Intensive Care Med (2012) 38:40–46
DOI 10.1007/s00134-011-2390-2

ORIGINAL

Lilian Minne
Saeid Eslami
Nicolette de Keizer
Evert de Jonge
Sophia E. de Rooij
Ameen Abu-Hanna

Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment

Received: 27 January 2011
Accepted: 28 June 2011
Published online: 28 October 2011

This article is discussed in the editorial available at:
doi:10.1007/s00134-011-2392-0.

Electronic supplementary material The online version of this article (doi:10.1007/s00134-011-2390-2) contains supplementary material, which is available to authorized users.

L. Minne (✉) · S. Eslami ·
N. de Keizer · A. Abu-Hanna
Department of Medical Informatics,
Academic Medical Center Amsterdam,
Room J1b-124, PO Box 22660,
1100 DD Amsterdam, The Netherlands
e-mail: L.Minne@amc.uva.nl
Tel.: +31-20-5666893
Fax: +31-20-6919840

E. de Jonge
Department of Intensive Care, Leiden
University Medical Center, PO Box 9600,
2300 RC Leiden, The Netherlands

S. E. de Rooij
Department of Geriatrics, Academic
Medical Center, PO Box 22660,
1100 DD Amsterdam, The Netherlands

Abstract *Purpose:* The aim of our study was to explore, using an innovative method, the effect of temporal changes in the mortality prediction performance of an existing model on the quality of care assessment. The prognostic model (rSAPS-II) was a recalibrated Simplified Acute Physiology Score-II model developed for very elderly Intensive Care Unit (ICU) patients. *Methods:* The study population comprised all 12,143 consecutive patients aged 80 years and older admitted between January 2004 and July 2009 to one of the ICUs of 21 Dutch hospitals. The prospective dataset was split into 30 equally sized consecutive subsets. Per subset, we measured the model's discrimination [area under the curve (AUC)], accuracy (Brier score), and standardized mortality ratio (SMR), both without and after repeated recalibration. All performance measures were considered to be stable if <2 consecutive points fell outside the green zone [mean ± 2 standard deviation (SD)] and none fell outside the yellow zone (mean ± 4SD) of pre-control charts. We compared proportions of hospitals with SMR>1 without and after repeated recalibration for the year 2009. *Results:* For all subsets, the AUCs were stable, but the Brier scores and SMRs were not. The SMR was downtrending, achieving levels significantly below 1. Repeated recalibration rendered it stable again. The proportions of hospitals with SMR>1 and SMR<1 changed from 15 versus 85% to 35 versus 65%. *Conclusions:* Variability over time may markedly vary among different performance measures, and infrequent model recalibration can result in improper assessment of the quality of care in many hospitals. We stress the importance of the timely recalibration and repeated validation of prognostic models over time.

Keywords Mortality prediction · Prognostic models · Intensive care · Temporal validation · Predictive performance · Elderly patients

Introduction

Prognostic models have been promoted as helpful tools to support health care professionals in the clinical management of their patients, both as individuals and as groups [1, 2]. In the field of intensive care, there is a long tradition of developing prognostic models of mortality for benchmarking among Intensive Care Units (ICU) for purposes of quality of care assessment [3, 4]. For the purpose of benchmarking, the model's predictions are adjusted for the severity of the illness of the patients in each unit and compared to the respective observed mortality rates.

The use of prognostic models in clinical practice requires that the user trusts the models. This trust is in turn dependent upon external validation of the model's performance. However, external validation studies are scarce, thereby jeopardizing the use of these models in clinical practice [5–7]. In addition, the external validation of a model's performance often relies on a single validation dataset, while changes in population and treatment over time may change the prognosis of patients and thereby limit the applicability of prognostic models. The aim of this paper is to explore the effect of variability of the performance measures over time of an earlier published model [8] on the quality of care assessment, with and without repeated recalibration of the model. The model, referred to as rSAPS-II, was obtained as a customization of the popular Simplified Acute Physiology Score (SAPS) II model for predicting mortality in ICU patients aged 80 years or older [8, 9].

Materials and methods

In this prospective cohort study we monitored, over the course of time, the predictive performance of the rSAPS-II model by partitioning the prospective data into 30 time-ordered equally sized groups and calculating the performance measures per group.

