Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom*

David A. Harrison, PhD; Anthony R. Brady, MSc; Gareth J. Parry, PhD; James R. Carpenter, DPhil; Kathy Rowan, DPhil

Objective: **To assess the performance of published risk prediction models in common use in adult critical care in the United Kingdom and to recalibrate these models in a large representative database of critical care admissions.**

Design: **Prospective cohort study.**

Setting: **A total of 163 adult general critical care units in England, Wales, and Northern Ireland, during the period of December 1995 to August 2003.**

Patients: **A total of 231,930 admissions, of which 141,106 met inclusion criteria and had sufficient data recorded for all risk prediction models.**

Interventions: **None.**

Measurements and Main Results: **The published versions of the Acute Physiology and Chronic Health Evaluation (APACHE) II, APACHE II UK, APACHE III, Simplified Acute Physiology Score (SAPS) II, and Mortality Probability Models (MPM) II were evaluated for discrimination and calibration by means of a combination of appropriate statistical measures recommended by an expert steering committee. All models showed good discrimination (the *c* index varied from 0.803 to 0.832) but imperfect calibration. Recalibration of the models, which was performed by both the Cox method and re-estimating coefficients, led to improved discrimination and calibration, although all models still showed significant departures from perfect calibration.**

Conclusions: **Risk prediction models developed in another country require validation and recalibration before being used to provide risk-adjusted outcomes within a new country setting. Periodic reassessment is beneficial to ensure calibration is maintained. (Crit Care Med 2006; 34:1378–1388)**

KEY WORDS: critical care; intensive care units; models; statistical; risk adjustment; severity of illness index; validation studies

Intensive care has developed over the past 50 yrs with little rigorous scientific evidence to guide those involved in intensive care on which patients benefit most, which treatments and procedures are best, and the optimal way to organize and deliver services. Randomized controlled trials (RCTs) are considered the "gold standard" design for detecting important effects of interventions and for cost-benefit analyses; however, there are questions that preclude evaluation by RCTs for ethical, logistic, or cost reasons (1). Where randomization is impractical, the optimal research design is prospective studies of outcome adjusting for variations in risk factors between the groups compared (2).

Current published risk prediction models proposed for use in adult intensive care are the Acute Physiology and Chronic Health Evaluation (APACHE) II (3, 4), APACHE III (5), Simplified Acute Physiology Score (SAPS) II (6), and Mortality Probability Models (MPM) II (7). All are based on physiologic data obtained after the patient has been admitted to intensive care and employ logistic regression techniques to provide a probability of hospital mortality.

The largest United Kingdom (UK)–based studies assessing and comparing these models to date have taken place in 22 critical care units in Scotland (8) and 17 critical care units in Southern England (9). Both studies showed that existing models were poorly calibrated, and the investigators recommended recalibration. Other studies have been set in single units or small numbers of units and have also been limited by a lack of both standardized data collection and standardized application of the models (10). Evaluation criteria are frequently based around the area under the receiver operating characteristic (ROC) curve (11) and the Hosmer-Lemeshow goodness-of-fit test (12), which may not be sufficient to guide decisions as to which model is optimal (13–15).

Consequently, there is a case for a substantial, prospective study to evaluate the established risk prediction models, which we report here. The ultimate aim was to identify a risk prediction model that will form the basis of the Intensive Care National Audit & Research Centre (ICNARC) Case Mix Programme (CMP), a national comparative audit of patient outcome in UK critical care units (16). This study sought to assess the proposed risk prediction models applied to a standardized database before and after recalibration, using robust measures of model performance appropriate for use in a large dataset.

***See also p. 1552.**

From the Intensive Care National Audit & Research Centre (DAH, ARB, KR), London; Health Services Research, Acute & Critical Care Research Group and School of Health & Related Research (GJP), University of Sheffield, Sheffield; and Medical Statistics Unit, London School of Hygiene & Tropical Medicine (JRC), London, UK.

Supported in part by the Medical Research Council, London, UK (G9813469).

All work was completed at the Intensive Care National Audit & Research Centre, London, UK. The authors do not have any financial interests to disclose.

DOI: 10.1097/01.CCM.0000216702.94014.75

MATERIALS AND METHODS

Data

The CMP Database (CMPD) contains data on the case mix, outcome, and activity for consecutive admissions to participating adult general critical care units. Data are collected prospectively, are abstracted by trained data collectors according to precise rules and definitions (17), and undergo extensive validation for completeness, illogicalities, and inconsistencies (18). The database contains sufficient raw physiologic data to calculate the APACHE II, APACHE III, SAPS II, and MPM II scores and probabilities. Some units elected not to collect the data for APACHE III, SAPS II, and/or MPM II. Scores and probabilities were calculated centrally with use of standardized algorithms based on the original publications and refined in consultation with the developers of the original methods. Missing physiologic data were assumed to take normal values.

Approval of the study by an institutional review board was not required. The CMP has received approval from the Patient Information Advisory Group (PIAG) to hold patient-identifiable information without consent (approval number PIAG 2–10(f)/2005).

Measures of Model Performance. An expert statistical steering committee advised on the best methods for assessing model performance (see Acknowledgements). Both calibration and discrimination were investigated for each model. A poorly calibrated model will erroneously predict that mortality is higher or lower than observed, and if the calibration differs according to predicted risk, then attempts to compare centers by ratios of observed to expected deaths will be biased. Discrimination of a model describes its ability to predict events from nonevents. A model with better discrimination will be a more powerful risk-adjustment tool. Model performance was assessed with use of the *c* index, Shapiro's *R*, Brier's score, Hosmer-Lemeshow goodness-of-fit, and Cox's calibration regression (Box 1) (19–25).

Box 1. Measures of model performance

c index

The *c* index (19) provides an indication of how well the model can discriminate between patients who die and patients who survive:

- Perfect discrimination, $c = 1$.
- Poor or no discrimination, $c = 0.5$.

The *c* index is the probability of concordance between outcomes and predictions. For binary outcomes (e.g., death), this is the probability that a randomly chosen individual with the event (a nonsurvivor) will have a higher predicted probability than a randomly chosen individual without the event (a survivor). This has been shown to be identical to the area under the receiver operating characteristics (ROC) curve (11, 20).

Shapiro's R

Shapiro's *R*, based on Shapiro's *Q* (21), is an overall measure of the accuracy of the model, reflecting both calibration and discrimination:

- Perfect prediction, $R = 1$.
- Poor prediction, when a constant of 0.5 is assigned to every individual, $R = 0.5$.

R is the geometric mean of the probability assigned to the event that occurred.

Brier's score

Brier's score, *B* (22), was developed in relation to meteorological forecasts; it is an overall measure of the accuracy of predictions:

- Perfect prediction, $B = 0$.
- Poor prediction, when a constant of 0.5 is assigned to every individual, $B = 0.25$.

B is the mean square error between outcomes and predictions.

Spiegelhalter's Z-statistic

Spiegelhalter's Z-statistic, z_c, is a normal test statistic derived from Brier's score to test for perfect calibration (23). Values above 1.64 indicate statistically significant departures from perfect calibration, with $p < .05$.

Accuracy of the average prediction

The accuracy of the average prediction, $(\bar{Y} - \bar{p})^2$, is the squared difference between the overall predicted mortality, $\bar{p}$, and the overall observed mortality, $\bar{Y}$:

- Perfect accuracy, $(\bar{Y} - \bar{p})^2 = 0$.

