Acta Anaesthesiol Scand 2014; **58:** 701–708
Printed in Singapore. All rights reserved

ACTA ANAESTHESIOLOGICA SCANDINAVICA
doi: 10.1111/aas.12327

A calibration study of SAPS II with Norwegian intensive care registry data

Ø. A. Haaland[1], F. Lindemark[1,2], H. Flaatten[3,4], R. Kvåle[3,4] and K. A. Johansson[1,2]
[1]*Department of Global Public Health and Primary Care, University of Bergen, Departments of* [2]*Research and Development and* [4]*Anesthesia and Intensive Care, Haukeland University Hospital and* [3]*Norwegian Intensive Care Registry, Helse Bergen HF, Bergen, Norway*

Background: Mortality prediction is important in intensive care. The Simplified Acute Physiology Score (SAPS) II is a tool for predicting such mortality. However, the original SAPS II is poorly calibrated to current intensive care unit (ICU) populations because it draws on data, which is more than 20 years old. We aimed to improve the calibration of SAPS II using data from the Norwegian Intensive Care Registry (NIR). This is the first recalibration of SAPS II for Nordic data.
Methods: A first-level customization was applied to improve calibration of the original SAPS II model (Model A). NIR data used covered more than 90% of adult patients admitted to ICUs in Norway from 2008 to 2010 ($n = 30712$).
Results: The modified SAPS II, Model B, outperformed the original Model A with respect to calibration. Model B gave more accurate predictions of mortality than Model A (Hosmer–Lemeshow's C: 22.01 vs. 689.07; Brier score: 0.120 vs. 0.131; Cox's calibration regression: $\alpha = -0.093$ vs. -0.747, $\beta = 0.921$ vs. 0.735, $(\alpha \mid \beta = 1) = -0.009$ vs. -0.630). The standardized mortality ratio was 0.73 [95% confidence interval (CI) of 0.70–0.76] for Model A and 0.99 (95% CI of 0.95–1.04) for Model B. Discrimination was good for both models (area under receiver operating characteristic curve = 0.83 for both models).
Conclusions: As expected, Model B is better calibrated than Model A, and both models have similar uniformity of fit and equal discrimination. Introducing Model B into Norwegian ICUs may improve precision in decision-making. Units will have a more realistic benchmark for the assessment of ICU performance. Mortality risk estimates from Model B are better than previous SAPS II estimates have been.

Accepted for publication 24 March 2014

Mortality prediction is important in clinical decision-making, benchmarking, policy making, priority setting at a population health level, and sometimes also for individual patients. Reliable information on risk is a key foundation for high-quality decisions. In intensive care, severity scores have been used for about 30 years and have been important decision-making tools for clinicians and planners.[1] However, such scoring systems have weaknesses, for example, they are not suitable for predicting individual survival probabilities, and they need continuous improvement and updating.[2]

Intensive care mortality prediction models are statistical methods designed to predict hospital mortality on the basis of patient data collected during admission to an intensive care unit (ICU). Often, the emphasis is on the first 24 h of care. Since the 1980s, several models have been developed [for example, Acute Physiology, Age, Chronic Health Evaluation (APACHE),[3] Mortality Prediction Model (MPM),[4] and Simplified Acute Physiology Score (SAPS)[5]], and they have applications in several domains. They are used to assess the severity of illness of individual patients in overall evaluations of the quality of care provided by the ICU, both between units and within a single unit.[6] In clinical trials, mortality prediction models are used to ensure that patients are assigned randomly to their respective groups.[6] In priority setting, mortality prediction models may also be applied when studying trade-offs between expected health benefits and severity of disease. SAPS II will be the focus of this paper.

