Intensive Care Med (2005) 31:1336–1344
DOI 10.1007/s00134-005-2762-6

ORIGINAL

# SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description

**Philipp G. H. Metnitz**
**Rui P. Moreno**
**Eduardo Almeida**
**Barbara Jordan**
**Peter Bauer**
**Ricardo Abizanda Campos**
**Gaetano Iapichino**
**David Edbrooke**
**Maurizia Capuzzo**
**Jean-Roger Le Gall**
**on behalf of the SAPS 3 Investigators**

Received: 8 April 2005
Accepted: 22 July 2005
Published online: 17 August 2005


Electronic Supplementary Material
Electronic supplementary material is included in the online fulltext version of this article and accessible for authorised users: http://dx.doi.org/10.1007/s00134-005-2762-6

P. G. H. Metnitz (✉)
Dept. of Anesthesiology
and General Intensive Care,
University Hospital of Vienna,
Vienna, Austria
e-mail: philipp@metnitz.biz
Fax: +43-1-9522843

R. P. Moreno
Unidade de Cuidados Intensivos
Polivalente,
Hospital de St. António dos Capuchos,
Centro Hospitalar de Lisboa (Zona Central),
Lisboa, Portugal

E. Almeida
Unidade de Cuidados Intensivos,
Hospital Garcia de Orta,
Pragal, Portugal

B. Jordan
Dept. of Medical Statistics,
University of Vienna, Vienna, Austria

R. A. Campos
Dept. of Intensive Care,
Hospital Universitario Asociado General
de Castelló, Castello, Spain

G. Iapichino
Dept. of Anesthesia
& Intensive Care Medicine,
Hospital San Paolo, Università degli Sudi,
Milano, Italy

D. Edbrooke
Critical Care Directorate,
Royal Hallamshire Hospital,
Sheffield, United Kingdom

M. Capuzzo
Dept. of Anesthesia
& Intensive Care Medicine,
Hospital of Ferrara,
Ferrara, Italy

J.-R. Le Gall
Dept. Réanimation Médicale,
Hôpital St. Louis, Université Paris VII,
Paris, France

**Abstract** *Objective:* Risk adjustment systems now in use were developed more than a decade ago and lack prognostic performance. Objective of the SAPS 3 study was to collect data about risk factors and outcomes in a heterogeneous cohort of intensive care unit (ICU) patients, in order to develop a new, improved model for risk adjustment. *Design:* Prospective multicentre, multinational cohort study. *Patients and setting:* A total of 19,577 patients consecutively admitted to 307 ICUs from 14 October to 15 December 2002. *Measurements and results:* Data were collected at ICU admission, on days 1, 2 and 3, and the last day of the ICU stay. Data included sociodemographics, chronic conditions, diagnostic information, physiological derangement at ICU admission, number and severity of organ dysfunctions, length of ICU and hospital stay, and vital status at ICU and hospital discharge. Data reliability was tested with use of kappa statistics and intraclass-correlation coefficients, which were >0.85 for the majority of variables. Completeness of the data was also satisfactory, with 1 [0–3] SAPS II parameter missing per patient. Prognostic performance of the SAPS II was poor, with significant differences between observed and expected mortality rates for the overall cohort and four (of seven) defined regions, and poor calibration for most tested subgroups. *Conclusions:* The SAPS 3 study was able to provide a high-quality multinational database, reflecting heterogeneity of current ICU case-mix and typology. The poor performance of SAPS II in this cohort underscores the need for development of a new risk adjustment system for critically ill patients.

**Keywords** Intensive care unit · Severity of illness · ICU mortality · Hospital mortality · Risk adjustment

## Introduction

Following the publication in the early 1980s of the Acute Physiology and Chronic Health Evaluation Score (APACHE [1]), Simplified Acute Physiology Score (SAPS [2]), and-some years later—APACHE II [3] systems, outcome prediction became an important topic among European intensivists. Ten years later, a new generation of these instruments was published: APACHE III [4], SAPS II [5], and Mortality Probability Model (MPM) II [6]. All of these newer systems were developed by using sophisticated statistical techniques in large multinational databases, and were found to perform better than their predecessors [7, 8].

The availability of such sophisticated methods for risk adjustment facilitated outcome research in critically ill patients, which became increasingly important over time. Risk adjustment systems now have a fixed place in critical care research for various purposes. At the patient level, the reporting of severity of illness and the use of risk-adjusted mortality rates to draw inferences from their results are a prerequisite for any study to be published. At the intensive care unit (ICU) level, observed-to-expected mortality ratios (or the use of direct standardisation techniques based on severity scores) have become standard for assessing the impact of ICU-related factors on outcome, such as the effects of organisation and management [9, 10].

However, a series of studies assessing the performance of risk adjustment systems unveiled a lack of prognostic performance of these systems: In most cases, lack of calibration was evident over several subgroups of patients, often accompanied by an underestimation of mortality in low-risk patients and an overestimation in high-risk patients. This pattern was observed for all published outcome prediction models in several countries [11, 12, 13, 14, 15, 16, 17, 18] and seemed to be worsening over time [19].

For this reason, several researchers tried to improve the prognostic performance of various systems through recalibration, using one of two possible approaches. A level 1 customization requires calculation of a new equation for the prediction of hospital mortality (without changing the weights of the constituent variables). A level 2 customization involves a reweighting of each variable contained in the model. Although recalibration was able to improve prognostic accuracy in some cases [13, 14], it generally did not solve the various problems inherent in the models.

These problems can be classified as either user-, patient-, or model-dependent. User-dependent problems include differences in the definitions and application criteria [20, 21]. Patient-dependent problems are mainly shifts in the baseline characteristics of the populations over time [22]: age distribution, distribution of illnesses, and the development of new treatments, all of which affect prognosis. Model-dependent problems have many different causes, such as the lack of important prognostic variables (e.g., diagnostic information [4, 23]) or the presence, location and aetiology of infection [24, 25, 26]. Confounding variables and statistically wrong assumptions [9, 27] also distort performance results.

If recalibration is not sufficient to improve the performance of the prognostic model, the only alternative is to develop a new model that takes into account the results of studies done since the original model was developed. This means incorporating missing variables that have been shown to affect outcome, minimizing problems with the application of the model, and reducing the possibility of other confounders.

The objective of the SAPS 3 project was to cope with the above-stated problems by developing a new model for improved risk adjustment in critically ill patients. Another important goal was to make the new model available free of charge for use in the scientific community.

In the SAPS 3 study (which took place at the end of 2002), data about risk factors and outcomes in an international multicentric cohort of critically ill patients were prospectively collected so that a high-quality database would be available for further analysis of the associations between risks and outcomes in our patients.

## Materials and methods

### Project Organization

The SAPS 3 project was conducted by the SAPS 3 Outcomes Research Group. The project was endorsed by the European Society for Intensive Care Medicine (ESICM, http://www.esicm.org) and conducted in cooperation with the Section on Health Services Research and Outcome of the ESICM. The SAPS 3 Outcomes Research Group consists of a project coordinator and a steering group. The steering group was responsible for the scientific conduct and consistency of the project. An additional advisory board integrated further scientists with special expertise who were asked for comments on the scientific content and for help in conducting the project. The complete board lists can be found in Appendix D of the Electronic Supplementary Material (ESM).

During the data collection phase, a coordination and communications centre (CCC) was installed. The CCC was responsible for the management and control of the project. This included the administration of all project tasks and implementation of actions and activities as necessary; communication between project partners (e.g., centres, researchers and institutions) through sampling and distribution of necessary information; and pooling and administration of the data provided by project participants. In addition, the CCC provided almost around-the-clock service to answer urgent questions and resolve problems during the phase of data collection.

In each country, a country coordinator was responsible for operational management and direct communication with the participating ICUs in that country, including giving specific help when necessary. The country coordinator was responsible for ensuring completion of the various tasks required of the participating ICUs. The list of country coordinators can be found in Appendix E of the ESM.

At the ICU level, an ICU coordinator was responsible for local activities, such as obtaining approval from the local ethics or data-

protection committees where applicable. In addition, the ICU coordinator was responsible for supervising the daily data collection, problem management, controlling the completeness of the data, data quality control, training medical and nonmedical staff for data collection, management of the data, and transmission of the data to the CCC or country coordinator. The list of ICU coordinators can be found in Appendix F of the ESM.

## Data collection

Patient data were recorded by using either online data collection software (provided by *i*MD*soft*, Tel Aviv, Israel) or the SAPS 3 stand-alone database system (provided by the CCC). The latter software used a Microsoft Access database (Microsoft Corporation, Redmond, WA, USA) for data storage and needed no Internet connection for data entry. Both systems maintained a variety of plausibility controls to ensure the quality of the recorded data. Each variable was precisely defined before the start of data collection (see Appendix C of the ESM). Detailed definitions of the variables were available to participants in both paper and electronic form. To facilitate plausibility checking, each variable was assigned a probability range, encompassing the range of probable values for that variable. In addition, a range of possible values (storage range) for that variable was defined (e.g., for $FiO_2$, no values <21% or >100% could be accepted). Thus, formal plausibility controls in the software systems were used wherever possible and ensured the maximum of data quality checking during data collection.

Participants who could not use one of the two software options were allowed to record the data on paper forms and submit them to the CCC ($n$=26 ICUs). Patient data were then entered into the SAPS 3 stand-alone software system and thus checked for plausibility. In cases of uncertainty, ICU coordinators were contacted for clarification.

In addition, each ICU received a questionnaire with detailed questions about ICU structures and about resources available in other areas of the hospital.

Data were collected at ICU admission, on days 1, 2 and 3, and on the last day of the ICU stay. Data from the day of admission (aside from sociodemographic data such as age and sex) were categorized into different levels: (i) data about the condition of the patient before ICU admission, such as chronic conditions and medical diseases; (ii) data about the patient's condition at ICU admission, such as the reason for admission, infection at admission, and surgical status; and (iii) data about the patient's physiologic derangement at ICU admission. These data were collected within an hour before or after ICU admission.

On the following days of the ICU stay, further information was collected: severity of illness, as measured by the SAPS II [5]; number and severity of organ dysfunction, as measured by the Sequential Organ Failure Assessment (SOFA) [28]; length of ICU and hospital stay; and outcome data, including vital status at ICU and hospital discharge. All patients were subjected to mandatory follow-up until hospital discharge, but not longer than 90 days after ICU admission. Patients still remaining in the hospital at 90 days were at that time classified as being "still in the hospital".

To record diagnoses, a three-level system was used. (i) An a*cute medical disease* was recorded for all patients, independent of surgical status, i.e., the acute (or acute on chronic) disease that best explained the ICU admission. If the reason for ICU admission was infectious disease, then this was recorded. (ii) *Surgical status* at admission and the anatomic site of surgery were recorded for all patients undergoing surgery during the hospital stay before ICU admission. (iii) A concrete *reason for admission* had to be selected. At least one reason for admission was required, but several selections were possible (one within each organ system). If no other reason was present, at least "basic and observational care" had to be selected.

All participants received detailed documentation of patient- and ICU-based data items as well as a detailed description of the data collection process. Moreover, specific forms to check the completeness of the patient-based documentation were provided. Additionally, a training session for ICU coordinators was organised at the 15th Annual Congress of the ESICM in Barcelona, Spain, before the start of data collection. Throughout the project, the project website (http://www.saps3.org) provided all necessary information. In addition, the CCC was available to answer questions by email, fax and phone. Data were to be collected from all consecutively admitted patients between 14 October and 15 December 2002. ICUs with a high number of beds (and thus also admissions) could stop patient enrolment after contributing 100 patients.

## Database

Data were collected and pooled by the CCC. The final database file was then imported into the SAS system, Version 8e (SAS Institute Inc., Cary, USA). Data cleaning was accomplished through the application of a variety of additional plausibility controls and cross-checking of information between redundant data fields.

A total of 22,791 admissions were recorded in the 309 participating ICUs during the study period. For patients who were admitted more than once ($n$=1,455), only the first admission was included, giving 21,336 admitted patients. Patients who were <16 years of age ($n$=628), those without ICU admission or discharge data ($n$=1,074), and those with records that lacked an entry in the field "ICU outcome" ($n$=57) were excluded. The *Basic SAPS 3 Cohort* thus comprises 19,577 patients from 307 ICUs.

For the development of a predictive model for hospital mortality as outcome, patients with a missing entry in the field of "vital status at hospital discharge" ($n$=2,540) or an entry of "still in the hospital" at the end of the follow-up period ($n$=253) were further excluded. The *SAPS 3 Hospital Outcome Cohort* thus comprises 16,784 patients from 303 ICUs.

Because the study was an observational study and no additional interventions were performed, the need for informed consent was waived by the institutional review board. Each ICU, however, was made responsible for obtaining local permissions as necessary.

## Data quality

Recorded data were evaluated for completeness of the documentation and reliability. Interrater quality control was performed through rescoring of the data from day 0 (the day of ICU admission) for three randomly selected patients in each ICU. From the rescored data, kappa coefficients and intra-class correlation coefficients were calculated, as appropriate. Availability of the variables necessary to calculate the SAPS II was used as an indicator for the completeness of the data.

## Statistical analysis

Statistical analysis was performed using the SAS system, version 8e (SAS Institute Inc., Cary, NC, USA). A $P$ value of <0.05 was considered significant. Unless otherwise specified, results are expressed as median and interquartile ranges (quartile). Observed-to-expected (O/E) mortality ratios were calculated by dividing the number of observed deaths per group by the number of expected deaths per group (as predicted by the SAPS II). To test for statistical significance, we calculated 95% confidence intervals (CI) according to the method described by Hosmer and Lemeshow [29]. The Hosmer-Lemeshow goodness-of-fit $\hat{H}$-statistic and $\hat{C}$-statistic [30] were used to evaluate the calibration of the SAPS II. Discrimination was tested by measuring the area under the receiver

operating characteristic (aROC) curve, as described by Hanley and McNeil [31]. Reliability of data collection was analysed using K-statistics or intra-class correlation coefficients, as appropriate. Statistical methods used for the development of the predictive model are described in Part 2 of this report.

## Results

### Data quality

Four hundred eighty-three rescored patients could be identified and linked to their original counterparts (2.5% of admitted patients). Data quality was found to be excellent, with the majority of coefficients being >0.85. Only two of the more than 50 tested variables had coefficients <0.80 (body weight, 0.79; positive end-expiratory pressure, 0.72), and only one was <0.70 (leukocytes [maximum], 0.57). For a detailed list of coefficients see Table E1 in the ESM. Data completeness was also found to be satisfactory, with 1 [0–3] SAPS II parameter missing per patient.

### Description of ICUs

The Basic SAPS 3 cohort includes 307 ICUs from 35 countries. On average each ICU contributed 50 (27–78) patients to the cohort. To assess heterogeneity of results between different geographic regions, seven regions were defined: Australasia, Central and South America, Central and Western Europe, Eastern Europe, North America, Northern Europe, and Southern Europe and Mediterranean countries. The allocation of countries to these regions can be seen from Table E10 of the ESM.

Seventy percent of the participating ICUs identified themselves as mixed medical-surgical (Table E2, ESM). Roughly half of the ICUs (46%) were located in university-affiliated or teaching hospitals. Eighty-four percent of ICUs (*n*=258) reported having a full-time medical director, and 272 (88.6%) reported having a full-time nursing director. On weekdays, 76.6% of ICUs reported having an intensivist available on the ICU 24 hours per day, whereas 6.2% had an intensivist available in the hospital. In 12.1% of ICUs, the intensivist was at home, on-call, during the daytime. During weekends, this proportion did not change much (74.3%, 5.5%, and 15.0% on the ICU, in the hospital, and on-call, respectively). None of the participating ICUs reported having no intensivists available during night or weekend shifts.

### Description of patients

The Basic SAPS 3 Cohort comprises 19,577 patients admitted to participating ICUs during the study period. More than 70% of patients were admitted from the same hospital as the ICU, with operating rooms, emergency departments and normal wards contributing most of the patients (Table 1). Almost two thirds of the admissions were classified as unplanned. The mean age of patients was 60.0±17.7 years (Fig. 1), and 39.2% were female. Comorbidities were recorded in 65% of admitted patients, with arterial hypertension, chronic obstructive pulmonary disease, and chronic heart failure being the most frequent (Table E3, ESM).