Excess variance of predictions

The excess variance of predictions, V_{exc}/V_{min}, represents the degree of unnecessary variation in the predictions:

- Perfect predictions, $V_{exc}/V_{min} = 0$.

The excess variance of predictions is calculated by decomposing the total variance of the predictions, $V(\mathbf{p})$, as $V_{min} + V_{exc}$, where V_{min} represents the minimum variance possible for predictions that would be just as good as the actual predictions, and V_{exc} the excess variance of the predictions above this minimum.

Covariance of outcome and prediction

The covariance of outcomes and predictions, $\mathrm{Cov}(\mathbf{Y}, \mathbf{p})$, is a measure of how accurately the predictions correspond to the outcomes:

- Perfect predictions, $\mathrm{Cov}(\mathbf{Y}, \mathbf{p}) = V(\mathbf{Y})$, the variance of the observed outcome.

The accuracy of the average prediction, excess variance of predictions, and covariance of outcome and prediction, all derive from a decomposition of Brier's score (24) as:

$$B = V(\mathbf{Y}) + (\bar{Y} - \bar{p})^2 + V_{min} + V_{exc} - 2\,\mathrm{Cov}(\mathbf{Y}, \mathbf{p}).$$

Hosmer-Lemeshow goodness-of-fit

Hosmer and Lemeshow proposed two goodness-of-fit statistics for binary outcome data, $\hat{C}_g$ and $\hat{H}_g$ (12), representing chi-squared test statistics for perfect calibration. Observations are grouped into *g* (typically 10) groups based on either quantiles of predicted probability ($\hat{C}_g$) or equally spaced cut-points ($\hat{H}_g$), and the observed and expected outcomes are compared. The tests are highly sensitive to sample size and can be directly compared only within the same sample.

Cox's calibration regression

Cox's calibration regression (25) provides a simple method to quantify the degree of miscalibration of a model. Cox suggested fitting the model

$$\text{true log odds} = \alpha + \beta \times \text{predicted log odds}$$

using logistic regression.

The value of α represents the calibration at a prediction of 0.5 when $\beta \neq 1$, or calibration more generally when $\beta = 1$. The value of β represents the degree of variability in the predicted probabilities. If $\beta > 1$, the "probabilities show the right general pattern of variation but do not vary enough." If $0 < \beta < 1$, the probabilities vary too much.

- Perfect prediction, $\alpha = 0$ and $\beta = 1$ (i.e., true log odds = predicted log odds)
- Perfect calibration, $\alpha = 0$ conditional on $\beta = 1$ ($\alpha = 0|\beta = 1$)
- Correct degree of variation, $\beta = 1$ conditional on the observed value of α ($\beta = 1|\alpha$).

Performance of Published Models. The risk prediction models included in this evaluation were the latest available published versions of the APACHE, SAPS, and MPM models: APACHE III, SAPS II, and MPM II. Widespread adoption of APACHE III has been hindered by its commercial nature, so APACHE II is still widely used and was included in this evaluation. The models are described in Box 2. Each model has different exclusion criteria that define patients for whom predictions may be made (26). We defined a *common cohort* of

Box 2. Risk prediction models

APACHE II

The Acute Physiology and Chronic Health Evaluation II (APACHE II) score (3) comprises an acute physiology score (APS) plus weights for age and a history of severe chronic health conditions. The APS is made up of weightings for 12 physiological variables: temperature, mean arterial pressure, heart rate, respiratory rate, oxygenation (Pao_2 or A-aDo_2), sodium, potassium, creatinine, hematocrit, and white blood cell count, and for admissions neither sedated nor paralyzed for the entire first 24 hrs of intensive care, neurologic status assessed by the Glasgow Coma Score (GCS). The physiological variables are weighted on the basis of the worst value (the measurement giving the highest weight) during the first 24 hours of intensive care. A mortality prediction is calculated from the APACHE II score plus coefficients for post-emergency surgery and 53 diagnostic categories. The APACHE II UK model (4) used the same underlying model, with no change to the APACHE II score, but estimated new coefficients. The UK model has a total of 74 diagnostic categories, based on the classification by system and precipitating factor from the original United States study. Admissions are excluded from the APACHE II model if they are aged <16 yrs, stay <8 hrs in the intensive care unit (ICU), or are admitted with burns or following coronary artery bypass grafting (CABG).

APACHE III

The Acute Physiology, Age, and Chronic Health Evaluation III (APACHE III) model (5) is also based on an APS. The APACHE III APS is made up of weightings for 17 physiological variables: heart rate, mean arterial pressure, temperature, respiratory rate, oxygenation (Pao_2 or A-aDo_2), hematocrit, white blood cell count, creatinine, urine output, urea, sodium, albumin, bilirubin, and glucose, plus interactions between pH and $Paco_2$, and between the eye, motor, and verbal components of the GCS. Weightings were objectively derived and are based on the most extreme measurement (furthest from a fixed value) during the first 24 hrs of intensive care. A mortality prediction is produced from the APS (modeled with restricted cubic splines) plus coefficients for age, severe chronic conditions, source of admission to ICU, post-emergency surgery, length of stay in hospital prior to ICU admission, and 94 diagnostic categories. Admissions are excluded from the APACHE III model if they are aged <16 yrs, stay <4 hrs in the ICU, or are admitted with burns or following transplant surgery. A separate model exists for admissions following CABG, but the variables for this model were not available in the Case Mix Programme Database (CMPD), so these admissions were also excluded.

SAPS II

The Simplified Acute Physiology Score II (SAPS II) (6) comprises weightings for age, heart rate, systolic blood pressure, temperature, oxygenation (Pao_2/Fio_2 only if ventilated), urine output, urea, white blood cell count, potassium, sodium, bicarbonate, bilirubin, GCS (including presedation GCS for sedated admissions), chronic diseases, and surgical status. Physiological weightings are based on the worst value during the first 24 hrs in the ICU. The SAPS II is transformed directly to a mortality prediction without any additional variables in the model, with use of a shrinking power transformation—log(SAPS II + 1)—to improve the fit. Admissions are excluded from the SAPS II model if they are aged <18 yrs; are admitted for coronary care, with burns, or following cardiac surgery; have insufficient data for calculation of surgical status or Pao_2/Fio_2; or are transferred to an ICU in another hospital.

MPM II

The Mortality Probability Models II (MPM II) make two mortality predictions: one using data collected within 1 hr of admission to the unit ($MPM\ II_0$) and an updated prediction at 24 hrs for patients staying 24 hrs or more ($MPM\ II_{24}$) (7). To compare the performance of MPM II with the other risk models, we have used the $MPM\ II_0$ prediction for patients staying <24 hrs and the updated $MPM\ II_{24}$ prediction otherwise. MPM II is not based on a score but instead makes direct mortality predictions. The variables in $MPM\ II_0$ are age, cardiopulmonary resuscitation within 24 hrs prior to admission, medical or unscheduled surgical admission, mechanical ventilation, coma or deep stupor not due to drug overdose, heart rate ≥150 beats min^{-1}, systolic blood pressure ≤90 mm Hg, three chronic diagnoses, and five acute diagnoses. The variables in $MPM\ II_{24}$ are age, medical or unscheduled surgical admission, mechanical ventilation, coma or deep stupor, creatinine value >2 mg dL^{-1}, confirmed infection, Pao_2 <60 mm Hg, prothrombin time >3 secs above reference, urine output <150 mL in 8 hrs, continuous intravenous vasoactive drug therapy for at least 1 hr, two chronic diagnoses, and one acute diagnosis. Admissions are excluded from MPM II if they are aged <18 yrs; are admitted for coronary care, with burns, or following cardiac surgery; or are transferred to an ICU in another hospital. Additionally, admissions staying <24 hrs in the unit are excluded from $MPM\ II_{24}$.

