ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 491–501

Journal of Clinical Epidemiology

External validation of prognostic models for critically ill patients required substantial sample sizes

N. Peek[a,*], D.G.T. Arts[b], R.J. Bosman[c], P.H.J. van der Voort[c], N.F. de Keizer[a]

[a]*Department of Medical Informatics, Academic Medical Center – Universiteit van Amsterdam, Amsterdam, the Netherlands*

[b]*Austrian Health Institute, Vienna, Austria*

[c]*Department of Intensive Care, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands*

Accepted 23 August 2006

Abstract

Objective: To investigate the behavior of predictive performance measures that are commonly used in external validation of prognostic models for outcome at intensive care units (ICUs).

Study Design and Setting: Four prognostic models (Simplified Acute Physiology Score II, the Acute Physiology and Chronic Health Evaluation II, and the Mortality Probability Models II) were evaluated in the Dutch National Intensive Care Evaluation registry database. For each model discrimination (AUC), accuracy (Brier score), and two calibration measures were assessed on data from 41,239 ICU admissions. This validation procedure was repeated with smaller subsamples randomly drawn from the database, and the results were compared with those obtained on the entire data set.

Results: Differences in performance between the models were small. The AUC and Brier score showed large variation with small samples. Standard errors of AUC values were accurate but the power to detect differences in performance was low. Calibration tests were extremely sensitive to sample size. Direct comparison of performance, without statistical analysis, was unreliable with either measure.

Conclusion: Substantial sample sizes are required for performance assessment and model comparison in external validation. Calibration statistics and significance tests should not be used in these settings. Instead, a simple customization method to repair lack-of-fit problems is recommended.

Keywords: Prognostic models; Validation studies; Sample size; SAPS II; APACHE II; MPM II

1. Introduction

Prognostic models are important tools to provide estimates of patient outcome probabilities. Within the field of intensive care (IC) medicine, prognostic models are often used for mortality predictions which enable, for example, the stratification of patients for enrollment in clinical trials and controlling for severity of illness in auditing quality of care [1,2]. Four well-known prognostic models in IC are the Simplified Acute Physiology Score II (SAPS II) [3], the Acute Physiology and Chronic Health Evaluation II (APACHE II) [4] and the Mortality Probability Models II (MPM_0 II and MPM_{24} II) [5]. All four models are logistic regression models to predict the probability of in-hospital mortality. They use slightly different sets of covariates describing the demography (e.g., age), admission type (e.g., medical, urgent surgical) comorbidity (e.g., chronic dialysis, respiratory insufficiency), and worst physiological status of the patient in the first 24 hours of IC admission (e.g., highest body temperature, lowest blood pressure), or, in case of the MPM_0 II, in the first hour of IC admission. In the Appendix, a more extensive description of the four models is given.

In many countries, regional or national registries have been established that use one or more of these four prognostic models to audit the quality of IC medicine [6,7]. One example is the National Intensive Care Evaluation (NICE) registry that aims to assess and improve the quality of intensive care units (ICUs) in the Netherlands [8]. Because the IC prognostic models were developed 20 years ago on other (American or European) patient populations than those to which they are applied now, their generalizability must be assessed before the models can be used in clinical practice [2,9]. Therefore, many studies have been published on validating and comparing these models in external settings, with the aim of choosing the best

* Corresponding author. Tel.: 31 20 5667872; fax: 31 20 6919840.
E-mail address: n.b.peek@amc.uva.nl (N. Peek).

0895-4356/07/$ – see front matter
doi: 10.1016/j.jclinepi.2006.08.011

performing model and to assess its performance. These studies commonly focus on measuring the models' discrimination using the area under the Receiver Operating Characteristic (ROC) Curve [10], and their calibration using the Hosmer–Lemeshow goodness-of-fit statistics [11].

The results of these studies vary considerably. Whereas one study [12] concludes that the discrimination of the SAPS II model is superior to that of the APACHE II model, another [13] does not find a difference in discriminative ability between the two models. Similarly, some studies conclude that based on the measured calibration the SAPS II model is insufficient [14,15], whereas another concludes that calibration of the SAPS II is sufficient [16]. The variation in these results might be caused by temporal or geographical differences between the data sets that were used. However, the variation in results might also be caused by random variation in the validation samples. Estimates of predictive performance in external data (i.e., sampled from a different population than the data that was used to derive the model) are known to be highly variable [17]. The numbers of observations in the data sets that were used in the studies mentioned above vary widely, from 300 to 16,000.

The goal of this study was to validate and compare the performance of the APACHE II, SAPS II, the MPM_0 II, and the MPM_{24} II models on a large data set from the Dutch NICE registry. To put the historical external validation studies of the four prognostic models into perspective, we also investigated the influence of sample size on the validation results. To this end, the validation process was repeated with smaller data sets that were randomly drawn from the NICE registry.

2. Methods

2.1. Data

In 1996, the Dutch NICE foundation has started the (voluntary) registration of data of admissions to Dutch ICUs. The NICE registry database contains for each ICU admission 108 demographic, diagnostic, and physiologic variables collected within the first 24 hours of ICU admission and outcome data, such as length of stay on ICU and in-hospital mortality.[1] Data collected include all raw data values necessary to calculate the original SAPS II [3], APACHE II [4], MPM_0 II, and MPM_{24} II [5] mortality probabilities. APACHE II, SAPS II, MPM_0 II, and MPM_{24} II mortality probabilities are calculated in the national database at the NICE data coordinating center. Stringent measures are taken to control the data quality and uniformity of data collection procedures in the participating ICUs [18,19].

[1] http://www.stichting-nice.org

The data set used in this study consisted of data from 83,824 admissions to 29 Dutch ICUs between January 1, 1999 and December 31, 2003 registered in the NICE database. The developers of the APACHE II, SAPS II, MPM_0 II, and MPM_{24} II models have defined criteria for populations on which the models can be applied. We combined the criteria of all four models to obtain one data set that satisfied all criteria. According to the combined criteria we excluded patients aged <18, patients with an ICU length of stay <8 hours, acute coronary care and cardiac surgery patients, burn patients, readmitted patients, patients with missing severity-of-illness scores, and patients with missing (hospital) survival status. The characteristics of the remaining data set were compared to those used to develop the original APACHE II, SAPS II, MPM_0 II, and MPM_{24} II models.

