Intensive Care Med (2000) 26: 1466–1472
DOI 10.1007/s001340000638

ORIGINAL

Philipp G. H. Metnitz
Thomas Lang
Herbert Vesely
Andreas Valentin
J. R. Le Gall

Ratios of observed to expected mortality are affected by differences in case mix and quality of care

Received: 20 December 1999
Final revision received: 10 July 2000
Accepted: 20 July 2000
Published online: 8 September 2000

P. G. H. Metnitz (✉)
Department of Anesthesiology and General Intensive Care,
University of Vienna,
Währinger Gürtel 18-20, 1090 Vienna, Austria
E-mail: philipp.metnitz@univie.ac.at
Fax: +43-1-533 91 26

T. Lang
Department of Medical Statistics,
University of Vienna, Austria

H. Vesely
Department of Anesthesiology and General Intensive Care, Hanusch KH,
Vienna, Austria

A. Valentin
Department of Internal Medicine II, KA Rudolfstiftung, Vienna, Austria

J. R. Le Gall
Department of Intensive Care Medicine,
Saint-Louis Hospital, Paris, France

Abstract *Objectives:* To validate SAPS II-AM, a recently customized version of the Simplified Acute Physiology Score II (SAPS II) in a larger cohort of Austrian intensive care patients and to evaluate the effect of the customization process on the ratio of observed to expected mortality.
Design: Prospective, multicentric cohort study.
Patients and setting: A total of 2901 patients consecutively admitted to 13 adult medical, surgical, and mixed intensive care units (ICUs) in Austria.
Measurements and results: After the database was divided randomly into a development sample (n = 1450) and a validation sample (n = 1451), logistic regression was used to develop a new model (SAPS II-AM2). The original SAPS II, the SAPS II-AM, and the newly developed SAPS II-AM2 were then compared by means of calibration, discrimination and O/E ratios. Differences in O/E ratios before and after customization (ΔO/E) were calculated. The Hosmer-Lemeshow goodness-of-fit $\hat{H}$ and $\hat{C}$ statistics revealed poor calibration of the original SAPS II on the database. The new model, SAPS II-AM2, performed better than the SAPS II-AM and excellent in the validation data set. However, mean O/E ratios varied widely among diagnostic categories (range 0.55–1.05 for the SAPS II). Moreover, the ΔO/E of the 13 ICUs ranged from –3.6 % to +25 %.
Conclusions: Today's severity scoring systems, such as the SAPS II, are limited by not measuring (and adjusting for) a profound part of what constitutes case mix. Changes in the distribution of patient characteristics (known and unknown) therefore affect prognostic accuracy. First-level customization was not able to solve all these problems. Using O/E ratios for quality of care comparisons one must therefore be critical when using these data and should search for possible confounding factors. In the case of unsatisfactory calibration, customized severity of illness models may be useful as an adjunct for quality control.

Key words Severity of illness · Outcome · Customization · Mortality prediction · Uniformity of fit · Simplified Acute Physiological Score II

Introduction

In recent years quality management has become an important issue in intensive care medicine as in other fields. Outcome research provides the tools to compare different outcomes and adjust for the variability of structures and processes. To compare data from different institutions one has to adjust for the heterogeneity

of patients in such categories as age, sex, chronic illness, underlying disease, and severity of the ongoing illness. This is called case-mix adjustment.

Severity scoring systems have been developed for use in such adjustments. To adjust for severity of illness, predicted hospital mortality (which is calculated by means of a severity of illness scoring system) is compared with the actual (observed) mortality. The ratio of observed to expected deaths (O/E ratio) has become standard for assessing the clinical performance of intensive care units (ICUs) on the assumption that scoring systems are able to adjust for the case mix of a specific population. If this were true, any differences in the risk-adjusted outcome data would be related to differences in the provided care.

Recent studies have shown that this assumption may not be valid, and that the performance of these predictive models may be affected by other factors [1, 2]. In a European multicenter study the Simplified Acute Physiology Score II (SAPS II) was found to overestimate hospital mortality and to give improper calibration across subgroups [1]. We found a similar pattern in Austrian ICU patients [3]. Customization of a general severity of illness score by deriving a new logistic regression equation has been advocated as being useful in these circumstances [2, 4, 5]. However, Moreno et al. [2] have demonstrated that, although customization can improve the overall predictive performance of a model, it does not solve other model-related problems, especially inhomogeneity across patient subgroups.