Table 1. Exclusion criteria for the published models

Exclusion Criterion	APACHE II	APACHE III	SAPS II	MPM II	Admissions (%)	Hospital Mortality, %
Age <18 yrs			X	X	6,600 (2.8)	8.8
Age <16 yrs	X	X			4,676 (2.0)	8.0
ICU stay <8 hrs	X				14,997 (6.5)	43.9
ICU stay <4 hrs		X			7,545 (3.3)	48.3
Cardiac surgery			X	X	2,650 (1.1)	10.5
CABG	X	X			1,905 (0.8)	7.4
Burns	X	X	X	X	338 (0.1)	33.0
Transplant surgery		X			43 (0.0)	20.9
Missing surgical status			X		602 (0.3)	32.3
Missing ventilation status or Pao_2/Fio_2			X		1,439 (0.6)	34.3
Transfer to an ICU in another hospital			X	X	8,104 (3.5)	0.0
Readmission within the same hospital stay	X	X	X	X	8,515 (3.7)	38.9
Any of the above					36,496 (15.7)	29.5

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Models; ICU, intensive care unit; CABG, coronary artery bypass graft.

Table 2. Comparison of the Case Mix Programme Database (CMPD) to the development databases for the published models

Database Location	CMPD UK	APACHE II USA	APACHE II UK UK	APACHE III USA	SAPS II Europe, North America	MPM II Europe, North America
Time period	1995–2003	1982	1988–1990	1988–1989	1991–1992	1989–1992
No. of admissions	231,930	5815	10,806	17,440	13,152	19,124
No. of critical care units	163	19	26	42	137	143
Mean age, yrs	61	55	56	59	57	57
Sex, % female	42	N/R	40	45	40	N/R
Surgical status, %						
Elective surgery	25	N/R	31	33	31	N/R
Emergency surgery	20	N/R	26	9	20	N/R
Nonsurgical	55	N/R	43	58	48	N/R
Mean APACHE II score	13	11	14	N/R	N/R	N/R
Mean APACHE II score	17	N/R	18	N/R	N/R	N/R
Mean APACHE III score	64	N/R	N/R	50	N/R	N/R
Mean SAPS II	38	N/R	N/R	N/R	N/R	N/R
Hospital mortality, %	30.7	19.7	27.1	17.3	21.8	20.8

APACHE, Acute Physiology and Chronic Health Evaluation; UK, United Kingdom; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Models; USA, United States; N/R, not reported in original publication.

Table 3. Performance of published models in common cohort (n = 141,106; observed mortality = 30.7%)

Model	Ideal Value	APACHE II	APACHE II UK	APACHE III	SAPS II	MPM II
Average predicted mortality probability, $\bar{p}$	0.307	0.256	0.267	0.225	0.279	0.276
c index (95% CI)	1	0.804 (0.802–0.806)	0.803 (0.801–0.805)	0.832 (0.830–0.834)	0.822 (0.820–0.824)	0.815 (0.813–0.817)
Shapiro's R (95% CI)	1	0.611 (0.609–0.613)	0.611 (0.609–0.613)	0.614 (0.611–0.616)	0.623 (0.620–0.625)	0.620 (0.618–0.622)
Brier's score and derivatives						
Brier's score, B (95% CI)	0	0.162 (0.161–0.164)	0.162 (0.161–0.163)	0.157 (0.156–0.159)	0.155 (0.154–0.156)	0.157 (0.156–0.158)
Spiegelhalter's Z-statistic, z_c[a]	0	57.5	46.0	133.3	74.7	66.1
Accuracy of the average prediction, $(\bar{Y} - \bar{p})^2$	0	0.0026	0.0016	0.0068	0.0008	0.0010
Excess variance of predictions, V_{exc}/V_{min}	0	3.02	3.06	2.42	2.59	2.76
Covariance of outcome and prediction, Cov(**Y**, **p**)	0.213	0.052	0.052	0.062	0.065	0.058
Hosmer-Lemeshow goodness-of-fit statistics						
20 equal-sized groups, $\hat{C}_{20}$[b]	0	2947	2321	10883	2664	1598
20 equally spaced cut-points, $\hat{H}_{20}$[b]	0	2957	1992	11256	2694	1585
Cox's calibration regression						
Intercept α (95% CI)	0	0.28 (0.26–0.30)	0.21 (0.20–0.23)	0.44 (0.43–0.46)	0.05 (0.04–0.07)	0.16 (0.14–0.17)
Slope β (95% CI)	1	0.93 (0.92–0.94)	0.94 (0.93–0.95)	0.83 (0.82–0.84)	0.81 (0.80–0.82)	0.93 (0.92–0.94)
Test hypothesis: $\alpha = 0, \beta = 1$[c]	0	2664	1628	8820	2169	1135
Test hypothesis: $\alpha = 0 \mid \beta = 1$[d]	0	2537	1539	7747	820	972
Test hypothesis: $\beta = 1 \mid \alpha$[d]	0	126	88	1073	1349	164

APACHE, Acute Physiology and Chronic Health Evaluation; UK, United Kingdom; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Models; CI, confidence interval.

[a]Z-statistic (one-tailed): $p < .05$ for values >1.64; $p < .01$ for values >2.33; $p < .001$ for values >3.09; [b]chi-square statistic on 20 degrees of freedom (*df*): $p < .05$ for values >31.4; $p < .01$ for values >37.6; $p < .001$ for values >45.3; [c]chi-square statistic on 2 *df*: $p < .05$ for values >5.99; $p < .01$ for values >9.21; $p < .001$ for values >13.8; [d]chi-square statistic on 1 *df*: $p < .05$ for values >3.84; $p < .01$ for values >6.63; $p < .001$ for values >10.8.

admissions who did not meet the exclusion criteria for any of the models. This common cohort was used in the primary comparisons between models. Each model was also assessed in its *own cohort* of all patients eligible for that particular model.

Changes in performance over time were assessed by evaluating the measures of model performance in each year from 1996 to 2002. Models were additionally assessed in subgroups defined by age, sex, surgical status (elective, emergency, or nonsurgical), and the five most common primary reasons for admission to the critical care unit, assessed by the ICNARC Coding Method (27) (pneumonia, bacterial, or no organism isolated; aortic or iliac dissection or aneurysm; large bowel tumor; septic shock; and esophageal or gastroesophageal tumor).

Recalibration of Published Models. Two of the major uses of risk prediction models in critical care are to enable comparisons between units and to assess performance within units over time. The models were recalibrated and their performance assessed for stability, both across critical care units and over time.