Model

In 2007 we reported on a prognostic model that predicts the mortality risk for ICU patients aged 80 years and older [8]. The dataset consisted of 6,867 ICU admissions between January 1997 and December 2003, of which two-thirds ($N = 4{,}578$) and one-third ($N = 2{,}289$) were randomly selected for with the aim of developing and validating, respectively, the model. The admissions originated from mixed medical and surgical ICU of 21 university, teaching, and non-teaching hospitals in the Netherlands that participated in the National Intensive Care Evaluation (NICE) registry [10].

This rSAPS-II model was obtained by recalibrating the SAPS-II model [9], which was developed for a general adult ICU population, using the developmental dataset. The model was recalibrated by refitting its coefficients [8]. The linear predictor of the rSAPS-II model is: $-3.623 + 0.073 \times \mathit{SAPS\text{-}II} - 0.089 \times \log(\mathit{SAPS\text{-}II} + 1)$ where *SAPS-II* indicates the severity score. The model's performance in terms of the area under the receiver operating characteristic curve (AUC) and Brier score [± standard deviation (SD)] on the validation set are 0.77 (±0.01) and 0.16 (±0.01), respectively.

Prospective data for temporal validation

Data for prospectively validating the model over time included all 12,143 consecutive admissions of patients aged 80 years and older between January 2004 and July 2009 (at which time the number of beds and of patient admissions had increased) (Table 1). These data were obtained from the 21 original ICUs. Because the model was already published, the timing of temporal validation was naturally set, and at the same time we had total access to the data in the previous period [11]. The NICE registry, which incorporates a framework for maintaining a high quality of data [10], includes demographic data and data necessary to calculate SAPS-II, as well as survival status in the ICU and hospital.

Performance measures

We used the AUC as a discrimination measure between survivors and non-survivors (the larger the AUC the better) and the Brier score as a measure of inaccuracy (the lower the score the better) [Electronic Supplementary Material (ESM)] [12]. Our third measure was the standardized mortality ratio (SMR) [13], which is the observed number of deaths divided by the case mix-adjusted predicted number of deaths. As the SMR indicates how close predicted mortality is to the observed mortality, it can be applied in two ways: (1) the overall SMR of all hospitals is a measure of model calibration (1.0 = perfect calibration; <1.0 = over-prediction of mortality; >1.0 = under-prediction of mortality), and (2) the individual SMR of each hospital is a measure of the hospital's delivered quality of care (1.0 = expected

Table 1 Patient characteristics in the developmental, temporal and external datasets

Patient characteristics	Internal validation set[a]	Temporal validation set
N	2,289	12,143
Age (range)	80–103	80–108
Age (mean ± SD)	83.5 ± 3.6	82.5 ± 13.7
Male (%)	48.0	50.2
Died (%)	30.5	32.0
APACHE II score (mean ± SD)	18.4 ± 7.2	19.1 ± 7.4
SAPS II score (mean ± SD)	41.6 ± 17.2	43.9 ± 17.4
LOS ICU [days; (median (IQR)]	1.1 (0.8–3.2)	1.4 (0.8–3.7)
Admission type (%)		
Medical	33.8	38.7
Unplanned surgery	19.3	20.0
Planned surgery	46.8	41.3

SD Standard deviation, *APACHE* Acute Physiology And Chronic Health Evaluation, *SAPS* Simplified Acute Physiology Score, *LOS ICU* Length of stay in the Intensive Care Unit, *IQR* inter-quartile range

[a] Used in de Rooij et al. [8]

quality of care; <1.0 = better care than expected; >1.0 = poorer care than expected). As time passes, the SMR can be influenced by changes over time in both the quality of care delivered and in patient mix. Therefore, the overall SMR will tend to be closer to 1 in the immediately prospectively collected validation sets than those collected later in time. However, if we use a model with an overall SMR that deviates from 1.0 (meaning the model is poorly calibrated), the individual SMRs calculated on the basis of this model will provide an incorrect view of the delivered quality of care.

