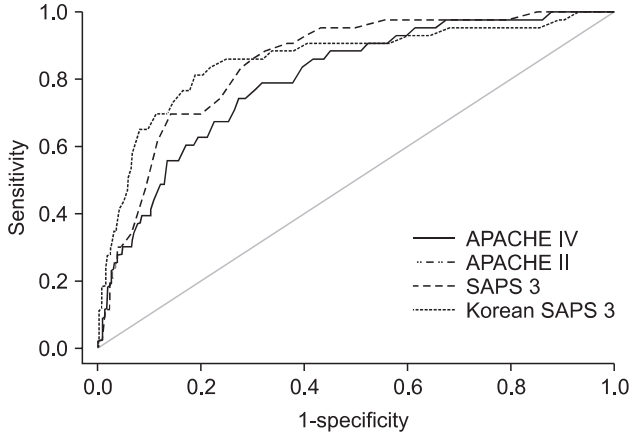

Fig. 4. Comparison of the area under the receiver operating characteristic curves (AUC) of APACHE IV, APACHE II, SAPS 3, and Korean SAPS 3. The AUCs are 0.80, 0.85, 0.86, and 0.86 in APACHE IV, APACHE II, SAPS 3, Korean SAPS3 models, respectively. APACHE: acute physiology and chronic health evaluation, SAPS: simplified acute physiology score.