The models were recalibrated in two ways. First, models were recalibrated by a simple "tilting" and "shifting" of the regression line relating observed log odds with predicted log odds, cor-

responding to using the model from Cox's calibration regression (25) as the recalibrated model. This has been recommended as the best approach for recalibrating logistic regression models when only a relatively small sample is available (28). Second, the model coefficients were re-estimated. For the APACHE II and SAPS II models, this re-estimation was performed at two levels. For APACHE II, coefficients were re-estimated first for the APACHE II score, diagnostic category for the primary reason for admission, and an indicator for emergency surgery, retaining the APACHE II score intact. This corresponds to the method used to establish the APACHE II UK model (4). Next, the APACHE II score was replaced by its components: Acute Physiology Score (APS), age, and medical history. Weightings for variables within the APS were not re-estimated. For SAPS II, coefficients were re-estimated, first retaining the complete SAPS II score and then for the individual components of the SAPS II score. For APACHE III, the total APACHE III score is not included in the risk prediction model; the individual components of APS, age, and medical history are. Weightings for variables within the APS were not re-estimated. The MPM II model consists of a probability only.

Recalibration was carried out in the common cohort. The database was split into development and validation samples in two ways. First, critical care units were randomly allocated to development or validation in a 2:1 ratio. To avoid chance differences in the characteristics of units allocated to the development and validation samples, the random splitting was repeated 200 times and the average calibration and discrimination statistics were obtained (median for skewed statistics; mean, otherwise). The variation in these statistics between samples was examined for evidence of models being more or less stable over repeated samples (29). Second, all admissions in 1999 or before were allocated to the development sample, with admissions from 2000 onward serving as the validation sample, split into three equal groups over time. This allowed us to measure any deterioration in the fit of models over time.

Final coefficients for recalibrated models were estimated on the entire database.

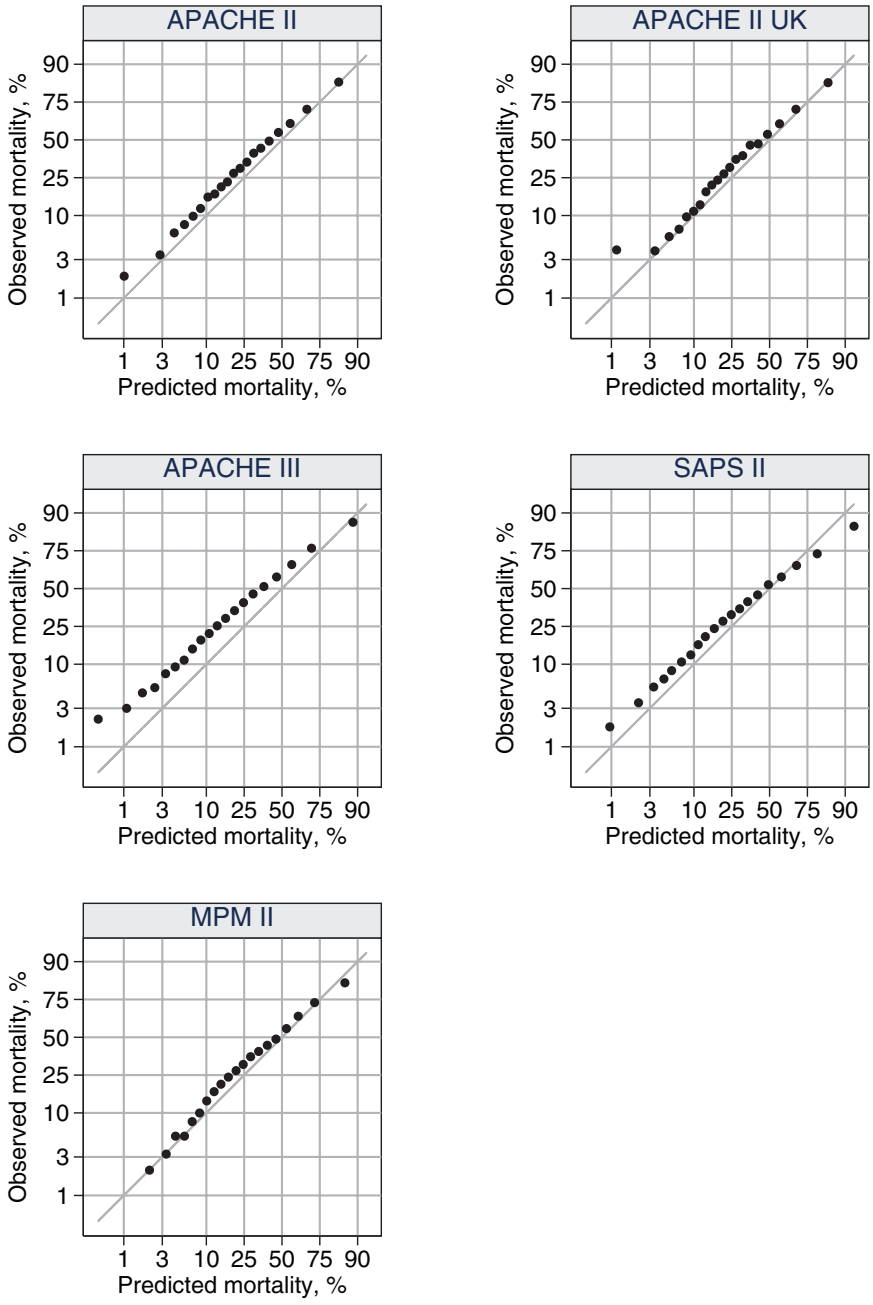

Figure 1. Calibration plots of published models in common cohort (*APACHE*, Acute Physiology and Chronic Health Evaluation; *SAPS*, Simplified Acute Physiology Score; *MPM*, Mortality Probability Models). Observed mortality vs. predicted mortality from model in 20 equal-sized groups, based on quantiles of predicted mortality. Diagonal line indicates perfect calibration. *Axes* drawn on a log odds scale.

RESULTS

Data

Validated data on 231,930 admissions to 163 critical care units between December 1995 and August 2003 were available for analysis. Of these, 49,042 admissions (21.1%) were from periods during which the variables for APACHE III, SAPS II, and/or MPM II were not collected in 44 units. Applying the exclusion criteria for all four models simultaneously excluded 36,496 admissions (15.7%) (Table 1), and further excluding 1,252 admissions (0.5%) with missing hospital outcome and 4,034 (1.7%) for whom an APACHE II or III probability could not be calculated (mostly because of missing or incomplete primary reason for admission) resulted in a common cohort of 141,106 admissions (60.8%) to 157 critical care units. Units contributed a median of 675 admissions (90% reference range, 142 to 2,365) to the common cohort. Because of units joining and leaving the CMP during the course of the study, units contributed a median of 997 days (90% reference range, 180 to 2,311) of data to the common cohort. This corresponds to a median rate of admissions per unit per year of 278 (90% reference range, 140 to 546). Table 2 compares the CMPD to the development databases for the published models.

