A new risk prediction model for critical care: The Intensive Care National Audit & Research Centre (ICNARC) model*

David A. Harrison, PhD; Gareth J. Parry, PhD; James R. Carpenter, DPhil; Alasdair Short, FRCP FRCA; Kathy Rowan, DPhil

Objective: **To develop a new model to improve risk prediction for admissions to adult critical care units in the UK.**

Design: **Prospective cohort study.**

Setting: **The setting was 163 adult, general critical care units in England, Wales, and Northern Ireland, December 1995 to August 2003.**

Patients: **Patients were 216,626 critical care admissions.**

Interventions: **None.**

Measurements and Main Results: **The performance of different approaches to modeling physiologic measurements was evaluated, and the best methods were selected to produce a new physiology score. This physiology score was combined with other information relating to the critical care admission—age, diagnostic category, source of admission, and cardiopulmonary resuscitation before admission—to develop a risk prediction model. Modeling interactions between diagnostic category and physiology score enabled the inclusion of groups of admissions that are frequently excluded from risk prediction models. The new model showed good discrimination (mean *c* index 0.870) and fit (mean Shapiro's *R* 0.665, mean Brier's score 0.132) in 200 repeated validation samples and performed well when compared with recalibrated versions of existing published risk prediction models in the cohort of patients eligible for all models. The hypothesis of perfect fit was rejected for all models, including the Intensive Care National Audit & Research Centre (ICNARC) model, as is to be expected in such a large cohort.**

Conclusions: **The ICNARC model demonstrated better discrimination and overall fit than existing risk prediction models, even following recalibration of these models. We recommend it be used to replace previously published models for risk adjustment in the UK. (Crit Care Med 2007; 35:1091–1098)**

KEY WORDS: critical care; hospital mortality; intensive care units; models, statistical; risk adjustment; severity of illness index

Studies of service and therapy provision in critical care frequently use observational designs (1, 2). These rely on robust, valid risk adjustment to reduce bias and ensure high-quality methodology (3–5).

There has been considerable work in risk prediction models in critical care. The original Acute Physiology and Chronic Health Evaluation (APACHE) model was published in 1981 (6), with three subsequent revisions (7–9). UK coefficients for the APACHE II model were published in 1990 (10). The Simplified Acute Physiology Score (SAPS) was devised as a simplification of APACHE (11) and has been revised twice (12, 13). The Mortality Prediction Model (MPM) produced a risk prediction based on categorical (yes/no) variables (14) and has been revised once (15).

Recent work has shown that even with sophisticated recalibration, these models display considerable lack of fit when evaluated in different critical care populations (16). Variations in calibration across the range of predictions may lead to significant bias if centers are compared based on the ratio of observed to expected deaths. The models also define exclusion criteria that may exclude up to 15% of admissions (17). These criteria are inconsistently applied and may introduce biases into risk-adjusted analyses.

We aimed to derive a new risk prediction model suitable for use in all admissions to UK critical care units, based on data from a large, multicenter, high-quality clinical database (18). To ensure continuity, we sought to build on the existing models rather than build a new model *de novo*. However, we aimed to improve on well-known limitations of the models, such as modeling of neurologic impairment (19), and to investigate the inclusion of interaction terms in the model, in particular to allow the effect of physiologic derangement on mortality to vary depending on diagnosis.

MATERIALS AND METHODS

Data. The Case Mix Programme (CMP) is a national comparative audit of patient outcome from adult, general (mixed medical/surgical) critical care units—intensive care units (ICUs) and combined intensive care and high-dependency units (HDUs)—in England, Wales, and Northern Ireland, coordinated by the Intensive Care National Audit & Research Centre (ICNARC). The CMP database contains raw physiologic and diagnosis data for the APACHE II, APACHE III, SAPS II, and MPM II models, together with demographic, outcome, and activity data, for consecutive admissions to units participating in the CMP. Data are collected prospectively, are abstracted by trained data collectors, and undergo extensive validation both locally and centrally (20). Additional physiologic variables not required by the published models were included in the original

***See also p. 1209.**

From the Intensive Care National Audit & Research Centre (ICNARC), London, UK (DAH, KR); Quality Measurement & Analysis, Department of Medicine, Children's Hospital Boston, Boston, MA (GJP); Medical Statistics Unit, London School of Hygiene & Tropical Medicine, London, UK (JRC); and Intensive Care Services, Broomfield Hospital, Chelmsford, Essex, UK (AS).

Supported, in part, by grant G9813469 from the Medical Research Council, London, UK.

The authors have not disclosed any potential conflicts of interest.