Cardiovascular, respiratory and neurologic diseases were the most frequent organ-specific reasons for admission (Table E4, ESM). The acute medical diseases

**Table 1** ICU admission data for the two cohorts (*Basic cohort:* SAPS 3 basic cohort; *HO cohort:* SAPS 3 Hospital Outcome Cohort; *n:* number of patients)

| | Basic cohort | | HO cohort | |
|---|---|---|---|---|
| | *n* | % | *n* | % |
| Number of patients | 19,577 | | 16,784 | 100.0 |
| Gender | | | | |
| Female | 7,678 | 39.2 | 6,610 | 39.4 |
| Male | 11,881 | 60.7 | 10,161 | 60.5 |
| Missing | 18 | 0.1 | 13 | 0.1 |
| Age, years (median, quartiles) | 63 | 49-74 | 64 | 49-74 |
| Origin | | | | |
| Home | 2,810 | 14.4 | 2,343 | 14.0 |
| Same hospital | 13,926 | 71.1 | 12,063 | 71.9 |
| Chronic care facility | 74 | 0.4 | 64 | 0.4 |
| Public place | 519 | 2.7 | 432 | 2.6 |
| Other hospital | 2,125 | 10.9 | 1,791 | 10.7 |
| Other | 80 | 0.4 | 59 | 0.4 |
| Missing | 43 | 0.2 | 32 | 0.2 |
| Intra-hospital location before ICU admission | | | | |
| Emergency room | 5,419 | 27.7 | 4,630 | 27.6 |
| Intermediate care unit/ High dependency unit | 562 | 2.9 | 475 | 2.8 |
| Operating room | 7,537 | 38.5 | 6,449 | 38.4 |
| Other | 552 | 2.8 | 413 | 2.5 |
| Other ICU | 698 | 3.6 | 611 | 3.6 |
| Recovery room | 482 | 2.5 | 400 | 2.4 |
| Ward | 3,411 | 17.4 | 3,036 | 18.1 |
| Missing | 916 | 4.7 | 770 | 4.6 |
| ICU admission status | | | | |
| Planned admission | 6,750 | 34.5 | 5,598 | 33.4 |
| Unplanned admission | 12,338 | 63.0 | 10,801 | 64.4 |
| Missing | 489 | 2.5 | 385 | 2.3 |
| Acute Infection at ICU admission | | | | |
| No infection | 15,254 | 77.9 | 12,968 | 77.3 |
| Clinically improbable/ colonization | 342 | 1.7 | 298 | 1.8 |
| Clinically probable/ documented | 2,761 | 14.1 | 2,422 | 14.4 |
| Microbiologically documented | 1206 | 6.2 | 1,083 | 6.5 |
| Missing | 13 | 0.1 | 13 | 0.1 |
| Surgical status | | | | |
| No surgical procedure | 8,437 | 43.1 | 7,305 | 43.5 |
| Scheduled surgery | 6,800 | 34.7 | 5,700 | 34.0 |
| Emergency surgery | 3,321 | 17.0 | 2,930 | 17.5 |
| Missing | 1,019 | 5.2 | 849 | 5.1 |

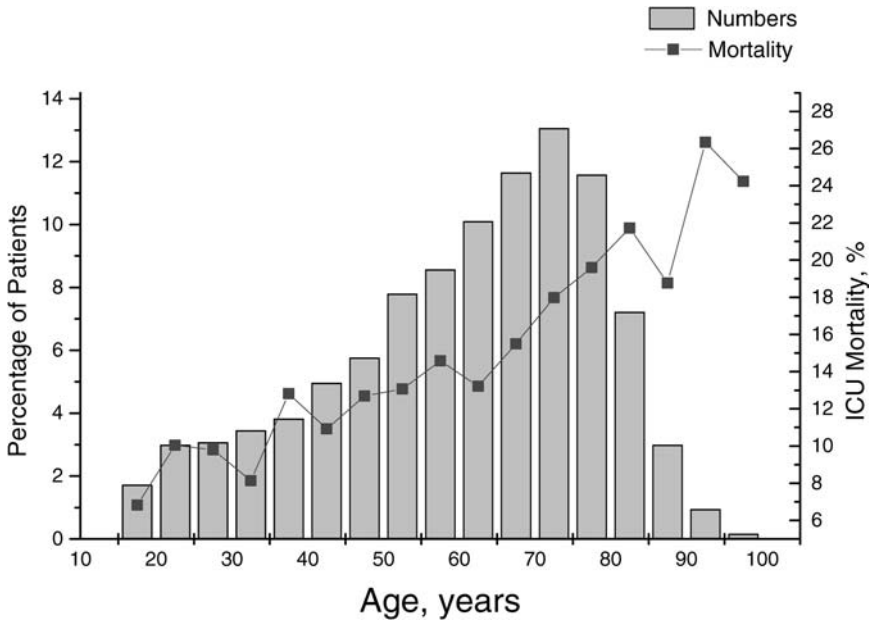


**Fig. 1** Age distribution and associated mortality. The figure shows the age distribution of the Basic SAPS 3 Cohort ($n$=19,577) and the corresponding ICU mortality rates for each stratum. *Columns:* Number of patients as percentages of the whole cohort; *squares:* ICU mortality rates for the corresponding stratum

**Table 2** ICU discharge and outcome data for the two cohorts (*Basic cohort:* SAPS 3 basic cohort; *HO cohort:* SAPS 3 Hospital Outcome Cohort; *n:* number of patients; *ICU LOS:* ICU length of stay; *IMCU/HDU:* intermediate care unit/high dependency unit; *Q1, Q3:* lower and upper interquartile range, respectively)

| | Basic cohort | | HO cohort | |
|---|---|---|---|---|
| | *n* | % | *n* | % |
| Number of patients | 19,577 | 100.0 | 16,784 | 100.0 |
| ICU LOS, days (median, quartiles) | 2 | 1–6 | 2 | 1–6 |
| ICU discharge–destination | | | | |
| Home | 438 | 2.2 | 361 | 2.2 |
| Same hospital | 14,946 | 76.3 | 12,477 | 74.3 |
| Other hospital | 1,029 | 5.3 | 852 | 5.1 |
| Missing | 3,164 | 16.2 | 3,094 | 18.4 |
| Intrahospital discharge | | | | |
| Emergency room | 58 | 0.3 | 50 | 0.3 |
| IMCU/HDU | 2,222 | 11.4 | 1,873 | 11.2 |
| Other | 303 | 1.5 | 257 | 1.5 |
| Other ICU | 583 | 3.0 | 479 | 2.9 |
| Recovery room | 306 | 1.6 | 218 | 1.3 |
| Ward | 12,250 | 62.6 | 10,291 | 61.3 |
| Missing | 3,855 | 19.7 | 3,616 | 21.5 |
| ICU discharge—status | | | | |
| Planned discharge | 14,872 | 76.0 | 12,262 | 73.1 |
| Unplanned discharge | 1,595 | 8.1 | 1,467 | 8.7 |
| Missing | 3,110 | 15.9 | 3,055 | 18.2 |
| Risk adjustment | | | | |
| SAPS II score (median, Q1–Q3) | 30 | 20–42 | 31 | 21–43 |
| SOFA score (median, Q1–Q3) | 9 | 6–11 | 9 | 6–11 |
| Outcome | | | | |
| ICU mortality (%) | | 15.2 | | 17.7 |

necessitating ICU admission included a broad spectrum of diagnoses (Table E5, ESM). Approximately one half of the patients underwent surgery before ICU admission, with abdominal, cardiac and vascular surgery being the most frequent procedures (Table E6, ESM).

Regarding discharge details (Table 2), it is notable that a high percentage of patients were discharged unplanned (8.15%), i.e., without at least a 12-hour planning window. 15.2% of patients from the SAPS 3 Basic cohort died within the ICU. As can be seen from Table 3, patient cohorts differed significantly between regions. Both, ICU and hospital mortality rates exhibited a broad spectrum between ICUs: hospital mortality was on average 28% (17–42%) in the SAPS 3 Hospital outcome cohort.

## Performance of the SAPS II

The performance of the original SAPS II model [5] (using data from the first 24 hours) was tested in the SAPS 3 Hospital Outcome Cohort ($n$=16,784). Discrimination was good with an aROC of 0.83 (95% CI: 0.824–0.838). SAPS II showed underestimation of hospital mortality: The O/E ratio of the overall cohort was 1.08 (1.06–1.10). O/E ratios significantly differed between regions: from 0.86 (0.81–0.91) for Central and Western Europe to 1.32 (1.25–1.38) for Central and South America, with four out of the seven defined regions exhibited O/E ratios significantly different from 1 (Table E7, ESM). Calibration, as assessed by the Hosmer-Lemeshow $\hat{H} + \hat{C}$ statistics, was poor for the overall cohort: $\hat{H}$ 227.21, $\hat{C}$ 184.70; both $p<0.0001$; This lack of calibration was present for all tested subgroups except for the region of North America (see Table E7, ESM).

**Table 3** ICU admission and discharge data for the seven defined geographic regions (SAPS 3 Basic Cohort; *n*=19,577)

| | Australasia | | Central & South America | | Eastern Europe | | Central and Western Europe | | Northern Europe | | Southern Europe and Mediterranean countries | | North America | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of patients | 2,235 | | 2,540 | | 1,084 | | 4,712 | | 355 | | 7,854 | | 797 | |
| Females, % | 38.0 | | 44.6 | | 42.2 | | 40.1 | | 42.5 | | 36.9 | | 38.0 | |
| Age, years (median, quartiles) | 59 | 45–71 | 31 | 46–74 | 62 | 48–72 | 65 | 52–75 | 66 | 51–76 | 64 | 59–74 | 64 | 49–75 |
| SAPS II score (median, Q1-Q3) | 28 | 19–40 | 30 | 20–42 | 27 | 18–43 | 29 | 21–40 | 35 | 25–46 | 32 | 22–44 | 29 | 20–39 |
| SOFA score (median, Q1-Q3) | 8 | 6–10 | 8 | 6–11 | 9 | 7–11 | 8 | 5–11 | 9 | 7–11 | 9 | 7–11 | 9 | 6–11 |
| ICU mortality, % | 12.7 | | 17.4 | | 16.9 | | 10.8 | | 20.6 | | 18.1 | | 8.5 | |

## Discussion

To the best of our knowledge, the SAPS 3 study is the largest prospective epidemiologic multicentre, multinational study conducted in health services and outcomes research in intensive care medicine to date.

The project was first intended to focus on Europe because it was believed such a strategy would produce a more homogeneous cohort of patients, which would in turn provide a more stable reference line for further comparisons. This idea was discussed during several investigator meetings and finally abandoned—first, because interest from outside Europe was enormous: 39% of ICUs that registered for the project were located outside Europe. The SAPS 3 board members thus agreed that such a high level of interest should not be ignored. Second, some investigators questioned whether a concentration on European ICUs would be successful in reducing heterogeneity anyway. Provision of intensive care in Europe is extremely variable, with enormous differences in severity of illness, provision of treatments and mortality from north to south and from west to east [32, 33].

For these reasons ICUs from regions outside Europe were invited to participate. Our results prove that we were right in our assumptions: First, one can easily see that the four European regions (as defined in our study) are hardly comparable: severity of illness as measured by the SAPS II varied from 27 to 35 points, and ICU mortality ranged from 10.8 to 20.6%—almost a doubling of mortality figures (Table 3). Second, almost a third of the patient cohort (28.5%) was contributed from regions outside Europe.

Although the decision to accept ICUs worldwide probably increased the heterogeneity of our sample, it also allowed the SAPS 3 database to better reflect important differences in patients' and health care systems' baseline characteristics that are known to affect outcome. These include, for example, different genetic makeups, different styles of living or a heterogeneous distribution of major diseases within different regions, as well as issues such as access to the health care system in general and to intensive care in particular, or differences in availability and use of major diagnostic and therapeutic measures within the ICUs [32, 34]. Although the integration of ICUs outside Europe and the U.S. surely increased it's representativeness, it must be acknowledged, that the extent to which the SAPS 3 database reflects case-mix on ICUs worldwide cannot be determined yet.

It should additionally be noted that allocation of countries to regions does not always follow geographic borders (Table 3; see also Table E10 in the ESM). Partitioning of the sample was done to adjust for some of the above-stated differences between different populations and to develop a system that uses several different reference lines to compare ICUs on a similar level. Thus,

patients are not necessarily representative of their respective regions.

To minimize possible seasonal influences, we chose late fall in the Northern Hemisphere for data collection. Thus, participants in both late fall/winter (Northern Hemisphere) and spring/summer (Southern Hemisphere) are represented in our cohort. A recent study [35] showed, moreover, that differences in seasonal mortality rates, at least in a sample of ICUs in the United Kingdom, were related to variations in case mix rather than to a specific impact of season on outcome.

Performance of the SAPS II was, not surprisingly, found to be similar to that in previous studies: acceptable discrimination but lack of calibration. Possible reasons for this have already been alluded to in the Introduction. In contrast to previous studies, however, we found an underestimation of hospital mortality, which contradicts the rationale that the shifting in calibration is due only to the development of new and possibly better therapies and thus to better ICU performance [19].

Analyzing the various geographic regions provides evidence that the underestimation of hospital mortality by the SAPS II might be partially attributable to the composition of the cohort: SAPS 3 is the first large international study on severity of illness systems to include patients from all continents. South America, for example, where provision of intensive care is much more limited than it is in Europe or North America, contributed extensively to the patient cohort. High O/E ratios have already been reported for this continent [36] and are probably linked to the limited availability of resources.

Data quality was one of our major concerns. Completeness of the documentation was found to be satisfactory: The amount of missing data is in fact smaller than reported from previous cohort studies on severity of illness systems [11, 12, 16]. With respect to reliability, intraclass-correlation coefficients and kappa coefficients were generally similar to or even better than those found in previous studies, showing a high degree of interrater agreement (see Table E1 in ESM) [37, 38].

We did, however, experience problems with the cohort of rescored patients: First, we had to exclude all rescored patients for whom the original counterpart was also excluded due to the application of any of the exclusion criteria. Second, in some cases the original patient identification was either missing or documented in such a way that a unique allocation was not possible. Both of these exclusions reduced the number of rescored patients available for analysis.

Two strategies to build up a cohort are available: first, to recruit only patients who meet well-documented inclusion criteria (such as documented vital status at hospital discharge) or, second, to document all patients and then exclude patients based on a predefined set of exclusion criteria. For the SAPS 3 study we chose the second option—to form two different cohorts—because we needed to provide a basic cohort for all further analyses of the SAPS 3 database. Since some studies will focus on different outcomes (e.g., ICU outcome rather than hospital outcome), we decided to use missing ICU outcome (and not hospital outcome) as an exclusion criterion for the basic cohort.

A possible limitation of the SAPS 3 database is that vital status at hospital discharge was not available for all admitted patients. Despite several efforts from the CCC and sufficient time to allow for a close follow-up, we did not succeed to receive all hospital outcomes documented. Recording of hospital outcome (or later outcomes) still poses major problems for ICUs in European and non-European hospitals, either because of technical problems or possibly because of data security algorithms in the hospitals. Exclusion of these patients did, however, not affect major criteria, such as geographic representation, ICU admission or discharge data, co-morbidities, or the distribution of the reasons for admission (Tables 1 and 2).

We conclude that the SAPS 3 database is within the above discussed limits of high quality and reflects the heterogeneity of current intensive care provision. As such, it provides an excellent basis for the development of a new risk adjustment system.

**Acknowledgements** The SAPS 3 project was endorsed in June 2002 by the European Society of Intensive Care Medicine (ESICM). It received support from the Austrian Centre for Documentation and Quality Assurance in Intensive Care Medicine (ASDI), the Portuguese Society of Intensive Care (SPCI), and the Medical Economics and Research Centre (MERCS) in Sheffield, U.K.. An unrestricted educational grant from Merck Sharp & Dohme Portugal to the SPCI made possible the installation of the CCC in Lisbon. *i*MD*soft* (Tel Aviv, Israel) developed and provided the Internet-based data collection software free of charge.