Severity of illness data and O/E ratios are being used increasingly by governmental and commercial institutions to assess the clinical performance of ICUs [6]. A recent study has even suggested that these data may be useful to classify ICUs into different levels of performance [7]. It is therefore of utmost importance to evaluate the performance of the statistical models used to generate these data and to detect any potential confounding factors.

We recently demonstrated that a first-level customized version of the SAPS II that we call SAPS II-AM (AM for "Austrian model") improved the accuracy of the prediction of hospital mortality in Austrian ICU patients [3]. The aim of the present study was to validate the customized model in a larger cohort of Austrian ICU patients and to assess further the effect of customization on O/E ratios.

Material and methods

Database

Data were collected by the Austrian Center for Documentation and Quality Assurance in Intensive Care Medicine, a nonprofit organization that is in the process of establishing a national intensive care database and external quality assurance program [8]. The collected data set includes sociodemographic data (age, sex, reason for admission, chronic condition); severity of illness as measured by the SAPS II; and outcome data, including vital status and date of ICU discharge and vital status and date of hospital discharge. Data were prospectively collected in 13 Austrian ICUs between 1 March 1997 and 30 June 1998. Ten ICUs were postoperative (operated by staff of the Department of Anesthesiology), and three were medical ICUs (operated by staff of the Medical Department). Four ICUs were located in teaching hospitals, four in central referral hospitals, and five in rural hospitals. Since no additional interventions were performed the need for informed consent was waived by the institutional review board.

Data quality

The quality of the data was tested on the first half of the data collection. To assess the reliability of the data we sent an independent observer to each unit to obtain SAPS II data from the histories of a random sample of ten patients. Variance-component analyses with the random factors "units," "patients within units," and "observers within units" were performed (SAS procedure "varcomp") as described previously [8]. To assess the completeness of the documentation we calculated the number of missing parameters for each SAPS II score.

Statistical analysis

To evaluate whether the recently developed SAPS II-AM model [3] remains stable in a larger cohort of patients we considered the following steps to be important: (a) to determine the performance of the original SAPS II and the SAPS II-AM, (b) to develop a new model (SAPS II-AM2) and validate it in an independent sample, and (c) to assess the stability of the customization by comparing the different models.

To test the prognostic performance of the various models, we used standard tests. The Hosmer-Lemeshow goodness-of-fit $\hat{C}$ statistic (with 10 d.f.) [9] was used to evaluate calibration which was considered satisfactory when the p value was > 0.05. Discrimination was tested by measuring the area under the receiver operating characteristic (ROC) curve [10]. O/E ratios were calculated by dividing the number of observed deaths per group by the number of expected deaths per group (as predicted by the three respective models). To test for statistical significance we calculated 95 % confidence intervals (CI) according to the formula described by Hosmer and Lemeshow [11].

After evaluating the calibration of SAPS II and SAPS II-AM on all patients we divided the database randomly into two samples: development ($n = 1450$) and validation ($n = 1451$). The two samples were comparable with respect to age, length of stay, severity of illness, and observed mortality (Table 1). The development sample was used to develop a new predictive model, using the same approach as for SAPS II-AM [3]: a logistic regression equation for the hospital mortality prediction was derived with the SAPS II score as the independent variable and outcome from hospital as the dependent variable. The new model (SAPS II-AM2) was then tested in the validation sample by means of the procedures described above. Furthermore, the effect of the customization process on O/E ratios was assessed: O/E ratios of patient subgroups and of ICUs (Table 3) were calculated, using the predicted mortality of the original SAPS II and the customized models. Differences between these O/E ratios were calculated and expressed as a percentage difference (ΔO/E) to the value, derived with the original SAPS II.

Statistical analysis was performed using the SAS system, version 6.11. For the comparison of development and validation samples Student's *t*test and Wilcoxon rank sum tests were used as appropriate. A p value less than 0.05 was considered significant. Unless otherwise specified, results are expressed as mean ± SD.