For information regarding this article, E-mail: david.harrison@icnarc.org

DOI: 10.1097/01.CCM.0000259468.24532.44

Table 1. Calibration and discrimination in the validation sample of a progressively simplified model for physiology

Model	*df*[a]	*c* Index	*C**[b]
Full	113	0.826	263
Simplified[c]	84	0.826	252
Variables removed[d]			
Hematocrit	82	0.823	252
Serum potassium	80	0.826	253
Serum bicarbonate	78	0.826	252
Prothrombin time	77	0.825	255
Serum glucose	73	0.825	251
$Paco_2$	68	0.824	267
Total serum bilirubin	65	0.824	264
Serum albumin	**62**	**0.823**	**279**
Serum creatinine	59	0.822	297
White blood count	54	0.821	307
Temperature	50	0.819	335
Heart rate	45	0.817	381
Respiratory rate	39	0.814	506
Serum sodium	35	0.811	591
Arterial pH	31	0.808	680
Pao_2/Fio_2	26	0.799	1,193
Urine output	20	0.788	2,019
Systolic blood pressure	11	0.754	4,687
Sedated, paralyzed or GCS	3	0.658	12,183
Serum urea	0	0.500	24,160

GCS, Glasgow Coma Scale score. Bold type indicates the variables included in the final physiology model and the fit of the final physiology model.

[a]Degrees of freedom (number of model parameters minus one); [b]Hosmer-Lemeshow chi-square statistic for 20 groups from quantiles of full model predictions; [c]the simplified model represents the full model with adjacent categories combined when not significant ($p > .1$); [d]least significant variable removed at each step until no variables remained.

data set specification to enable investigation of alternative approaches to modeling.

Admissions were excluded from the study if they were missing the outcome variable (mortality at ultimate discharge from an acute hospital) or were readmissions of the same patient within the hospital stay.

The CMP has received approval from the Patient Information Advisory Group to hold patient identifiable information without consent (approval number PIAG 2-10(f)/2005). Approval by an Institutional Review Board was not required.

Measures of Model Performance. Model performance was assessed using a combination of techniques recommended by an expert statistical steering committee (as listed in the Acknowledgments). Discrimination was assessed with the *c* index (21), equivalent to the area under the receiver operating characteristic curve (22). Overall accuracy was summarized by the *R* statistic from Shapiro's *Q* (23) (geometric mean probability assigned to the event that occurred) and Brier's score (24) (mean square error between outcome and prediction) and its associated calibration test (25) and decomposition (26). Calibration was assessed by the Hosmer-Lemeshow goodness-of-fit statistic (27) and Cox's calibration regression (28) (linear recalibration of the predicted log odds). These techniques have previously been used to evaluate the published risk prediction models in the same cohort of admissions, where full details of each technique were presented (16).

Model Development. Full details of the model development have been published in an online supplement (29). A brief outline is given next.

Different approaches to modeling blood pressure, respiratory rate, oxygenation, acid base disturbance, creatinine, and neurologic status were compared in a development sample (random two thirds of ICUs) and assessed in a validation sample (remaining one third of ICUs). The best model for each was included in the full physiology model, together with all other physiologic variables from APACHE II, APACHE III, SAPS II, and MPM II. A parsimonious physiology score was developed by simplification of the full model. Adjacent categories within each variable were combined if the difference in risk was not significant ($p > .1$). Variables were then removed from this simplified model in a stepwise manner until none remained. At each step, the model was fitted in the development sample and assessed in the validation sample, with the "best" model chosen to balance parsimony against predictive ability. The full stepwise selection procedure was repeated in 1,000 bootstrap samples to assess the stability of the selected variables (30). For variables recorded as both a highest and lowest value, either the highest or lowest was selected when the effect on predictive ability was minimal. Coefficients for the final physiology model were estimated in the full data, and the model was converted to an integer score by multiplying the predicted log odds of mortality by a constant and rounding, an approach that has been used in previous models (8, 12, 13). The purpose of this is to provide a simple summary of the degree of physiologic derangement, and the constant was chosen to ensure that the final score would have a possible range of zero to 100.

The following additional factors with a well-established association with mortality were investigated to find the best modeling approach: age, past medical history, and source of admission to ICU/surgical urgency. The full model consisted of the selected models for these factors, plus physiology score, gender, cardiopulmonary resuscitation within 24 hrs before ICU admission, and the body systems from the ICNARC Coding Method for primary reason for admission (31) plus interactions with surgical status. Each individual diagnostic category from the ICNARC Coding Method and an interaction with the physiology score was entered one by one into the full model and retained if $p < .001$. This stringent cutoff was chosen to ensure that diagnostic categories were only included if the coefficient could be accurately estimated. The full model was simplified by stepwise backward elimination in the same manner as the physiology score, and a final model was selected based on predictive ability in the validation sample.

Model Validation. The final model was refitted in 200 repeated development samples and assessed in the corresponding validation samples. This cross-validation ensures that the model is fitted and validated in separate data sets while avoiding the potential for misleading results based on a single random split into development and validation samples.

The model was compared with the best recalibrated versions of APACHE II, APACHE III, SAPS II, and MPM II, refitted in the 200 development samples, in the subset of admissions that were not excluded from any of the models.