Statistical analysis was supported by a grant from the Fund of the Austrian National Bank, Project # 10995 ONB.

Statistical analysis was further supported by Lorenz Dolanski and Johanna Einfalt, both from the Department of Medical Statistics, University of Vienna, Vienna, Austria.

Our thanks to the participants from all over the world who dedicated a significant amount of their time and effort to this project, proving that it is still possible to conduct a worldwide academic study. The SAPS 3 is primarily their study, and we are deeply indebted to them for the honour of conducting it.

## References


1. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. (1981) APACHE—acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9:591–597
2. Le Gall J-R, Loirat P, Alperovitch A. (1983) Simplified acute physiological score for intensive care patients. Lancet ii:741
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. (1985) APACHE II: a severity of disease classification system. Crit Care Med 13:818–829
4. Knaus WA, Wagner DP, Draper EA et al. (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100:1619–1636
5. 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
6. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. (1993) Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270:2478–2486
7. Castella X, Artigas A, Bion J, Kari A, The European / North American Severity Study Group. (1995) A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. Crit Care Med 23:1327–1335
8. Bertolini G, D'Amico R, Apolone G et al. Predicting outcome in the intensive care unit using scoring systems: is new better? (1998) A comparison of SAPS and SAPS II in a cohort of 1.393 patients. Med Care 36:1371–1382
9. Organization and management of Intensive Care: a prospective study in 12 European countries. Berlin Heidelberg: Springer-Verlag, 1997 (Reis Miranda D, Ryan DW, Schaufeli WB, Fidler V, eds. vol 29)
10. Moreno R, Miranda DR, Matos R, Fevereiro T. (2001) Mortality after discharge from intensive care: the impact of organ system failure and nursing workload use at discharge. Intensive Care Med 27:999–1004
11. Apolone G, D'Amico R, Bertolini G et al. (1996) The performance of SAPS II in a cohort of patients admitted in 99 Italian ICUs: results from the GiViTI. Intensive Care Med 22:1368–1378
12. Moreno R, Morais P. (1997) Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. Intensive Care Med 23:177–186
13. Moreno R, Reis Miranda D, Fidler V, Van Schilfgaarde R. (1983) Evaluation of two outcome predictors on an independent database. Crit Care Med 1998;26:50–61
14. Metnitz PG, Valentin A, Vesely H et al. (1999) Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Intensive Care Med 25:192–197
15. K Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. (1993) Intensive Care Society's APACHE II study in Britain and Ireland—II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. Br Med J 307:977–981
16. Bastos PG, Sun X, Wagner DP, Knaus WA, Zimmerman JE, The Brazil APACHE III Study Group. (1996) Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. Intensive Care Med 22:564–570
17. Rivera-Fernandez R, Vazquez-Mata G, Bravo M et al. (1998) The Apache III prognostic system: customized mortality predictions for Spanish ICU patients. Intensive Care Med 24:574–581
18. Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA. (1998) Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Crit Care Med 26:1317–1326
19. Popovich MJ. (2002) If most intensive care units are graduating with honors, is it genuine quality or grade inflation? Crit Care Med 30:2145–2146
20. Rowan K. The reliability of case mix measurements in intensive care. (1996) Curr Opin Crit Care 2:209–213
21. Fery-Lemmonier E, Landais P, Kleinknecht D, Brivet F. (1995) Evaluation of severity scoring systems in the ICUs: translation, conversion and definitions ambiguities as a source of inter-observer variability in APACHE II, SAPS, and OSF. Intensive Care Med 21:356–360
22. Zhu B-P, Lemeshow S, Hosmer DW, Klarm J, Avrunin J, Teres D. (1996) Factors affecting the performance of the models in the mortality probability model and strategies of customization: a simulation study. Crit Care Med 24:57–63
23. Knaus WA, Sun X, Nystrom PO, Wagner DP. (1992) Evaluation of definitions for sepsis. Chest 101:656–662
24. Alberti C, Brun-Buisson C, Burchardi H, Martin C, Goodman S, Artigas A, Sicignano A, Palazzo M, Moreno R, Boulme R, Lepage E, Le Gall R. (2002) Epidemiology of sepsis and infection in ICU patients from an international multicentre cohort study. Intensive Care Med; 28:108–121
25. Azoulay E, Alberti C, Legendre I, Buisson CB, Le Gall JR. (2005) Post-ICU mortality in critically ill infected patients: an international study. Intensive Care Med; 31 (1):56–63
26. Fagon J-Y, Chastre J, Novara A, Medioni P, Gilbert C. (1993) Characterization of intensive care unit patients using a model based on the presence or absence of organ dysfunctions and/or infection: the ODIN model. Intensive Care Med 19:137–44
27. Pollack MM, Alexander SR, Clarke N et al. (1990) Improved outcomes from tertiary center pediatric intensive care: a statewide comparison of tertiary and nontertiary care facilities. Crit Care Med 19:150–159
28. Vincent JL, de Mendonca A, Cantraine F et al. (1983) Use of the SOFA score to assess the incidence of organ dysfunction/ failure in intensive care units: results of a multicentric, prospective study. Crit Care Med 1998 26:1793–1800
29. Hosmer DW, Lemeshow S. (1995) Confidence interval estimates of an index of quality performance based on logistic regression models. Statistics in Medicine 14:2161–2172
30. Lemeshow S, Hosmer DW (1982) A review of Goodness-of-fit Statistics for Use in the development of logistic regression models. American-J-Epidemiology 115:92–106
31. Hanley JA, McNeil BJ (1981) The Meaning and Use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
32. Miranda DR, Ryan DW, Schaufeli WB, Fidler V (Eds.) Organization and Management of Intensive Care. In: Vincent JL (Ed.) Update in Intensive Care Medicine, Vol. 29. Springer-Verlag Berlin-Heidelberg 1998
33. Vincent JL. (1999) Forgoing life support in western European intensive care units: the results of an ethical questionnaire. Crit Care Med. 27:1626–33

34. Metnitz PhGH, Kopp A, Jordan B, Lang Th. (2004) More interventions do not necessarily improve outcome in critically ill patients. Intensive Care Med 30:1586–1593
35. Harrison DA, Lertsithichai P, Brady AR, Carpenter JR, Rowan K. (2004) Winter excess mortality in intensive care in the UK: an analysis of outcome adjusted for patient case mix and unit workload. Intensive Care Med 30:1900–1907
36. Bastos PG, Knaus WA, Zimmerman JE, Magalhaes A Jr, Sun X, Wagner DP. (1996) The importance of technology for achieving superior outcomes from intensive care. Brazil APACHE III Study Group. Intensive Care Med 22:664–669
37. Damiano AM, Bergner M, Draper EA, Knaus WA, Wagner DP. (1992) Reliability of a measure of severity of illness: acute physiology and chronic health evaluation. II. J Clin Epidemiol 45:93–101
38. Chen LM, Martin CM, Morrison TL, Sibbald WJ. (1999) Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. Crit Care Med. 27:1999–2004

Intensive Care Med (2005) 31:1345–1355
DOI 10.1007/s00134-005-2763-5

ORIGINAL

# SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission

**Rui P. Moreno**
**Philipp G. H. Metnitz**
**Eduardo Almeida**
**Barbara Jordan**
**Peter Bauer**
**Ricardo Abizanda Campos**
**Gaetano Iapichino**
**David Edbrooke**
**Maurizia Capuzzo**
**Jean-Roger Le Gall**
**on behalf of the SAPS 3 Investigators**

Received: 8 April 2005
Accepted: 22 July 2005
Published online: 17 August 2005


Electronic Supplementary Material
Electronic supplementary material is included in the online fulltext version of this article and accessible for authorised users: http://dx.doi.org/10.1007/s00134-005-2763-5

R. P. Moreno (✉)
Unidade de Cuidados Intensivos Polivalente,
Hospital de St. António dos Capuchos,
Centro Hospitalar de Lisboa (Zona Central),
Lisbon, Portugal
e-mail: r.moreno@mail.telepac.pt
Fax: +351-21-3153784

P. G. H. Metnitz
Department of Anaesthesiology and General Intensive Care,
University Hospital of Vienna,
Vienna, Austria

E. Almeida
Unidade de Cuidados Intensivos,
Hospital Garcia de Orta,
Pragal, Portugal

B. Jordan · P. Bauer
Department of Medical Statistics,
University of Vienna,
Vienna, Austria

R. A. Campos
Department of Intensive Care,
Hospital Universitario Asociado General de Castelló,
Castello, Spain

G. Iapichino
Department of Anesthesia and Intensive Care Medicine,
Hospital San Paolo, Università degli Sudi,
Milan, Italy

D. Edbrooke
Critical Care Directorate,
Royal Hallamshire Hospital,
Sheffield, UK

M. Capuzzo
Department of Anesthesia and Intensive Care Medicine,
Hospital of Ferrara,
Ferrara, Italy

J.-R. Le Gall
Department Réanimation Médicale,
Hôpital St. Louis, Université Paris VII,
Paris, France

**Abstract** *Objective:* To develop a model to assess severity of illness and predict vital status at hospital discharge based on ICU admission data. *Design:* Prospective multicentre, multinational cohort study. *Patients and setting:* A total of 16,784 patients consecutively admitted to 303 intensive care units from 14 October to 15 December 2002. *Measurements and results:* ICU admission data (recorded within ±1 h) were used, describing: prior chronic conditions and diseases; circumstances related to and physiologic derangement at ICU admission. Selection of variables for inclusion into the model used different complementary strategies. For cross-validation, the model-building procedure was run five times, using randomly selected four fifths of the sample as a development- and the remaining fifth as validation-set. Logistic regression methods were then used to reduce complexity of the model. Final estimates of regression coefficients were determined by use of multilevel logistic regression. Variables selection and weighting were further checked by bootstraping (at patient level and at ICU level). Twenty variables were selected for the final model, which exhibited good discrimination (aROC curve 0.848), without major differences across patient typologies. Calibration was also satisfactory (Hosmer-Lemeshow goodness-of-fit test $\hat{H}$=10.56, $p$=0.39, $\hat{C}$=14.29, $p$=0.16). Customised equations for major areas of the world were computed and demonstrate a good overall goodness-of-fit. *Conclusions:* The SAPS 3 admission score is able to predict vital status at hospital discharge with use of data recorded at ICU admission. Furthermore, SAPS 3 conceptually dissociates evaluation of the individual patient from evaluation of the ICU and thus allows them to be assessed at their respective reference levels.

**Keywords** Intensive care unit · Severity of illness · ICU mortality · Hospital mortality · Risk adjustment

## Introduction

One of the crucial steps in the evaluation of risk-adjusted outcomes is the choice of the reference database for estimating adequate reference lines for the analyzed variables. For the SAPS 3 to reflect the standard of practices and outcome in intensive care at the beginning of the 21st century, we decided to collect data from a large sample of intensive care units (ICUs) worldwide. Other models have restricted data collection to large ICUs in Europe or North America—SAPS II [1], MPM II [2], APACHE II [3] and APACHE III [4], a strategy that minimizes the heterogeneity of the sample but restricts the generalization of the results.

At the statistical level, there is also a need for change, in order to take into account the hierarchic nature of our data [5, 6]. Current general outcome prediction models do not consider the existence of clinical and nonclinical factors, aggregated at the ICU level, that can have an important impact on prognosis. Instead, they assume that these factors are either not important or are randomly distributed throughout large samples and that the variation between ICUs is small. This assumption is not likely to be borne out at the ICU level for either nonclinical factors (e.g. organization and management, organizational culture) or clinical factors (e.g. clinical management, diagnostic and therapeutic strategies). If the variation between ICUs is not negligible, it will compromise the stability of the equations used to compute predicted mortality. Furthermore, the published models consider the relation between performance and severity of illness to be constant, and that may not be the case, since performance can vary within ICUs according to the severity of illness of the patients [7, 8]. To overcome this problem, we chose to adopt a new strategy for the development of the SAPS 3 score and to apply statistical modelling techniques that control for the clustering of patients within ICUs instead of assuming the independence of observations. Conceptually, the SAPS 3 admission score comprises the following parts:

First, the *SAPS 3 ADMISSION SCORE*, represented by the arithmetic sum of three subscores, or boxes:

- **Box I:** What we know about the patient characteristics before ICU admission: age, previous health status, comorbidities, location before ICU admission, length of stay in the hospital before ICU admission, and use of major therapeutic options before ICU admission.
- **Box II:** What we know about the circumstances of ICU admission: reason(s) for ICU admission, anatomic site of surgery (if applicable), planned or unplanned ICU admission, surgical status and infection at ICU admission.
- **Box III:** What we know about the presence and degree of physiologic derangement at ICU admission (within 1 h before or after admission).

Second, the *SAPS 3 PROBABILITY OF DEATH* during a certain period of time (in the case of the main model, the probability of death at hospital discharge).

Given our objective of evaluating not only individual patient outcome but also the effectiveness of ICU practices, we focused the model on *data available at ICU admission or shortly thereafter*. This model will be completely open and available free of any direct or indirect charges to the scientific community.

## Methods and statistical analysis

### Primary variable selection

Based on the SAPS 3 Hospital Outcome Cohort as described in Part 1 of this report, continuous predictive variables were categorized in mutually exclusive categories based on smoothed curves such as LOWESS [9], showing the univariate dependence of hospital mortality on the predictive variables. Classes of categorical variables were also collapsed according to their univariate hospital mortality levels using multidimensional tables and clinical judgment as appropriate, depending on the nature of the data. Additively, regression trees (MART) [10] were applied to check the cutoffs.

Missing values were coded as the reference or "normal" category for each variable. When dual data collection was used—maximum and minimum values recorded during a certain time period—missing maximum values of a variable were replaced by the minimum, if documented, and vice versa. Some regression imputations were performed if noticeable correlations to available values could be exploited. For a detailed description of data collection and handling, see Part 1 of this report.

Selection of variables was done according to their association with hospital mortality, together with expert knowledge and definitions used in other severity of illness scoring systems. The objective of using this combination of techniques rather than regression-based criteria alone was to reach a compromise between over-sophistication of the model and knowledge from sources beyond the sample with its specific case mix and ICU characteristics.

### Cross validation

For being able to cross-validate the model, we randomly extracted five roughly equal-sized parts based on number of patients from the database, as suggested previously [11]. In a second approach, partitioning was based on ICUs and not on patients. It was thus possible to run the model-building procedure five times in each of the two approaches, each time taking four parts of the sample as a development set and the remaining one as the validation set. This allowed to estimate the variability of prediction

resulting from the construction process of the prognostic score. A further check of the stability of the predictions was made by partitioning the sample according to major patient characteristics, such as surgical status and infection status.

The quality of predictions in the validation sets was assessed by looking at the goodness-of-fit in terms of the $p$ values for the Hosmer-Lemeshow tests $\hat{C}$ and $\hat{H}$ [13] and the discriminative capability of the models by the use of the area under the receiver operating characteristic (aROC) curve [14, 15]. Another criterion to judge the appropriateness of the model was the fit in certain subsamples, defined according to major patient typologies [16].

### Reducing model complexity

To reduce the complexity of the model classes, we concentrated on logistic regression. In the first step a stepwise logistic regression was used to identify the significant predictors in each of the five subsamples. A threshold of 0.01 for the $p$ value was generally applied for inclusion in the model to separate irrelevant predictors [12]. At this stage we also evaluated if interactions among these predictors would influence results. Interactions, however, did not make a valuable contribution for the prediction.

Significant predictors (n=70) were in a second step entered into a logistic regression model. The criterion for a predictor to enter the model was homogeneity across the five model-building processes: in principle, predictors should enter the model in all five development sets, but depending on the frequency of the predictor in the samples, the magnitude of the effect, and medical reasoning, some predictors were included if they appeared in the model in at least three subsamples. An example is the presence of Acquired Immunodeficiency Syndrome (AIDS): it was selected as a comorbidity in only 81 patients (0.48%), but the mortality—without controlling for other variables—in these patients was 42%. By taking all the above steps to identify the set of predictors, although deliberately not using any formal numeric criterion, we reduced the complexity of the model to minimize the amount of overfitting: This process resulted in 61 item classes (representing 20 variables) remaining in the final model.