Results

Data from 13 ICUs were included in the study, representing 3536 consecutive admissions. From these we excluded readmissions ($n = 56$), patients aged under 18 years ($n = 149$), those whose records were missing hospital outcome data ($n = 348$), and those without SAPS II score ($n = 82$). The remaining patients ($n = 2901$) were included in the analysis. Basic demographic data of the patients are shown in Table 1.

Completeness of the documentation was satisfactory. On average 1.3 ± 1.9 SAPS II parameters were missing. Interrater quality control was performed by variance-component analysis. The results, which have been published elsewhere [3], indicated an overall excellent grade of agreement: interobserver variability on most of the variables was less than 1 %. In the case of only a few variables (temperature, urine output, serum bicarbonate, Glasgow Coma Scale score) up to 15 % of the total variance was attributed to interobserver variability.

Evaluation of the original SAPS II model

As previously demonstrated [3], calibration of the original SAPS II model on all patients was poor (Table 2). SAPS II underestimated mortality in the lowest deciles of risk (0–20 % mean risk of predicted hospital mortality) and overestimated mortality in the higher deciles of risk (> 30 % mean risk of predicted hospital mortality). Discrimination of the SAPS II was good, with an area under the ROC of 0.83. Calibration was poor when patients were grouped according to the types of admission (medical, scheduled surgical, and unscheduled surgical).

Evaluation of the SAPS II-AM model

Evaluation of the customized model SAPS II-AM showed a better, although not ideal, calibration (Table 2). Application of this model improved the fit in subgroups of scheduled and unscheduled surgical admissions, but not in medical admissions.

Evaluation of the SAPS II-AM2 model

After development of a new predictive model (SAPS II-AM2) using the development sample ($n = 1450$), validation was done in the independent validation sample ($n = 1451$) and showed excellent calibration (Table 2). Calibration improved significantly (compared with the previous models) when patients were grouped by type of admission. Discrimination was similar, with areas under the ROC curve of 0.83 for both development and validation samples.

Evaluation of the ratios of observed to expected deaths

Mean predicted hospital mortality for all patients differed significantly between the three models (SAPS II, 23.3 %, 95 % CI 22.34–24.29 %; SAPS II-AM, 19.9 %, 95 % CI 19.13–20.57 %; SAPS II-AM2, 21.5 %, 95 % CI

Table 1 Basic demographic data

	All patients ($n = 2901$)	Development sample ($n = 1450$)	Validation sample ($n = 1451$)	p[a]
Age (years; mean ± SD)	61.5 ± 16.8	61.6 ± 16.8	61.4 ± 16.8	0.802
ICU length of stay (days; median and range)	4 (0–121)	4 (0–121)	4 (0–105)	0.390
SAPS II score (median and range)	29 (0–115	29 (0–115	29 (0–98)	0.976
SAPS II predicted risk (median and range)	10 (0–100)	10 (0–100)	10 (0–98)	0.531
Observed ICU mortality	477 (16.5 %)	245 (16.9 %)	232 (16.0 %)	0.255
Observed hospital mortality	612 (21.1 %)	312 (21.5 %)	300 (20.7 %)	0.289
Type of admission				
Medical	1319 (45.6 %)	641 (44.3 %)	678 (46.9 %)	0.914
Scheduled surgical	920 (31.8 %)	470 (32.5 %)	450 (31.1 %)	0.209
Unscheduled surgical	655 (22.6 %)	336 (23.2 %)	319 (22.0 %)	0.222
TISS 28 score per patient per day (median and range)[c]	24.1 (0–65)	24.6 (0–62.2	23.5 (0–65)	0.147
Mechanical ventilation[b]	48.0 %	49.6 %	46.3 %	0.001
Renal replacement therapy[b]	4.8 %	4.4 %	5.2 %	0.003

[a]Student's *t* test or Wilcoxon rank sum test between development and validation samples
[b]Percentage of days when these activities were performed
[c] Simplified Therapeutic Intervention Scoring System

Table 2 Uniformity of fit for the original SAPS II, the SAPS II-AM, and the SAPS II-AM2 (*n* number of admissions, *Ĉ test* Hosmer-Lemeshow goodness-of-fit Ĉ test, *ROC* area under the receiver operating characteristic curve)