The final coefficients for the ICNARC model were estimated in the full data set and shrunk by Efron's .632 bootstrap method to adjust for overfitting (32).

Analyses were performed in Stata 8.2 (StataCorp LP, College Station, TX). Models were fitted with logistic regression using robust (Huber-White) standard errors clustered by ICU due to the hierarchical nature of the data (30).

RESULTS

Data. Validated data on 231,930 admissions to 163 critical care units between December 1995 and August 2003 were available for analysis. Excluding 4,857 (2.1%) admissions missing hospital outcome and 10,447 (4.6%) readmissions of the same patient during the hospital stay, a total of 216,626 admissions were analyzed. No further exclusions were applied. Details of the case mix of this population have been reported previously (16).

Model Development. The following modeling approaches were selected: categories of extreme systolic blood pressure, categories of respiratory rate from APACHE III (removing the condition that ventilated respiratory rates between 6 and 13 are not weighted), categories of Pao_2/Fio_2 from SAPS II and interaction with ventilation status, categories of arterial pH and associated $Paco_2$, categories of creatinine from APACHE II (without doubling the weighting for acute renal failure). Neurologic status was modeled with 13 categories for individual Glasgow Coma Scale values from 3 to 15 (assessed during the first 24 hrs following admission to ICU) and two additional categories for patients who were sedated or paralyzed and sedated for the entire of the first 24 hrs. This approach considerably outperformed all methods from the pub-

Table 2. Intensive Care National Audit & Research Centre physiology score

Highest heart rate									
min^{-1}	≤39	40–109	110–119	120–139	≥140				
Score	14	0	1	2	3				
Lowest systolic BP									
mm Hg	≤49	50–59	60–69	70–79	80–99	100–179	180–219	≥220	
Score	15	9	6	4	2	0	7	16	
Highest temperature									
°C	≤33.9	34–35.9	36–38.4	38.5–40.9	≥41				
Score	12	7	1	0	5				
Lowest respiratory rate									
min^{-1}	≤5	6–11	12–13	14–24	≥25				
score	1	0	1	2	5				
Pao_2/Fio_2 ratio[a] (ventilation)									
mm Hg	≤99 (NV)	100–199 (NV)	≥200 (NV)	≤99 (V)	100–199 (V)	≥200 (V)			
Score	6	3	0	8	5	3			
Lowest arterial pH									
pH	≤7.14	7.15–7.24	7.25–7.32	7.33–7.49	≥7.50				
Score	4	2	0	1	4				
Highest serum urea									
mmol L^{-1}	≤6.1	6.2–7.1	7.2–14.3	≥14.4					
Score	0	1	3	5					
Highest serum creatinine									
mg dL^{-1}	≤0.5	0.6–1.4	≥1.5						
Score	0	2	4						
Highest serum sodium									
mmol L^{-1}	≤129	130–149	150–154	155–159	≥160				
Score	4	0	4	7	8				
Urine output (24 hrs[b])									
mL	≤399	400–599	600–899	900–1499	1500–1999	≥2000			
score	7	6	5	3	1	0			
Lowest WBC									
$\times 10^{-9}$ L^{-1}	≤0.9	1–2.9	3–14.9	15–39.9	≥40				
Score	6	3	0	2	4				
Sedated, paralyzed or GCS									
Value	S	P	3	4	5	6	7–13	14	15
Score	5	6	11	9	6	4	2	1	0

BP, blood pressure; NV, not ventilated at any time; V, ventilated at some time, during first 24 hrs or entire stay if <24 hrs; WBC, white blood cell count; GCS, Glasgow Coma Scale score; S, sedated; P, paralyzed and sedated, for whole of first 24 hrs or entire stay if <24 hrs.

[a]From arterial blood gas with lowest Pao_2; [b]for admissions staying <24 hrs, urine output from entire stay scaled to represent a 24-hr measurement.

lished models (*c* index 0.694 vs. 0.568 for APACHE II and III, 0.611 for SAPS II, 0.588 for MPM II_0, 0.598 for MPM II_{24}) and also outperformed the model using presedation Glasgow Coma Scale for sedated patients (*c* index 0.626).

The full physiology model had a *c* index of 0.826 and Hosmer-Lemeshow *C** statistic (for 20 groups based on quantiles of predicted mortality) of 263 in the validation sample (Table 1). Simplification of the full model by combining adjacent categories resulted in no loss of discrimination and a slight improvement in calibration. Following backward elimination, the "best" model contained 12 physiologic variables (vital signs, urine output, Glasgow Coma Scale score, and laboratory values) and 63 variables (*c* index 0.823, Hosmer-Lemeshow *C** 279). On bootstrapping the backward elimination, 10 of the 12 selected physiologic variables were included in the last 12 variables in all 1,000 bootstrap samples with the other two found in around 75% of samples. All physiologic variables recorded as both a lowest and highest value were replaced with either the lowest or highest value with little loss of performance (*c* index 0.822, Hosmer-Lemeshow *C** 333). The final physiology score resulting from this modeling is shown in Table 2. The mean (SD) physiology score in the full data set was 18.5 (10.0), with median (interquartile range) 17 (11–24).