Time-series analysis

Performance measures were scrutinized using statistical process control (SPC; ESM). We split our dataset into 30 consecutive subsets of equal size N_g and calculated the performance measures described above for each set. The number of groups was chosen to be 30 because (1) the groups were still of sufficient size (≥400 admissions per group) and (2) the series was large enough to allow performance measures to be scrutinized over time. Group sizes were fixed such that their means were based on the same numbers of patients (time periods were still similar—between 2 and 3 months in each group). We used pre-control charts (also known as zone charts) which allow users to pre-specify the limits of three zones and which are intuitive by their "traffic light" design: the process is said to be stable if all points fall within the green (safe) zone, no two consecutive points fall within the yellow (warning) zone, and no points fall within the red (critical) zone [14]. An unstable process in SPC refers to a statistically significant change at the 0.05 level [14]. We defined our zones by mean values ± 2 SD (green zone), 2–4 SD (yellow zone), and >4 SD (red zone). Mean values and standard deviations of the performance measures were calculated based on the internal validation set ($N = 2{,}289$) from the previous period. Specifically, the performance measure statistics were calculated by taking 3,000, possibly overlapping, samples of size N_g from this dataset. For each sample we first calculated the performance measure of interest and then calculated the mean and standard deviation of these 3,000 values. These values were used to determine the central and control limits in the pre-control charts. In addition, to obtain insight into changes in the case-mix over time we plotted a regular graph of the original SAPS-II score, mean age, and mortality rates in the 30 groups.

Effect on quality of care assessment and recalibration

In the NICE registry [10], hospitals are annually ranked based on their individual SMR, which is used as a key measure of their delivered quality of care. To estimate the effect of prognostic model performance on the quality of care assessment, we calculated the individual SMR of each hospital for the last year, 2009, based on several rSAPS-II models that had been recalibrated on data of earlier years. Specifically, we obtained five models by using first-level recalibration (ESM) of the rSAPS-II model on the following datasets, but now used as developmental sets: 1997–2004, 1997–2005, 1997–2006, 1997–2007, and 1997–2008.

To explore the effect of *repeated* recalibration on the overall SMR, we recalibrated rSAPS-II for each time point p in the 30-point time-series on the dataset from 1997 until the period just preceding p. The model was then prospectively evaluated on the dataset at time point p.

Results

Figures 1 and 2 show the behavior of the performance measures over the course of time. The AUC and Brier graphs in Figs. 1 and 2 both show no signs of instability: with the exception of two non-consecutive measurements in the yellow zone, all measurements remain inside the green zone. In the SMR graph shown at the top of Fig. 3, all measurements until the 23rd measurement remain inside the green zone; thereafter, however, the SMR is clearly unstable: four consecutive measurements fall within the lower yellow zone and after two consecutive measurements in the green zone, two consecutive measurements once again fall within the yellow zone. A downward trend is clearly visible, and eight measurements are significantly <1.0, as illustrated by their 95% confidence intervals. Although fluctuating around their mean values, AUCs, Brier scores, and SMRs all show some relatively large jumps in consequent values. For example, the AUC changed from 0.75 at measurement 25 to 0.81 at the following measurement.

The graph at the bottom of Fig. 3 shows that after repeated recalibration of rSAPS-II, all overall SMRs from the 23rd measurement onward remain inside the green zone, with none significantly <1. The graph at the middle of Fig. 3 shows the behavior of the original SAPS-II score over time. On the whole, the series of severity of illness score shows a positive trend, while mortality and mean age do not show any trend over time (Figs. 4, 5). The implications of repeated recalibration can be shown by measuring its effects on the SMRs of individual hospitals. Based on the rSAPS-II model without any further recalibration, three of 20 hospitals in 2009 (15%) were assessed to deliver poorer care than expected according to their SMR (SMR >1), while the other 17 (85%) were