2.2. Validation measures

2.2.1. Discrimination

The term *discrimination* refers to a model's ability to distinguish survivors from nonsurvivors. As a measure of discrimination we calculated the area under the ROC Curve [10]. This Area under the Curve (AUC; sometimes called C-index) is a normalized Mann–Whitney U statistic applied to the predictions by the model, grouped by observed outcomes. It represents the probability that an arbitrary patient who died had a higher predicted risk than an arbitrary patient who survived. An AUC of 0.5 indicates that the model does not predict better than chance. An AUC of 1 indicates that the model discriminates perfectly. Under the assumption that the distribution of AUCs is approximately Normal, we can compute the standard error of an estimated AUC [10].

For each pair of models (six in total), the difference in AUC was statistically tested with the nonparametric method of DeLong et al. [20]. The main problem in testing the difference between two AUC values that were computed on the same data set is that these values are highly correlated. The method of DeLong et al. solves this problem by estimating the correlation between the two values using the theory of generalized U statistics.

The AUC of a model depends only on the order of observations induced by its predictions and provides no indication of how close, on average, the predicted probabilities are to the observed outcomes. To take this aspect of a model's performance into account, we have to look at the *accuracy* and *calibration* of a model.

2.2.2. Accuracy

Accuracy refers to the difference between predicted risks and observed outcomes at the level of individuals. In this study, we applied the Brier inaccuracy score. The (*mean*) *Brier inaccuracy score*, also known as *mean squared error* or *mean probability score* [21,22], is calculated as

$$\overline{B} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \pi_i)^2, \quad (1)$$

where y_i is the observed outcome, and π_i is the risk of death predicted by the model for patient i, for each of n independent observations. So the score assigns a penalty to each individual prediction, based on the disagreement with the observed outcome, and computes the mean of the assigned penalties. If there is good agreement, the penalty is close to 0. If the agreement is very poor, the penalty is close to 1. Low values of $\overline{B}$ therefore indicate high accuracy of the predictions made by the model.

Also for the Brier score, a standard error can be computed using assumptions of normality [22]. Statistical comparison of the Brier scores from different models, however, raises even more problems than the comparison of AUC values. It is hardly ever performed in practice, and was neither done here.

2.2.3. Calibration

The concept of *calibration* refers to the agreement between observed and predicted risks. Because we cannot observe risks directly (we only know whether a patient died or not), calibration can only be measured indirectly. Two different approaches to measure calibration were used in this study; each of the approaches allows us to test the hypothesis that the model is well calibrated (''fits'') to the data.

The first approach was proposed by D. Cox [23,24]. It uses logistic regression to verify the agreement between predicted and observed risks, by fitting the equation

$$y_i \sim \beta_0 + \beta_1 \log\left(\frac{\pi_i}{1-\pi_i}\right) \quad (2)$$

to the data set. Again y_i is the observed outcome, and π_i is the risk of death predicted by the model, for patient i. In case of perfect fit of the model to the data, we will find that $\beta_0 = 0$ and $\beta_1 = 1$. Otherwise, the parameters $\beta_0$and β_1 indicate whether the π_i vary not enough ($\beta_1 > 1$) or too much ($\beta_1 < 1$), and when β_1 is fixed to 1, whether the average predicted risk is too low ($\beta_0 > 0 | \beta_1 = 1$) or too high ($\beta_0 < 0 | \beta_1 = 1$) [24]. The hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$ can be tested by comparing the fitted model of equation (2) to the original model using the likelihood ratio test [11]. The test statistic, $\tilde{F}$, follows a χ^2 distribution with two degrees of freedom; the test result states whether the fitted model is significantly better calibrated to the data than the original model.

The second approach that we used was proposed by D. Hosmer and S. Lemeshow [11], and builds on the calibration statistic $\hat{C}$. This statistic was designed to assess whether a given logistic regression model fits to a particular data set; it is commonly used during the model building phase, but also frequently in external validations. To compute the $\hat{C}$ statistic, the data set is partitioned into 10 equally sized subsets based on deciles of predicted risks, and the predicted and observed numbers of deaths are compared per subset:

$$\hat{C} = \sum_{g=1}^{10} \frac{(O_g - n_g\overline{\pi}_g)^2}{n_g\overline{\pi}_g(1-\overline{\pi}_g)}, \quad (3)$$

where O_g is the number of deaths in subset g, n_g is the number of observations in group g, and $\overline{\pi}_g$ is the mean of the predicted risks in group g. It has been shown that $\hat{C}$ follows a χ^2 distribution with eight degrees of freedom on the training sample (apparent performance), with nine degrees of freedom on an independent test set from the same population (internal validation), and with 10 degrees of freedom in external data sets [11]. A high value of $\hat{C}$ relates to a small P-value, implying rejection of the Null hypothesis that the model fits to the data.

Both calibration tests described above were applied to all four prognostic models, using a significance level of $\alpha = 0.001$. One disadvantage of statistical testing is that the decision of whether to reject the model will partially depend on the size of the data set: with (very) small data sets, the power to detect poor fit is lacking, whereas with (very) large data sets, even the smallest calibration problem will lead to a significant result. In addition to performing the statistical tests, we therefore compared calibration statistics to each other to determine which of the four models fitted best to the data set. As with the Brier score, statistical testing of differences in calibration of different models raises substantial methodological problems, and was therefore not performed here.