The important properties of any mortality prediction model are discrimination, uniformity of fit, and calibration. Discrimination measures the degree to which the model is able to assign high mortality probabilities to patients who die, and low probabilities to those who survive. SAPS II has been reported to provide very good discrimination.[7] Uniformity of

fit is good when patients with similar SAPS II scores also have similar hospital mortalities across a broad range of subgroups, such as age groups, gender, or types of admission. SAPS II has been found to have poor uniformity of fit across diagnostic groups.[5,8] Finally, a model is well calibrated if the predicted mortality rate is close to the observed mortality rate. Calibration for SAPS II has repeatedly been reported as inadequate.[9–13]

If calibration is poor, the quality of mortality predictions will be reduced. For example, policy decisions relying on such low-quality information may assume that ICU mortality is higher (or lower) than what it actually is. This may skew resource allocation away from the optimum. Also, benchmarking becomes difficult, which may influence how ICUs perceive their performance indicators. For example, ICUs are expected to perform better with time, so a poor ICU in 2010 may still perform much better than an excellent ICU did in 1990. Thus, customization of the SAPS II model should be performed regularly, preferably as often as every 2–3 years.[14] It should also be performed within as many settings as possible. The disease burden and access to technology vary across populations. For example, lack of respirators in low-income countries may highly impact ICU mortality. Also, if admissions due to drug and alcohol intoxication are more frequent in one ICU than in others, we may expect substantial differences in SAPS II adjusted mortalities. The calibration of mortality prediction models is therefore highly dependent on context.

Customization can be done at two levels. In second-level customization, the underlying variables are altered or differently weighted, giving entirely new SAPS II scores for each patient. In first-level customization, the SAPS II scores remain unchanged, but the equation converting these scores to mortality probabilities is modified. First-level customization is the most common approach. The results at this level are easy to use, as all one needs is a set of SAPS II scores and the new coefficients of Equation 1. A third approach to customization is to add new variables to the underlying model, or to use SAPS II scores as one of several explanatory variables in a different model. Investigators have recognized that the original SAPS II model was far too pessimistic in estimating the mortality rates of patients suffering from drug or alcohol intoxication.[15,16] As a consequence, new models were constructed where 'intoxication' was added as a variable. Other new variables, such as 'sex' and 'sequential organ failure assessment score' have also been added.

SAPS II scores are routinely recorded in the ICUs of many European countries, including Norway. Other prediction tools, such as APACHE and MPM, are not as commonly used in this region. Recently, SAPS 3 was developed as an attempt to improve SAPS II. However, it is not clear that SAPS 3 is indeed superior to SAPS II.[12,13,17] To allow for tracking of ICU performance over time, and keeping old research comparable with new results, we therefore focus on SAPS II. As mentioned, it is important to calibrate this prediction model using data from a local and contextualized cohort. The Norwegian Intensive Care Registry (NIR) consists of data from patients admitted to ICUs in Norway. We have used the most recent cohorts available in NIR in this study (the 2008–2010 cohorts). This is the first time SAPS II is recalibrated for NIR data. SAPS II has previously never been recalibrated for data from any Nordic country.

The aim of this study was to recalibrate the original SAPS II model by performing a first-level customization based on the NIR data set.

Material and methods

Patients

The NIR data formed the basis of this customization. The registry consists of data from more than 90% of adult patients (18+) admitted to ICUs in Norway. NIR includes 42 surgical, medical, and mixed ICUs in 38 hospitals at primary, secondary, and tertiary level. Because participation in NIR is not mandatory for Norwegian hospitals, there still are a few hospitals not delivering data to NIR. This is similar to most other national intensive care registries. Some specialized ICUs, one burn unit and some post-cardiac surgery combined-recovery ICUs, have also systematically not been included. We do not consider this a severe bias. Patients were not given a SAPS II score if they were younger than 18 years, had undergone post-cardiac surgery, or suffered burns.

Our source population was data collected from 38,257 adult patients admitted during the period 2008–2010. SAPS II scores were reported for each patient, along with age, ICU length of stay (LOS), vital status at hospital discharge, time on respiratory support, type of admission (planned surgery, acute surgery, or acute medical), gender, and hospital category. Of the source population, 3970 were excluded because of missing SAPS II scores, and 29 were excluded because the LOS was missing. A further 2552 patients were excluded because they were

Table 1

Characteristics of study population.