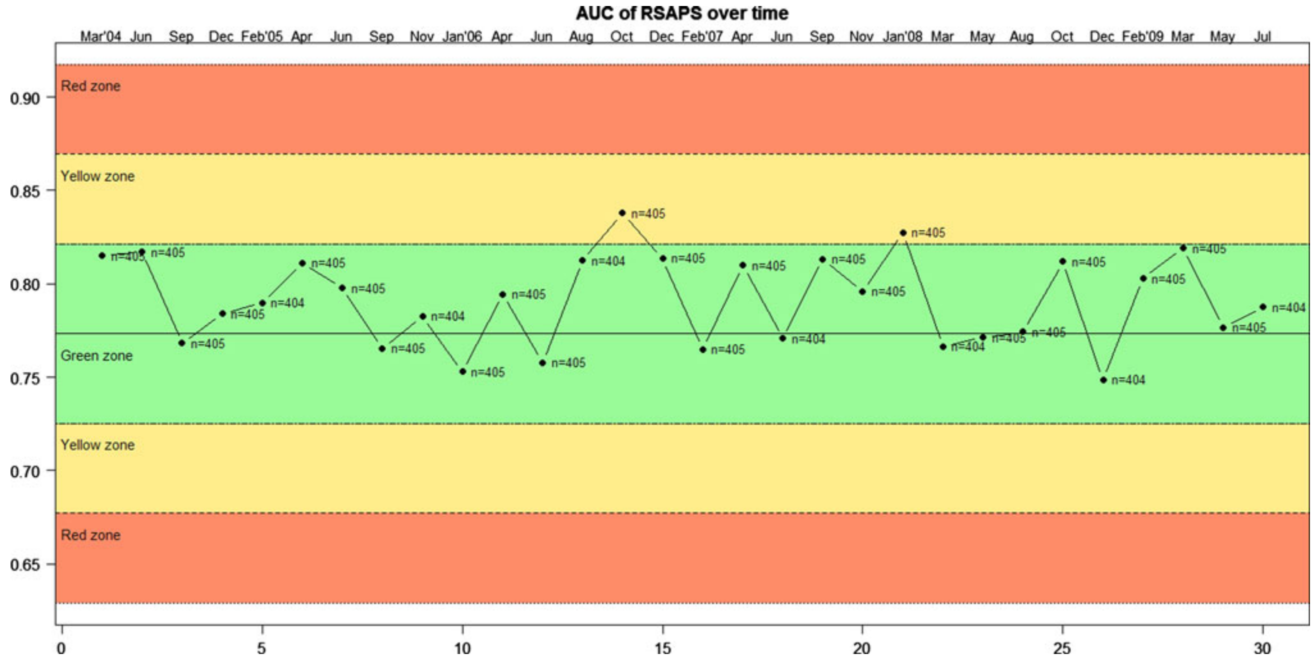

Fig. 1 Area under the receiver operating characteristic curve (*AUC*) of the rSAPS-II over time. Means and standard deviations (SD) are based on the bootstrap sample distribution obtained by taking 3000 samples of size 405 from the internal validation set. *rSAPS-II* Model developed for assessing the quality of care based on the Simplified Acute Physiology Score (*SAPS*)