			Original SAPS II				SAPS II-AM				SAPS II-AM2 (validation sample)					
	n	ROC	Ĉ test	*p*	O/E	CI	Ĉ test	*p*	O/E	CI	*n*	ROC	Ĉ test	*p*	O/E	CI
All patients	2901	0.83	69.1	< 0.001	0.90	0.85–0.96*	20.7	0.02	1.06	1.00–1.13	1451	0.83	14.8	0.14	0.96	0.88–1.05
Types of admission																
Medical	1319	0.85	23.1	0.01	1.02	0.95–1.08	44.0	< 0.001	1.22	1.14–1.31	678	0.85	12.0	0.29	1.09	0.98–1.20
Scheduled surgical	920	0.74	43.3	< 0.001	0.85	0.70–1.00	11.7	0.30	0.87	0.70–1.03	450	0.77	13.0	0.22	0.84	0.61–1.06
Unscheduled surgical	655	0.80	58.1	< 0.001	0.72	0.63–0.82*	13.0	0.22	0.89	0.77–1.01	319	0.78	18.9	0.04	0.81	0.66–0.96*

* CI ≠ 1

Table 3 O/E ratios of participating ICUs (*O/E* O/E ratio, *CI* confidence interval, *ICU type* description of the associated specialty, *medical* ICUs associated with a medical department, *postoperative/trauma* ICUs staffed by department of anesthesiology)

	SAPS II		SAPS II-AM		SAPS II-AM2		
Unit	O/E	95 % CI	O/E	95 % CI	O/E	95 % CI	ICU type
A	0.56	0.32–0.80*	0.62	0.33–0.90*	0.58	0.31–0.85*	Medical
B	0.58	0.45–0.72*	0.70	0.54–0.87*	0.65	0.49–0.80*	Postoperative
C	0.58	0.21–0.94*	0.69	0.24–1.13	0.63	0.22–1.04	Postoperative
D	0.60	0.46–0.74*	0.75	0.57–0.94*	0.68	0.52–0.85*	Postoperative
E	0.72	0.48–0.95*	0.75	0.48–1.01	0.71	0.46–0.96*	Postoperative
F	0.81	0.48–1.14	0.93	0.51–1.35	0.86	0.48–1.25	Postoperative
G	1.02	0.64–1.40	1.18	0.74–1.62	1.09	0.68–1.51	Postoperative
H	1.03	0.87–1.19	1.17	0.99–1.36	1.09	0.91–1.27	Postoperative
I	1.05	0.85–1.26	1.15	0.91–1.38	1.08	0.85–1.30	Trauma
J	1.11	1.02–1.20*	1.38	1.26–1.50*	1.26	1.15–1.37*	Medical
K	1.11	0.70–1.52	1.12	0.68–1.56	1.07	0.65–1.49	Medical
L	1.12	0.92–1.32	1.28	1.04–1.53*	1.19	0.97–1.42	Postoperative
M	1.29	0.97–1.60	1.54	1.14–1.94*	1.41	1.05–1.78*	Postoperative

* CI different from 1

20.71–22.32 %). The mean O/E ratio for the 13 ICUs included in the study changed from 0.89 ± 0.26 (95 % CI 0.75–1.03) in the original SAPS II model to 1.02 ± 0.30 (95 % CI 0.86–1.18) in the SAPS II-AM model and to 0.95 ± 0.27 (95 % CI 0.80–1.10) for the SAPS II-AM2 model. The ΔO/E for the 13 ICUs ranged from +0.9 % to +25.0 % (average 14.8 ± 7.2 %) for the SAPS II-AM and from –3.6 % to +13.3 % (average 6.4 ± 5.2 %) for the SAPS II-AM2 (Table 3, Fig. 1).

The O/E ratios in the eight most frequent reasons for admission categories for the original SAPS II model ranged from 0.55 to 1.05 (median 0.85), reflecting a different performance of this model in each of the categories (Table 4). This variability remained in the customized models. For several surgical patients (abdominal, general, trauma, and transplant surgery), the O/E ratio was significantly different from 1. Although the customization process corrected this for abdominal and trauma surgery, it did not do so for general and transplant surgery.