Variable	Characteristic	Sample
n	Total number	30,712
Age (years)	Mean (SD)	63.2 (18.2)
	Median (IQR)	66.0 (52.4,77.3)
Sex, %	Male	56.7
	Female	43.3
SAPS II	Mean SAPS II (SD)	36.8 (18.2)
	Median SAPS (IQR)	34.0 (24,47)
Length of stay, days	Mean (SD)	4.3 (6.8)
	Median (IQR)	2.0 (1.1,4.3)
Type of admission, %	Medical	55.8
	Acute surgery	31.7
	Planned surgery	12.6
Hospital category, %	Primary	36.7
	Secondary	39.8
	Tertiary	23.5
Survival status, %	Died ICU	12.7
	Died ward	6.7
	Survived hospital	80.6

SD, standard deviation; IQR, inter-quartile range.

readmitted to the ICU several times during the same hospital stay or transferred to other hospitals. Transfers and readmissions were excluded to avoid counting the same individual more than once. Finally, 994 patients were excluded because of missing values for: vital status, duration of respiratory support, type of admission, or gender. This left a study population of $n = 30{,}712$ patients. Patient characteristics are showed in Table 1.

Statistical analysis

We compared two models in this study. Model A was the original SAPS II model that is based on a multicenter study with international data from the early development.[5] Model B was a first-level customization of Model A, i.e., a modification of the equation used to describe the relationship between SAPS II scores and predicted hospital mortalities. Hence, the predicted risk of death for a given SAPS II score was

$$\mathrm{PRD} = \frac{e^{\mathrm{logit}}}{1 + e^{\mathrm{logit}}}$$

where

$$\mathrm{logit} = \beta_0 + \beta_1 \times (\mathrm{SAPS\ II}) + \beta_2 \times ln(\mathrm{SAPS\ II} + 1) \qquad (1)$$

The β's represent the weights assigned to each term in the equation and were estimated from the NIR data. Thus, both models were fit using this logistic regression approach where an extra term was added to adjust for non-linearity of the logit function. For Model A, the equation was as follows:[5]

$$\mathrm{logit}_A = -7.7631 + 0.0737 \times (\mathrm{SAPS\ II}) + 0.9971 \times ln(\mathrm{SAPS\ II} + 1)$$

To evaluate model performance reliably, different data sets should be used when designing and validating the model. Therefore, the NIR data set was divided into a training set, used for model design, and a validation set. The training set consisted of two thirds of the NIR patients, chosen at random, and the remaining third was used for validation. To assess how variability in the data affected the outcome, we also applied a fivefold cross-validation approach.[18] This involved splitting the training set into five equal parts, and fitting Model B onto data consisting of four of those parts. The remaining part was used for validation. Repeating this procedure five times, so that all parts were used for validation, we were able to evaluate how variation within the training set could affect the performances of Model A and Model B.

In order to evaluate the performance of Model A and Model B, three aspects were considered: discrimination, calibration, and uniformity of fit. Model discrimination is commonly evaluated by calculating the area under the receiver operating characteristic (ROC) curve.[19] This number, here referred to as aROC, is the probability that a random patient who died had a higher SAPS II score than a random patient who survived. Probabilities close to 1 indicate good discrimination. Because the aROC is based on the untransformed SAPS II scores, Model A and Model B were identical regarding model discrimination.

A model is well calibrated if the predicted proportion of deaths among patients within different SAPS II strata is close to the observed proportion. To evaluate the calibration of Model A and Model B, we used Hosmer–Lemeshow's C statistic,[20] the Brier score,[21] and Cox's calibration regression.[22] In order to calculate C, patients were sorted according to SAPS II scores and divided into 10 deciles. The 10% with the lowest SAPS II scores were in the first decile, the next 10% were in the second decile, and so on. Now, we had

$$C = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{N_g p_g (1 - p_g)}$$

where O_g and E_g were the observed and predicted number of deaths in decile g, respectively. N_g was the number of patients in decile g, p_g was the predicted

risk of death in decile g, and $\sum_{g=1}^{10}$ indicated that the sum was taken over all 10 deciles. As C has a χ^2 distribution with 8 degrees of freedom, each value of C corresponds to a *P*-value. Low values of C indicate well-calibrated models and yield high *P*-values.