Using the parameter estimates from the logistic regression as starting values, a multilevel model was applied in the next step, using patient characteristics as fixed effects and ICUs as a random effect. Estimates were again calculated for the five development sets (for both, patient and ICU -based development subsamples).

At this stage it was checked if rounding of coefficients (which allows for an easier manual computation of the score) would influence results, which was found not to be the case. Consequently, this was the approach chosen for the final construction of the SAPS 3 admission score sheet.

The stability of the processes of variable selection and reducing complexity was further checked by bootstraping with replacement the total sample 100 times, both at patient level and at ICU level.

### Predicting hospital mortality

After this step was completed, a shrinking power transformation was applied. This procedure uses log-transformation of the score to reduce the influence of extreme score values (outliers) on the mortality prediction. For this purpose, the SAPS 3 score and the transformed log (SAPS 3 + $g$) score were used to predict hospital mortality. Conventional logistic regression was used in the evaluation of this step because of convergence problems for the corresponding multilevel model in a few subsamples. The best shrinkage model then was selected (excluding the trivial model with the SAPS 3 score as the single predictor) by checking which of the terms in the model contributed best to the prediction and was moreover stable over the respective validation sets and specific subsamples. This procedure was applied on both, patient and ICU -based subsamples.

After finishing these steps of cross-validation, the final estimates for the selected predictors of the SAPS 3 score as well as the selected shrinkage procedure were then calculated from the total sample of patients.

To arrive at the customised models for each major geographic region, specific customised equations were calculated, relating, by logistic regression, the transformed log (SAPS 3 + $g$) admission scores computed as described above to the vital status at hospital discharge. This process allows both the intercept and the slope of the curve relating the SAPS 3 admission score to change across different regions. The goodness-of-fit of these equations was evaluated by means of the same methodology used for the global sample.

SAS for Windows, version 8.02 (SAS Institute Inc., Cary, NC, USA) and MLwiN version 1.10.0007 (Centre for Multilevel Modelling, Institute of Education, London, UK) and the R Software Package (http://www.r-project.org) were used for the development of the model.

## Results

Based on the methodology described, 20 variables were selected for the SAPS 3 admission score (Tables 1 and 2):

- Five variables for evaluating Box I: age, co-morbidities, use of vasoactive drugs before ICU admission, intrahospital location before ICU admission, and length of stay in the hospital before ICU admission;

– Five variables for evaluating Box II: reason(s) for ICU admission, planned/unplanned ICU admission, surgical status at ICU admission, anatomical site of surgery, and presence of infection at ICU admission and place acquired;
– Ten variables for evaluating Box III: lowest estimated Glasgow coma scale, highest heart rate, lowest systolic blood pressure, highest bilirubine, highest body temperature, highest creatinine, highest leukocytes, lowest platelets, lowest hydrogen ion concentration (pH), and ventilatory support and oxygenation.

An estimation of the variability of the coefficients in the overall sample and in the five disjoint subsamples is given in Table E8 of the Electronic Supplementary Material (ESM), together with their respective coefficients (unrounded and rounded) and *p* values. The SAPS 3 admission score can thus, in theory, vary from a minimum of 0 points to a maximum of 217 points. The distribution of the SAPS 3 admission score in our sample is presented in Fig. 1. The minimum value observed was 5, and the maximum value was 124, with a mean of 49.9±16.6 (mean ± SD) and a median of 48 (38–60). The highest explanatory power came from Box I, with Box II and Box III being less important for the outcome; the three boxes represent 50%, 22.5% and 27.5%, respectively, of the total Nagelkerke's R-Square. The relationship between the SAPS 3 and vital status at hospital discharge is given by the equation:

$$\text{Logit} = -32.6659 + \ln(\text{SAPS 3 score} + 20.5958) \times 7.3068$$

and the probability of mortality by the equation:

$$\text{Probability of death} = e^{\text{logit}}/(1+e^{\text{logit}}).$$

The relationship between the SAPS 3 admission score and the respective probability of death in the hospital is described in Fig. 2. Overall, no combined discrepancy between observed and expected outcomes across all of the strata was outside sampling variability as demonstrated a Hosmer-Lemeshow goodness-of-fit test $\hat{H}$ of 10.56 ($p$=0.39) and a Hosmer-Lemeshow goodness-of-fit test $\hat{C}$ of 14.29 ($p$=0.16) (Figs. 3, 4 and Table E9, ESM). The overall discriminatory capability of the model, as measured by aROC curve, was 0.848. The goodness-of-fit according to major patient typologies (surgical status, trauma, and infection) can be found in Table 3. Calibration and discrimination presented differences across different geographic areas: the best predictive results were achieved in patients from Northern Europe (observed-to-expected [O/E] mortality ratio 0.96 [0.83–1.09]) and the worst predictive results were obtained in patients from Central and South America (O/E mortality ratio, 1.30 [1.23–1.37]); see also Table 4 and Fig. 5 and Appendix B in the ESM.

For a more precise estimation of the probability of death in the hospital across the different geographic re-

**Table 1** SAPS 3 admission scoresheet—Part 1

| Box I | 0 | 3 | 5 | 6 | 7 | 8 | 9 | 11 | 13 | 15 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Age, years | <40 | | >=40<60 | | | | >=60<70 | | >=70<75 | >=75<80 | >=80 |
| Co-Morbidities | | Cancer [2] therapy | | Chron, HF (NYHA IV), Haematological cancer [3,4] | | Cirrhosis, AIDS [3] | | Cancer [5] | | | |
| Length of stay before ICU admission, days [1] | <14 | | | >=14<28 | >=28 | | | | | | |
| Intra-hospital location before ICU admission | | | Emergency room | | Other ICU | Other [6] | | | | | |
| Use of major therapeutic options before ICU admission | | Vasoactive drugs | | | | | | | | | |

| Box II | 0 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| ICU admission: Planned or Unplanned | Unplanned | | | | |
| Reason(s) for ICU admission | please see Part 2 of the scoresheet | | | | |
| Surgical status at ICU admission | Scheduled surgery | | | No surgery [7] | Emergency surgery |
| Anatomical site of surgery | please see Part 2 of the scoresheet | | | | |
| Acute infection at ICU admission | | | Nosocomial [8] | Respiratory [9] | |

**Table 1** continued

| Box III | 15 | 13 | 11 | 10 | 8 | 7 | 5 | 3 | 2 | 0 | 2 | 4 | 5 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimated Glasgow Coma Scale (lowest), points | 3–4 | | | 5 | | 6 | | | 7–12 | >=13 | | | | | |
| Total bilirubine (highest), mg/dL | | | | | | | | | | <2 | | >=2<6 | >=6 | | |
| Total bilirubine (highest), µmol/L | | | | | | | | | | <34.2 | | >=34.2 <102.6 | >=102.6 | | |
| Body temperature (highest), Degrees Celsius | | | | | | <35 | | | | >=35 | | | | | |
| Creatinine (highest), mg/dL | | | | | | | | | | <1.2 | >=1.2<2 | | | >=2<3.5 | >=3.5 |
| Creatinine (highest), µmol/L | | | | | | | | | | <106.1 | >=106.1< 176.8 | | | >=176.8 <309.4 | >=309.4 |
| Heart rate (highest), beats/minute | | | | | | | | | | <120 | | | >=120 <160 | >=160 | |
| Leukocytes (highest), G/L | | | | | | | | | | <15 | >=15 | | | | |
| Hydrogen ion concentration (lowest), pH | | | | | | | | <=7.25 | | >7.25 | | | | | |
| Plateletes (lowest), G/L | | <20 | | | >=20<50 | | >=50<100 | | | >=100 | | | | | |
| Systolic blood pressure (lowest), mm Hg | | | <40 | | >=40<70 | | | >=70<120 | | >=120 | | | | | |
| Oxygenation [10);11)] | | | PaO2/ FiO2 <100 and MV | | | PaO2/ FiO2> =100 and MV | PaO2<60 and no MV | | | PaO2> =60 and no MV | | | | | |

The definition for all variables can be found in detail in Appendix C of the ESM. For names and abbreviations which are differing from those in the ESM, explanations are given below. Generally, it should be noted that no mutually exclusive conditions exist for the following fields: Comorbidities, Reasons for ICU admission, and Acute infection at ICU admission. Thus, if a patient has more than one condition listed for a specific variable, points are assigned for all applicable combinations.

[1] This variable is calculated from the two data fields: ICU Admission date and time—Hospital admission date and time (see Appendix C of the ESM)

[2] Cancer Therapy refers to the data definitions in Appendix C of the ESM: Co-Morbidities: Chemotherapy, Immunosupression other, Radiotherapy, Steroid treatment

[3] If a patient has both conditions he/she gets double points.

[4] Chronic HF (NYHA IV)/Haematological cancer refer both to the data definitions in Appendix C of the ESM: Co-Morbidities: Chronic heart failure class IV NYHA, Haematological cancer.

[5] Cancer refers to the data definitions in Appendix C of the ESM: Co-Morbidities: Metastatic cancer.

[6] Other refers to the data definitions in Appendix C of the ESM: Intra-hospital location before ICU admission: Ward, Other.

[7] No surgery refers to the data definitions in Appendix C of the ESM: Surgical Status at ICU Admission: Patient not submitted to surgery.

[8] Nosocomial refers to the data definitions in Appendix C of the ESM: Acute infection at ICU admission—Acquisition: Hospital-acquired.

[9] Respiratory refers to the data definition in Appendix C of the ESM: Acute infection at ICU admission—Site: Lower respiratory tract: Pneumonia, Lung asbcess, other.

[10] $PaO_2$, $FIO_2$ refer to the data definitions in Appendix C of the ESM: Arterial oxygen partial pressure (lowest), Inspiratory oxygen concentration.

[11] MV refers to the data definition in Appendix C of the ESM: Ventilatory support and mechanical ventilation.

**Table 2** SAPS 3 admission scoresheet – Part 2

| Box II – continued | |
|---|---|
| ICU admission [12)] | 16 |
| **Reason(s) for ICU admission** | |
| Cardiovascular: Rhythm disturbances [13)] | –5 |
| Neurologic: Seizures [13)] | –4 |
| Cardiovascular: Hypovolemic hemorrhagic shock, Hypovolemic non hemorrhagic shock. / Digestive: Acute abdomen, Other [3)] | 3 |
| Neurologic: Coma, Stupor, Obtuned patient, Vigilance disturbances, Confusion, Agitation, Delirium | 4 |
| Cardiovascular: Septic shock. / Cardiovascular: Anaphylactic shock, mixed and undefined shock | 5 |
| Hepatic: Liver failure | 6 |
| Neurologic: Focal neurologic deficit | 7 |
| Digestive: Severe pancreatitis | 9 |
| Neurologic: Intracranial mass effect | 10 |
| All others | 0 |
| **Anatomical site of surgery** | |
| Transplantation surgery: Liver, Kidney, Pancreas, Kidney and pancreas, Transplantation other | –11 |
| Trauma – Other, isolated: (includes Thorax, Abdomen, limb); Trauma – Multiple | –8 |
| Cardiac surgery: CABG without valvular repair | –6 |
| Neurosurgery: Cerebrovascular accident | 5 |
| All others | 0 |

[12)] Every patient gets an offset of 16 points for being admitted (to avoid negative SAPS 3 Scores).
[13)] If both reasons for admission are present, only the worse value (–4) is scored.

gions, specific customised equations were calculated (Table 5). This customised approach allows each ICU to choose its own reference line for the prediction of hospital mortality: either the overall SAPS 3 hospital mortality sample or its own regional subsample. This approach can be supplemented in the future by customised equations at the country level if data are available and if a more precise estimation of outcome in a specific setting is needed.

The overall goodness-of-fit of these customised equations for each region is presented in Table 5. A complete list of the number of patients and the respective O/E mortality ratios by country, according to the global equation and the regional equations, are presented in Tables E10 and E11 of the ESM, with point estimates varying at the global level from 0.68 (0.56–0.80) to 2.05 (1.27–2.82). Most O/E ratios are close to the identity line, as expected for a stable model.

## Discussion

We have presented the results of a large multicentric, multinational study aimed at updating the SAPS II model. This study was necessary for several reasons. First, the reference line used by SAPS II was derived from a database collected in the early 1990s; since that time, there have been changes in the prevalence of major diseases and in the availability and use of major diagnostic and therapeutic methods that are associated with a shift toward poor calibration of older models such as SAPS II and APACHE III [17, 18]. Second, SAPS II was developed from a database built exclusively from patients in Europe and North America. This sample may not be representative of the case mix and medical practices that constitute the reality of intensive care medicine in the rest of the world (e.g. Australasia or South America), where variability in structures and organization is probably related to outcome [19].

Third, since computation of predicted mortality is based on a reference database, the user should be able to choose between them, i.e., a global database, which provides a broader comparison at the potential cost of less relevance to local conditions, and a regional database, which provides a better comparison with ICUs in geo-