Table 4. Model performance in 200 repeated validation samples (random 1/3 of units) of models recalibrated by re-estimating coefficients in the corresponding development sample (mean sample size, n = 47,648)

Model	Ideal Value	APACHE II	APACHE III	SAPS II	MPM II	p Value[a]
Average predicted mortality probability, $\bar{p}$	0.308	0.307 (0.008)	0.308 (0.009)	0.307 (0.008)	0.308 (0.008)	.22
c Index (area under the ROC curve)	1	0.832 (0.004)	0.845 (0.004)	0.840 (0.004)	0.824 (0.003)	<.001
Shapiro's R	1	0.633 (0.005)	0.644 (0.005)	0.640 (0.005)	0.629 (0.004)	<.001
Brier's score and derivatives						
Brier's score, B	0	0.150 (0.003)	0.143 (0.003)	0.145 (0.003)	0.152 (0.003)	<.001
Spiegelhalter's Z-statistic, z_c[b], median	0	1.89	1.59	0.47	1.23	
Accuracy of the average prediction, $(\bar{Y} - \bar{p})^2$	0	1.1×10^{-4} (1.6×10^{-4})	1.3×10^{-4} (2.2×10^{-4})	1.0×10^{-4} (1.4×10^{-4})	7.7×10^{-5} (1.0×10^{-4})	.001
Excess variance of predictions, V_{exc}/V_{min}	0	2.36 (0.07)	2.05 (0.06)	2.12 (0.06)	2.49 (0.07)	<.001
Covariance of outcome & prediction, Cov(**Y**, **p**)	0.213	0.064 (0.002)	0.071 (0.002)	0.069 (0.001)	0.062 (0.001)	<.001
Hosmer-Lemeshow goodness-of-fit statistics						
20 equal-sized groups, $\hat{C}_{20}$[c], median	0	140.2	80.0	52.0	92.7	
20 equally spaced cut-points, $\hat{H}_{20}$[c], median	0	146.9	85.2	46.7	91.3	
Cox's calibration regression						
Intercept α	0	0.00 (0.08)	0.00 (0.07)	0.00 (0.07)	0.00 (0.06)	.14
Slope β	1	0.99 (0.03)	0.99 (0.03)	1.00 (0.02)	1.00 (0.02)	.015
Test hypothesis: $\alpha = 0$, $\beta = 1$[d], median	0	23.7	23.5	20.0	18.6	
Test hypothesis: $\alpha = 0 \mid \beta = 1$[e], median	0	19.1	15.9	13.8	13.5	
Test hypothesis: $\beta = 1 \mid \alpha$[e], median	0	3.10	3.24	2.15	2.00	

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Models; ROC, receiver operating characteristic.

[a]P value for a difference between the models (repeated-measures analysis of variance); [b]Z-statistic (one-tailed): $p < .05$ for values >1.64; $p < .01$ for values >2.33; $p < .001$ for values >3.09; [c]chi-square statistic on 20 degrees of freedom (df): $p < .05$ for values >31.4; $p < .01$ for values >37.6; $p < .001$ for values >45.3; [d]chi-squared statistic on 2 df: $p < .05$ for values >5.99; $p < .01$ for values >9.21; $p < .001$ for values >13.8; [e]chi-square statistic on 1 df: $p < .05$ for values >3.84; $p < .01$ for values >6.63; $p < .001$ for values >10.8. Values are mean (SD), unless otherwise stated.

Performance of Published Models. On average, all models estimated lower risk of hospital death than observed in both the common cohort (Table 3) and their own cohorts. Inspection of calibration plots (Fig. 1) confirms there is a tendency for more deaths to occur than predicted for lower-risk patients.

All models displayed moderate discrimination, with a slight advantage by APACHE III and SAPS II. The variation in the c index across models was statistically significant ($p < .001$). Shapiro's R, reflecting overall accuracy, was around 0.6 for all models, far from the ideal value of 1.0. The APACHE II models produced marginally the most variable predictions ($V_{exc}/V_{min} = 3.02$ and 3.06).

All measures of goodness-of-fit (z_c, $\hat{C}_{20}$, $\hat{H}_{20}$, χ^2 test statistic for $\alpha = 0$, $\beta = 1$) confirmed highly significant departures of observed deaths from predictions of risk, but these measures were smallest for MPM II in the common cohort and for APACHE II UK in its own cohort. APACHE III had by far the worst calibration of the models considered. All models also failed tests of incorrect calibration given appropriate refinement ($\alpha = 0|\beta = 1$) and refinement given correct calibration ($\beta = 1|\alpha$).

There was a small reduction in overall mortality ratio from all models except MPM II over the period 1996 to 2002 (APACHE II, 1.27 to 1.17; APACHE II UK, 1.25 to 1.13; APACHE III, 1.57 to 1.31; SAPS II, 1.17 to 1.09; MPM II, 1.14 to 1.13); however, there was little corresponding change in discrimination or accuracy. Model performance varied markedly by patient age, with better discrimination and calibration for younger patients than older patients (Brier's score for age <45 yrs, 0.09 to 0.10 across models; for age 85+ yrs, 0.20 to 0.23). This suggests the relationship between age and risk is stronger in the CMPD than in the databases used to develop these models.

There was no difference between men and women in model performance but considerable variation in both discrimination and accuracy by surgical status and primary reason for admission. Risk was underestimated for nonsurgical admissions, particularly by APACHE II (mortality ratio, 1.29) and APACHE III (mortality ratio, 1.38). APACHE III also underestimated the risk of death following emergency surgery (mortality ratio, 1.50). Discrimination tended to be worse for surgical admissions than for nonsurgical admissions (c index for surgical admissions, 0.73 to 0.79; for nonsurgical admissions, 0.79 to 0.82).

Recalibration of Published Models. Ninety-seven of the 157 critical care units in the common cohort were randomly assigned to development sample I, with the remaining 60 units assigned to validation sample I (repeated 200 times). Development sample II included all admissions from December 1995 to December 1999, with validation samples II-1, II-2, and II-3 consisting of admissions from January 2000 to January 2001, February

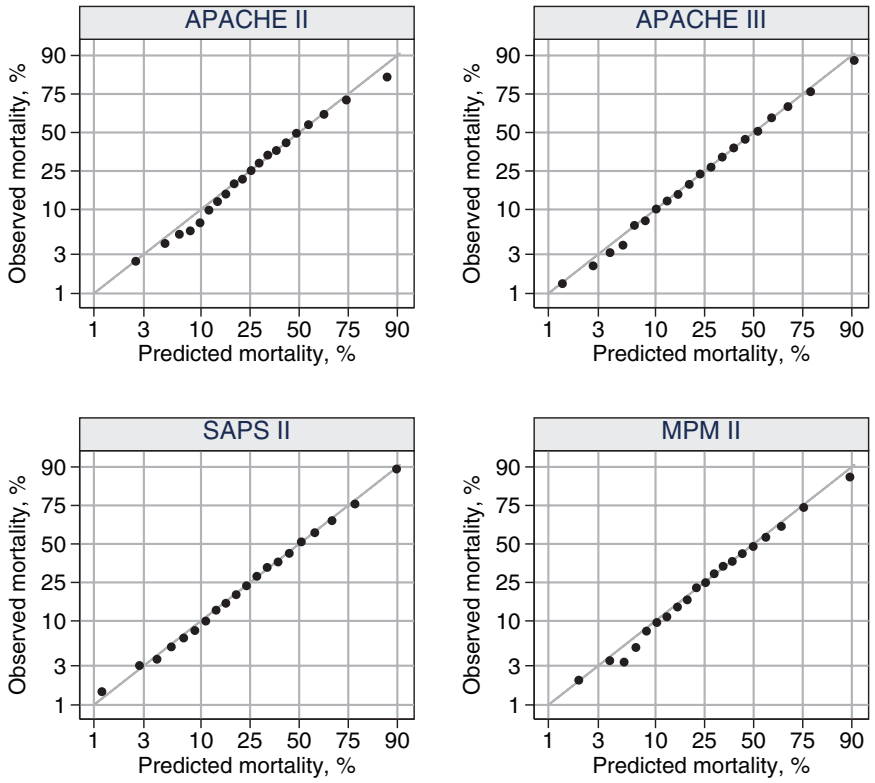

Figure 2. Calibration plots in validation sample I of models recalibrated by re-estimating coefficients in development sample I. Observed mortality vs. predicted mortality from model in 20 equal-sized groups, based on quantiles of predicted mortality. Diagonal line indicates perfect calibration. Axes drawn on a log odds scale. *APACHE*, Acute Physiology and Chronic Health Evaluation; *SAPS*, Simplified Acute Physiology Score; *MPM*, Mortality Probability Models.