The Brier score, B, is calculated as follows,

$$B = \frac{1}{n}\sum_{i=1}^{n}(PRD_i - y_i)^2$$

where PRD_i is the predicted risk of death of individual i, and y_i is the observed outcome (1 if death or 0 if survival). The sum is over all n individuals. B is always between 0 and 1, where 0 denotes perfect prediction. According to Redelmeier et al.,[23] the standard deviation of B is

$$SD(B) = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} PRD_i(1-PRD_i)(1-2PRD_i)^2}$$

allowing us to calculate 95% confidence intervals for B. Note that the Brier score does not evaluate the calibration alone, but is a measure of overall accuracy of predictions. Still, in our scenario, because discrimination is the same for both models, and their performances are compared using the same data set, an improved Brier score implies improved calibration.

Performing Cox's calibration regression means fitting the model

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

using logistic regression. Perfect prediction yields $\alpha = 0$ and $\beta = 1$. This would indicate that the true logit is equal to the predicted logit. Also, if we condition on $\beta = 1$, so that

$$\text{true logit} = \alpha + \text{predicted logit}$$

the two logits are separated by a shift of α. Hence, if $\alpha = 0$ conditioned on $\beta = 1$ ($\alpha = 0 \mid \beta = 1$), calibration is perfect. A negative α indicates that the predicted mortality is too high, and a positive α that the predicted mortality is too low.

A good mortality prediction model should perform similarly across different subgroups of patients. To evaluate uniformity of fit, we considered the standardized mortality rate,

$$SMR = \frac{O}{E} \qquad (2)$$

where O and E are the observed and predicted hospital mortalities, respectively. Hence, a standardized mortality ratios < 1 suggests that the predicted mortality is higher than that actually observed. SMRs were obtained across age groups (< 40, 40–59, 60–69, 70–79, and 80+), types of admission (planned surgery, acute medical, acute surgery), lengths of stay at the ICU (< 1 day, 1–3 days, 4–29 days, and 30+ days), and hospital categories (primary, secondary, and tertiary). An SMR close to one indicated good fit. Using Byar's approximation,[24] 95% confidence intervals for the SMRs could be calculated. The lower limit was

$$SMR_L = \frac{O\left(1 - \frac{1}{9O} - \frac{1.96}{3\sqrt{O}}\right)^3}{E}$$

and the upper limit was

$$SMR_U = \frac{(O+1)\left(1 - \frac{1}{9(O+1)} + \frac{1.96}{3\sqrt{O+1}}\right)^3}{E}$$

Uniformity of fit can be assessed by dividing the SMR for Model B on that of Model A. If this ratio is equal across subgroups, the uniformity of fit has not changed.

All analyses were conducted using R version 3.0.2 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria).

Ethics and prior publication of data

NIR is one of the Norwegian national medical quality registries, and the data used were routinely and anonymously collected. Therefore, the regional ethics committee (Western Norway Regional Health Authority, Norway) waived approval. No consent was needed.

Data have not been previously published.

Results

Figure 1 (left) shows the observed mortality in the validation set and the relationship between SAPS II scores and predicted hospital mortalities for the old and the revised prediction models. For Model B, fitted on the training set, the equation corresponding to (1) was

$$\text{logit}_B = -9.0917 + 0.0325 \times (\text{SAPS II}) + 1.6698 \times ln(\text{SAPS II} + 1)$$

Model A was generally too pessimistic regarding the likelihood of survival. The black line in Fig. 1 (left) represents Model B, where predicted mortalities are