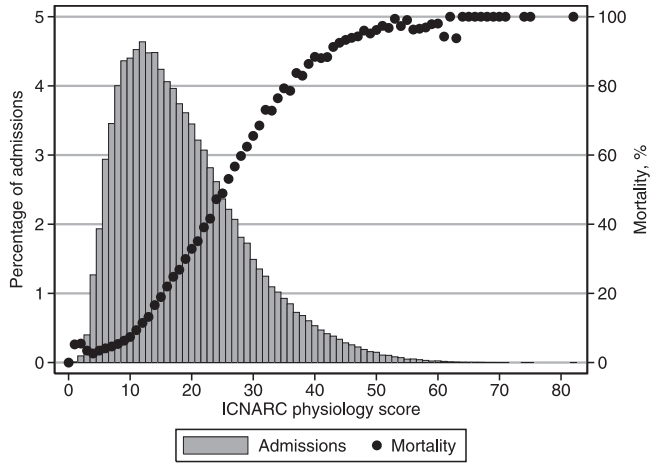

Figure 1. Distribution of Intensive Care National Audit & Research Centre (*ICNARC*) physiology score and observed mortality in the full data set (n = 216,626).

Although the maximum possible physiology score is 100, the highest observed was 82 (Fig. 1). The relative weightings of physiologic variables therefore remained fixed in all further models.

A linear model was selected for age. Past medical history was modeled in five categories: liver, cardiovascular, respiratory, renal, and immunocompromised. The original nine categories for source of admission were combined into the following six:

1. Elective surgery: Admissions from theater in the same hospital (direct or via accident and emergency) with a classification of surgery (according to the definitions of the National Confidential Enquiry into Perioperative Death) of elective or scheduled
2. Emergency surgery: Admissions from theater in the same hospital with a classification of surgery of emergency or urgent
3. Ward: Admissions from the ward or an intermediate care area (where the level of care is greater than the normal ward but not an ICU or HDU) in the same hospital
4. Critical care transfer: Admissions from an ICU or HDU in the same or another hospital (direct or via accident and emergency)
5. Accident and emergency/other hospital: Admissions from accident and emergency in the same hospital (not following surgery or transfer from an ICU or HDU) or from any location in another hospital except ICU or HDU
6. Clinic or home: Admissions from an outpatient or other clinic or directly from the community without being admitted to any other part of the hospital

For admissions from recovery (not having undergone a surgical procedure) or from diagnostic/imaging areas, the prior location was used to allocate the patient to one of the preceding categories.

Of 709 individual diagnostic categories from the ICNARC Coding Method represented in the data, 67 nonsurgical diagnoses (plus 19 interactions with the physiology score) and 34 surgical diagnoses (plus four interactions) were included in the model. These diagnoses accounted for 54.7% of admissions, with the remaining 45.3% allocated coefficients according to body system.

The full model had a *c* index of 0.874 and a Hosmer-Lemeshow *C** statistic of 74.3 in the validation sample. Table 3 shows the process of backward elimination: Gender and all variables relating to past medical history were eliminated with a minimal decrease in discrimination (*c* index 0.872) and a small improvement in calibration (Hosmer-Lemeshow *C** 68.2) in the validation sample. The final model included physiology score, age, diagnostic category coefficients and interactions with the physiology score, cardiopulmonary resuscitation within 24 hrs before admission, and source of admission.

Model Validation. The performance of the final model in 200 repeated validation samples is summarized in Table 4 and compared with the recalibrated APACHE II, APACHE III, SAPS II, and MPM II (16) in 141,106 admissions eligible for all models in Table 5. The ICNARC model outperformed all other models in terms of discrimination (*c* index), accuracy (Shapiro's *R*, Brier's score), unnecessary variability in predictions (lowest excess variance and test statistic for $\beta = 1|\alpha$ in Cox's calibration regression), and highest covariance between outcomes and predictions. SAPS II performed better than the ICNARC model in tests of perfect calibration (Spiegelhalter's *Z* and Hosmer-

Table 3. Calibration and discrimination in the validation sample of a progressively simplified full model

Model	*df*[a]	*c* Index	*C**[b]
Full	154	.874	74.3
Variables removed			
Chronic renal replacement	153	.874	74.3
Gender	152	.874	74.3
Chronic cardiovascular disease	151	.874	74.0
Chronic respiratory disease	150	.874	72.1
Chronic liver disease	149	.873	67.9
Immunocompromised	148	.872	68.2
Source of admission	143	.870	87.8
CPR prior to admission	142	.868	104
Diagnostic category interactions	119	.867	162
Diagnostic category coefficients	2	.840	1519
Age	1	.822	3289
Physiology score	0	.500	31,250

CPR, cardiopulmonary resuscitation. Bold type indicates the variables included in the final model and the fit of the final model.