2001 to January 2002, and February 2002 to August 2003, respectively.

Calibration of all models was markedly improved by simple Cox recalibration. Although goodness-of-fit statistics indicated statistically significant (if reduced) lack of fit, qualitative predictions from the recalibrated models were remarkably good (indicated by values of α and β from Cox's calibration regression close to 0 and 1, respectively, and by inspection of calibration plots). Under repeated random selection of the development and validation samples, APACHE II demonstrated superior calibration, with the lowest $\hat{C}_{20}$ statistic in 73% of validation samples. However, MPM II tended to make the least-biased predictions in the validation samples: in 50% of samples, the test statistic for $\alpha = 0, \beta = 1$ was lower than that of other models.

Table 4 and Figure 2 present the results of the most detailed level of recalibration by re-estimation of model coefficients. Fit of the APACHE II model seemed to benefit little from further recalibration beyond the Cox recalibration, whereas the re-estimated APACHE III was superior and SAPS II had progressively better calibration when coefficients were re-estimated and when the score was re-estimated. Re-estimating MPM II reduced the Hosmer-Lemeshow statistics but moved α further from its ideal value of zero. Discrimination improved for all models following re-estimation of coefficients and further improved for APACHE II and SAPS II by re-estimation of the components of the scores. However, improvements were modest. Under repeated random selection of the development and validation samples, SAPS II had the best calibration, with the lowest $\hat{C}_{20}$ statistic in 62% of validation samples. APACHE III consistently had the best discrimination, with the highest *c* index in 199 of the 200 samples.

In validation samples II-1, II-2, and II-3, discrimination slightly improved over time for all models fitted in development sample II (Table 5). This may reflect the increased frequency over time of nonsurgical admissions, for whom better discrimination was noted. There was some deterioration in calibration over the 3.5 yrs covering validation samples II-1, II-2, and II-3 after model development in development sample II (Table 5), suggesting recalibration of models every 2 to 3 yrs would be beneficial.

DISCUSSION

This study showed that all published risk prediction models required recalibration for use with current UK data. All the models were miscalibrated: not only those fitted entirely in other countries (APACHE II, APACHE III) or in multinational cohorts including UK centers (SAPS II, MPM II), but also the previous recalibration of APACHE II to earlier UK data. Recalibrated models showed greatly improved calibration and slight improvements to discrimination when assessed in validation samples independent from the development samples used to fit the models. There was little difference in performance between the recalibrated models; no model consistently outperformed the others across the different methods of assessing model performance. This suggests that the choice of risk prediction model in a particular situation could reasonably be based on pragmatic considerations such as a requirement for comparability with other studies or over time, consistency with what is already collected locally, or the time and cost burden associated with data collection. Further investigation of model performance within a specific patient group or setting may enable a more evidence-based choice of model for that particular situation.

The relative importance of different aspects of model performance may also depend on the use for which the model is being employed. In many situations, good discrimination may be the most desirable property; however, if comparisons are to be made between different groups of patients, good calibration is also essential.

The study used a panel of expert statisticians to review current statistical methodology for the assessment of risk prediction models. The panel recommended that current statistical methods must be supplemented with the more robust and appropriate methods used in this study, in particular, the use of the Cox's calibration regression in addition to the Hosmer-Lemeshow goodness-of-fit test. The additional methods help in interpreting the fit of the models by providing quantitative measures of the degree of miscalibration of the models (α and β from Cox's calibration regression, Shapiro's *R*, Brier's score). They also allow a more detailed analysis of the nature of the miscalibra-

Table 5. Model performance over time in validation samples II-1, II-2, and II-3 of models recalibrated in development sample II by re-estimation of coefficients

Variable	Validation Period 1 Jan 2000–Jan 2001	Validation Period 2 Feb 2001–Jan 2002	Validation Period 3 Feb 2002–Aug 2003
Number of admissions, n	23,772	25,063	26,844
Observed mortality, $\bar{Y}$	0.316	0.311	0.297
APACHE II			
Predicted mortality, $\bar{p}$	0.326	0.328	0.314
c Index	0.829	0.834	0.837
Shapiro's *R*	0.629	0.633	0.641
Brier's score, *B*	0.152	0.150	0.145
Hosmer-Lemeshow $\hat{C}_{20}$	61.1	111.4	124.7
Hosmer-Lemeshow $\hat{H}_{20}$	64.6	119.8	123.6
Cox's calibration: α, β	−0.07, 0.99	−0.11, 1.00	−0.10, 1.03
χ^2 for $\alpha = 0, \beta = 1$	15.7	44.1	60.6
APACHE III			
Predicted mortality, $\bar{p}$	0.335	0.341	0.324
c Index	0.840	0.849	0.852
Shapiro's *R*	0.637	0.644	0.651
Brier's score, *B*	0.147	0.143	0.139
Hosmer-Lemeshow $\hat{C}_{20}$	105.9	183.8	172.1
Hosmer-Lemeshow $\hat{H}_{20}$	102.2	189.0	169.3
Cox's calibration: α, β	−0.14, 0.96	−0.21, 1.00	−0.18, 1.02
χ^2 for $\alpha = 0, \beta = 1$	64.9	152.1	144.3
SAPS II			
Predicted mortality, $\bar{p}$	0.324	0.326	0.309
c Index	0.833	0.842	0.845
Shapiro's *R*	0.633	0.641	0.648
Brier's score, *B*	0.149	0.145	0.141
Hosmer-Lemeshow $\hat{C}_{20}$	40.5	60.2	48.8
Hosmer-Lemeshow $\hat{H}_{20}$	38.8	62.0	58.8
Cox's calibration: α, β	−0.06, 0.97	−0.10, 1.01	−0.08, 1.02
χ^2 for $\alpha = 0, \beta = 1$	12.7	39.3	32.3
MPM II			
Predicted mortality, $\bar{p}$	0.321	0.314	0.300
c Index	0.823	0.826	0.824
Shapiro's *R*	0.626	0.631	0.633
Brier's score, *B*	0.154	0.151	0.150
Hosmer-Lemeshow $\hat{C}_{20}$	28.2	46.1	57.6
Hosmer-Lemeshow $\hat{H}_{20}$	32.1	51.2	55.2
Cox's calibration: α, β	−0.03, 1.00	−0.00, 1.03	0.00, 1.03
χ^2 for $\alpha = 0, \beta = 1$	3.86	4.98	5.12

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Models.

tion (tests of $\alpha = 0|\beta = 1$ and $\beta = 1|\alpha$, decomposition of Brier's score).