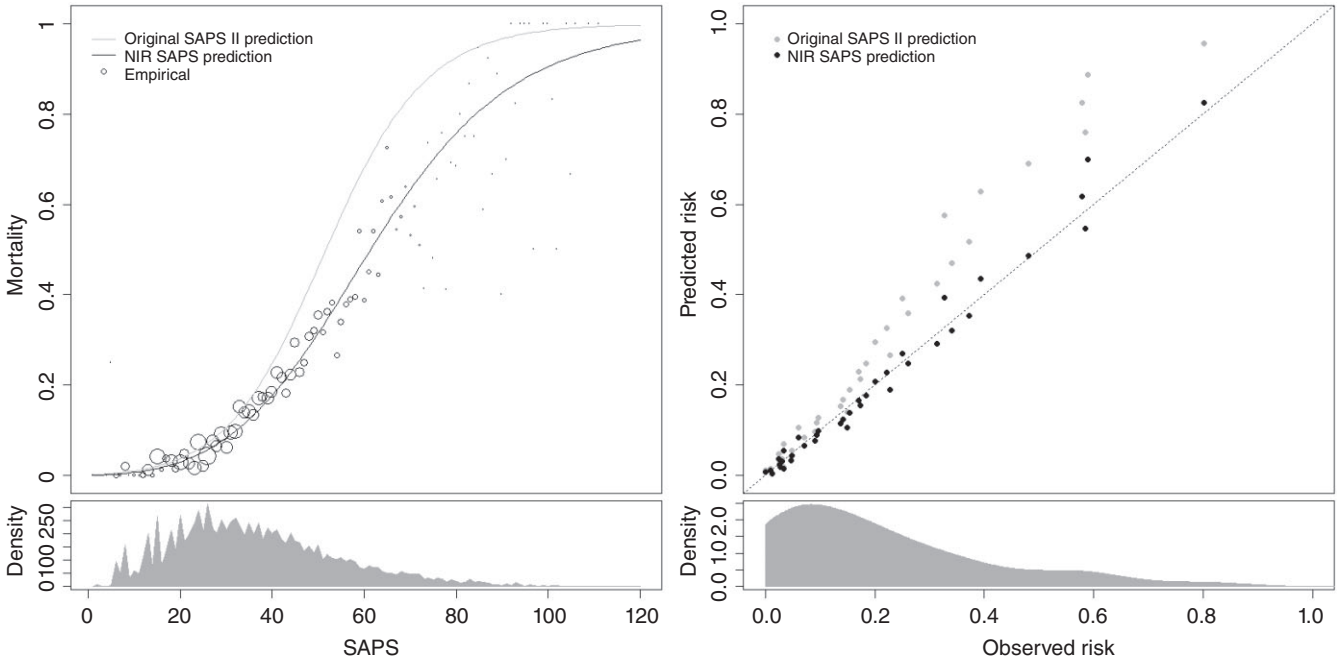

Fig. 1. Left: SAPS vs. mortality. The grey line represents the mortality predicted by Model A, while the black line is the mortality predicted by Model B. Circles are observed mortalities for each SAPS II score. Circle size is proportional to the number of patients with that SAPS II score. The mountain-like grey shape at the bottom shows the distribution of patients according to SAPS II score. Right: calibration plot for Model A and Model B. The grey shape is the distribution of patients according to observed risks.

lower than those of Model A (grey line) for all SAPS II scores. Model B is closer to the observed mortality, particularly for SAPS II scores between 40 and 100 where Model A clearly overestimates mortality. These findings are also supported by Fig. 1 (right), which shows calibration plots for both models.

Note that in Fig. 1 (left), the observed mortality appears to vary more for patients with the highest SAPS II scores. This is because the observed mortalities of patients with extreme SAPS II scores are averages of only a few observations. For example, only two patients had a SAPS II score of 100, and both died. Hence, the observed mortality was 1. If one had survived, the observed mortality would have been 0.5, and if both had survived, it would have been 0.

The aROC was 0.83 for both Model A and Model B (Table 2). Values above 0.80 indicate that discrimination is very good.[25,26]

Table 2 presents the results of the validation procedures. As seen, both the fivefold cross-validation and the validation on the validation set of Model A suggested poor calibration, with values of C corresponding to *P*-values lower than 0.001. For Model B, the fivefold cross-validation was better than the validation on the validation set, which gave a *P*-value of 0.005. For the cross-validation, the mean of the five *P*-values was reported in Table 2.

Table 2

Validating the calibration of Model A and Model B.