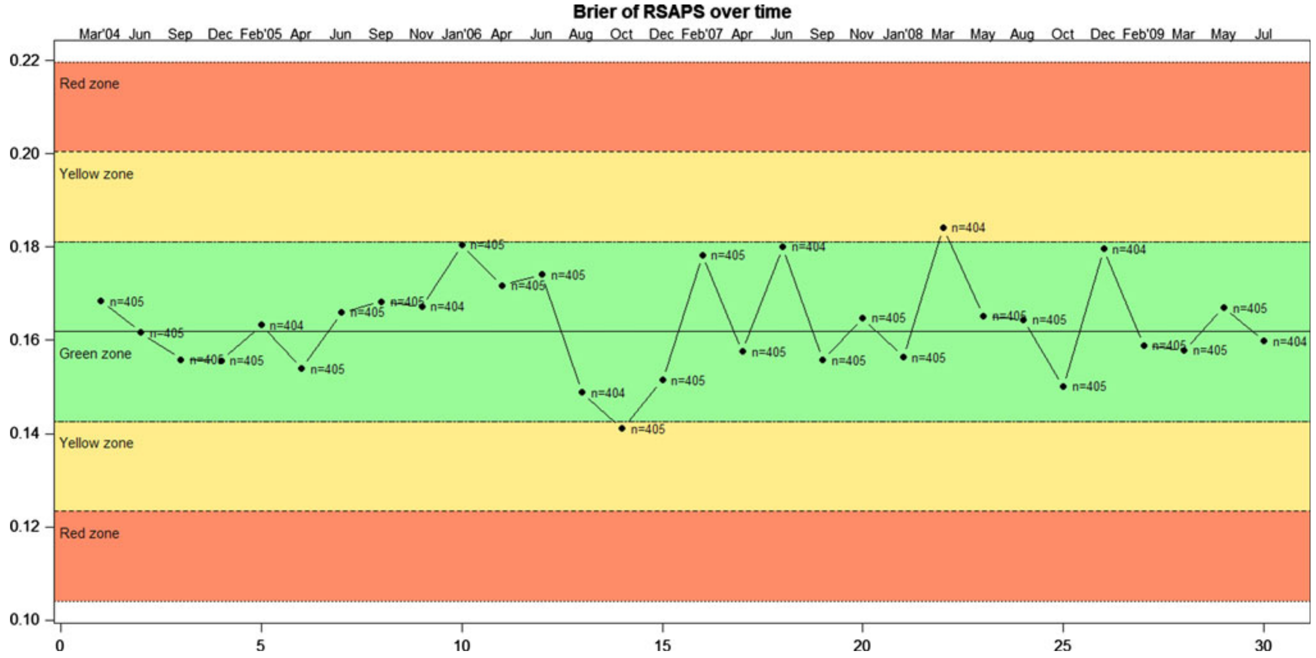

Fig. 2 Brier scores of rSAPS-II over time. Means and standard deviations (SD) are based on the bootstrap sample distribution obtained by taking 3000 samples of size 405 from the internal validation set

assessed to deliver better care than average (SMR <1). After recalibration of the rSAPS-II model on data until 2004, 2005, 2006, 2007, and 2008, the percentage of hospitals with SMR >1 and SMR <1 gradually changed to 35 and 65%, respectively (Table 2).

Discussion

The temporal validation of an earlier published model for predicting mortality risk in elderly ICU patients revealed

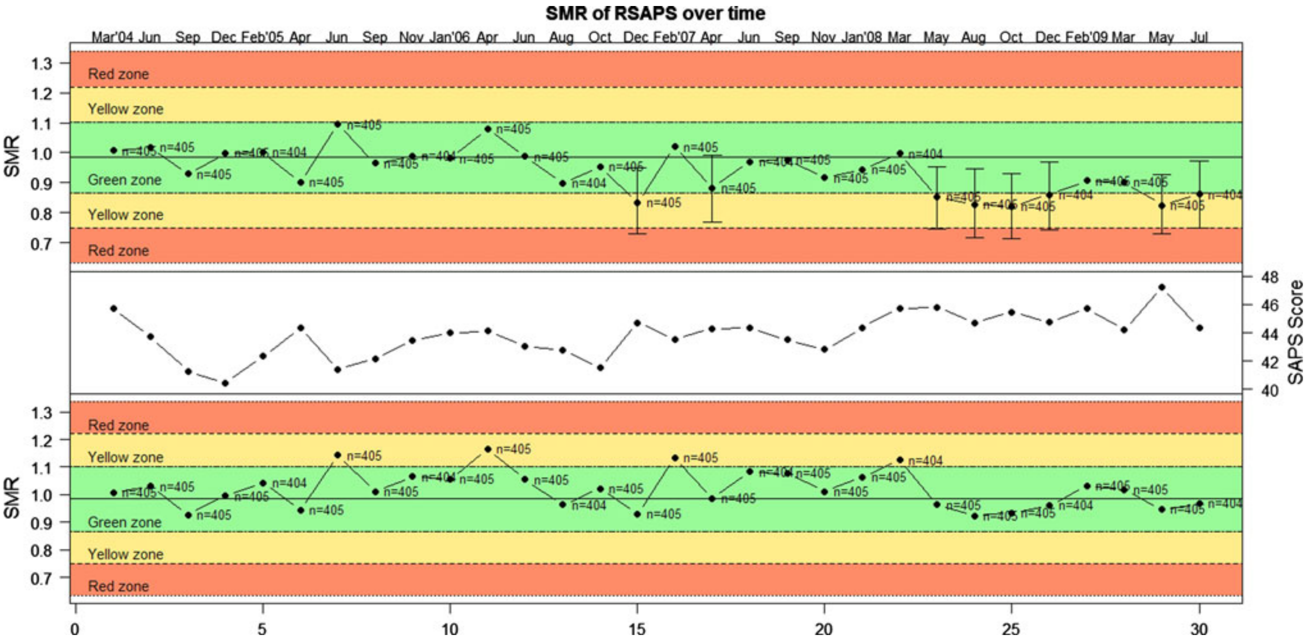

Fig. 3 Standardized mortality ratio (*SMR*) of rSAPS-II over time. Means and standard deviations (SD) are based on the bootstrap sample distribution obtained by taking 3000 samples of size 405 from the internal validation set