2.3. Customization of the models

When a predictive model calibrates poorly on an external data set, one may try to improve its performance by customizing the model to the data. Several strategies exist to do so [25,26]. For instance, one may choose to re-estimate a model's coefficients on the new data, and to add or remove terms from the model. A simpler customization strategy is to re-estimate the intercept and slope of the linear predictor, by fitting a new logistic regression equation with observed outcome as the dependent variable and the logit-transformed original predictions as the independent variable, as in equation (2); this has been termed ''logistic recalibration'' [27]. After estimating the coefficients β_0 and β_1, the new prediction for observation i becomes

$$\pi_i^{\text{new}} = \frac{e^{\text{Li}}}{(1+e^{\text{Li}})}, \quad (4)$$

where

$$L_i = \beta_0 + \beta_1 \log\left(\frac{\pi_i}{1-\pi_i}\right). \quad (5)$$

Of course, this strategy will only be effective when the model shows structural errors on the external data set, that is, when the average predicted risk is either to high or too low, or when the predicted probabilities either vary not enough or too much, and not when the calibration problems are more subtle. Nevertheless, this customization strategy was chosen here because of its parsimony and because it was earlier shown to be effective in the context of ICU prognosis [28]. We do note, however, that this type of customization will not change the order of the predictions, thus leaving AUC values unaffected.

All four models were customized to the data set, and their performance was assessed both before and after the customization using the measures of discrimination, accuracy, and calibration described above, with the exception of Cox calibration measure $\tilde{F}$; this measure has become useless after customization, as it will, by definition, consider a customized model to be perfectly calibrated. The Hosmer–Lemeshow $\hat{C}$ statistic was tested with 10 degrees before and nine degrees after customization.

Below, we will distinguish *original models* from *customized models* when referring the four models before and after customization.

2.4. Sampling procedure

It is known from various studies (e.g., Ref. [17]) that the precision of statistics for quantifying predictive performance is highly dependent on sample size, especially in external validation. With small sample sizes, the standard error of measured performance may be too large to draw reliable conclusions. To assess the required sample size in external validation studies in the IC domain, the following experiment was carried out. After assessing and comparing the performance of the four original and the four customized models on the entire data set, this procedure was repeated on a large number of subsamples of smaller sizes that were randomly drawn from the data set. Reflecting the different sizes of validation samples used in other external validation studies on IC prognosis, we drew subsamples containing data of 250, 500, 750, 1,000, 2,500, and 5,000 ICU admissions. ICU admissions for these data sets were randomly selected from the entire data set. The sampling process and the model evaluation and customization procedure were repeated 500 times for each sample size.

2.5. Statistical analyses

The results that were obtained on the entire data set acted as reference values in the interpretation and analysis of results that were obtained on subsamples; we will therefore speak of these results as the ''gold standard''. For each sample size, the following statistical analyses were performed on the results of the 500 iterations of the model evaluation procedure.

From the sampling results with respect to discrimination, we determined at each iteration whether the gold standard AUC for each model was contained in the 95% confidence interval that was estimated from the subsample in question. Furthermore, at each iteration, it was both determined which model had the highest AUC (direct comparison without statistical verification), and all six pair wise differences in AUC were statistically tested with the method of the DeLong et al. [20]. With the latter results we computed, for each pair of models, the frequency with which a wrong conclusion was drawn (significant result, but favoring the wrong model) and type II errors (lack of significant result) occurred. The results from comparisons that were made on the entire data set acted as gold standards to assess these errors. The significance level that was used to determine to this standard was 0.001; only those differences in AUC value that were significant at this extremely low level were considered to be ''real''. In the sampling procedure, differences were tested with a significance level of $\alpha = 0.05$.

With respect to accuracy, we again determined whether the gold standard Brier scores were contained in the 95% confidence intervals estimated from the subsamples. Furthermore, for each subsample, it was assessed which model had the lowest Brier score. From the 500 iterations, we computed the frequency with which the ''correct'' models had these superior scores, as compared to the gold standard.

From the sampling results with respect to calibration, we computed the frequency with which models were rejected and accepted according to both calibration tests, again using a significance level of 0.05. Furthermore, we computed the frequency with which the two calibration tests agreed on this decision. Finally, for each calibration statistic we computed the frequency with which the ''correct'' (gold standard) models displayed superior performance over the others.

All experiments and statistical analyses were conducted in S-plus Professional 6.2 (Insightful Corporation 2003).

3. Results

We first describe the results that were obtained on the entire data set, and present the results from the sampling procedure thereafter.

3.1. Results based on entire data set

After applying the combined exclusion criteria of the four prognostic models, a data set consisting of 42,139 ICU admissions remained. The number of ICU deaths within this data set was 5,478 (13.0%), and the number of hospital deaths was 8,354 (19.8%). Table 1 compares the data set used in this study to those used to develop the original models [3–5].

Table 1
Basic demographic data

	NICE	SAPS II	APACHE II	MPM II
No. of admissions	42,139	12,997	5,030	19,124
Hospital mortality (%)	19.8	21.8	19.7	20.8
Medical (%)	43.2	48.4	46.0	73.0
Unscheduled surgery (%)	17.4	19.6	54.0	73.0
Scheduled surgery (%)	39.4	31.2	54.0	27.0
Mean age survivors (years ± SD)	59.5 ± 16.9	57.2 ± 18.5		55.4
Mean age nonsurvivors (years ± SD)	66.86 ± 14.9			62.9
Male (%)	59.5	59.6		
Mean LOS ICU (days ± SD)	4.7 ± 9.6	6.6 ± 9.5		
Mean LOS hospital (excl preIC)	18.4 ± 26.1	19.1 ± 18.9		
SAPS II				
Mean score ± SD	33.3 ± 18.3			
Mean probability ± SD	0.22 ± 0.26			
APACHE II				
Mean score ± SD	16.0 ± 8.2			
Mean probability ± SD	0.24 ± 0.24			
MPM_0 II (mean probability ± SD)	0.20 ± 0.21			
MPM_{24} II (mean probability ± SD)	0.22 ± 0.22			

NICE data set compared to data sets used for development and evaluation of the original APACHE II, SAPS II, and MPM II prognostic models.

Table 2 shows the results of all six performance measures as applied on the original models and on their customized counterparts. As expected, customization did not affect discrimination of the models (AUC), and improved performance according to all other measures. For all performance measures except AUC, the results are optimistically biased after customization because customization and performance measurement were conducted on the same data set. Furthermore, because of its tight relation to the customization method, the Cox test considers calibration to be optimal after customization. In the table, we highlighted the best result for each performance measure, both for the original and the customized models. The models with the best results are henceforth called the ''gold standard best models''.