	Model A	Model B
HL-test		
Fivefold cross-validation		
Mean *P*-value	< 0.001	0.306
Standard deviation	< 0.001	0.250
Validation set		
HL's C	689.07	22.01
P-value	< 0.001	0.005
Brier score		
B	0.131	0.120
95% confidence interval	0.127–0.134	0.116–0.123
Cox's calibration regression		
α	–0.747	–0.093
β	0.735	0.921
$\alpha \mid \beta = 1$	–0.630	–0.009
aROC	0.83	0.83

HL-test: *P*-values > 0.05 indicate good calibration.
HL-test: Hosmer–Lemeshow's C. C is χ^2-distributed with 8 degrees of freedom.
Brier score: Lower values indicate better calibration.
Cox's calibration regression: α should be close to 0 and β should be close to 1.
aROC: Area under receiver operating characteristic curve.

More important is the fact that C dropped from the enormous 689.07 for Model A to the more moderate 22.01 for Model B. Further, the confidence intervals of the Brier scores did not overlap (0.116–

Table 3

Standardized mortality ratios (SMR) across different groups of patients for Model A and Model B in validation set. The column 'Ratio' contains SMRs for Model B divided by SMRs for Model A. Age was measured in years, and LOS in days.

	Model A	Model B		
	SMR (95% CI)	SMR (95% CI)	Ratio	*n*
Total	0.73 (0.70,0.76)	0.99 (0.95,1.04)	1.36	10,237
Age				
18–39	0.62 (0.50,0.76)	0.85 (0.69,1.04)	1.37	1392
40–59	0.59 (0.52,0.66)	0.80 (0.71,0.90)	1.37	2379
60–69	0.65 (0.58,0.72)	0.88 (0.79,0.97)	1.36	2218
70–79	0.74 (0.68,0.81)	1.01 (0.93,1.10)	1.36	2340
80+	0.93 (0.86,1.01)	1.26 (1.17,1.37)	1.36	1908
Type of admission				
Planned surgery	0.65 (0.55,0.76)	0.88 (0.75,1.04)	1.36	1367
Acute medical	0.77 (0.73,0.81)	1.04 (0.98,1.10)	1.35	5589
Acute surgery	0.68 (0.63,0.74)	0.94 (0.86,1.02)	1.37	3281
Length of stay (LOS)				
< 1	0.94 (0.87,1.02)	1.25 (1.15,1.35)	1.33	2304
1–3	0.66 (0.62,0.71)	0.91 (0.84,0.97)	1.37	5306
4–29	0.66 (0.60,0.72)	0.91 (0.83,0.99)	1.37	2484
30+	0.78 (0.56,1.05)	1.07 (0.78,1.44)	1.38	143
Hospital category				
Primary	0.78 (0.72,0.84)	1.06 (0.98,1.15)	1.36	3699
Secondary	0.74 (0.69,0.79)	1.00 (0.94,1.07)	1.35	4112
Tertiary	0.66 (0.60,0.72)	0.90 (0.82,0.99)	1.37	2426
Sex				
Female	0.74 (0.69,0.79)	1.01 (0.94,1.08)	1.36	4462
Male	0.72 (0.68,0.77)	0.98 (0.93,1.04)	1.36	5775

CI, confidence interval; LOS, length of stay.

0.123 for Model B, 0.127–134 for Model A). Finally, applying Cox's calibration regression α was closer to 0 and β was closer to 1 for Model B (Table 2). All this suggests that Model B was much better calibrated than Model A. This finding is confirmed by visual inspection of Fig. 1.

Figure 1 shows that Model B mortality estimates were much closer to the observed mortalities. This is also confirmed in Table 3, where Model A had SMR values < 1 all over, and overestimated the mortality in most subgroups, whereas Model B tended to be more in line with observed data. Notable exceptions were for patients aged 80 years and older, and for those with length of state (LOS) < 1 day, where 1 was contained in the SMR confidence intervals for Model A, but not for Model B (Table 3). The 80+ patients comprised 19% of the patient population in the validation set, and the patients with the shortest LOS comprised 23%. Dividing SMR for Model B with that of Model A yielded ratios of 1.33–1.38 for all the different subgroups. Hence, the uniformity of the fit did not change after recalibration of Model A.