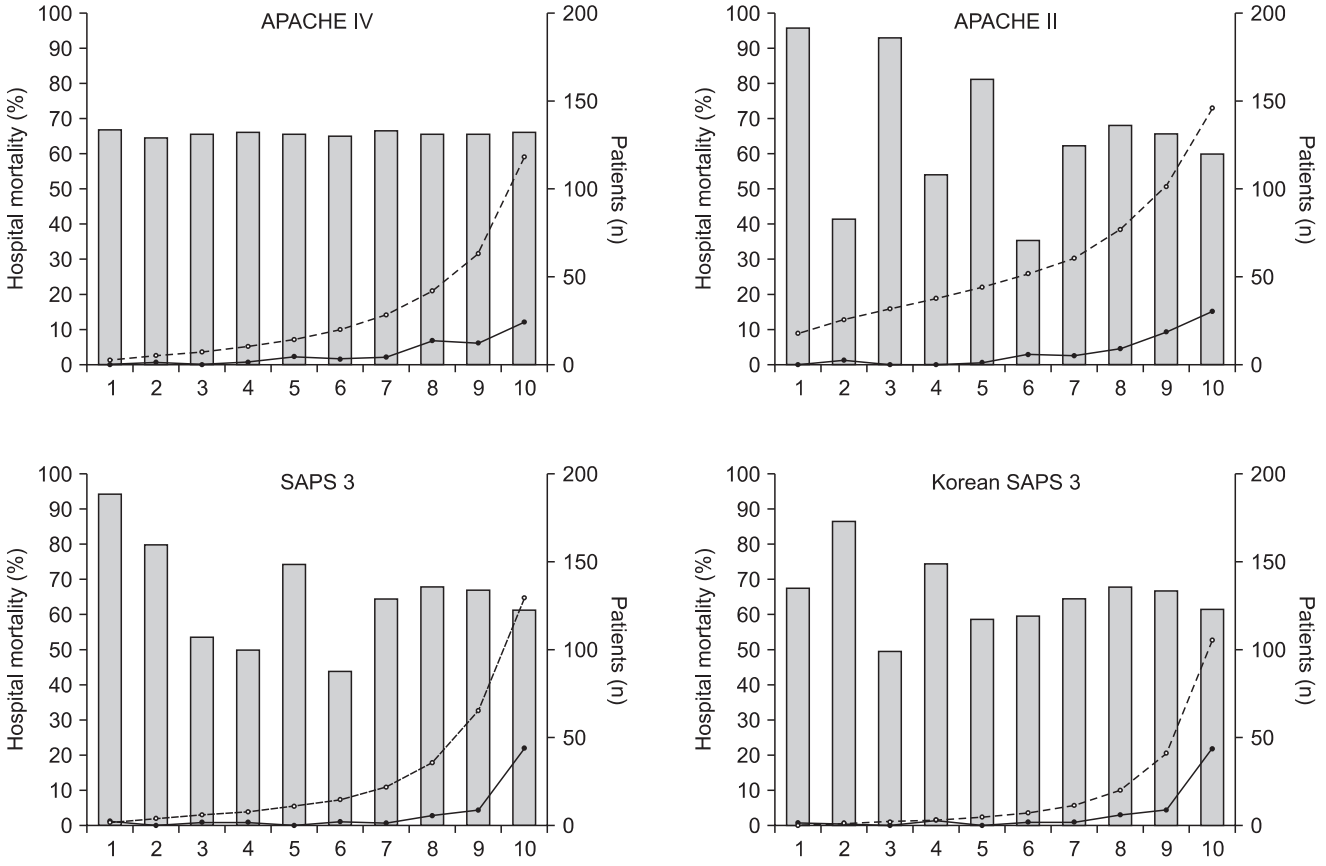

Fig. 3. Calibration plots of four different models (C-statistics). X-axis represents an approximate decile patient of 10 groups. The bars represent the number of patients. The line with open circles represents mean predicted mortality. The line with closed circles represents mean observed mortality. APACHE: acute physiology and chronic health evaluation, SAPS: simplified acute physiology score.

discriminatory power (AUC = 0.86) of APACHE IV [9].

The SAPS 3 model, the latest version of the SAPS system, was developed in 2005 [24]. One of the advantages of this model is its short data collection window (1 h), which may be useful in triage and could save the abstraction time. This model was assessed frequently in worldwide ICUs showing good discrimi-

nation but poor calibration [10,12,25-27]. Similar to our study, other studies revealed that SAPS 3 and Australasia SAPS 3 had good discrimination but poor calibration, a feature that was improved after customization. In Korean ICUs, SAPS 3 also has been validated and customized [28]. In the present study, the Korean SAPS 3 demonstrated good discrimination but poor calibration. Such findings can be explained by the difference in the proportions of elective surgical patients and admission diagnoses between the customization cohort of Korean SAPS 3 and our study population.

Previous studies concerning external validation of other various prediction models showed patterns similar to those reported here, with good discrimination but imperfect calibration [8,25-27]. Nassar et al. suggested that the pattern can be explained by differences in the study population, regional variability of end-of-life decisions, and temporal bias—the time interval between the development of the prognostic models and study enrollment [29]. Other studies have suggested disparities in case presentation, mortality rates among countries, and differences in sample size between the study population and the original cohort used in the development of the scoring systems as explanations for those differing patterns [8,20,30].

Of previous studies comparing the APACHE IV and SAPS 3 prognostic models, two regarding general ICU patients showed that the APACHE IV model had better discriminatory capability than SAPS 3 [9,29], but other studies for acute kidney injury and acute coronary artery syndrome patients showed that the two models have similar discriminatory performance [10,11]. Additionally, one study reported that the discrimination and calibration of APACHE II are similar to those of SAPS 3 [26]. Our study also showed that the discrimination and calibration of APACHE II were similar to those of SAPS 3 and APACHE IV, but APACHE II significantly overestimated hospital mortality than APACHE IV and SAPS 3.

All of the prognostic models in our study overestimated mortality. The observed mortality rate was 3.3% in the present study, whereas the mortality rates of each prognostic model were 15.6, 28.7, 14.3, and 9.6% for APACHE IV, APACHE II, SAPS 3, and Korean SAPS 3, respectively. Additionally, the proportion of elective surgical patients was 70.6% in the current study, whereas the portion was 30.9% for APACHE IV and 34.7% for SAPS 3 when each prognostic model was applied. We assumed that the low mortality rate and higher proportion of elective surgical patients contributed to the difference between the actual mortality and predicted mortality derived from each prognostic model. Therefore, our findings suggest that customization of each prognostic model should be required when the model is applied to a different ICU because differences in mortality and study population, type of ICUs, and other ICU environmental factors contribute to the discrepancy between the actual mortality and predicted mortality derived from each prognostic model.

Among the admission diagnoses subgroup, thoracotomy for lung cancer showed poor discrimination in SAPS 3 and Korean SAPS 3. A previous study reported that the SMR is lowest in patients with thoracotomy for lung cancer [8]. Such a finding suggests that other factors, such as size and location of the cancer, not included in the prognostic model, might influence the patient outcome. Elective surgery patients had the lowest SMR, whereas no surgery patient had a relatively higher SMR in our study. The prior study suggested that, for ICUs, which have higher proportions of elective surgical patients, a higher level of customization could be considered [8].