Both calibration tests deemed the fit of all four models to the data insufficient ($P < 0.001$), and the Hosmer–Lemeshow test still rejects all four models after customization. Nevertheless, there is considerable variation in measured calibration before and after customization. Before customization, calibration of the MPM_{24} II was considered superior to the other three models by both methods, whereas after customization, the SAPS II model performs better. The improvement in calibration of the MPM_0 II model due to customization was relatively poor compared to the other models.

As appears from Table 2, the SAPS II model was superior in terms of discriminatory performance with an AUC of 0.831, followed by the MPM_{24} II (0.822), APACHE II (0.818), and MPM_0 II models (0.796). Table 3 displays the results from six pair wise statistical comparisons of AUC values with the method of DeLong et al. [20]. All differences in discriminatory performance, except the difference between the APACHE II and MPM_{24} II models, were statistically significant at the 0.001 level.

3.2. Results from the sampling procedure

For reasons of space restriction Figs. 1 and 2 display only the results for the SAPS II model; for the other three models, similar results were found. Fig. 1 displays the means and 95% empirical ranges of AUC and Brier score that were obtained in the sampling experiment; results are shown for both the original and customized versions of the model. The average number of events (deaths) in the subsamples increased from 20 ($n = 100$) to 990 ($n = 5000$). With small sample sizes, considerable variation is found for both performance statistics. After customization, there is a small optimistic bias in the estimated Brier score, which is higher with small samples; such a bias is not found for the AUC, because this performance measure is invariant under the type of customization used in this study.

For both AUC and Brier score, at each iteration, a 95% confidence interval was estimated from the subsample in question, and it was verified whether the gold standard value was contained in this interval (results not shown in the figure). For the AUC, the estimated interval contained the gold standard value in more than 98% of the cases (e.g., SAPS II: 98.4% at sample size 100 and 100% at sample size 5,000). For the Brier score, however, this occurred only in 85–95% of the cases, increasing with sample size (e.g., SAPS II: 87.2% at sample size 100 and 94.0% at sample size 5,000).

It was also determined, at each iteration, which model performed best according the AUC and Brier score, disregarding confidence bounds. In this direct comparison, the

Table 2
Results of the performance measurements of all models before and after customization, based on the entire data set ($n = 42{,}139$)

Validation measure		Model	Original model	Customized model
Discrimination	AUC ± S.D.	APACHE II	0.818 ± 0.005	0.818 ± 0.005
		SAPS II	0.831 ± 0.005	0.831 ± 0.005
		MPM_0 II	0.796 ± 0.005	0.796 ± 0.005
		MPM_{24} II	0.822 ± 0.005	0.822 ± 0.005
Accuracy	Brier score ± S.D.	APACHE II	0.125 ± 0.001	0.121 ± 0.001
		SAPS II	0.120 ± 0.001	0.117 ± 0.001
		MPM_0 II	0.129 ± 0.001	0.127 ± 0.001
		MPM_{24} II	0.120 ± 0.001	0.119 ± 0.001
Calibration	Cox $\tilde{F}$	APACHE II	859.1 *	0
		SAPS II	850.3 *	0
		MPM_0 II	276.6 *	0
		MPM_{24} II	169.4 *	0
	Hosmer-Lemeshow $\hat{C}$	APACHE II	881.3 *	38.2 *
		SAPS II	879.4 *	28.0 *
		MPM_0 II	371.1 *	128.1 *
		MPM_{24} II	206.1 *	44.7 *

* $p < 0.001$, ▫ = best performance.

superior discrimination (AUC) of the SAPS II model was found in less than 50% of the cases with sample sizes of 100 or 250 observations. As the sample size increased, the SAPS II model was more often correctly identified as the best discriminating model. With a sample size of 5,000 the SAPS II model was identified as the best model in 97% of the 500 runs. A similar pattern was found for the Brier score.

For the five pairs of models with a statistically significant difference in AUC value, these statistical comparisons were also repeated on subsamples (using a significance level of 0.05). Fig. 2 shows the results of the repeated comparisons of the APACHE II and SAPS II models. A wrong conclusion was favored by the data in 0.8% of the cases with a sample size of 100 observations, in 0.2% of the cases with sample sizes of 250, 500, and 750 observations, and in less than 0.05% of the cases with larger sample sizes. The percentage of correct results (i.e., where the SAPS II model could be proved to perform significantly better) improved from 6.0% (100 observations) to 73.6% (5,000 observations). In all remaining cases (decreasing from 93.2% at 100 observations to 26.4% at 5,000 observations), no significant difference in discriminatory performance was found, and therefore a type II error occurred.

Similar results were obtained for the other four comparisons (results not shown): the frequency of wrong conclusions never exceeded 1.6%, but the power to detect a difference was always below 80% with sample sizes smaller than 2,500, and still below 80% for two out of five comparisons (including the one shown in Fig. 1) with a sample size of 5,000 observations.

Figure 3 displays median values and 95% empirical ranges of the two calibration statistics, computed for the SAPS II model in the sampling procedure. Calibration statistics represent the strength of evidence against a proper fit of the model, and therefore increases with larger sample sizes. As appears from the figure, with

Table 3
Results from statistical comparisons of discriminatory performance measured by the AUC, performed on the entire data set ($n = 42{,}139$), for all six pairs of models

Model 1	AUC 1	Model 2	AUC 2	Difference	*P*-value
MPM_0 II	0.796	APACHE II	0.818	0.022	<0.0001
MPM_0 II	0.796	MPM_{24} II	0.822	0.026	<0.0001
MPM_0 II	0.796	SAPS II	0.831	0.035	<0.0001
APACHE II	0.818	MPM_{24} II	0.822	0.004	0.053
APACHE II	0.818	SAPS II	0.831	0.013	<0.0001
MPM_{24} II	0.822	SAPS II	0.831	0.009	<0.0001