[a]Degrees of freedom (number of model variables minus one); [b]Hosmer-Lemeshow chi-square statistic for 20 equal-sized groups from quantiles of full model predictions.

Table 4. Performance of the final model in 200 repeated validation samples (random one third of units) following refitting in the corresponding development sample

	Ideal Value	Observed Value
Average predicted mortality probability, mean (SD)	0.313	0.313 (0.008)
c index, mean (SD)	1	0.870 (0.003)
Shapiro's *R,* mean (SD)	1	0.665 (0.004)
Brier's score and derivatives		
Brier's score, mean (SD)	0	0.132 (0.002)
Spiegelhalter's Z statistic,[a] median (IQR)	0	1.58 (−0.39 to 3.73)
Accuracy of the average prediction, mean (SD)	0	5.6×10^{-5} (8.6×10^{-5})
Excess variance of predictions, mean (SD)	0	1.57 (0.04)
Covariance of outcome and prediction, mean (SD)	0.215	0.084 (0.002)
Hosmer-Lemeshow goodness-of-fit		
*C** statistic (20 equal size groups),[b] median (IQR)	0	82.6 (65.8 to 108)
*H** statistic (20 equal width groups),[b] median (IQR)	0	80.9 (63.5 to 105)
Cox's calibration regression		
Error in intercept (α), mean (SD)	0	-2.1×10^{-3} (0.060)
Error in slope (β-1), mean (SD)	0	-4.6×10^{-3} (0.020)
Test hypothesis: $\alpha = 0$, $\beta = 1$,[c] median (IQR)	0	20.1 (8.04 to 41.9)
Test hypothesis: $\alpha = 0 \mid \beta = 1$,[d] median (IQR)	0	12.3 (3.32 to 37.1)
Test hypothesis: $\beta = 1 \mid \alpha$,[d] median (IQR)	0	3.31 (0.64 to 7.28)

IQR, interquartile range.

[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 *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. Mean (SD) sample size of validation samples was 72,706 (9717).

Table 5. Performance of the final model compared with the best recalibrated versions of existing models assessed in admissions eligible for all models (n = 141,106)

Model	Ideal Value	ICNARC Model	APACHE II	APACHE III	SAPS II	MPM II
Mortality, n (%)	31.1 (0.9)	31.1 (0.8)	31.0 (0.8)	31.2 (1.0)	31.1 (0.8)	**31.1 (0.8)**
c index	1	**0.863 (0.003)**	0.832 (0.004)	0.845 (0.004)	0.840 (0.004)	0.824 (0.004)
Shapiro's *R*	1	**0.656 (0.005)**	0.633 (0.005)	0.644 (0.005)	0.640 (0.005)	0.629 (0.004)
Brier's score and derivatives						
Brier's score	0	**0.136 (0.003)**	0.150 (0.003)	0.143 (0.003)	0.145 (0.003)	0.152 (0.003)
Spiegelhalter's Z-statistic,[a] median	0	1.38	1.89	1.59	**0.47**	1.23
Accuracy of the average prediction	0	8.2×10^{-5} (9.6×10^{-5})	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})**
Excess variance of predictions	0	**1.67 (0.05)**	2.34 (0.08)	2.03 (0.07)	2.10 (0.06)	2.45 (0.07)
Covariance of outcome and prediction	0.213	**0.082 (0.002)**	0.064 (0.002)	0.071 (0.002)	0.069 (0.001)	0.062 (0.001)
Hosmer-Lemeshow goodness-of-fit statistics						
*C** (20 equal-sized groups),[b] median	0	64.2	140.2	80.0	**52.0**	92.7
*H** (20 equally spaced cut points),[b] median	0	62.4	146.9	85.2	**46.7**	91.3
Cox's calibration regression						
Error in intercept (α)	0	-3.7×10^{-3} (0.07)	-1.0×10^{-3} (0.08)	-9.5×10^{-3} (0.07)	**-5.4×10^{-4} (0.07)**	-2.7×10^{-3} (0.06)
Error in slope (β-1)	0	-7.7×10^{-3} (0.02)	-9.1×10^{-3} (0.03)	-1.1×10^{-2} (0.03)	-3.6×10^{-3} (0.02)	**-3.2×10^{-3} (0.02)**
Test hypothesis: $\alpha = 0$, $\beta = 1$,[c] median	0	20.5	23.7	23.5	20.0	**18.6**
Test hypothesis: $\alpha = 0 \mid \beta = 1$,[d] median	0	15.8	19.1	15.9	13.8	**13.5**
Test hypothesis: $\beta = 1 \mid \alpha$,[d] median	0	**1.61**	3.10	3.24	2.15	2.00

ICNARC, Intensive Care National Audit & Research Centre; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Prediction Model. Bold type indicates closest to ideal value.