Discussion

This study confirms that mortality predictions can be improved by customizing the original SAPS II model.[5] The Brier score was lower for Model B than for Model A, and the confidence intervals for Brier scores did not overlap (Table 2). Also, the coefficients of Cox's calibration regression were closer to $\alpha = 0$ and $\beta = 1$ for Model B than for Model A (Table 3). Hosmer–Lemeshow's C was reduced from 689.07 (Model A) to 22.01 (Model B), which is substantial despite the *P*-value being less than 0.05. When analyzing large data sets, such as NIR, the C is expected to yield low *P*-values regardless of model fit.[27,28] In fact, some of the reason that the fivefold cross-validation yielded high *P*-values (Table 2) may be that they were based on smaller data sets than the validation set. Discrimination was good using both models. With respect to uniformity of fit, both models performed similarly. The ratios of SMRs were approximately the same across all subgroups. This is as expected for calibrations based on first-level customization, because the underlying SAPS II scores remain unchanged. Model B outperformed Model A regarding model fit, although Model B tended to overestimate the mortality for patients younger than 70 years, and underestimate the mortality for those older than 80 years.

In accordance with the general literature, we found that Model A was poorly calibrated.[9–13] This is as expected, given the advances in medical care in

general, and intensive care in particular, since Model A was developed.[5] To overcome the problem of poor calibration, customization of Model A has become a popular exercise.[2,12,14,29] This is the first time Model A has been calibrated to NIR data, and we now have the tools handy to perform recalibration more often. It would be better to base the calibration on newer data, but the cohorts from 2008 to 2010 were the most recent in NIR available to us. Still, this is a clear improvement of Model A, which has been the standard in Norway so far.

Other studies that applied first-level customization to SAPS II, also reported differences in SMR across age groups.[5,8,30] Aegerter et al.,[8] Le Gall et al.,[15] and Apolone et al.[30] all underestimated the risk of dying for the oldest patients they studied. However, for Aegerter et al., the SMR was not significantly different from 1. Further, both Aegerter et al. and Le Gall et al. significantly overestimated the mortality for the youngest patients, which is similar to the results presented in Table 3.

A limitation of this study is that second-level customization was not performed. Second-level customization is when the underlying variables are changed or differently weighted. This approach requires more work and access to other types of data and is therefore hard to generalize. However, models based on second-level customization tend to perform better than their first-level counterparts.[8,15,16] For example, the variation in SMR across age groups could be rectified by weighting age differently before obtaining the SAPS II scores.

We focused on first-level customization in order to keep our results as general as possible, although the case-mix will influence the performance of the model across settings. Model B will be more applicable in the Nordic countries and less in countries with different epidemiological profiles. However, if there is no locally calibrated SAPS II model available, Model B may be used, with caution, as a rough prognostic indicator.

Local calibrations of SAPS II are regularly performed all over Europe.[2,10,12,29–31] This is necessary in order for SAPS II-based mortality predictions to remain meaningful and to keep up with medical advances. As mentioned, it has been argued that SAPS II should be recalibrated every second or third year.[14] As an attempt to improve SAPS II and create a new international benchmark, SAPS 3 was created. However, several validation studies have questioned whether SAPS 3 is indeed an improvement of SAPS II regarding discrimination, uniformity of fit, and calibration.[12,13,17] We therefore argue that recalibrating SAPS II in an international multicenter study could be a better alternative than changing to SAPS 3. This has the advantage of keeping previous results comparable with new research and allows for better tracking of changes in ICU performance over time. For example, it may be of interest to know how the mortality risk has changed for a patient with a SAPS II score of 50 for the last 20 years, and if the change is similar across settings. Data are already routinely collected in many countries. An international benchmark will allow for different settings to be evaluated according to an international standard. Variations in ICU performance between countries are inevitable, but a regularly recalibrated international benchmark may shed a light on some of the reasons for this variation and allow for improvements to be made.

Conclusion

The updated version of SAPS II, Model B, had the same discrimination and uniformity of fit as Model A, but was better calibrated. Introducing Model B into Norwegian ICUs may improve precision in decision-making at multiple levels. It will yield a more realistic benchmark in the assessment of ICU performance. If SAPS II is used in clinical decisions, with caution and good clinical judgment, these decisions will have a more reliable risk estimate than previous versions of SAPS II could provide.

Conflicts of interest: None of the authors has any conflicts of interests to declare.