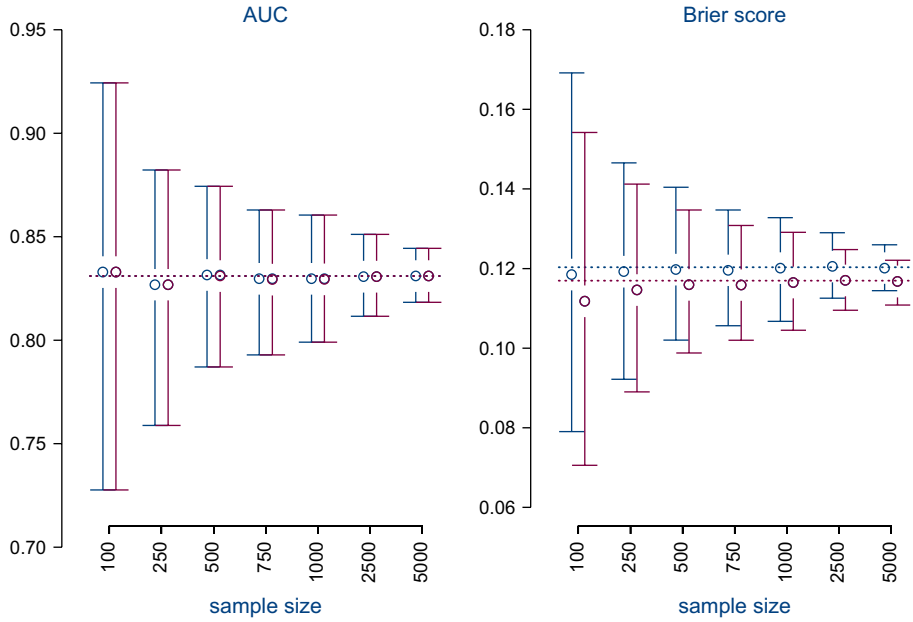

Fig. 1. Mean value and 95% empirical confidence interval of AUC and Brier score of the SAPS II model, across 500 random subsamples, at different sample sizes. Results are shown both before (left bars) and after (right bars) customization. The dotted horizontal lines display gold standard values, obtained from the entire data set ($n = 42{,}139$).

both calibration tests there is a strong tendency to reject the model when is has not been customized to the data. Cox's method, for instance, rejects the model in 50% of the cases with sample sizes of 250 observations (50 deaths), and in 99% of the cases with samples sizes of 1,000 observations (198 deaths); similar results are obtained with the Hosmer–Lemeshow test. Remarkably, the agreement in the decisions of both calibration tests is moderate, with frequencies of agreement rising from 78% (250 observations) to 86% (750 observations) and 93% (1,000 observations). After customization, the Hosmer–Lemeshow test accepts the model in the majority of cases (99% with a sample size of 250, 89% with a sample size 5,000).

Figure 4 shows per sample size the frequencies by which each of the four original models showed superior calibration over the others, in a direct comparison of Hosmer–Lemeshow $\hat{C}$ after customization on the subsamples. In the gold standard, the SAPS II model obtained the best (lowest) value for this statistic, closely followed by the APACHE II model. As appears from the figure, it is largely a matter of chance which model shows the best performance in the subsamples. Even the MPM_0 II model, whose calibration is clearly inferior to that of the other three models in the gold standard, has the best value in 15–20% of the cases when 1,000 observations (198 events) or less are used.

4. Discussion and conclusion

Several studies have shown that external validation of predictive models is necessary to verify their transportability to new sites [29,30]. In the field of intensive care, a considerable number of external validation studies have been performed on the four prognostic models that were validated here, for example, see Refs. [12–15], with varying results. In the current study, the geographical and temporal confounders causing variation were eliminated by drawing samples from one large database. Therefore, the effects measured in this study are merely caused by variation due to random sampling.

In our study, we compared the results that were obtained with random samples of different numbers of observations, ranging from one hundred to a few thousands, to reference results that were computed on a very large data set from the NICE registry, containing tens of thousands of observations. Considering the large number of observations in this reference data set and the small confidence intervals resulting from it, we believe that the reference results are highly precise estimates of prognostic performance on the Dutch ICU population, and that we may therefore consider them as a gold standard in the simulation procedure.

According to our gold standard, the differences in performance between the four models are small. The SAPS II model appears to be slightly superior over the others, especially when it is customized to the data set, and performance of the MPM_0 II model was clearly below the other three models. This is probably due to the fact that this model is based only on data that are available 1 hour after ICU admission, whereas the other models are all based on data that are available after the first 24 hours of ICU stay.

The APACHE II, SAPS II, MPM_{24} II, and MPM_0 II models were validated and compared using samples of varying sizes. With small sample sizes ($n = 100$, 250,

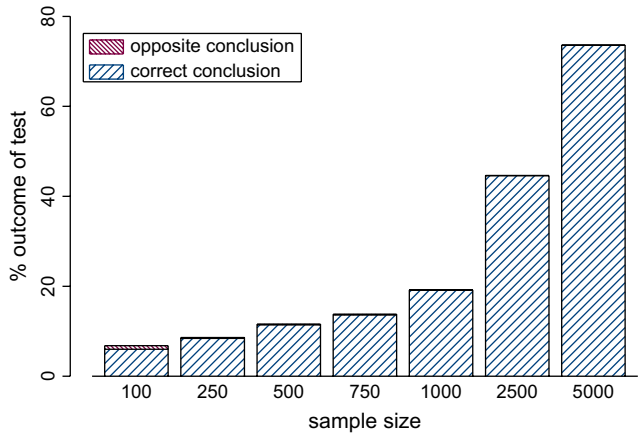

Fig. 2. Results from repeated statistical comparisons of discriminatory performance of the APACHE II and SAPS II models in the sampling procedure. The difference in AUC between both models was 0.013 on the entire data set, in favor of the SAPS II model, and this difference was significant at the 0.001 level. Each bar displays, for a given sample size, the percentage of cases (from 500 iterations) where this procedure favored SAPS II (correct conclusion) or APACHE II (opposite conclusion), when tested at a 0.05 significance level. For each sample size, the remaining fraction of cases represents the prevalence of type II errors (lack of power to detect a significant difference).