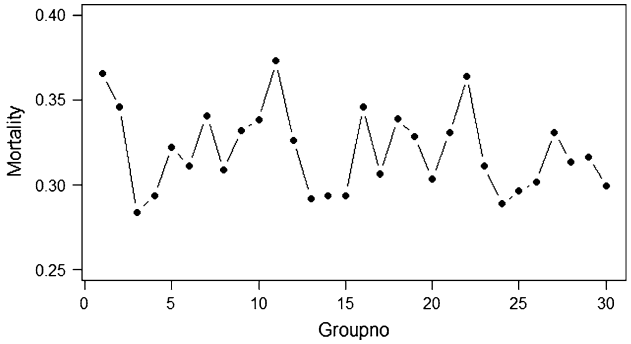

Fig. 4 Mortality over time

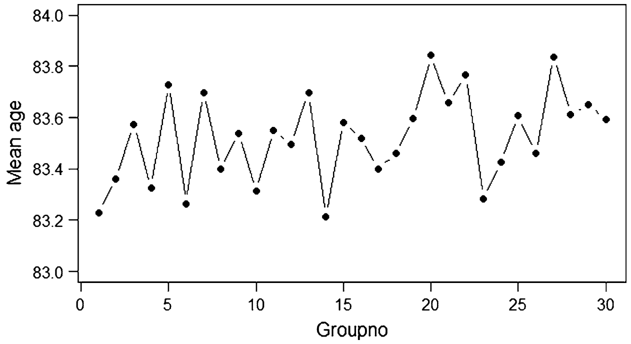

Fig. 5 Mean age over time

Table 2 Percentage of hospitals with a SMR above 1 or below 1 after recalibration

Data used for recalibration	% of hospitals with SMR >1.0	% of hospitals with SMR <1.0
1997–2003	15	85
1997–2004	20	80
1997–2005	25	75
1997–2006	25	75
1997–2007	30	70
1997–2008	35	65

SMR Standardized mortality rate

that significant differences appeared over time between the expected and observed values of the SMR, while the AUC and Brier score did not show this behavior. One may expect a gradual deterioration of model performance, but this trend was only visible in the overall SMR [15, 16]. There was a slight increase in illness severity, as demonstrated by the original SAPS-II score, but in an ideal model the overall SMR should not be affected. A steady decrease of the overall SMR is caused by the overestimation of mortality, indicating that the model is outdated. When the quality of care provided by ICUs is assessed by such a model, a significant proportion of the ICUs would appear to be performing better (SMR <1) than the norm (SMR = 1), whereas in practice they may actually be underperforming. Repeated recalibration of models adequately alleviates these problems.

To the best of our knowledge, this is the first published study exploring a prognostic model's prospective performance repeatedly over the course of time. In an earlier study [17], Harrison et al. measured the prospective performance of SAPS-II, Acute Physiology Score And

Chronic Health Evaluation (APACHE) II and III, and the Mortality Prediction Model (MPM) II at three different moments; however, they did not repeat any measurements nor did they address the policy implications of using the SMR for assessing the quality of care. We scrutinized the behavior of three different measures (AUC, Brier score, and SMR) covering discrimination, accuracy, and calibration performance aspects, but used fewer measures than Harrison et al. [17]. Mean values and standard deviations (for setting the zone limits) were obtained from the previous period, such that all of the prospective data can be used. Our sampling approach showed no marked deviations from normality in the performance measure distributions. Pseudo R^2 measures [1 − Brier score/overall mortality × (1 − overall mortality)], which adjust for mortality percentage per group, yielded the same pattern as the Brier score (data not shown). We used 30 prospective groups for analysis, which are firmly within the acceptable range of 12–36 groups [14]. We validated the robustness of our approach by using different numbers (and hence sizes) of groups, which all yielded the same patterns (data not shown). The model we investigated was already published so the start of the temporal validation was not subject to arbitrary choices. The number of patients in the temporal validation set was large, and prospective data collection covered a period of 6 years.

A main limitation of our study is the restriction of the population to elderly ICU patients. However, this is an important and growing population, and the rSAPS-II model was already recalibrated for this population and had good performance when it was internally validated. Moreover, our work addresses the applicability of prognostic models in general. The robustness of our approach could have also been validated more extensively by other means. For example, it is possible to investigate the model's behavior on different proportions of randomly selected patients in each time interval to inspect the sensitivity of the results to the specific case-mixes. These kinds of sensitivity analyses merit future research.