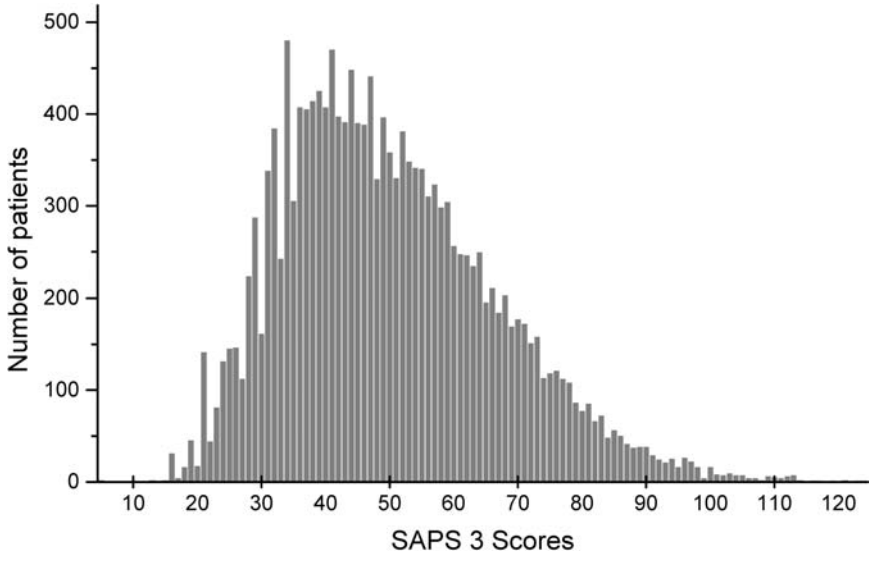


**Fig. 1** Distribution of the SAPS 3 admission score in the SAPS 3 database

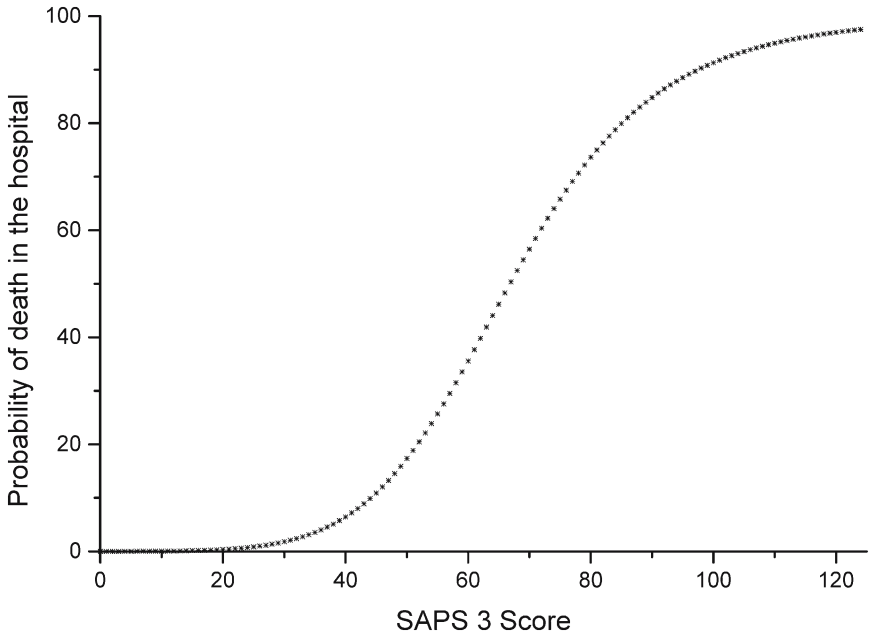


**Fig. 2** Relationship between the SAPS 3 admission score and the respective probabilities of hospital mortality

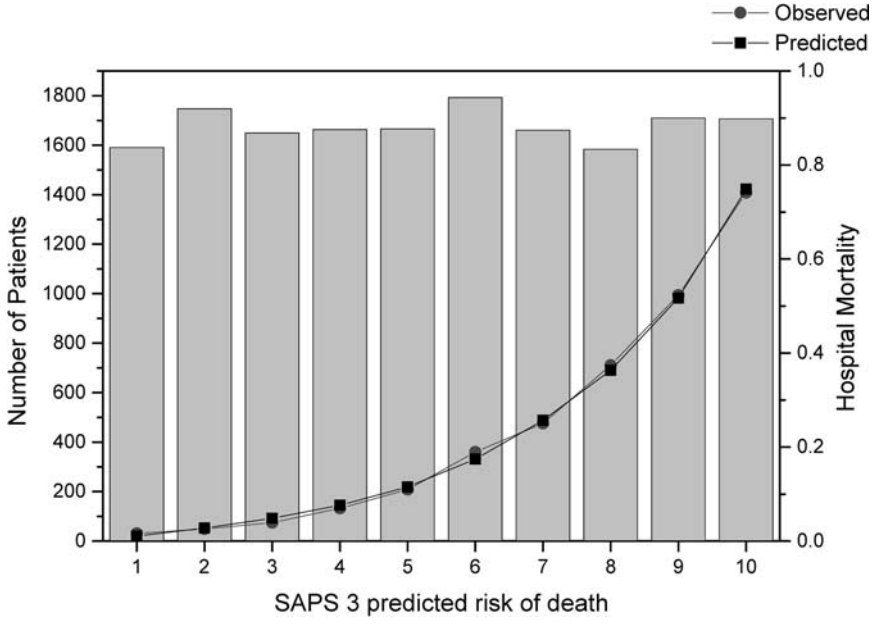


**Fig. 3** Hosmer-Lemeshow goodness-of-fit test $\hat{C}$ in the overall sample. Predicted risk of hospital death, observed hospital mortality rate, and the corresponding number of patients per decile are shown. *Columns:* Number of patients; *squares:* mean SAPS 3-predicted mortality per decile; *circles:* mean observed mortality per decile

graphic proximity but at the cost of losing comparability with ICUs in other parts of the world. A third possibility could be added—a country-representative database—but such a database would raise the problem of whether the ICUs selected were representative of a certain country.

Fourth, the development of computers in recent years has created easy access to strong computational power. One of the implications of this is that it is now possible to develop a new outcome prediction model, based on digital data acquisition and analysis, with minimal differences in definitions and application criteria. These advances were coupled with extensive automatic logical and error-checking capabilities and the availability of data collection manuals online. Moreover, developers of the SAPS 3 model could take advantage of computer-intensive methods of data selection and analysis, such as the use of additive partition trees and logistic regression with random effects. Several new statistical techniques have been used in recent years to allow a more stable prediction of outcome, such as genetic algorithms and artificial neural

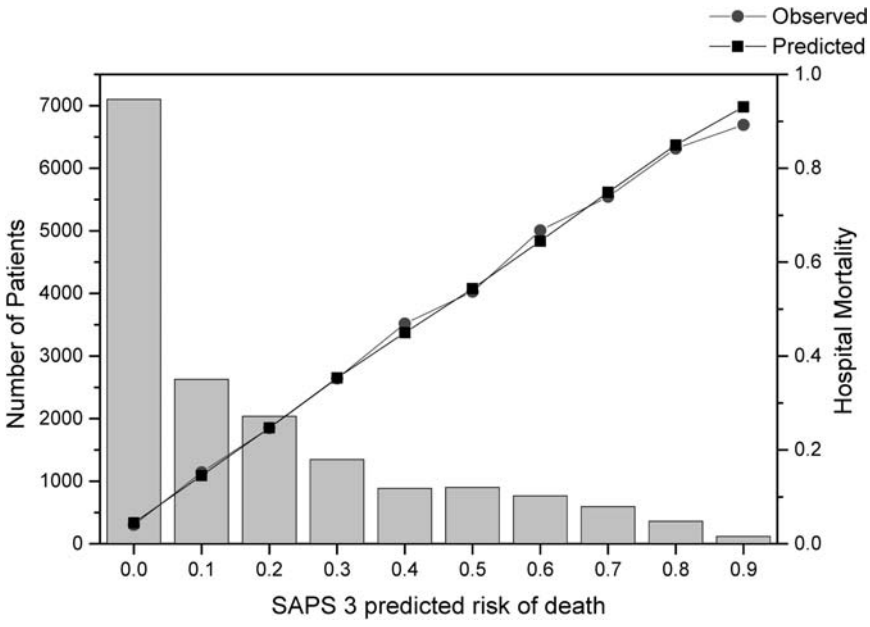


**Fig. 4** Hosmer-Lemeshow goodness-of-fit test $\hat{H}$ in the overall sample. Predicted risk of hospital death, observed hospital mortaliy rate, and the corresponding number of patients per decile are shown. *Columns:* Number of patients; *squares:* mean SAPS 3-predicted mortality per decile; *circles:* mean observed mortality per decile

**Table 3** Performance of the model across major patient typologies

| Patient characteristics | GOF test $\hat{H}$ | *p* | GOF test $\hat{C}$ | *p* | O/E ratio | 95% CI | aROC |
|---|---|---|---|---|---|---|---|
| Trauma patients | 19.92 | 0.03 | 9.03 | 0.53 | 1.03 | 0.93–1.12 | 0.854 |
| Non-operative admissions[a] | 14.86 | 0.14 | 17.8 | 0.06 | 1.01 | 0.98–1.04 | 0.825 |
| Scheduled surgery[a] | 11.5 | 0.32 | 27.39 | <0.01 | 0.97 | 0.90–1.03 | 0.825 |
| Emergency surgery[a] | 4.97 | 0.89 | 12.88 | 0.23 | 1.00 | 0.95–1.05 | 0.809 |
| No infection[b] | 8.57 | 0.57 | 14.77 | 0.14 | 1.00 | 0.97–1.02 | 0.846 |
| Community-acquired infection[c] | 8.4 | 0.59 | 11.76 | 0.3 | 1.00 | 0.96–1.05 | 0.786 |
| Hospital-acquired infection[d] | 15.21 | 0.12 | 7.11 | 0.72 | 1.02 | 0.97–1.07 | 0.77 |

*GOF*: Hosmer-Lemeshow goodness-of-fit; *O/E*: observed-to-expected mortality; *CI*: 95% confidence interval; *aROC*: area under receiver operating characteristic (curve)

[a] Non-operative admissions, scheduled surgery emergency surgery: see data definitions appendix C, ESM

[b] *No infection*: Patients not infected at ICU admission

[c] *Community-acquired infection*: Patients with community-acquired infection at ICU admission

[d] *Hospital-acquired infection*: Patients with hospital-acquired infection at ICU admission

**Table 4** Performance of the model in the global sample and in different geographic areas

| Regions | GOF test $\hat{H}$ | *p* | GOF test $\hat{C}$ | *p* | O/E ratio | 95% CI | aROC |
|---|---|---|---|---|---|---|---|
| Australasia | 15.25 | 0.12 | 8.09 | 0.62 | 0.92 | 0.85–0.99 | 0.839 |
| Central, South America | 78.01 | <0.01 | 80.82 | <0.01 | 1.30 | 1.23–1.37 | 0.855 |
| Central, Western Europe | 56.45 | <0.01 | 47.89 | <0.01 | 0.84 | 0.79–0.90 | 0.861 |
| Eastern Europe | 19.45 | 0.03 | 18.69 | 0.04 | 1.09 | 1.00–1.19 | 0.903 |
| North Europe | 2.44 | 0.99 | 2.34 | 0.99 | 0.96 | 0.83–1.09 | 0.814 |
| Southern Europe, Mediterranean countries | 14.18 | 0.16 | 20.78 | 0.02 | 1.02 | 0.98–1.05 | 0.834 |
| North America | 10.57 | 0.39 | 9.63 | 0.47 | 0.91 | 0.78–1.04 | 0.812 |
| Global database | 10.56 | 0.39 | 14.29 | 0.16 | 1 | 0.98–1.02 | 0.848 |

*GOF*: Hosmer-Lemeshow goodness-of-fit; *O/E*: observed-to-expected mortality; *CI*: 95% confidence interval; *aROC*: area under the receiver operating characteristic (curve).

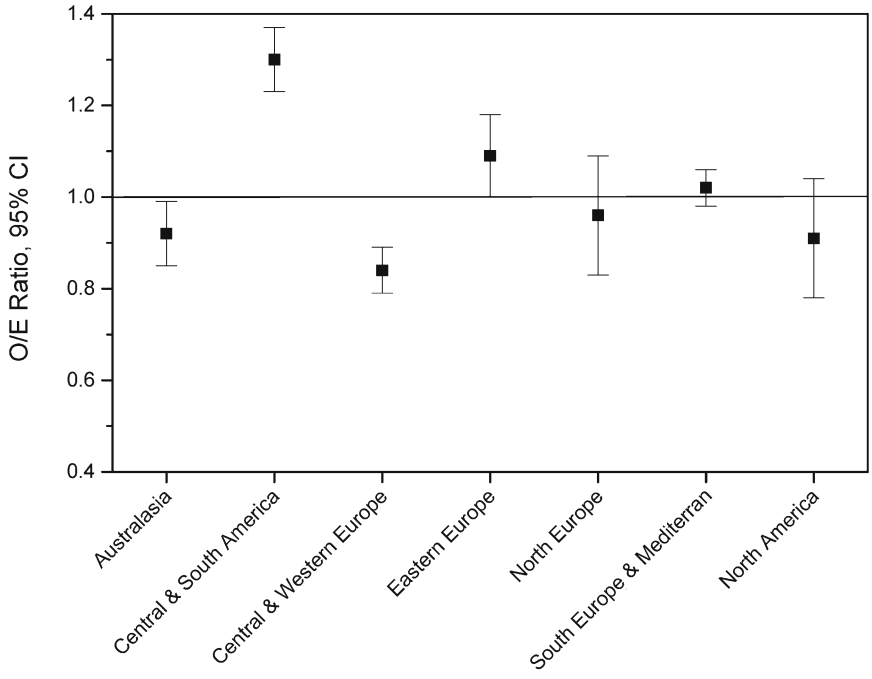


**Fig. 5** Observed-to-expected (O/E) mortality ratios by region. Observed-to-expected (O/E) mortality ratios are shown by region. Bars indicate 95% confidence intervals

**Table 5** Customized SAPS 3 admission equations for the different geographic areas

| Area | Equation | GOF $\hat{H}$ | *p* | GOF $\hat{C}$ | *p* | O/E | CI |
|---|---|---|---|---|---|---|---|
| Australasia | Logit=−22.5717 + ln (SAPS 3 score + 1) ×5.3163 | 10.43 | 0.40 | 2.20 | 0.99 | 1.00 | 0.93–1.07 |
| Central, South America | Logit=−64.5990 + ln (SAPS 3 score + 71.0599) ×13.2322 | 8.94 | 0.54 | 7.03 | 0.72 | 1.00 | 0.94–1.06 |
| Central, Western Europe | Logit=−36.0877 + ln (SAPS 3 score + 22.2655) ×7.9867 | 15.13 | 0.13 | 12.15 | 0.27 | 1.00 | 0.94–1.06 |
| Eastern Europe | Logit=−60.1771 + ln (SAPS 3 score + 51.4043) ×12.6847 | 10.13 | 0.43 | 7.12 | 0.71 | 1.00 | 0.92–1.08 |
| North Europe | Logit=−26.9065 + ln (SAPS 3 score + 5.5077) ×6.2746 | 3.45 | 0.97 | 2.22 | 0.99 | 1.00 | 0.86–1.14 |
| Southern Europe, Mediterranean countries | Logit=−23.8501 + ln (SAPS 3 score + 5.5708) ×5.5709 | 5.28 | 0.87 | 13.12 | 0.22 | 1.00 | 0.97–1.03 |
| North America | Logit=−18.8839 + ln (SAPS 3 score + 1) ×4.3979 | 4.22 | 0.93 | 4.47 | 0.92 | 1.00 | 0.86–1.14 |

*GOF* $\hat{H}$: Hosmer-Lemeshow goodness-of-fit $\hat{H}$ test; *GOF* $\hat{C}$: Hosmer-Lemeshow goodness-of-fit $\hat{C}$ test; *p*: respective p-values; *O/E*: observed-to-expected mortality ratio; *CI*: 95% confidence interval

networks [20, 21], dynamic microsimulation techniques [22], and first- and second-level customization strategies [23–25]. However, the value of these techniques is for the moment limited, usually because they are based on regional databases [24–26] that prevent extrapolation to other settings; moreover, their superiority in even the regional setting still needs to be established.

Finally, the SAPS 3 conceptually dissociates evaluation of the individual patient from evaluation of the ICU. Thus, for individual patient assessment, the system separates the relative contributions to prognosis of (i) chronic health status and previous therapy, (ii) the circumstances related to ICU admission, and (iii) the presence and degree of physiologic dysfunction. It is interesting to note that one half of the predictive power of the model is achieved with Box I, i.e., with the information that is available before ICU admission. The prognostic capabilities of the model can be further improved by 22.5% by using data related to the circumstances of the ICU admission (Box II), and by another 27.5% by the incorporation of physiologic data (Box III). These numbers are different from those published by Knaus et al. [4] but are based on what we have learned in the last years about prognostic determinants in the critically ill patient.

For performance evaluation, several reference lines should be used, with risk-adjusted mortality in different

patient typologies and not only O/E mortality ratios at hospital discharge in the overall ICU population [27]. The results of the SAPS 3 study showing that different O/E ratios were observed in different regions of the world should be explored further, since, apart from regional differences in case mix (not taken into account by the model), they can also be related to regional variations in structures and organization of acute medical care, to different lifestyles (e.g., prevalence of obesity, or alcohol and tobacco use) and/or—though less likely—to genetic differences among populations.

We would like to re-emphasize that the model presented here is based exclusively on data (including physiologic data) available within 1 h of ICU admission and calibrated for manual data acquisition; consequently, it should be expected to overestimate mortality when an automatic patient data management system with a high sampling rate is used [28, 29]. Limiting acquisition of physiologic data to the hour of ICU admission should minimise the impact of this factor when compared with models based on the most deranged data from the first 24 h after ICU admission, probably at the expense of a small decrease in the ROC curve, a greater sensitivity to the exact time point at which admission to ICU occurs, and therefore more reliant on the assumption that measured physiology alone (as opposed to changes in physiology) predict outcome. It also allows the prediction of mortality to be done before ICU interventions take place. This gives the SAPS 3 admission model a major advantage over existing systems, such as the SAPS II or the APACHE II and III, since all these systems can be affected by the so-called Boyd and Grounds effect: the occurrence of more abnormal physiologic values during the first 24 h in the ICU, leading to an increase in computed severity of illness and a corresponding increase in predicted mortality. These increases may, however, be due not to a greater intrinsic severity of illness of the patient but to the provision of suboptimal care in the first 24 h of ICU admission, when a stable patient may be allowed to deteriorate [30].

Further studies should be done of factors occurring after ICU admission that influence risk-adjusted mortality. We should keep however in mind that this approach comes with one potential pitfall: a possible decrease in the amount of data available for the computation of the model; also, the shorter time period for data collection can eventually increase the likelihood of missing physiological data and the reliance on the assumption that missing physiological data are normal. This effect should be small, considering the widespread availability of monitoring and point-of-case analysers.