This study used data from 231,930 admissions to 163 critical care units within a single healthcare system, making it the largest and most powerful dataset ever used to assess these models. The finding that models require recalibration for use in the UK is consistent with the two smaller, regional multicenter UK-based studies (8, 9). However, only one of these databases has been used to recalibrate a risk prediction model (30), and the recalibration was performed only for SAPS II. In addition, the current study uses data from a representative sample of critical care units in three countries of the UK, improving the scope for generalizing the results to the entire UK vs. results from smaller geographical regions.

A number of large (≥1,000 admissions) multicenter (≥5 critical care units) studies have validated one or more of these models in independent populations (Table 6). Investigators in the EURICUS-I study, involving 89 critical care units in 13 European areas, reported poor calibration for both SAPS II and MPM II_0 (31). An analysis in Portugal (32) showed that SAPS II outperformed APACHE II, but both models had poor calibration. Further validation studies of SAPS II have been carried out in Italy (33, 34) and Austria (35), revealing good discrimination but poor calibration. SAPS II was recalibrated in all three datasets, reporting acceptable calibration.

The original authors of the APACHE III model performed a validation in an independent sample of 37,668 admissions to 285 critical care units in the United States (36). Discrimination in this sample was excellent (*c* index, 0.89). They concluded that "APACHE III accurately predicted aggregate hospital mortality," but there were highly significant departures from perfect calibration (Hosmer-Lemeshow test, $p < .0001$). Good discrimination of APACHE III was also found in Brazil (37) and Spain (38), but the model showed a significant lack of calibration in both populations. APACHE III was recalibrated in the same cohort of Spanish patients, leading to improved discrimination and calibration (39).

These international studies consistently show that the published risk prediction models provide good discrimination when applied in new populations, but that they should not be used without appropriate recalibration. Interpretation of these studies is limited by the overreliance on the Hosmer-Lemeshow statistic for measuring calibration. Differing reports of "poor," "acceptable," or "good" calibration of the models may be the result of differences in sample size rather than true differences in performance. This reinforces the need to use measures that provide a quantitative indication of miscalibration.

There is considerable debate about how to interpret the lack of calibration when a scoring system derived from patients in one country or healthcare system in a given time period is applied to patients admitted for care in other settings or in other time periods. For example, the current study could be interpreted as showing that the odds of

Table 6. Comparison of the Case Mix Programme Database (CMPD) with other independent multicenter validation studies

Variable	CMPD	Beck (9,29)	Livinston (8)	Moreno (30)	Moreno (31)	Apolone (32)
Location	UK	Southern England	Scotland	Europe	Portugal	Italy
Time period	1995–2003	1993–1996	1995–1996	1994–1995	1994–1995	1994
No. of admissions	231,930	17,210	13,291	16,060	1094	2202
No. of critical care units	163	17	22	89	19	99
Mean age, yrs	61	61	59	59	55	60
Sex, % female	42	41	45	N/R	32	38
Surgical status, %						
Elective surgery	25	25	21	24	12	16
Emergency surgery	20	16	27	20	20	12
Nonsurgical	55	59	51	56	68	71
Hospital mortality, %	30.7	26.5	29.4	20.0	32.0	34.1
APACHE II						
Mean (SD) score	17 (7)	15 (7)	N/R	N/R	20 (10)	N/R
Predicted hospital mortality, %	25.6	22.4	30.0	N/R	33.5	N/R
c index (recalibrated)	0.804 (0.832)	0.835	0.805	N/R	0.787	N/R
APACHE II UK						
Predicted hospital mortality, %	26.7	N/R	35.9	N/R	N/R	N/R
c index	0.803	N/R	0.809	N/R	N/R	N/R
APACHE III						
Mean (SD) score	64 (29)	57 (25)	N/R	N/R	N/R	N/R
Predicted hospital mortality, %	22.5	21.5	24.0	N/R	N/R	N/R
c index (recalibrated)	0.832 (0.845)	0.867	0.845	N/R	N/R	N/R
SAPS II						
Mean (SD) score	38 (18)	34 (17)	N/R	34	41 (21)	39
Predicted hospital mortality, %	27.9	22.7	30.4	22.3	32.6	29.9
c index (recalibrated)	0.822 (0.840)	0.852 (0.845)	0.843	0.822	0.817	0.80
MPM II						
Predicted hospital mortality, %	27.6	N/R	29.3/30.1[a]	23.6[b]	N/R	N/R
c index (recalibrated)	0.815 (0.824)	N/R	0.741/0.791[a]	0.785[b]	N/R	N/R

survival for a patient admitted for intensive care have worsened in comparison with the United States in 1985, when the APACHE II system was developed. Other than being of limited use for guiding current and future improvements in intensive care, this is also a conclusion that should be approached with caution. For a tool to calibrate well in a new setting, all factors that influence outcome (including patient factors and quality of care) must either be included in the model or have the same distribution in the new setting as in the sample used to develop the model.

Differences between countries and over time make this second condition unlikely. Critical care units vary considerably in their provision, structure, organization and staffing across countries. Furthermore, differences in critical care admission policies will change not only the populations that are included in the models but also the relationships between prognostic variables and outcome. For this reason, it is likely that any model seeking to compare risk-adjusted outcomes for critical care between different health care systems would need to be derived in a larger, more objectively defined population than only those patients admitted to a critical care unit. Because of differences in management after discharge from critical care, it would also be advisable for such a model to evaluate mortality at a fixed time point (e.g., 90 days) rather than at discharge from hospital.

We have shown that significant gains in discrimination and calibration can be made by recalibrating these models for the population in which they are to be used. This does not, however, tell us whether any of the models represent the best possible model for making risk-adjusted outcome comparisons in this population. Further work should investigate whether additional improvements can be made by selecting the best features from the different models. Some authors have proposed modifications to the published models, such as the use of presedation Glasgow Coma Score (40), and any new model should investigate these modifications. The exclusion criteria have been shown to exclude large numbers of admissions, and they are applied inconsistently (26); a new model should seek to minimize exclusion criteria.

Because it is impossible to account for all variability among patients with these models, there is likely to be some limit beyond which the models cannot be improved further; perfect calibration and discrimination will never be achievable.

CONCLUSIONS

This study confirms that adult intensive care risk prediction models primarily developed in other countries require validation and recalibration before being used to provide risk-adjusted outcomes for units within a new country setting. It is beneficial to assess the performance of these tools periodically in order to ensure calibration is maintained. We recommend that this periodic assessment be carried out with use of the methodology presented here, in particular the use of Cox's calibration regression to provide a quantitative measure of model calibration. We would also recommend that investigators use this methodology when undertaking evaluations of risk prediction models in other countries, by other healthcare providers, or in other settings (41, 42).