[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 *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. Values are mean (SD) over 200 repeated validation samples, unless otherwise stated. Mean (SD) sample size of validation samples was 46,589 (6924).

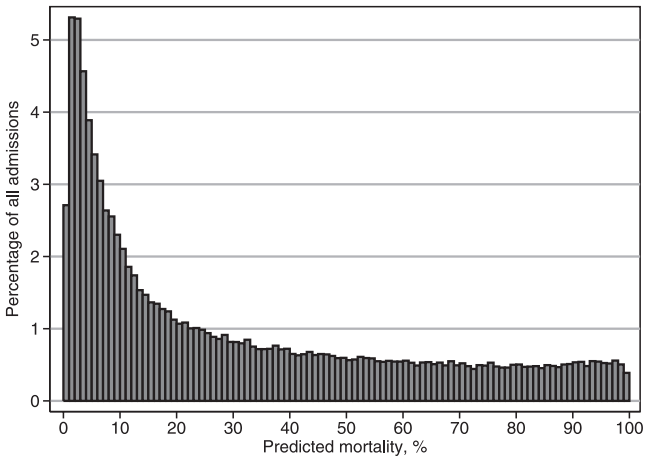

Figure 2. Distribution of predicted mortality from the final model in validation sample 1 (n = 79,526).

Lemeshow *C** and *H**); however, the median *Z* statistic for the ICNARC model did not indicate a significant lack of calibration. MPM II had the best accuracy of the average prediction; however, this measure was close to zero for all models. Similarly, both SAPS II and MPM II had mean values of α and β from Cox's calibration regression closer to the ideal values of 0 and 1, but the values for the ICNARC model did not indicate significant bias.

Figure 2 shows the distribution of predicted mortality, and Figure 3 shows the calibration curve for the final model in the validation sample. Coefficients for the ICNARC model are freely available for research purposes via the ICNARC Web site (33).

Worked Examples. Table 6 shows two worked examples of applying the ICNARC model, together with comparative mortality predictions from the other models (requiring additional data not shown).

DISCUSSION

We have developed a new risk prediction model for UK critical care, the

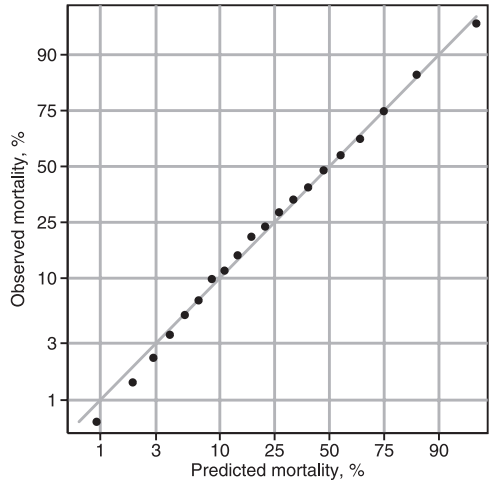

Figure 3. Calibration plot for the final model in validation sample 1 (n = 79,526). Observed mortality vs. predicted mortality from the final model in 20 equal-sized groups based on quantiles of predicted mortality. Diagonal line indicates perfect calibration. Axes drawn on a log odds scale.

Table 6. Worked examples

	Case 1		Case 2	
	Measurement	Score	Measurement	Score
Highest heart rate	97 min^{-1}	0	135 min^{-1}	2
Lowest systolic BP	78 mm Hg	4	66 mm Hg	6
Highest temperature	39.4°C	0	38.0°C	1
Lowest respiratory rate	17 min^{-1}	2	9 min^{-1}	0
Mechanical ventilation?	No		Yes	
Lowest Pao_2	98 mm Hg		61 mm Hg	
Associated FIo_2	0.30		0.42	
Pao_2/FIo_2	327 mm Hg	0	145 mm Hg	5
Lowest pH	7.34	1	6.92	4
Highest serum urea	7.9 mmol L^{-1}	3	Not recorded	0
Highest serum creatinine	1.6 mg/100 mL	4	1.3 mg/100 mL	2
Highest serum sodium	137 mmol L^{-1}	0	154 mmol L^{-1}	4
Urine output (24 hrs)	1983 mL	1	648 mL	5
Lowest WBC	6.7×10^9 L^{-1}	0	2.0×10^9 L^{-1}	3
Paralyzed/sedated?	No		Sedated	
Lowest GCS	15	0	N/A	5
ICNARC physiology score		15		37
Age, yrs	74		54	
Source of admission	Elective surgery		A&E/other hospital	
Diagnostic category	Surgical: aortic aneurysm		Nonsurgical: respiratory	
CPR	No		Yes	
Predicted mortality, %				
ICNARC model	10.5		87.6	
APACHE II	15.5		52.7	
APACHE III	6.2		56.6	
SAPS II	11.7		91.9	
MPM II	3.9		27.5	

BP, blood pressure; WBC, white blood cell count; GCS, Glasgow Coma Scale score; N/A, not applicable; A&E, accident and emergency; CPR, cardiopulmonary resuscitaton; ICNARC, Intensive Care National Audit & Research Centre; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Prediction Model.