Having demonstrated the internal validity of the SAPS 3 admission model by the extensive use of cross-validation techniques, we should stress that external validation is also necessary. The fact that the overall database was not collected to be representative of the global case-mix (and especially the case-mix of specific regional areas or patient typologies such as specific diseases) should be empirically tested. Furthermore, the rate of deterioration of our estimates over time should be followed by the appropriate use of temporal validation, especially to avoid what Popovich called grade inflation [18].

The SAPS 3 system was developed to be used free of charge by the scientific community; no proprietary information regarding the scientific content is retained. All the coefficients needed for the computation of outcome probabilities are available in the published material. The SAPS 3 can even be computed manually, using a simple scoresheet, although it was designed to be integrated into computerised data acquisition and storage systems that allow the automatic check of the quality of the registered data.

In conclusion, we can say that at the end of this stage of the project, we have been able to overcome some of the problems inherent in current risk-adjustment systems. We have minimized *user-dependent problems* through the publication of careful, detailed definitions and criteria for data collection [31]. We have also addressed the *patient-dependent problems* by expanding the reference database and making it more representative of reality, in order to include the maximum possible range of variations for patient-centred variables and resulting patient-centred outcomes. This approach was complemented by the development of specific customised equations for major areas of the world, allowing ICUs to choose a reference line for outcome prediction—the global database or the regional database for their own area.

Users of these models should keep in mind that benchmarking is a process of comparing an ICU with a reference population. The appropriate choice of reference population is difficult, and we cannot simply change it because the observed-to-predicted mortality rate is not the one we want. For this reason, the choice should depend on the objective of the benchmark: more precise estimation will need local or regional equations, developed from a more homogeneous case mix. A generalisable estimation will, on the other hand, need more global equations developed from a more representative case mix.

Last but not least, we have successfully addressed some of the problems of prognostic model development, especially those related to the underlying statistical assumptions for the use of specific methods for selection and weighting of variables and the conceptual development of outcome prediction models. In the future, multilevel modelling with varying slopes (and not just random intercepts) might be able to give a better answer to researchers but for the moment they would make the models to complex to be managed outside a research environment.

**Acknowledgements** The SAPS 3 project was endorsed in June 2002 by the European Society of Intensive Care Medicine (ESICM). It received support from the Austrian Centre for Documentation and Quality Assurance in Intensive Care Medicine (ASDI), the Portuguese Society of Intensive Care (SPCI), and the Medical Economics and Research Centre in Sheffield (UK). An unrestricted educational grant from Merck Sharp & Dohme Portugal to the SPCI allowed for the installation of the Coordination and Communication Centre in Lisbon. *i*MD*soft* (Tel Aviv, Israel) developed and provided free of charge the Internet-based data collection software.

Statistical analysis was supported by a grant from the Fund of the Austrian National Bank, Project # 10995 ONB.

Statistical analysis was further supported by Lorenz Dolanski and Johanna Einfalt, both: Dept. of Medical Statistics, University of Vienna, Vienna, Austria.

Our thanks to the participants from all over the world who dedicated a significant amount of their time and effort to this project, proving that it is still possible to conduct a worldwide academic study. The SAPS 3 is primarily their study, and we are deeply indebted to them for the honour of conducting it.

## References

1. 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
2. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270:2478–2486
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13:818–829
4. Knaus WA, Wagner DP, Draper EA et al (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100:1619–1636
5. Goldstein H (1995) Multilevel statistical models. Arnold, London
6. Goldstein H, Spiegelhalter DJ (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc A 159:385–443
7. Teres D, Lemeshow S (1993) Using severity measures to describe high performance intensive care units. Crit Care Clin 9:543–554
8. Reis Miranda D, Ryan DW, Schaufeli WB, Fidler V (1997) Organization and management of intensive care: a prospective study in 12 European countries. Springer, Berlin Heidelberg New York
9. Cleveland WS (1981) LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Stat 35:54
10. Ridgeway G (1999) The state of boosting. Comput Sci Stat 31:172–181
11. Hastie T, Tibshirani R, Friedman J (eds) (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin Heidelberg New York
12. Bauer P, Pötscher BM, Hackl P (1988) Model selection by multiple test procedures. Statistics 19:39–44
13. 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
14. Hanley J, McNeil B (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
15. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–77 (published erratum 39:1589)
16. 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
17. Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA (1998) Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Crit Care Med 26:1317–1326
18. Popovich MJ (2002) If most intensive care units are graduating with honors, is it genuine quality or grade inflation? Crit Care Med 30:2145–2146
19. Bastos PG, Knaus WA, Zimmerman JE, Magalhães Jr A, Wagner DP, the Brazil APACHE III Study Group (1996) The importance of technology for achieving superior outcomes from intensive care. Intensive Care Med 22:664–669
20. Engoren M, Moreno R, Reis Miranda D (1999) A genetic algorithm to predict hospital mortality in an ICU population. Crit Care Med 27:A52
21. Nimgaonkar A, Karnad DR, Sudarshan S, Oho-Machado L, Kohane I (2004) Prediction of mortality in an indian intensive care medicine. Comparison between APACHE II and artificial neural networks. Intensive Care Med 30:248–253
22. Clermont G, Kaplan V, Moreno R et al (2004) Dynamic microsimulation to model multiple outcomes in cohorts of critically ill patients. Intensive Care Med 30:2237–2244
23. Moreno R, Apolone G (1997) The impact of different customization strategies in the performance of a general severity score. Crit Care Med 25:2001–2008
24. Metnitz PG, Valentin A, Vesely H et al (1999) Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Intensive Care Med 25:192–197
25. Aegerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B (2005) SAPS II revisited. Intensive Care Med 31:416–423
26. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP (1993) Intensive Care Society's APACHE II study in Britain and Ireland. II. Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. BMJ 307:977–981
27. Moreno R, Matos R (2001) Outcome prediction in intensive care. Solving the paradox. Intensive Care Med 27:962–964
28. Bosman RJ, Oudemane van Straaten HM, Zandstra DF (1998) The use of intensive care information systems alters outcome prediction. Intensive Care Med 24:953–958
29. Suistomaa M, Kari A, Ruokonen E, Takala J (2000) Sampling rate causes bias in APACHE II and SAPS II scores. Intensive Care Med 26:1773–1778
30. Boyd O, Grounds M (1994) Can standardized mortality ratio be used to compare quality of intensive care unit performance?. Crit Care Med 22:1706–1708
31. Rowan K (1996) The reliability of case mix measurements in intensive care. Curr Opin Crit Care 2:209–213

Poncet et al. Critical Care (2017) 21:85
DOI 10.1186/s13054-017-1673-6

Critical Care

RESEARCH Open Access

# Determinants of the calibration of SAPS II and SAPS 3 mortality scores in intensive care: a European multicenter study

Antoine Poncet[1,2], Thomas V. Perneger[1,2], Paolo Merlani[3,4], Maurizia Capuzzo[5] and Christophe Combescure[1,2*]

## Abstract

**Background:** The aim of the Simplified Acute Physiology Score (SAPS) II and SAPS 3 is to predict the mortality of patients admitted to intensive care units (ICUs). Previous studies have suggested that the calibration of these scores may vary across countries, centers, and/or characteristics of patients. In the present study, we aimed to assess determinants of the calibration of these scores.

**Methods:** We assessed the calibration of the SAPS II and SAPS 3 scores among 5266 patients admitted to ICUs during a 4-week period at 120 centers in 17 European countries. We obtained calibration curves, Brier scores, and standardized mortality ratios. Points attributed to SAPS items were reevaluated and compared with those of the original scores. Finally, we tested associations between the calibration and center characteristics.

**Results:** The mortality was overestimated by both scores: The standardized mortality ratios were 0.75 (95% CI 0.71–0.79) for the SAPS II score and 0.91 (95% CI 0.86–0.96) for the SAPS 3 score. This overestimation was partially explained by changes in associations between some items of the scores and mortality, especially the heart rate, Glasgow Coma Scale score, and diagnosis of AIDS for SAPS II. The calibration of both scores was better in countries with low health expenditures. The between-center variability in calibration curves was much greater than expected by chance.

**Conclusions:** Both scores overestimate current mortality among European ICU patients. The magnitude of the miscalibration of SAPS II and SAPS 3 scores depends not only on patient characteristics but also on center characteristics. Furthermore, much between-center variability in calibration remains unexplained by these factors.

**Trial registration:** ClinicalTrials.gov identifier: NCT01422070. Registered 19 August 2011.

**Keywords:** Calibration, SAPS II, SAPS 3, Determinants

## Background

Scores that predict in-hospital survival of patients admitted to the intensive care unit (ICU) can be used for the assessment of ICU performance [1–4], to measure patient case mix, and to make statistical adjustments for between-group comparisons. Several predictive scores have been developed for this purpose, including the Simplified Acute Physiology Score (SAPS) II and SAPS 3 [5, 6].

Desirable characteristics of predictive scores are the capacity to distinguish between patients who will experience the studied outcome and patients who will not (i.e., discrimination) and the agreement between the observed occurrence of the outcome and the risk predicted by the score (i.e., calibration) [7]. If the discrimination is poor, the predictive score is useless in clinical practice, and calibration is irrelevant. When the discrimination is acceptable, it is necessary to investigate the quality of the calibration. Researchers in various studies have assessed the calibration of the SAPS II and SAPS 3 scores and, on the whole, found a poor calibration in European countries, especially for SAPS II. Whereas some researchers have reported that the SAPS II overestimated mortality [8–10], others have found the opposite [4, 11, 12]. The calibration

* Correspondence: christophe.combescure@hcuge.ch
[1]Clinical Research Center, Faculty of Medicine, University of Geneva, Geneva 4, 1211 Geneva, Switzerland
[2]Division of clinical epidemiology, Department of health and community medicine, University Hospitals of Geneva, Rue Gabrielle Perret-Gentil 4, 1211 Geneva, Switzerland
Full list of author information is available at the end of the article

of predictive scores can change over time because ICU populations change and new diagnostic, therapeutic and prognostic techniques become available [3]. Additionally, calibration of scores can vary across countries and even between centers within a country. Villers et al. reported a high level of heterogeneity in calibration of the SAPS II between French centers [12]. Indeed, it is possible that the reasons for admission to an ICU differ between centers, such that risk factors for mortality that are important in one center will not be useful in another, thus reducing discriminative ability. It is also possible that the general level of care differs between centers, which would influence the background risk of dying and therefore affect the calibration of the score [13]. Ethical issues such as limitation or withdrawal of therapies can also change between geographic regions and probably between centers [14].

In this study, we assessed the calibration of the SAPS II and SAPS 3 in patients admitted to ICUs in 17 European countries and sought to identify sources of miscalibration. We hypothesized that the magnitude of the association between some items of the scores and death might have decreased since the development of the scores, especially for the SAPS II, which was developed 20 years ago. We reevaluated points attributed to SAPS items and compared them with those of the original scores. We investigated the impact of the modification of scoring on calibration curves. In addition, we explored whether characteristics of centers contributed to miscalibration.

## Methods

### ELOISE study and subset of analyzed data

The primary objective of the European Mortality & Length of Intensive Care Unit Stay Evaluation (ELOISE) study was to estimate the effect on hospital mortality of the presence of an intermediate care unit (IMCU) in the hospital [15]. The analysis presented in this paper is an ancillary study. The ELOISE study included 5834 patients admitted during one of two 4-week periods (either in November 2011 or in February 2012) to 167 ICUs from 17 European countries. Excluding from our analysis ICUs that recruited fewer than 20 patients for the ELOISE study, so as to have enough observations to estimate a calibration curve for each center and enough centers to explore heterogeneity, we analyzed data of 5266 patients from 120 centers located in 17 countries. Data collection is detailed in Additional file 1.

### Calculation of SAPS II and SAPS 3 scores

The scores and the predicted mortality were calculated following the original equations for both SAPS scores [5, 6]. The risk predicted by the SAPS 3 score was assessed with equations customized for geographical area (Central/Western, Eastern, Northern, and Southern Europe) [6].

### Assessment of calibration of SAPS II and SAPS 3 scores

The calibration curves of the SAPS II and SAPS 3 scores for the prediction of in-hospital death were obtained to show the relationship between the observed and the predicted mortality. The observed risk function of the predicted mortality was assessed using smooth kernel functions [16] and was plotted against the predicted mortality. The identity line represents a perfect calibration of the score. If the curve is below (above) the identity line, the score overestimates (underestimates) the mortality. The greater the deviation from the identity line, the greater the miscalibration. Additionally, we calculated the Brier score and the standardized mortality ratio (SMR) of the scores [7]. The Brier score is the mean squared difference between the probability of death and the actual outcome (0 if the patient survives, 1 if the patient dies); a smaller value is better [17]. An SMR greater (or lower) than 1 indicates an underestimation (or overestimation) of the mortality by the predictive score.

### Calibration and patient characteristics

We reassessed the points attributed to each item in the SAPS II following the methodology used in the original work [5]. The associations between the components and mortality were based on a multivariable logistic regression model, and the number of points of an item were the nearest integer of ten times the estimated regression coefficient. If the associations obtained with data from the ELOISE study changed from the original work, the number of attributed points would also change. A greater difference between original and attributed points reflects a greater impact on calibration. Similar analyses were conducted for the SAPS 3, but using a logistic regression model with mixed effects (with patients' characteristics as fixed effects and centers as random effects on the intercept) to reproduce the methodology followed in the in the original work [6]. A post hoc analysis was conducted to assess the calibration curves, the SMRs, and the Brier scores according to the reasons for admission to the ICU. Only reasons with more than 200 admissions were investigated (cardiovascular reason, digestive reason, neurological reason, respiratory reason, severe trauma, basic observation).

### Calibration and center characteristics

We also hypothesized that some centers' characteristics may influence the calibration. First, we verified whether the variability in the calibration across centers is compatible with the variability caused by random sampling. For this purpose, we fitted a calibration curve for each of the 120 centers. The variances of the center-specific Brier scores and SMRs reflected the between-center variability in calibration. A permutation test was conducted to determine if the observed value of these variances was compatible with the hypothesis that the calibration is the same for all centers. The permutation test consisted in

attributing patients at random to centers, computing their Brier scores and SMRs, then obtaining the variances of these quantities, and repeating this procedure 1000 times. The resulting distribution of the variances of Brier scores and SRMs reflects between-center variance that is attributable only to chance; the actual observed values were compared with these distributions. To evaluate if center characteristics have an effect on the calibration of the SAPS II score, we modeled the calibration curve using the approach proposed by Finazzi et al., and we introduced interaction terms between the centers' characteristics and the logit values of the predicted mortality [18]. This analysis was conducted for each of the following characteristics: 2012 national health expenditure in percentage of gross domestic product (GDP), number of hospital beds (<500, 500–1000, >1000 beds), presence of an IMCU, presence of IMCU beds inside the ICU, number of ICU adjusted beds (two IMCU beds inside the ICU equal one ICU bed [15]), possibility of allocating additional beds inside the ICU, and the nurse/patient daytime ratio (<0.5, 0.5–1, >1). The same analyses were conducted for the SAPS 3 score.

Statistical methods are detailed in Additional file 2. All statistical analyses were performed with the R statistical software package (https://www.r-project.org/; R Foundation for Statistical Computing, Vienna, Austria). The significance level was set at 0.05, and all statistical tests were two-sided.

## Results

The characteristics of the 120 participating centers and 5266 participating patients are described in Table 1. Most hospitals had a capacity of 500–1000 beds, were located in countries with annual health expenditures greater than 8% of GPD, had an IMCU, had a daytime nurse/patient ratio between 0.5 and 1, and had a number of ICU adjusted beds greater than 12. Patients were 62.4 years old, on average (range 18–98), at ICU admission, and 60% were men. Admissions to the ICU were unplanned for 69% of patients, and 49% were admitted following surgery.