One limitation of the present study is that it was conducted at a single surgical ICU, limiting the ability to generalize our results to other ICUs because admission diagnoses, patient populations such as medical or medico-surgical patients, and other environments are diverse. Although our data included 1,314 patients, the overall hospital mortality rate was very low, which might have affected the performance of each prognostic model. Another limitation is associated with retrospective data collection. Although selected residents were trained to collect the data, such retrospective data abstraction carries a risk of error. Finally, the equation in each prognostic model for the prediction of hospital mortality was not customized in the current study.

In summary, the discriminatory performance of the APACHE IV model was very good and similar to those of the APACHE II, SAPS 3, and Korean SAPS 3 models. All of the models, however, showed poor calibration, although some subgroups with a relatively high mortality rate showed good calibration. To improve the calibration performance, all of the original prognostic models in the present study setting should be customized.

Acknowledgments

We express our gratitude to J.W. Park in Medical Research Collaborating Center for assisting with data analysis.

References

1. Gunning K, Rowan K. ABC of intensive care: outcome data and scoring systems. BMJ 1999; 319: 241-4.
2. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981; 9: 591-7.

3. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. Crit Care Med 1984; 12: 975-7.
4. Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit Care Med 1985; 13: 519-25.
5. 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: 1297-310.
6. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985; 13: 818-29.
7. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993; 270: 2957-63.
8. Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, de Jonge E, Bosman RJ, Peelen L, et al. External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. J Crit Care 2011; 26: 105.e11-8.
9. Keegan MT, Gajic O, Afessa B. Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. Chest 2012; 142: 851-8.
10. Costa e Silva VT, de Castro I, Liaño F, Muriel A, Rodríguez-Palomares JR, Yu L. 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: 3894-901.
11. Nassar Junior AP, Mocelin AO, Andrade FM, Brauer L, Giannini FP, Nunes AL, et al. SAPS 3, APACHE IV or GRACE: which score to choose for acute coronary syndrome patients in intensive care units? Sao Paulo Med J 2013; 131: 173-8.
12. Soares M, Silva UV, Teles JM, Silva E, Caruso P, Lobo SM, et al. Validation of four prognostic scores in patients with cancer admitted to Brazilian intensive care units: results from a prospective multicenter study. Intensive Care Med 2010; 36: 1188-95.
13. Metnitz PG, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, et al. 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: 1336-44.
14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29-36.
15. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148: 839-43.
16. Afessa B, Gajic O, Keegan MT. Severity of illness and organ failure assessment in adult intensive care units. Crit Care Clin 2007; 23: 639-58.
17. 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: 92-106.
18. Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. Methods Inf Med 1978; 17: 238-46.
19. Wagner DP. What accounts for the difference between observed and predicted? Crit Care Med 2006; 34: 1552-3.
20. 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: 965-80.
21. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003; 56: 826-32.
22. Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. J Clin Epidemiol 2003; 56: 721-9.
23. Kuzniewicz MW, Vasilevskis EE, Lane R, Dean ML, Trivedi NG, Rennie DJ, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 2008; 133: 1319-27.
24. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, et al. 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: 1345-55.
25. Khwannimit B, Bhurayanontachai R. The performance and customization of SAPS 3 admission score in a Thai medical intensive care unit. Intensive Care Med 2010; 36: 342-6.
26. Sakr Y, Krauss C, Amaral AC, Réa-Neto A, Specht M, Reinhart K, et al. Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit. Br J Anaesth 2008; 101: 798-803.
27. Poole D, Rossi C, Anghileri A, Giardino M, Latronico N, Radrizzani D, et al. External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units. Intensive Care Med 2009; 35: 1916-24.
28. Lim SY, Koh SO, Jeon K, Na S, Lim CM, Choi WI, et al. Validation of SAPS3 admission score and its customization for use in Korean intensive care unit patients: a prospective multicentre study. Respirology 2013; 18: 989-95.
29. Nassar AP Jr, Mocelin AO, Nunes AL, Giannini FP, Brauer L, Andrade FM, et al. Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. J Crit Care 2012; 27: 423.e1-7.
30. 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: 1968-73.