Discussion

We report on the validation of a customized model of a general severity of illness score for ICU patients in a European country. The prognostic accuracy of the original SAPS II model over several subgroups of patients was not satisfactory. This has been demonstrated both for Austrian ICU patients [3] and for a larger cohort of patients in several European countries [1].

There are several possible explanations for these results. First, SAPS II does not take into account all the factors that are known to affect outcome. For example, diagnostic coding: two patients with similar physiological derangements may have different prognoses depending on the underlying disease that led to the ICU admission [12]. This may partially explain the variability in O/E ratios in the eight most frequent reasons for admission (Table 4). Second, our results demonstrate that the lack in the uniformity of fit of the SAPS II was also attributable to factors that are included in the model. Patients who were classified as "medical admissions" had consistently higher O/E ratios than those classified

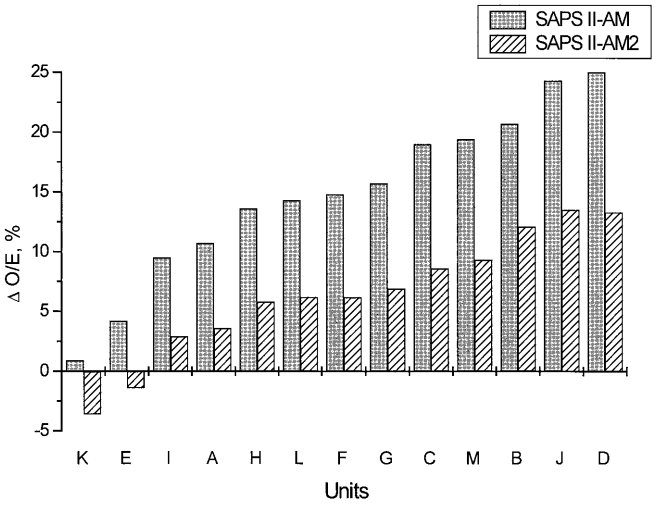

Fig. 1 ΔO/E for the SAPS II-AM and the SAPS II-AM2. Differences in O/E ratios of the 13 participating ICUs before and after customization (ΔO/E). Values are shown as percentage change from the O/E ratios calculated from the original SAPS II model

as scheduled or unscheduled surgical admissions (Table 2). An ICU with a high proportion of medical patients may therefore have a higher mean predicted mortality and O/E ratio than a unit with a high proportion of surgical patients. Third, there exist a variety of factors (known and unknown) which are not included in the SAPS II, but which contribute to the phenomenon of unmeasured case mix. For example, it is not known to what extent the diversity of organization, education, and practice of intensive care medicine affects the outcome of patients [13].

We recently demonstrated an improvement in calibration of the SAPS II through first-level customization [3]. However, this model, SAPS II-AM, had not yet been validated. The SAPS II-AM2 was therefore developed on a randomized sample of the database. Validation of this model showed excellent performance in the validation sample; there was a significant improvement in the goodness of fit (Table 2) compared to the SAPS II and also to the SAPS II-AM. This improvement was seen not only for the cohort of all patients but also for the subgroups of admission types (medical, scheduled surgical, and unscheduled surgical). The remaining difference between the SAPS II-AM and the SAPS II-AM2, however, shows that even minor changes in case mix affect the performance of a prognostic model. The optimal performance can be found only in a sample with a case mix identical to that of the development sample.

Performance of the SAPS II differed significantly between diagnostic categories, as shown by the respective O/E ratios (reasons for admission, Table 4). Should we therefore use reasons for admission to predict outcome? There has been a long discussion about this issue. As yet there is no internationally accepted classification of ICU admissions. It is especially unclear how to deal with patients with multiple diseases. This may also explain why the reason for admission was the parameter that was most often missing in our study: 186 patients (6.4 %) had no documented reason for admission, which supports the usefulness of a severity of illness system without the need for coding a disease at admission (as e.g., the SAPS II) for the assessment of the whole ICU population. O/E ratios between different disease groups, however, should possibly be compared with a system that takes this factor into account, but only if the definitions of the diseases are precise enough.