### Calibration of SAPS II and SAPS 3 scores

The SAPS II and SAPS 3 scores were collected for 5209 (98.9%) and 5206 (98.9%) patients. The number of deaths expected by the SAPS II score was 1568 (30.1%), whereas the number of observed deaths was 1194 (22.7%), resulting in an SMR of 0.75 (95% CI 0.71–0.79). The calibration curve (Fig. 1) below the identity line confirmed that the SAPS II score globally overestimated the mortality. The magnitude of the overestimation varied with the level of the mortality predicted by the SAPS II score. The predicted mortality was reasonably accurate for low-risk patients: the overestimation was less than 0.04 up to a predicted mortality of 0.20. The overestimation became important for patients with intermediate and high levels of predicted mortality (between 0.50 and 0.90): The overestimation reached 0.25 for a predicted mortality around 0.75. The Brier score for the prediction by SAPS II was 0.132 (95% CI 0.127–0.137). If the score was not able to discriminate between deceased patients and survivors (i.e., if the observed risk of death of 0.227 was used for all patients), the Brier score would be 0.175.

The number of deaths expected by the SAPS 3 score was 1322 (25.4%), resulting in an SMR of 0.91 (95% CI 0.86–0.96). The calibration curve was closer to the identity line than for the SAPS II score (Fig. 1). However, the mortality predicted by the SAPS 3 score was higher than the observed mortality for patients with a predicted risk between 0.50 and 0.90. The overestimation did not exceed 0.13. The Brier score was 0.131 (95% CI 0.126–0.136).

### Predictive value of individual items on miscalibration?

To determine if the miscalibration of the score was uniform or specific to certain score items, we compared the points attributed to each item according to the original work and the point weights derived from ELOISE data (Table 2). Items of the SAPS II score with a lowered association with mortality are extreme heart rate (<70 or >160 beats/minute), a Glasgow Coma Scale (GCS) score less than 6, a diagnosis of AIDS, a systolic blood pressure (SBP) less than 70 mm Hg and a serum sodium level less than 125 mmol/L. The SMR was 0.68 (95% CI 0.63-0.73) in patients with at least one of these items ($n = 2230$, 42.8%) and 0.89 (95% CI 0.81-0.97) in other patients.

For the SAPS 3 score, items with a decreased association with mortality were the presence of metastatic cancer, intrahospital location before ICU admission, cardiac surgery, and a heart rate greater than 160 beats/minute (Additional file 3: Table S1). The SMRs were 0.82 (95% CI 0.76–0.88) in patients with at least one of these items ($n = 2751$ [52.8%]) and 1.10 (95% CI 1.00–1.21) in other patients.

### Reasons for admission to ICU and calibration

The calibration curves and the SMRs were assessed by reason for admission to ICU (Fig. 2 and Additional file 4: Table S2). For both SAPS scores, the overestimation of mortality was especially high in patients admitted to the ICU for a basic observation for SAPS II score (SMR 0.44, 95% CI 0.34–0.57) and for SAPS 3 score (SMR 0.68, 95% CI 0.52–0.88). In this subpopulation, the calibration curves deviated from the identity line even for low predicted risks. A similar but less marked trend was observed in patients admitted to the ICU for a severe trauma for SAPS II score (SMR 0.56, 95% CI 0.39–0.78) and for SAPS 3 score (SMR 0.73, 95% CI 0.51–1.02). For other reasons for admission, the miscalibration was less pronounced or even low. For instance, the SAPS 3 score was well calibrated in patients admitted to the ICU for a cardiovascular reason (SMR 0.94, 95% CI 0.86–1.03).

**Table 1** Centers and patients characteristics

| Center characteristics | Centers ($n = 120$) | Patients ($n = 5266$) |
|---|---|---|
| Number of patients/ICU, median [range] | 32 [20–89] | |
| Number of hospital beds[a], *n* (%) | | |
| < 500 | 39 (33.6%) | 1403 (27.5%) |
| 500–1000 | 54 (46.6%) | 2630 (51.5%) |
| > 1000 | 23 (19.8%) | 1072 (21.0%) |
| Health expenditure (% of GDP[b]), *n* (%) | | |
| < 8% | 19 (15.8%) | 961 (18.2%) |
| 8% to 10% | 51 (42.5%) | 2107 (40.0%) |
| > 10% | 50 (41.7%) | 2198 (41.7%) |
| IMCU (intermediate care unit), *n* (%) | | |
| Yes | 103 (85.8%) | 4563 (86.7%) |
| Daytime nurse/patient ratio, *n* (%) | | |
| < 0.5 | 25 (20.8%) | 1150 (21.8%) |
| 0.5–1 | 58 (48.3%) | 2536 (48.2%) |
| > 1 | 37 (30.8%) | 1580 (30.0%) |
| ICU adjusted beds, *n* (%) | | |
| < 8 | 19 (15.8%) | 571 (10.8%) |
| 8–12 | 49 (40.8%) | 1901 (36.1%) |
| > 12 | 52 (43.3%) | 2794 (53.1%) |
| Possibility of extra beds inside ICU, *n* (%) | | |
| Yes | 24 (20.0%) | 1114 (21.2%) |
| Patient characteristics | | |
| Male sex, *n* (%) | | 3143 (59.7%) |
| Age, years, mean ± SD | | 62.4 ± 16.9 |
| SAPS II[c], mean ± SD | | |
| Score | | 39.3 ± 21.3 |
| Predicted mortality | | 30.1% ± 30.2 |
| SAPS 3[d], mean ± SD | | |
| Score | | 35.0 ± 17.2 |
| Predicted mortality | | 25.4% ± 24.5 |
| Hospital mortality, *n* (%) | | |
| Death | | 1194 (22.7%) |
| ICU admission, *n* (%) | | |
| Unplanned | | 3613 (68.7%) |
| Surgery, *n* (%) | | |
| Emergency surgery | | 983 (18.7%) |
| No surgery | | 2663 (50.6%) |
| Scheduled surgery | | 1612 (30.7%) |
| Reason for admission[e], *n* (%) | | |
| Basic observation | | 1111 (21.1%) |
| Cardiovascular | | 1252 (23.8%) |
| Digestive | | 526 (10.0%) |
| Hematological | | 77 (1.5%) |
| Hepatic | | 62 (1.2%) |
| Metabolic | | 195 (3.7%) |
| Neurological | | 800 (15.2%) |
| Renal | | 200 (3.8%) |
| Respiratory | | 980 (18.6%) |
| Severe trauma | | 255 (4.8%) |

*Abbreviations: GDP* Gross domestic product, *ICU* Intensive care unit, *IMCU* Intermediate care unit, *SAPS* Simplified Acute Physiology Score

[a]The total number of hospitals giving information on the number of hospital beds was 116

[b]Health expenditure in the country of the center expressed in percentage of GDP. (*Source*: World Bank [http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS].)

[c]The medians (interquartile ranges) were 35 [23–52] for the SAPS II score and 16.7% [5.2% to 50.7%] for the mortality predicted by the SAPS II score

[d]The medians (interquartile ranges) were 33 [22–46] for the SAPS 3 score and 15.9% [5.1% to 39.8%] for the mortality predicted by the SAPS 3 score

[e]Reasons for admission were not exclusive (except "basic observation," which is exclusive of all other reasons)

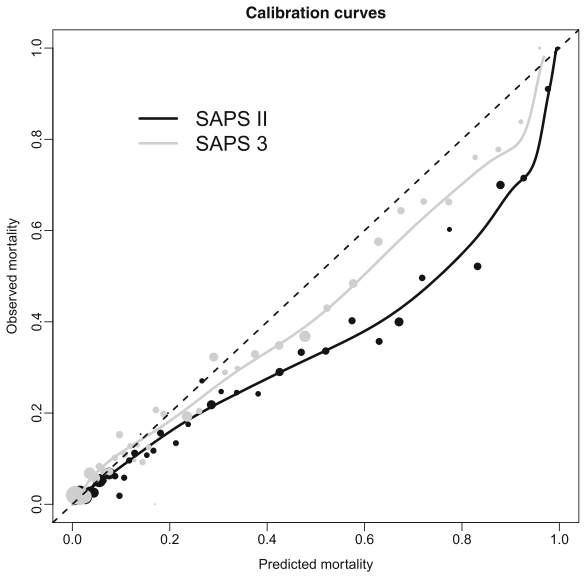


**Fig. 1** Calibration curves for the Simplified Acute Physiology Score (SAPS) II (*dark line*) and the SAPS 3 (*gray line*) obtained with a kernel function. The calibration curve represents the relationship between the mortality predicted by the score (*x*-axis) and the observed mortality (*y*-axis). The identity line (*dashed line*) represents a perfect calibration. A calibration curve below the identity line indicates that the score overestimates the mortality. The *black* and *gray circles* represent the estimates of the observed mortality in sample, stratified by levels of predicted mortality (by step of 0.01 up to a predicted mortality of 0.20, by step of 0.025 for a predicted mortality from 0.20 to 0.35, and by step of 0.05 for a predicted mortality greater than 0.35). The size of the *circles* is proportional to the number of patients in categories of predicted mortality

**Table 2** Reassessment of the points allocated to each item of Simplified Acute Physiology Score II items

| Items of SAPS II score | Points[a] (original/ELOISE study) | Difference |
|---|---|---|
| Age, years | | |
| 20–39 | 0/0 | 0 |
| 40–59 | 7/7 | 0 |
| 60–69 | 12/11 | 1 |
| 70–74 | 15/14 | 1 |
| 75–79 | 16/15 | 1 |
| ≥ 80 | 18/19 | −1 |
| Heart rate, beats/minute | | |
| < 40 | 11/4 | 7 |
| 40–69 | 2/−5 | 7 |
| 70–119 | 0/0 | 0 |
| 120–159 | 4/3 | 1 |
| ≥ 160 | 7/−5 | 12 |
| SBP, mmHg | | |
| ≥ 200 | 2/3 | −1 |
| 100–199 | 0/0 | 0 |
| 70–99 | 5/3 | 2 |
| < 70 | 13/7 | 6 |
| $PaO_2$, mmHg/$FiO_2$ | | |
| No ventilation | 0/0 | 0 |
| ≥ 200 | 6/3 | 3 |
| 100–199 | 9/6 | 3 |
| < 100 | 11/11 | 0 |
| Urinary output, L/day | | |
| ≥ 1.000 | 0/0 | 0 |
| 0.500–0.999 | 4/0 | 4 |
| < 0.500 | 11/8 | 3 |
| Serum urea level, mmol/L | | |
| < 10.0 | 0/0 | 0 |
| 10.0–29.9 | 6/4 | 2 |
| ≥ 30.0 | 10/5 | 5 |
| Body temperature | | |
| < 39 °C | 0/0 | 0 |
| ≥ 39 °C | 3/−2 | 5 |
| WBC count, $\times 10^3/mm^3$ | | |
| < 1.0 | 12/8 | 4 |
| 1.0–19.9 | 0/0 | 0 |
| ≥ 20.0 | 3/2 | 1 |
| Serum potassium, mmol/day | | |
| ≥ 3 and <5 | 0/0 | 0 |
| < 3 or ≥5 | 3/2 | 1 |
| Serum sodium level, mmol/L | | |
| < 125 | 5/−1 | 6 |
| ≥ 125 and <145 | 0/0 | 0 |
| ≥ 145 | 1/5 | −4 |
| Serum bicarbonate level, mEq/L | | |
| ≥ 20 | 0/0 | 0 |
| 15–19 | 3/4 | −1 |
| < 15 | 6/9 | −3 |
| Bilirubin level, μmol/L | | |
| < 68.4 | 0/0 | 0 |
| 68.4–102.5 | 4/2 | 2 |
| ≥ 102.6 | 9/10 | −1 |
| Glasgow Coma Scale score | | |
| 14–15 | 0/0 | |
| 11–13 | 5/5 | 0 |
| 9–10 | 7/9 | −2 |
| 6–8 | 13/10 | 3 |
| < 6 | 26/16 | 10 |
| Chronic disease | | |
| No | 0/0 | 0 |
| Metastatic cancer | 9/8 | 1 |
| Hematologic malignancy | 10/9 | 1 |
| AIDS | 17/9 | 8 |
| Type of admission | | |
| Scheduled surgical | 0/0 | 0 |
| Medical | 6/11 | −5 |
| Unscheduled surgical | 8/9 | −1 |

*Abbreviations: ELOISE* European Mortality & Length of Intensive Care Unit Stay Evaluation study, *$FiO_2$* Fractional inspired oxygen, *$PaO_2$* Partial pressure of arterial oxygen, *SAPS* Simplified Acute Physiology Score, *SBP* Systolic blood pressure, *WBC* White blood cell

[a] Points proposed in the original SAPS II score and the points derived from the association between the items of the SAPS II score and the mortality reassessed with data from the ELOISE study

### Between-center variability

We fitted a calibration curve of the SAPS II score separately in each of the 120 centers (Fig. 3a). The calibration curves varied considerably, but it was unclear if the variance was greater than what would be expected by chance alone. A typical pattern of calibration curves expected under the assumption that calibration is the same in all centers was obtained by randomly permuting the patients between centers (Fig. 3b). These figures suggest that the observed between-center variability in calibration is higher than the variability expected by chance. Figure 3c represents the distribution of the SD of the SMRs expected under the null hypothesis of absence of center effect on calibration. The observed SD of the SMR, represented by a vertical line, falls on the right-hand side of the distributions; the *p* value from

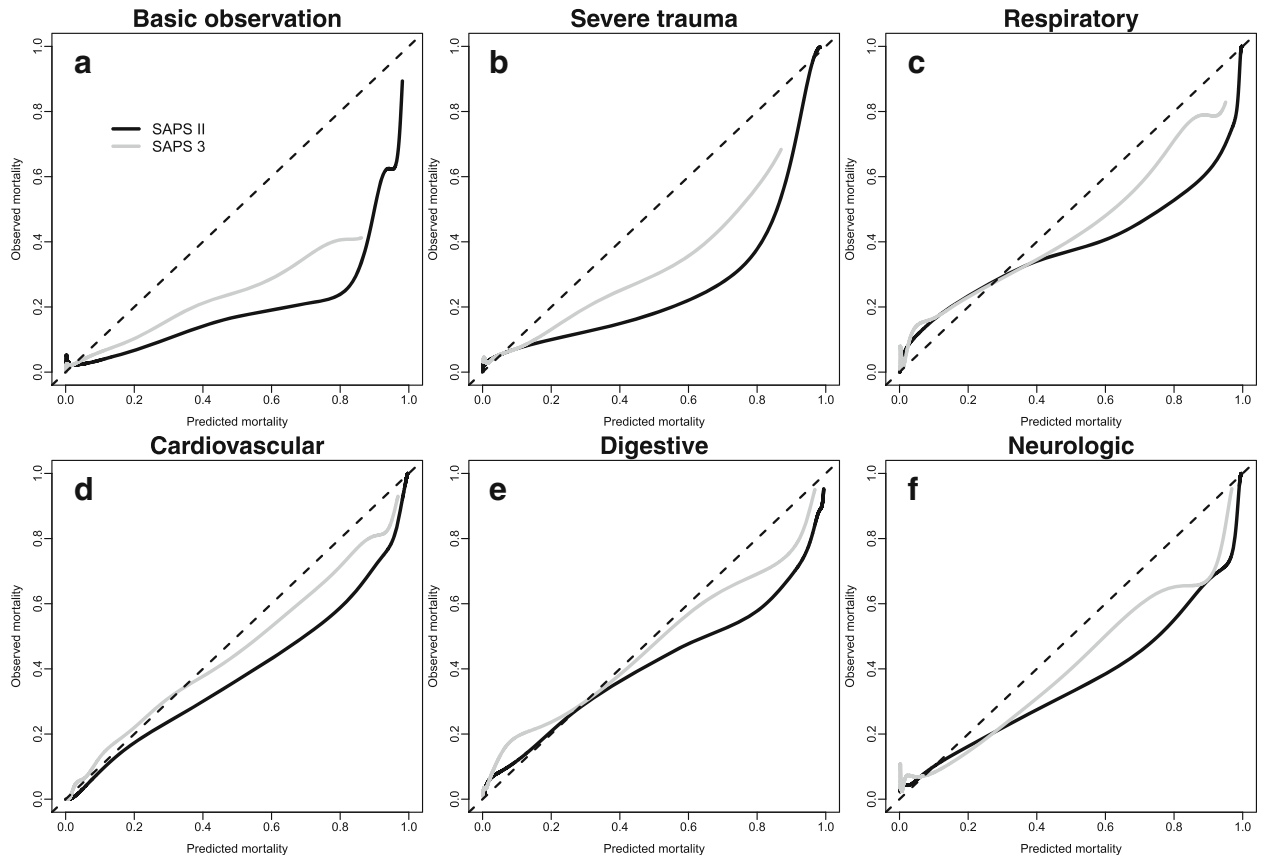


**Fig. 2** Calibration curves of the Simplified Acute Physiology Score (SAPS) II and SAPS 3 obtained by kernel function by reason for admission to the intensive care unit. **a** Basic observation. **b** Severe trauma. **c** Respiratory reason. **d** Cardiovascular reason. **e** Digestive reason. **f** Neurological reason

the permutation test was less than 0.001. These findings show that the between-center variability in calibration of the SAPS II score is not well explained by random variability and suggest that center characteristics may add to this variability. The same findings were observed with the Brier score (Additional file 5: Figure S1A).