Our findings signify the caveats of not timely calibrating prognostic models and the importance of assessing prognostic model performance over the course of time. Routine recalibration is imperative to adjust for a changing environment [16, 18, 19]. In their evaluation of risk models, Harrison et al. [17] found a "shelf life" of about 3 years before recalibration was required, which is similar to the approximately 4 years between recalibration in our study. We advise for the continuous monitoring of risk models and subsequent recalibration when an overall recalibration is observed to be worsening. Pre-control charts provide a comprehensive way to distinguish genuine worsening from considerable noise by using yellow (warning) and red (critical) zones. Although we succeeded in rendering our model stable again by using the simplest form of recalibration, more rigorous techniques may sometimes be needed, from recalculating the coefficients of each individual variable to removing or adding new variables [18, 19]. While refraining from repeated model recalibration may not extensively change the relative ranking among hospitals [20], we have demonstrated that the percentage of hospitals with poorer care than expected (SMR >1) changed markedly after recalibration from 15 to 35%.

An increased number of temporal (and external) validation studies are needed for users to acquire an understanding of a model's behavior over the course of time in various domains. Important questions to address include deciding on the required frequency for recalibrating models, various approaches for recalibration, and ways to determine the extent of influence that older observations should exert when recalibrating models in a dynamically changing environment. In all of these efforts, policy implications of model's performance should play a central role because the use of these models requires trust and, in turn, this trust requires an extensive understanding of the effect of models' performance on policy decisions.

Conclusion

The Brier scores and SMR of rSAPS-II showed statistically significant differences between expected and observed values over time, but the AUC did not show this behavior. Thus, variability patterns over time may markedly vary among different performance measures, thereby illustrating the importance of using a set of measures covering both aspects of model discrimination and calibration. The worsening of the overall SMR resulted in an improper assessment of quality of care for many hospitals. Repeated recalibration of models adequately alleviated these problems. Our findings stress the importance of timely recalibrating prognostic models and the assessment of its performance repeatedly over the course of time. More temporal (and external) validation studies are needed to understand models' behavior over the course of time in various domains.

Conflict of interest The authors declare that they have no competing interests. The authors have full control of all primary data and agree to allow the journal to review their data if requested.

References

1. Lucas PJ, Abu-Hanna A (2009) Prognostic methods in medicine. Artif Intell Med 15:105–119
2. Abu-Hanna A, Lucas PJ (2001) Prognostic models in medicine. AI and statistical approaches. Methods Inf Med 40:1–5
3. Zimmerman D (1999) Benchmarking: measuring yourself against the best. Trustee 52:22–23
4. Zimmerman JE, Alzola C, Von Rueden KT (2003) The use of benchmarking to identify top performing critical care units: a preliminary assessment of their policies and practices. J Crit Care 18:76–86
5. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? Br Med J 338:b375
6. Wyatt JC (1995) Prognostic models: clinically useful or quickly forgotten? Br Med J 311:1539–1541
7. Mallet S, Royston P, Waters R, Dutton S, Altman DG (2010) Reporting performance of prognostic models in cancer: a review. BMC Med 8:21
8. de Rooij SE, Abu-Hanna A, Levi M, de Jonge E (2007) Identification of high-risk subgroups in very elderly intensive care unit patients. Crit Care 11:R33
9. Le Gall JR, Lemeshow S, Saulnier F (1993) A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA 270:2957–2963
10. de Jonge E, Bosman RJ, van der Voort PH, Korsten HH, Scheffer GJ, de Keizer NF (2003) Intensive care medicine in the Netherlands, 1997–2001. I. Patient population and treatment outcome. Ned Tijdschr Geneeskd 147:1013–1017
11. Altman DG, Vergouwe Y, Royston P, Moons KG (2009) Prognosis and prognostic research: validating a prognostic model. Br Med J 338:b605
12. Altman DG (1990) Practical statistics for medical research. Chapman and Hall, London
13. Fleis J, Levin B, Paik M (2003) Statistical methods for rates and proportions. J Wiley, New York
14. Wheeler JW (2004) Advanced topics in statistical process control. SPC Press, Knoxville
15. Kramer AA (2005) Predictive mortality models are not like fine wine. Crit Care 9:636–637
16. Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B et al (2005) Mortality prediction using SAPS II: an update for French intensive care units. Crit Care 9:R645–R652
17. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K (2006) 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 34:1378–1388
18. Steyerberg EW (2009) Clinical Prediction Models. A practical approach to development, validation, and updating. Springer, New York
19. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD (2004) Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 23:2567–2586
20. Bakhshi-Raiez F, Peek N, Bosman RJ, de Jonge E, de Keizer NF (2007) The impact of different prognostic models and their customization on institutional comparison of intensive care units. Crit Care Med 35:2553–2560