Funding: None.

References

1. Flaatten H. Prognostic scoring systems in the ICU. Acta Anaesthesiol Scand 2006; 50: 1175–6.
2. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. Intensive Care Med 2012; 38: 40–6.
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985; 13: 818–29.
4. Teres D, Lemeshow S, Avrunin JS, Pastides H. Validation of the mortality prediction model for ICU patients. Crit Care Med 1987; 15: 208–13.
5. 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.
6. Barnato AE, Angus DC. Value and role of intensive care unit outcome prediction models in end-of-life decision making. Crit Care Clin 2004; 20: 345–62, vii–viii.
7. Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, de Jonge E, Bosman RJ, Peelen L, de Keizer NF. External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. J Crit Care 2011; 26: e11–8.

8. Aegerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B. SAPS II revisited. Intensive Care Med 2005; 31: 416–23.
9. Moreno R, Morais P. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. Intensive Care Med 1997; 23: 177–86.
10. Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med 2003; 29: 249–56.
11. Strand K, Flaatten H. Severity scoring in the ICU: a review. Acta Anaesthesiol Scand 2008; 52: 467–78.
12. Sakr Y, Krauss C, Amaral AC, Rea-Neto A, Specht M, Reinhart K, Marx G. Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit. Br J Anaesth 2008; 101: 798–803.
13. Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G, GiViTi. Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? Intensive Care Med 2012; 38: 1280–8.
14. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. 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: 1378–88.
15. Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, Gouzes C, Lepage E, Moine P, Villers D. Mortality prediction using SAPS II: an update for French intensive care units. Crit Care 2005; 9: R645–52.
16. Reinikainen M, Mussalo P, Hovilehto S, Uusaro A, Varpula T, Kari A, Pettila V, Finnish Intensive Care C. Association of automated data collection and data completeness with outcomes of intensive care. A new customised model for outcome prediction. Acta Anaesthesiol Scand 2012; 56: 1114–22.
17. Strand K, Soreide E, Aardal S, Flaatten H. A comparison of SAPS II and SAPS 3 in a Norwegian intensive care unit population. Acta Anaesthesiol Scand 2009; 53: 595–600.
18. Kohavi R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Proceedings IJCAI-95 1995: 1137–43.
19. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29–36.
20. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982; 115: 92–106.
21. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950; 78: 1–3.
22. Cox D. Two further applications of a model for binary regression. Biometrika 1958; 45: 562–5.
23. Redelmeier DA, Bloch DA, Hickam DH. Assessing predictive accuracy: how to compare Brier scores. J Clin Epidemiol 1991; 44: 1141–6.
24. Rothman K, Boice J Jr. Epidemiologic analysis with a programmable calculator. Bethesda, MD & Washington: US Dept. of Health, Education, and Welfare, Public Health Service, National Institutes of Health, 1979. NIH Pub, 79: 31–32.
25. Lloyd-Jones DM. Cardiovascular risk prediction: basic concepts, current status, and future directions. Circulation 2010; 121: 1768–77.
26. Siontis GC, Tzoulaki I, Ioannidis JP. Predicting death: an empirical evaluation of predictive tools for mortality. Arch Intern Med 2011; 171: 1721–6.
27. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer–Lemeshow test revisited. Crit Care Med 2007; 35: 2052–6.
28. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. Stat Med 2013; 32: 67–80.
29. Suistomaa M, Niskanen M, Kari A, Hynynen M, Takala J. Customized prediction models based on APACHE II and SAPS II scores in patients with prolonged length of stay in the ICU. Intensive Care Med 2002; 28: 479–85.
30. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G, Melotti RM. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva. Intensive Care Med 1996; 22: 1368–78.
31. Metnitz PG, Lang T, Vesely H, Valentin A, Le Gall JR. Ratios of observed to expected mortality are affected by differences in case mix and quality of care. Intensive Care Med 2000; 26: 1466–72.

Address:
Øystein Ariansen Haaland
Department of Global Public Health and Primary Care
University of Bergen
IGS PO Box 7804
NO-5020 Bergen
Norway
e-mail: oystein.haaland@igs.uib.no