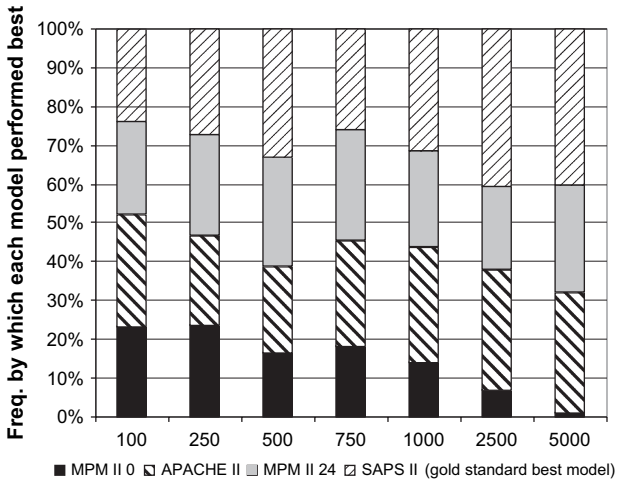

Fig. 4. Variation in apparent superior calibration, as measured by the Hosmer–Lemeshow $\hat{C}$ statistic, across 500 random subsamples. At each iteration, all four models were customized to the subsample. For each sample size, the frequency (y-axis) is shown by which each of the models (APACHE II, SAPS II, MPM_0 II, and MPM_{24} II) appeared to be the best calibrated model (lowest $\hat{C}$ statistic) after customization.

500, 750, 1,000; approximately 200 events or less), the estimated AUCs and Brier scores showed a large variation. This is in concordance with earlier findings on variability of performance estimates in external validation by Steyerberg et al. [17]. We therefore believe that performance assessment in terms of discrimination and accuracy is difficult to accomplish with these small numbers of observations. Furthermore, the estimated standard error of the Brier score often turns out to be too small. This means that we cannot trust that the true Brier score is contained in the estimated 95% confidence interval with 95% certainty. We conjecture that this problem is caused by the assumption of normality, which is known to be dubious for the Brier score. In contrast to the Brier score, the estimated standard error of the AUC seems larger than is actually needed.

Also the statistical test of differences in AUC by DeLong et al. [20] appears to be too conservative. As a result, the test is highly reliable, but its power to detect

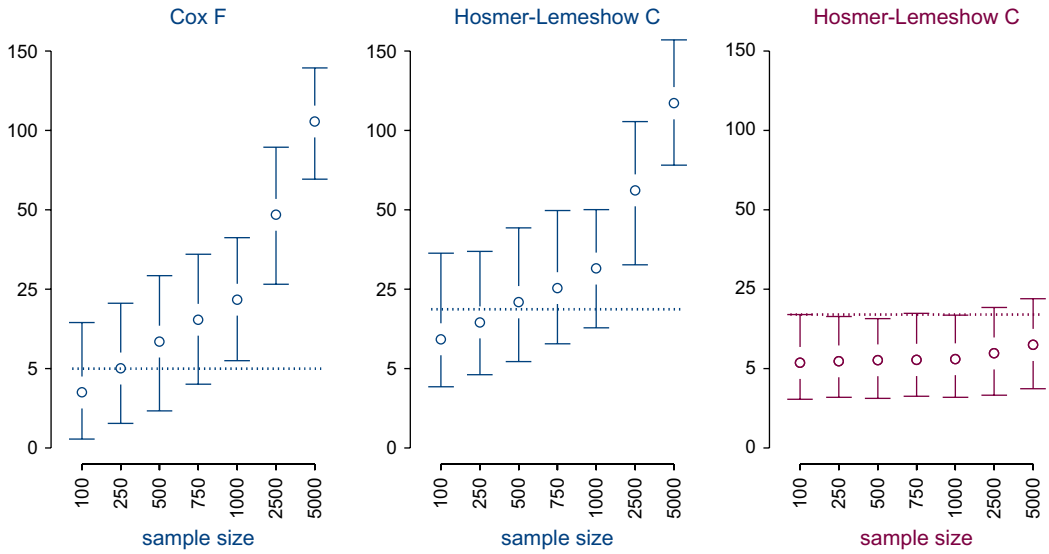

Fig. 3. Median value and 95% empirical confidence intervals of Cox's calibration statistic $\tilde{F}$ and Hosmer–Lemeshow $\hat{C}$ statistic for the SAPS II model, across 500 random subsamples, at different sample sizes. Both calibration statistics follow a χ^2 distribution; for ease of reference, a square-root scale was used on the y-axis. Cox's $\tilde{F}$ was computed only before customization (left panel); Hosmer–Lemeshow $\hat{C}$ was computed both before (middle panel) and after (right panel) customization of the model to the subsample. The dotted horizontal line displays the value above which the model is rejected ($\alpha = 0.05$).

differences in performance is small. In the field of IC prognosis, where the prevailing models are highly competitive, there seems to be no point in conducting this test with less than 2,500 observations (approximately 500 events) in external validations.

Because of the large sample size that is required in testing differences in AUC, one might consider comparing estimated AUCs directly, without statistical verification of the difference. This is also an option for Brier scores and calibration statistics, for which no statistical comparison methods exist in external settings (in internal settings this problem can be solved by bootstrapping). The results from our simulation indicate, however, that such a direct comparison is unlikely to yield a reliable answer when the performance differences are as small as is the case here. In our comparison, 5,000 observations (approximately 1,000 events) are needed to obtain 95% certainty of a correct answer for the AUC and Brier score. The reliability of comparing the Hosmer–Lemeshow $\hat{C}$ is even worse, especially after customizing the models to the data.

In contrast to the discrimination and accuracy measurements, the values of calibration statistics are highly sensitive to sample size. This is because calibration statistics represent evidential strength for a lack of fit of the model. In our study, calibration statistics appeared to increase quickly with increasing sample sizes. This was also found earlier by Zhu et al. [28] who examined the impact of hospital mortality on measured performance of the MPM_0 II and the MPM_{24} II models. Nevertheless, tests such as the one proposed by Hosmer and Lemeshow are often used instruments to detect calibration problems of logistic regression models. From the results that were found in this study, however, we tend to advice against using these methods in external validations. For both, the methods of Cox and Hosmer–Lemeshow, we found that the *P*-values are too closely related to sample size to provide reliable decisions on calibration, and test statistics vary too much to enable a reliable comparison between models.