O/E ratios have commonly been used to compare the quality of care between different ICUs. The standardization of a unit's mortality through case-mix adjustment has been said to allow for an objective approach to these data [14]. Our results show that customization changed both predicted hospital mortality and O/E ratios to varying degrees (Fig. 1). Moreover, O/E ratios differed widely across subgroups in reason for admission (Table 4). Although it is in general be more difficult to establish the significance of the difference from 1 for the O/E ratio in smaller subgroups, we found various clear deviations. Neither the number of deviations nor

Table 4 Mean predicted and observed mortality for the most frequent reasons for admission (frequency data do not apply to the SAPS II-AM2; *n* number of admissions, *OM* observed mortality, *PM* predicted mortality, *O/E* O/E ratio)

	Mortality		SAPS II				SAPS II AM			SAPS II AM2		
Reason for admission	*n*	%	OM (%)	PM (%)	O/E	CI	PM (%)	O/E	CI	PM (%)	O/E	CI
Cardiovascular disease	463	16.0	24.6	27.0	0.94	0.83–1.04	23.1	1.10	0.96–1.23	25.0	1.01	0.89–1.14
Trauma surgery	404	13.9	14.6	18.0	0.81	0.64–0.98*	15.8	0.92	0.72–1.12	17.0	0.86	0.67–1.04
Abdominal surgery	387	13.3	19.9	23.4	0.85	0.71–0.99*	19.9	1.00	0.82–1.17	21.6	0.92	0.76–1.08
General surgery	318	11.0	7.2	13.2	0.55	0.31–0.78*	12.6	0.57	0.31–0.84*	13.3	0.54	0.29–0.79*
Cardiovascular surgery	256	8.8	14.8	16.4	0.91	0.68–1.13	15.0	0.99	0.73–1.25	16.0	0.93	0.68–1.18
Respiratory disease	200	6.9	31.5	30.1	1.05	0.89–1.21	25.0	1.26	1.06–1.46*	27.3	1.16	0.97–1.34
Neurological disease	143	4.9	24.5	28.7	0.85	0.64–1.06	23.0	1.06	0.79–1.33	25.4	0.96	0.72–1.21
Transplant surgery	114	3.9	11.4	20.4	0.56	0.25–0.87*	17.4	0.66	0.29–1.03	18.8	0.61	0.26–0.95*

* CI different from 1

their extent can be fully explained by arguments of multiplicity of testing. For example, the six significant deviations of O/E ratios among the 13 ICUs (Table 3) and the very low O/E ratio for the subgroup of general surgery in Table 4 ($p < 0.001$ for testing against O/E ratio = 1).

These results demonstrate that, to an unknown extent, differences in O/E ratios are attributable to differences that are not related to the quality of provided care. It is inevitable that this behavior will affect the overall O/E ratio of ICUs: An ICU with a high proportion of a specific patient group could accordingly exhibit lower or higher O/E ratios (Table 4). One can hypothesize an ICU, for example, with only postoperative trauma or transplant patients. Since SAPS II overestimates mortality in these patients grossly, this ICU would therefore exhibit a lower overall O/E ratio. We found no correlation, however, between ICU type (medical vs. postoperative) or hospital type (teaching vs. nonteaching) and O/E ratios (Table 3) in this study. This suggests that other, unmeasured case-mix factors affect O/E ratios as well.

It seems, ideally, that at least two criteria must be met before O/E ratios can be used for interunit comparisons: (a) the distribution of patients between units should be identical, and (b) the calibration of the severity of illness model should be satisfactory. The first of these two criteria may be met by grouping similar ICUs and comparing them, if possible, within each group (e.g., neurosurgical ICUs, trauma units). Regular assessment of the calibration of the model used (as recommended by the consensus conference on outcome prediction of the European Society for Intensive Care Medicine [15]) may be necessary to meet the second criterion