ICNARC model, based on the best elements of the existing models and further research into modeling techniques. When restricted to a cohort of admissions eligible for all models, the ICNARC model outperformed existing models on measures of model performance including discrimination and overall fit. In addition, the ICNARC model has no exclusions, allowing it to be applied to all critical care admissions regardless of age, diagnosis, or length of stay. Performance of the ICNARC model was better when applied to all critical care admissions than when restricted to those eligible for all models.

The great strength of this study lies in the large, representative database in which the model was developed. The CMP database has been assessed according to published criteria for high-quality clinical databases and scored highly (20). It includes patients from all geographical regions of the target population (England, Wales, and Northern Ireland), with a representative spread of units in teaching and nonteaching hospitals and a representative, wide variation in the size of units. We can therefore expect the ICNARC model to perform well within this target population. There is no evidence that the ICNARC model would perform well outside of its target population; indeed, the evidence suggests that models fitted in one healthcare setting require recalibration before use in a different setting (16), and we would advise anyone wishing to use the ICNARC model outside of the UK to first validate the model and to recalibrate it if necessary. Statistical techniques, including robust variance estimation, cross-validation, and bootstrapping, were applied to ensure that the estimates of model performance are as accurate as possible. Nevertheless, some overfitting may remain as it was not possible to validate the model in an entirely independent sample.

A major advantage of the ICNARC model over previous risk prediction models is the elimination of exclusion criteria. Excluding groups of patients has the potential to bias risk-adjusted analyses (17). The relationship between age and outcome (adjusted for physiology score) was found to remain linear below the traditional cutoffs of 16 or 18 yrs, so this exclusion was considered unnecessary. However, our data and model are for admissions to adult ICUs; thus, we would not recommend that this model be used for analyses of admissions to pediatric units, where specific models exist. Rather, our model enables children admitted to adult units to be scored. Removal of exclusions based on particular diagnoses (e.g., cardiac surgery/burns) has been made possible by introducing interactions between diagnostic categories and the physiology score. The relationship between physiology and outcome was found to be stronger for certain groups, including coronary artery bypass graft surgery and burns, so that a small change in physiology could lead to a larger change in predicted outcome. For other groups (e.g., acute renal failure, chronic obstructive pulmonary disease),

a weaker relationship between physiology and outcome was found.

It was surprising that no variables relating to past medical history improved the model performance. It may be that the effects of these chronic conditions are reflected through the physiology or diagnosis. Alternatively, it may be that although specific chronic conditions are important prognostic factors for admissions with certain diagnoses, these effects are not consistent across all diagnostic groups. The variables relating to past medical history represented very severe chronic conditions (e.g., New York Heart Association functional class IV for cardiovascular disease) (34) and were therefore rare, ranging from 0.07% for acquired immunodeficiency syndrome to 3.2% for severe respiratory disease. There was also considerable variation across units in the reporting of past medical history conditions, which may reflect either genuine differences such as geographical variation or varying interpretation of the definitions. These differences across units may affect the ability of a model fitted in one group of units to validate in different units.

The burden of data collection was considered to be important when developing the model. For this reason, physiologic measurements were based on either the highest or the lowest value over 24 hrs, but not both. This simplifies the data collection by comparison with APACHE II, APACHE III, and SAPS II, which require the worst value (most accurately collected as the lowest and highest values) for certain physiologic variables.

Due to the nature of the data set, it was not possible to deviate significantly from the approach of previous models in basing predictions on extreme physiology measurements and diagnoses. We decided to retain the traditional approach of basing predictions on a physiology score produced by allocating points for physiologic derangement. Advances in computation techniques since the original models were developed mean that more sophisticated "black-box" methods are now possible (35–38), although these have often been found not to improve significantly on the logistic-regression-based approach, especially when the number of events far exceeds the number of covariates. Also, discussion with clinicians established that it was useful to have a score that can be calculated simply, summarizes the patient's physiologic condition, and reflects a methodology that is widely accepted and trusted. Investigation of these newer techniques in the CMP database may in the future produce a less transparent but more accurate model for use in research and comparative risk adjustment.

CONCLUSIONS

The ICNARC model is a more accurate model for predicting the risk of hospital mortality for admissions to adult, general critical care units in the UK than any of the published models. Even following recalibration of the published models, the ICNARC model demonstrated better discrimination and overall fit in the cohort of patients eligible for all models. The elimination of all exclusion criteria makes the ICNARC model a more reliable tool to use as the basis for risk-adjusted comparisons between critical care units. We recommend that it be used to replace previously published models for risk adjustment in the UK.

ACKNOWLEDGMENTS

Steering committee: Doug Altman, James Carpenter, Harvey Goldstein, Jon Nicholl, Gareth Parry, Patrick Royston, David Spiegelhalter. We thank Tony Brady for his early work on this project and Mike Kenward for his comments on the manuscript.