The models that were validated in this study were developed in the late 1980s and early 1990s. Since then, the mortality among ICU patients has decreased considerably and all four models therefore over predict death in our data set. To eliminate the effects of over prediction, we applied a simple customization strategy to the models, which had a substantial effect on measured performance. After customization, all performance measures agreed on the superiority of the SAPS II model over the others, whereas the inferior performance of the MPM_0 II model became more obvious. We therefore believe that performance assessment of customized models is more valuable than that of the original models.

Our results confirm the findings of Vergouwe et al. [31] that calibration problems, such as over prediction of mortality, are difficult to detect in external validation, especially with small samples. In their study, it was found that testing the intercept and slope of the linear predictor separately is more powerful than the joint test (Cox $\tilde{F}$) that was applied here. Furthermore, they concluded that the Hosmer–Lemeshow goodness-of-fit test is much less powerful, and may be better used for model development than for validation purposes.

There are a number of limitations to our study. First, the validation was carried out on models for the IC domain, where the average mortality is around 20%. Although the models that were validated here are still commonly used in clinical practice, there exist models (e.g., APACHE III [32]) that were more recently developed for this domain. Data for these models are however not recorded in the NICE registry, and were therefore not contained in our data set. A second, more serious, limitation is that the differences in performance between the models, especially between the APACHE II, SAPS II, and MPM_{24} II models are small. Had these differences been more pronounced, then we would probably have obtained better (i.e., more powerful and reliable) results for comparisons with small samples. We do note, however, that the mere fact that these models are so competitive has led to ongoing discussions about their relative merits in the literature.

Third, we applied a relatively simple strategy, logistic recalibration, to customize the models to our data. A number of different customization strategies exist, some of which are more flexible than this strategy [25]. Logistic recalibration assumes that the relative strengths that were assigned to the model's covariates during its development are correct in the external setting. Nevertheless, various studies have shown this strategy to be very robust, and therefore preferable over more flexible model revision methods [26,28,33]. The fourth and final restriction is that customization and validation were carried out on the same sets, causing the performance estimates to be optimistically biased after customization. Because the number of estimated parameters in the customization is small (two), we believe that the resulting bias is small and does not influence our conclusions regarding the behavior of performance measurement in external validation.

To summarize, we conclude that the four models perform similarly on Dutch ICU data. Substantial sample sizes are however required for performance assessment and model comparison in external validation of these models. We believe that methods for calibration measurement and testing are not helpful in these settings. Instead, we recommend to apply the customization method that was used here as a standard procedure before applying a model in external settings.

Acknowledgment

Niels Peek receives a grant from the Netherlands Organisation of Scientific Research (NWO) under project number 634.000.020.

References

[1] Gunning K, Rowan K. ABC of intensive care: outcome data and scoring systems. BMJ 1999;319:241–4.

[2] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.

[3] Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270:2957–63.

[4] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985;13: 818–29.

[5] Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993;270:2478–86.

[6] Rowan K, Kerr J, Major E, McPherson K, Short A, Vessey M. Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. Crit Care Med 1994;22:1392–401.

[7] Project IMPACT CCM, Inc. http://www.cerner.com/piccm/. Accessed Aug 12, 2006.

[8] de Keizer NF, de Jonge E. National IC Evaluation (NICE): a Dutch quality control system. J ICU Management 2005;3:62–4.

[9] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.

[10] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143: 29–36.

[11] Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 2000.

[12] Moreno R, Morais P. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. Intensive Care Med 1997;23:177–86.

[13] Katsaragakis S, Papadimitropoulos K, Antonakis P, Strergiopoulos S, Konstadoulakis MM, Androulakis G. Comparison of Acute Physiology and Chronic Health Evaluation II (APACHE II) and Simplified Acute Physiology Score II (SAPS II) scoring systems in a single Greek intensive care unit. Crit Care Med 2000;28:426–32.

[14] Metnitz PG, Valentin A, Vesely H, Alberti C, Lang T, Lenz K, et al. Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Simplified Acute Physiology Score. Intensive Care Med 1999;25:192–7.

[15] Bertolini G, D'Amico R, Apolone G, Cattaneo A, Ravizza A, Iapichino G, et al. Predicting outcome in the intensive care unit using scoring systems: is new better? A comparison of SAPS and SAPS II in a cohort of 1,393 patients. GiViTi Investigators (Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva). Simplified Acute Physiology Score. Med Care 1998;36:1371–82.

[16] Auriant I, Vinatier I, Thaler F, Tourneur M, Loirat P. Simplified acute physiology score II for measuring severity of illness in intermediate care units. Crit Care Med 1998;26:1368–71.

[17] Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003;56:441–7.

[18] Arts D, de Keizer N, Scheffer GJ, de Jonge E. Quality of data collected for severity of illness scores in the Dutch National Intensive Care Evaluation (NICE) registry. Intensive Care Med 2002;28:656–9.

[19] Arts DG, Bosman RJ, de Jonge E, Joore JC, de Keizer NF. Training in data definitions improves quality of intensive care data. Crit Care 2003;7:179–84. Epub Feb 18, 2003.

[20] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.

[21] Brier G. Verification of forecasts expressed in terms of probability. Monthly Weather Rev 1950;78:1–3.

[22] Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis. III—Methods based on continuous functions of the diagnostic probabilities. Methods Inf Med 1978;17: 238–46.

[23] Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562–5.

[24] Miller ME, Hui SL, Tierney WL. Validation techniques for logistic regression models. Stat Med 1991;10:1213–26.

[25] Van Houwelingen HC. Validation, calibration, revision, and combination of prognostic survival models. Stat Med 2000;19:3401–15.

[26] Steyerberg EW, Borsboom GJJM, Van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23:2567–86.

[27] Harrell FE Jr. Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.

[28] Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study. Crit Care Med 1996;24:57–63.

[29] Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003;56:826–32.

[30] Terrin N, Schmid CH, Griffith JL, D'Agostino RB Sr, Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. J Clin Epidemiol 2003;56:721–9.

[31] Vergouwe Y, Steyerberg EW, Eijkemans R, Habbema D. Sample size considerations for the performance assessment of predictive models: a simulation study. Control Clin Trials 2003;24(2 Suppl):S43–4.