A recent study has suggested that ICUs with O/E ratios that are statistically significant different from 1 should be further evaluated for possible problems in the process of care [9]. Our study identified several ICUs whose O/E ratios were significantly different from 1 (Table 3). The question arises as to whether we can use these data still for the detection of differences in the quality of care, although the above mentioned "ideal" criteria are not completely fulfilled (e.g., the distribution of patients was quite heterogeneous in the present study). Take, for example, the two extremes on the scale of O/E ratios from Table 3: ICU A (a highly specialized medical ICU in a tertiary referral center) with an O/E ratio of 0.58 (0.31–0.85, with the SAPS II-AM2) and ICU M (a postoperative ICU in a rural setting) with an O/E ratio of 1.41 (1.05–1.78, with the SAPS II-AM2). Although ICU A has a high proportion of patients with medical admissions, it has a low O/E ratio. In contrast, ICU M has a high proportion of surgical patients but a high O/E ratio. In these circumstances we can conclude that these differences are likely related to differences in the quality of care and should therefore lead to further investigation.

Customized severity of illness models could be used in this setting as an instrument in the quality control of O/E ratios. For example, O/E ratios derived from the original model may be compared with those from the customized model. As a possibility, only O/E ratios that are significantly different from 1 in both models would then be evaluated further. The customized model would thus be used to find and filter out possible artifacts resulting from inadequate calibration of the original model.

In conclusion, our results demonstrate that today's severity scoring systems, such as the SAPS II, are limited by not measuring (and adjusting for) a profound part of what constitutes case mix. Changes in the distribution of patient characteristics (known and unknown) therefore affect prognostic accuracy. First-level customization was not able to solve all these problems. Using O/E ratios for quality of care comparisons, one must be critical when using these data and should search for possible confounding factors. In the case of unsatisfactory calibration, customized severity of illness models may be useful as an adjunct for quality control.

References

1. Moreno R, Apolone G, Reis Miranda D (1998) Evaluation of the uniformity of fit of general outcome prediction models. Intensive Care Med 24: 40–47
2. Moreno R, Apolone G (1997) Impact of different customization strategies in the performance of a general severity score. Crit Care Med 25: 2001–2008
3. Metnitz PGH, Valentin A, Vesely H, Alberti C, Lang T, Lenz K, Steltzer H, Hiesmayr M (1999) Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Intensive Care Med 25: 192–197
4. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G, Melotti RM (1996) The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Intensive Care Med 22: 1368–1378
5. Le Gall JR, Lemeshow St, Leleu G, Klar J, Huillard J, Rue M, Teres D, Artigas A, for the Intensive Care Scoring Group (1995) Customized probability models for early severe sepsis in adult intensive care patients. JAMA 273: 644–650
6. Green J, Wintfeld N. Report cards on cardiac surgeons (1995) Assessing New York State's approach. N Engl J Med 332: 1229–32
7. Rapoport J, Teres D, Lemeshow St, Gehlbach St (1994) A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. Crit Care Med 22: 1385–1391

8. Metnitz PGH, Vesely H, Valentin A, Popow C, Hiesmayr M, Lenz K, Krenn CG, Steltzer H (1999) Evaluation of an interdisciplinary data set for national ICU assessment. Crit Care Med 27: 1486–1491
9. Lemeshow S, Hosmer DW (1982) A review of goodness-of-fit statistics for use in the development of logistic regression models. Am J Epidemiol 115: 92–106
10. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36
11. Hosmer DW, Lemeshow S (1995) Confidence interval estimates of an index of quality performance based on logistic regression models. Stat Med 14: 2161–2172
12. Knaus WA, Draper EA, Wagner DP, Zimmermann JE (1995) APACHE II: a severity of disease classification system. Crit Care Med 13: 818–829
13. Miranda DR, Ryan DW, Schaufeli WB, Fidler V (1998) Organisation and management of intensive care. Update in intensive care and emergency medicine, vol 29. Springer, Berlin Heidelberg New York
14. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP (1993) Intensive care society's APACHE II study in Britain and Ireland – variations in case mix of adult admissions to general intensive care units and impact on outcome. BMJ 307: 972–977
15. Suter P, Armanigidis A, Beaufils F, Bomnfill X, Cook D, Fagot-Largeault, Thijs L, Vesconi S, Williams A (1994) Predicting outcome in ICU patients. Consensus conference organized by the ESICM and the SRLF. Intensive Care Med 20: 390–397