ACKNOWLEDGMENTS

The authors acknowledge the Steering Committee: Doug Altman, James

Table 6.—(*Continued*)

Sicignano (33)	Metnitz (34)	Zimmerman (35)	Bastos (36)	Vazquez Mata (37)/ Rivera-Fernandez (38)
Italy	Austria	USA	Brazil	Spain
1995–1996	1997	1993–1996	1990–1991	1992–1995
9185	1733	37,668	1856	12,174
24	9	285	10	86
62	59	60	52	58
N/R	N/R	44	38	32
32	39	22	N/R	14
25	22	6	N/R	10
44	40	72	N/R	76
31.1	19.5	12.4	34.3	21.2
N/R	N/R	N/R	N/R	N/R
N/R	N/R	N/R	N/R	N/R
N/R	N/R	N/R	0.79	N/R
N/R	N/R	N/R	N/R	N/R
N/R	N/R	N/R	N/R	N/R
N/R	N/R	45 (27)	55	54
N/R	N/R	12.3	20.6	19.8
N/R	N/R	0.89	0.82	0.808
36	N/R	N/R	N/R	N/R
30.7	23.0	N/R	N/R	N/R
0.87 (0.87)	0.81	N/R	N/R	N/R
N/R	N/R	N/R	N/R	N/R
N/R	N/R	N/R	N/R	N/R

N/R, not reported; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Models.

[a]MPM II_0/MPM II_{24}; [b]MPM II_0.

Carpenter, Harvey Goldstein, Jon Nicholl, Gareth Parry, Patrick Royston, and David Spiegelhalter. They also thank Dr. Alasdair Short for comments on an early draft of this article.

REFERENCES

1. Black N: Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; 312:1215–1218
2. UK Neonatal Staffing Study Group: A prospective evaluation of patient volume, staffing and workload in relation to risk-adjusted outcomes in a random, stratified sample of all UK neonatal intensive care units. *Lancet* 2002; 359:99–107
3. Knaus WA, Draper EA, Wagner DP, et al: APACHE II: A severity of disease classification system. *Crit Care Med* 1985; 13:818–829
4. Rowan KM: Outcome Comparisons of Intensive Care Units in Great Britain and Ireland Using the APACHE II Method [DPhil thesis]. Oxford, University of Oxford, 1992
5. Knaus WA, Wagner DP, Draper EA, et al: The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100:1619–1636
6. 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; 270:2957–2963
7. Lemeshow S, Teres D, Klar J, et al: Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993; 270:2478–2486
8. Livingston BM, MacKirdy FN, Howie JC, et al: Assessment of the performance of five intensive care scoring models within a large Scottish database. *Crit Care Med* 2000; 28: 1820–1827
9. Beck DH, Smith GB, Pappachan JV, et al: External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: A multicentre study. *Intensive Care Med* 2003; 29:249–256
10. Fery-Lemonnier E, Landais P, Loirat P, et al: Evaluation of severity scoring systems in ICUs: Translation, conversion and definition ambiguities as a source of inter-observer variability in APACHE II, SAPS and OSF. *Intensive Care Med* 1995; 21:356–360
11. Hanley JA, McNeil BJ: The meaning and use of the area under the receiver operating characteristics (ROC) curve. *Radiology* 1982; 143:29–36
12. Hosmer DW, Lemeshow S: Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics* 1980; A9:1043–1069
13. Van Houwelingen JC, Le Cessie S: Predictive value of statistical models. *Stat Med* 1990; 9:1303–1325
14. Miller ME, Hui SL, Tierney WM: Validation techniques for logistic regression models. *Stat Med* 1991; 10:1213–1226
15. Hosmer DW, Hosmer T, Le Cessie S, et al: A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997; 16:965–980
16. Rowan K: Intensive Care Society has set up a centre for national audit. *BMJ* 1996; 313: 1007–1008
17. Intensive Care National Audit & Research Centre (ICNARC). ICNARC Case Mix Programme Dataset Specification, Version 2.0. London, ICNARC, 1997
18. Harrison DA, Brady AR, Rowan K: Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit & Research Centre Case Mix Programme Database. *Crit Care* 2004; 8:R99–R111; available online at http://ccforum.com/content/8/2/R99 (accessed June 20, 2005)
19. Harrell FE, Califf RM, Pryor DB, et al: Evaluating the yield of medical tests. *JAMA* 1982; 247:2543–2546
20. Bamber D: The area above the ordinal dom-

inance graph and the area below the receiver operating characteristic graph. *J Math Psychol* 1975; 12:387–415
21. Shapiro AR: The evaluation of clinical predictions. *N Engl J Med* 1977; 296:1509–1514
22. Brier GW: Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 1950; 75:1–3
23. Spiegelhalter DJ: Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986; 5:421–433
24. Yates JF: External correspondence: Decomposition of the mean probability score. *Organ Behav Hum Perform* 1982; 30:132–156
25. Cox DR: Two further applications of a model for binary regression. *Biometrika* 1958; 45: 562–565
26. Wunsch H, Brady AR, Rowan K: Impact of exclusion criteria from severity of disease scoring methods on outcomes in intensive care. *J Crit Care* 2004; 19:67–74
27. Young JD, Goldfrad C, Rowan K: Development and testing of a hierarchical method to code the reason for admission to intensive care units: The ICNARC Coding Method. *Br J Anaesth* 2001; 87:543–548
28. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, et al: Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat Med* 2004; 23:2567–2586
29. Efron B, Tibshirani RJ: An Introduction to the Bootstrap. London, Chapman & Hall, 1993
30. 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; 57:785–793
31. Moreno R, Miranda DR, Fidler V, et al: Evaluation of two outcome prediction models on an independent database. *Crit Care Med* 1998; 26:50–61
32. Moreno R, Morais P: Outcome prediction in intensive care: Results of a prospective, multicentre, Portuguese study. *Intensive Care Med* 1997; 23:177–186
33. Apolone G, Bertolini G, D'Amico R, et al: The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: Results from GiViTI: Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva. *Intensive Care Med* 1996; 22:1368–1378
34. Sicignano A, Giudici D: Customization of SAPS II for the assessment of severity in Italian ICU patients. ARCHIDIA: Archivio Diagnostico. *Minerva Anestesiol* 2000; 66: 139–145
35. Metnitz PG, Valentin A, Vesely H, et al: Prognostic performance and customisation of the SAPS II: Results of a multicenter Austrian study: Simplified Acute Physiology Score. *Intensive Care Med* 1999; 25:192–197
36. Zimmerman JE, Wagner DP, Draper EA, et al: Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. *Crit Care Med* 1998; 26:1317–1326
37. Bastos PG, Sun X, Wagner DP, et al: Application of the APACHE III prognostic system in Brazilian intensive care units: A prospective multicenter study. *Intensive Care Med* 1996; 22:564–570
38. Vázquez Mata G, del Mar Jiménez Quintana M, Rivera Fernández R, et al: Objetivación de la gravedad mediante el sistema APACHE-III aplicado en España. *Med Clin (Barc)* 2001; 117:446–451
39. Rivera-Fernández R, Vázquez-Mata G, Bravo M, et al: The APACHE III prognostic system: Customized mortality predictions for Spanish ICU patients. *Intensive Care Med* 1998; 24:574–581
40. Livingston BM, Mackenzie SJ, MacKirdy FN, et al., on behalf of the Scottish Intensive Care Society Audit Group: Should the presedation Glasgow Coma Scale value be used when calculating Acute Physiology and Chronic Health Evaluation scores for sedated patients? *Crit Care Med* 2000; 28:389–394
41. Brady AR, Harrison D, Black S, et al., on behalf of the UK PICOS study group: Assessment and optimization of mortality prediction tools for admissions to pediatric intensive care in the United Kingdom. *Pediatrics* (In Press)
42. The SAPS III Outcome Research Group. Available online at http://www.saps3.org (accessed June 20, 2005)