[32] Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA. Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Crit Care Med 1998;26:1317–26.

[33] Moreno R, Apolone G. Impact of different customization strategies in the performance of a general severity score. Crit Care Med 1997;25: 2001–8.

Appendix

Table 4 (viewable on the journal's website at www.elsevier.com/locate/jclinepi) shows all variables used in the four IC prognostic models. The APACHE II and SAPS II both use a number of variables to calculate a severity-of-illness score representing the imbalance of the patient's condition. The SAPS II severity-of-illness score is used in the SAPS II logistic regression model to predict in-hospital mortality as follows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -7.7631 + 0.0737(\text{SAPSscore}) + 0.9971[\log(\text{SAPSscore}+1)].$$

In the APACHE II model, this severity-of-illness score is used together with admission type and one of 54 main reasons for admission (see Table 5 for coefficients (viewable on the journal's website at www.elsevier.com/locate/jclinepi)) in a logistic regression model to predict the risk of in-hospital mortality:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -3.517 + 0.146(\text{APACHEscore}) + 0.603\text{emerg} + \beta^{\text{reason for admission}},$$

where *emerg* is a binary variable that indicates whether the patient had to undergo emergency surgery.

The MPM_0 II and MPM_{24} II models do not use a severity-of-illness score but use dichotomous covariates to predict the risk of in-hospital mortality; they are regular logistic regression models. The coefficients of their covariates are represented in Table 4.

Table 4
Description of covariates used in APACHE II, SAPSII, MPM_0 II, and MPM_{24} II

Variables	APACHE II	SAPS II	MPM_0 II	MPM_{24} II
Acute Diagnoses				
Acute renal failure	S		1.4821	
Cardiac dysrhytmia			0.28095	
Cerebrovascular incident			0.21338	
Gastrointestinal bleeding			0.39653	
Intracranial mass effect			0.86533	0.91314
Confirmed infection				0.49742
Chronic Diagnoses				
Chronic renal insufficiency			0.91906	
Chronic dialysis	S			
Metastatic neoplasm	S	S	1.9979	1.16109
AIDS	S	S		
Hematologic malignancy	S	S		x
Cirrhosis	S		1.13681	1.08745
Cardiovascular insufficiency	S			
Respiratory insufficiency	S			
Immune insufficiency	S			
Physiology				
Heart rate (⩾150)	S	S	0.45603	
Respiratory rate	S			
Systolic blood pressure (⩽90 mmHg)		S	1.06127	
Mean blood pressure	S			
Temperature	S	S		
Prothrombin time (>3 seconds)				0.55352
Urine output (<150 ml in 8 hours)		S		0.82286
PaO_2 (<60 mmHg)				0.46677
PaO_2/FIO_2		S		
PaO_2 or A-aDO_2	S			
PH	S			
White blood cell count	S	S		
Serum creatinine (2.0 mg/dL)	S			0.72283
Serum potassium	S	S		
Serum sodium	S	S		
Serum bicarbonate	S	S		
Serum urea		S		
Bilirubin		S		
Hematocrite	S			
Platelets				
Glasgow Coma Score (3−5)	S	S	1.48592	1.6879
Other variables				
Age (10-years periods)	S	S	0.03057	0.03268
Type of admission (nonelective surgery)	S	S	1.19098	0.83404
CPR prior to ICU admission			0.56995	
Mechanical ventilation			0.79105	0.80845
Vasoactive drug ⩾1 hour				0.71628

An ''S'' means that the covariate is part of the severity-of-illness score. In the columns of MPM_0 II and MPM_{24} II, the numbers represented are the regression coefficients. The dichotomization condition of continuous variables, used in the MPM_0 II and MPM_{24} II models, is mentioned between brackets.

Table 5
Regression coefficients for APACHE II reasons for admission

Reason for admission	Coefficient
Cardiovascular–Surgical	−0.797
Multiple trauma	−1.684
Heart valve surgery	−1.261
Peripheral vascular surgery	−1.315
Hemorrhagic shock	−0.682
Chronic cardiovascular disease	−1.376
Sepsis	0.113
After cardiac arrest	0.393
Gastrointestinal–Surgical	−0.613
Neoplasm	−0.248
Bleeding	−0.617
Perforation or obstruction	0.060
Hematological–Surgical	−0.196
Renal–Surgical	−0.196
Neoplasma	−1.204
Transplant	−1.042
Metabolic–Surgical	−0.196
Neurological–Surgical	−1.150
Craniotomy for neoplasm	−1.245
Head trauma	−0.955
Craniotomy for intracerebral subdural or subarachnoid hemorrhage	−0.788
Laminectomy and other spinal cord surgery	−0.699
Respiratory–Surgical	−0.610
Thoracic surgery for neoplasm	−0.802
Respiratory insufficiency after surgery	−0.140
After respiratory arrest	−0.168
Cardiovascular–Nonsurgical	0.470
Multiple trauma	−1.228
Coronary artery disease	−0.191
Thoracic or abdominal aneurysm	0.731
Congestive heart failure	−0.424
Hypertension	−1.798
Rhythm disturbance	−1.368
Cardiogenic shock	−0.259
Sepsis	0.113
Hemorrhagic shock or hypovolemia	0.493
After cardiac arrest	0.393
Gastrointestinal–Nonsurgical	0.501
Bleeding	0.334
Hematological–Nonsurgical	−0.885
Renal	−0.885
Metabolic–Nonsurgical	−0.885
Diabetic ketoacidosis	−1.507
Neurological–Nonsurgical	−0.759
Head trauma	−0.517
Drug overdose	−3.353
Intracerebral subdural or subarachnoid hemorrhage	0.723
Seizure disorder	−0.584
Respiratory–Nonsurgical	−0.890
Infection	0.0
Neoplasm	0.891
Pulmonary embolus	−0.128
Noncardiogenic lung edema	−0.251
After respiratory arrest	−0.168
Asthma or allergy	−2.108
Chronic obstructive lung disease	−0.367
Aspiration poisoning or toxic	−0.142
Laminectomy and other spinal cord surgery	−0.699

(Continued)