REFERENCES

1. Padkin A, Rowan K, Black N: Using high quality clinical databases to complement the results of randomized controlled trials: The case of recombinant human activated protein C. *BMJ* 2001; 323:923–926
2. Cook D, Heyland D, Marshall J, on behalf of the Canadian Critical Care Trials Group: On the need for observational studies to design and interpret randomized trials in ICU patients: A case study in stress ulcer prophylaxis. *Intensive Care Med* 2001; 27:347–354
3. Randolph AG, Guyatt GH, Carlet J: Understanding articles comparing outcomes among intensive care units to rate quality of care. *Crit Care Med* 1998; 26:773–781
4. Rubenfeld GD, Angus DC, Pinsky MR, et al and The Members of the Outcomes Research Workshop: Outcomes research in critical care: Results of the American Thoracic Society Critical Care Assembly Workshop on Outcomes Research. *Am J Respir Crit Care Med* 1999; 160:358–367
5. Wunsch H, Linde-Zwirble WT, Angus DC: Methods to adjust for bias and confounding in critical care health services research involving observational data. *J Crit Care* 2006; 21:1–7
6. Knaus WA, Zimmerman JE, Wagner DP, et al: APACHE—Acute Physiology and Chronic Health Evaluation: A physiologically based classification system. *Crit Care Med* 1981; 9:591–597
7. Knaus WA, Draper EA, Wagner DP, et al: APACHE II: A severity of disease classification system. *Crit Care Med* 1985; 13: 818–829
8. 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
9. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
10. Rowan KM: *Outcome Comparisons of Intensive Care Units in Great Britain and Ireland Using the APACHE II Method* [DPhil thesis]. Oxford, UK, University of Oxford, 1992
11. Le Gall JR, Loirat P, Alperovitch A, et al: A simplified acute physiology score for ICU patients. *Crit Care Med* 1984; 12:975–977
12. 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
13. Moreno RP, Metnitz PG, Almeida E, 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; 31:1345–1355
14. Lemeshow S, Teres D, Pastides H, et al: A method for predicting survival and mortality of ICU patients using objectively derived weights. *Crit Care Med* 1985; 13:519–525
15. 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
16. Harrison DA, Brady AR, Parry GJ, et al: Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med* 2006; 34: 1378–1388
17. 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
18. DoCDat: Directory of clinical databases. Available at: http://www.docdat.org. Accessed March 14, 2005
19. Livingston BM, Mackenzie SJ, MacKirdy FN, et al on behalf of the Scottish Intensive Care Society Audit Group: Should the pre-sedation 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
20. 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
21. Harrell FE, Califf RM, Pryor DB, et al: Evaluating the yield of medical tests. *JAMA* 1982; 247:2543–2546
22. Hanley JA, McNeil BJ: The meaning and use of the area under the receiver operating characteristics (ROC) curve. *Radiology* 1982; 143: 29–36
23. Shapiro AR: The evaluation of clinical predictions. *N Engl J Med* 1977; 296:1509–1514
24. Brier GW: Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 1950; 75:1–3
25. Spiegelhalter DJ: Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986; 5:421–433
26. Yates JF: External correspondence: Decomposition of the mean probability score. *Organ Behav Hum Perform* 1982; 30:132–156
27. Hosmer DW, Lemeshow S: Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics* 1980; A9:1043–1069
28. Cox DR: Two further applications of a model for binary regression. *Biometrika* 1958; 45: 562–565
29. Harrison DA, Parry GJ, Carpenter JR, et al: A new risk prediction model for critical care: The Intensive Care National Audit & Research Centre (ICNARC) model. Supplemental material: Details of modelling. Available at: http://www.icnarc.org/research/risk-prediction/risk-adjustment/icnarc-model. Accessed January 22, 2007
30. Harrell FE Jr: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, Springer-Verlag, 2001
31. 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
32. Efron B: Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc* 1983; 78:316–331
33. Coefficients for the Intensive Care National Audit & Research Centre (ICNARC) model. Available at: http://www.icnarc.org/audit/cmp/icmpds-resources/coefficients. Accessed January 22, 2007
34. The Criteria Committee of the New York Heart Association. Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels. Ninth Edition. Boston, Little, Brown, 1994
35. Dybowski R, Weller P, Chang R, et al: Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet* 1996; 347:1146–1150
36. Wong LS, Young JD: A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. *Anaesthesia* 1999; 54:1048–1054
37. Neumann A, Holstein J, Le Gall JR, et al: Measuring performance in health care: Case-mix adjustment by boosted decision trees. *Artif Intell Med* 2004; 32:97–113
38. Clermont G, Kaplan V, Moreno R, et al: Dynamic microsimulation to model multiple outcomes in cohorts of critically ill patients. *Intensive Care Med* 2004; 30:2237–2244