In regression models, the health expenditure and the number of hospital beds were significantly associated with the shape of the calibration curve of the SAPS II score ($p < 0.001$ and $p = 0.004$, respectively). The calibration curves according to these factors are shown in Fig. 4a and b. The SAPS II score was well calibrated in centers located in countries with health expenditures in 2012 less than 8% of the GDP, but the fit gets progressively worse as health expenditures grow. Furthermore, the overestimation of the risk of death was lower in ICUs in hospitals with 500–1000 beds than in centers in either smaller or larger hospitals. Other center characteristics were not significantly associated with the shape of the calibration curve (presence of IMCU $p = 0.91$, presence of an IMCU beds inside ICU $p = 0.20$, number of ICU adjusted beds $p = 0.73$, possibility of allocating extra beds inside the ICU $p = 0.99$, ICU nurse/patient ratio in daytime $p = 0.10$).

For SAPS 3, excess between-center variability in SMRs (Fig. 3d–f) ($p < 0.001$) and in Brier scores (Additional file 5: Figure S1B) was also found. In regression models, the health expenditure and ICU nurse/patient ratio in daytime were significantly associated with the shape of the calibration curve ($p < 0.001$ and $p = 0.036$, respectively). The corresponding calibration curves are shown in Fig. 4c and d. For centers located in countries with health expenditures less than 8% of the GDP, the SAPS 3 score underestimated the mortality. The other center characteristics were not significantly associated with the shape of the calibration curve (number of hospital beds $p = 0.09$, presence of IMCU $p = 0.68$, presence of an IMCU beds inside ICU $p = 0.80$, number of ICU adjusted beds $p = 0.62$, possibility of allocating extra beds inside the ICU $p = 0.96$). Data for SMR and Brier score by level of health expenditure are shown in Additional file 6 (Table S3) for both SAPS scores.

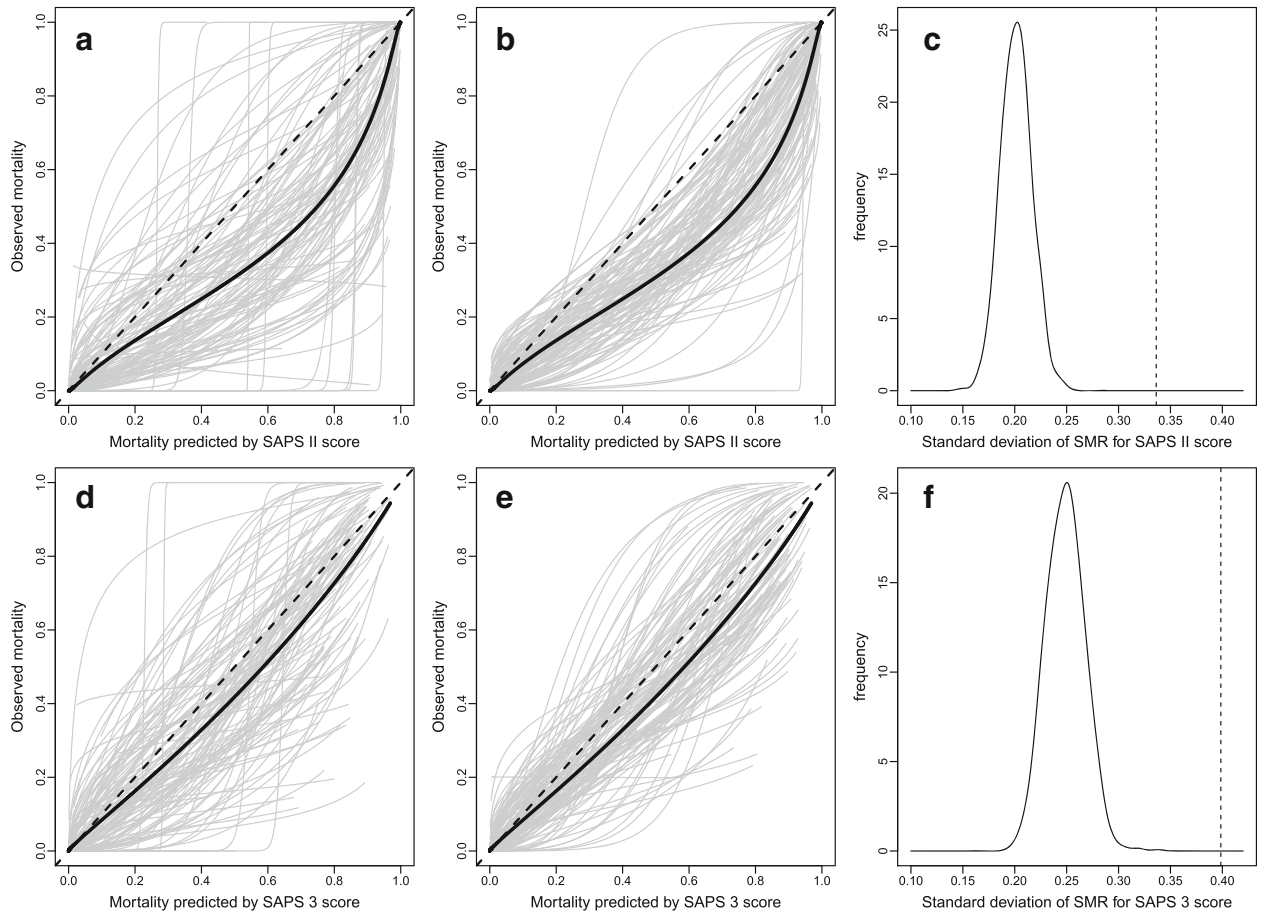


**Fig. 3** Observed and expected calibration curves for the Simplified Acute Physiology Score (SAPS) II score (*top*) and SAPS 3 score (*bottom*) in 120 centers and between-center variability in standardized mortality ratio (SMR). **a** Calibration curves of SAPS II in each of the 120 centers fitted with a logistic regression model. The *black line* represents the overall calibration curve. **b** Expected calibration curves of SAPS II under the assumption that the calibration is the same in all centers. The represented between-center variability is the random (sampling) variability. **c** Distribution of the SD of the center-specific SMRs under the assumption that the calibration of SAPS II is the same in all centers. The *vertical line* represents the observed value of the SD. **d**–**f** are the same figures shown in **a**–**c** repeated for the SAPS 3 score

## Discussion

The SAPS II and SAPS 3 scores globally overestimated mortality, with an overestimation more marked for the SAPS II (SMR 0.75) than for the SAPS 3 (SMR 0.91). Although overestimation of mortality has been reported by others [10, 19–22], we show that this miscalibration does not affect all patients and all ICUs similarly. First, the miscalibration depended on the level of the predicted risk in each patient and on the specific items of the scores presented by the patients. Second, the calibration varied across centers; the miscalibration was more important in countries with high health expenditures, as well as in small and large hospitals than in hospitals of medium size.

The scores calibrated well when the predicted risk was low (below a predicted risk of 0.30 approximately), and the overestimation increased up to 0.25 for the SAPS II score (0.13 for the SAPS 3) at around 0.75 predicted mortality. The points originally attributed to some items of the score do not capture correctly the increase of mortality anymore, owing to the magnitude of the associations changed since the development of the score. The main items of predictive scores with a lowered association were heart rate, GCS (<6), and chronic disease (AIDS) for the SAPS II score and anatomical site of surgery (transplantation, trauma–other), intrahospital location before ICU admission, comorbidities (metastatic cancer), heart rate (≥160 beats/minute) for the SAPS 3 score. Some of these decreased associations (heart rate, SBP) may be explained by modern automatic or semiautomatic data collection methods that have been shown to find more “pathological” elements, thereby inflating the assigned SAPS scores [23]. The decreased association of AIDS may be explained by the introduction of highly effective therapies against HIV. The decreased predictive capacity of GCS for SAPS II may be caused by

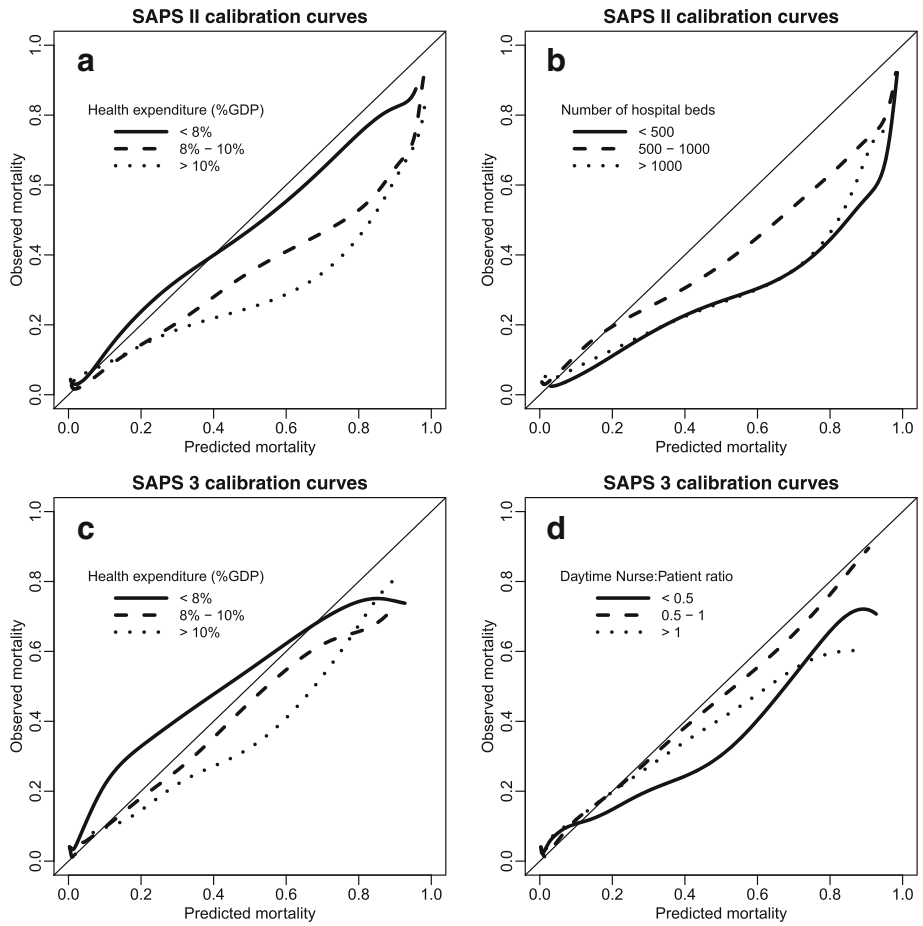


**Fig. 4** Calibration curves of the Simplified Acute Physiology Score (SAPS) II score obtained by kernel function according to (**a**) health expenditure expressed in percentage of gross domestic product (GDP) and (**b**) number of hospital beds, and calibration curves of the SAPS 3 score by (**c**) health expenditure expressed as a percentage of GDP and (**d**) daytime nurse/patient ratio

a common misconception about the evaluation rules [24]: A sedated patient is sometimes mistakenly attributed the worst score (3), whereas the score should reflect the state in which we believe the patients would be without sedation. Another possible explanation is that data are of lower quality in real life than in research validation studies, and random errors would also dilute the associations.

The calibration of the scores varied across the reason for admission to ICU. Especially, the mortality predicted by SAPS II and SAPS 3 scores was too high when the scores were applied to patients admitted to the ICU for a basic observation or for a severe trauma. Possibly, the relationship between the mortality and biological parameters involved in the predictive scores is different in patients admitted to the ICU for any traumatic injuries responsible for a strong physiological stress reaction and in patients admitted for another reason. The biological values may capture well the stage of medical diseases but poorly the effects of the homeostatic mechanisms favoring recovery after trauma. In addition to the influence of characteristics of patients on calibration, we detected a large heterogeneity across centers. The variability of the calibration was too large to be explained only by random sampling. Some characteristics of centers were associated with the miscalibration of the SAPS scores: the country's health expenditure (SAPS II and SAPS 3), number of hospital beds (SAPS II), and the daytime nurse/patient ratio (SAPS 3). If we have no reasonable explanation for the variation by hospital size, the effect of health expenditure may be explained by the amount of resources available in the ICU to treat patients. In low-expenditure countries, lifesaving medical technologies may be underused or rationed, which may cause higher mortality more comparable to mortality rates that existed 25 years ago, when the SAPS II score was developed. Any new effective medical treatment is bound to reduce the predictive value of the medical condition it treats; for example, survival after a myocardial infarction has improved since the introduction of percutaneous transluminal coronary angioplasty and thrombolytic therapies.

This study has several limitations. Analyzed data were collected as part of the ELOISE study, in which researchers sought to detect an effect on mortality of the presence of an IMCU in the hospital. Because the ELOISE study was not designed to assess the determinants of the calibration of the SAPS II and SAPS 3 mortality scores, some determinants of the calibration were not collected, such as the policy for end-of-life care. Moreover, ICUs participated on a voluntary basis, and they may not represent all European ICUs.

## Conclusions

This study suggests that the prognostic significance of SAPS II and SAPS 3 scores is not uniform across Europe, because it depends on both patient-specific and center-specific characteristics. Another important part of variability remains unexplained. This suggests that users of these scores should proceed with caution, especially if ICUs that serve different patient populations and that are located in countries with different levels of health expenditures are being compared. More generally, our results suggest that the external validity of prognostic scores developed in a given context should not be taken for granted, as well as that local revalidation is a useful precaution. Furthermore, it may be prudent to reassess periodically the predictive capacity of even well-established scores because changes in medical treatments may alter the value of such instruments.

## Additional files

**Additional file 1:** Additional details on data collection. (DOCX 14 kb)

**Additional file 2:** Additional details on statistical methods. (DOCX 16 kb)

**Additional file 3:** Original and reassessed points of the items of SAPS 3 score. (DOCX 18 kb)

**Additional file 4:** SMRs and Brier scores of the SAPS II and SAPS 3 scores, by reason for admission to ICU. (DOCX 11 kb)

**Additional file 5:** Distribution of the SD of the center-specific Brier scores under the assumption that the calibration is the same in all centers for (**a**) the SAPS II score and (**b**) the SAPS 3 score. The *vertical lines* represent the observed SD of Brier score. (DOCX 24 kb)

**Additional file 6:** SMRs and Brier scores of the SAPS II and SAPS 3 scores, by categories of health expenditure (percentage of GDP). (DOCX 12 kb)

**Additional file 7:** Ethics committees. (DOCX 33 kb)

### Abbreviations

ELOISE: European Mortality & Length of Intensive Care Unit Stay Evaluation study; $FiO_2$: Fractional inspired oxygen; GCS: Glasgow Coma Scale; GDP: Gross domestic product; ICU: Intensive care unit; IMCU: Intermediate care unit; $PaO_2$: Partial pressure of arterial oxygen; SAPS: Simplified Acute Physiology Score; SBP: Systolic blood pressure; SMR: Standardized mortality ratio; WBC: White blood cell

### Acknowledgements

We thank the steering committee members, country coordinators, and study unit coordinators of the ELOISE study, endorsed by the European Society of Intensive Care Medicine. Lists of these individuals can be found in the appendix of the original study [15]. This project was funded by the University Hospitals of Geneva (Projet de recherche et développement PRD 15-I-2012).

### Funding

This project was funded by the University Hospitals of Geneva (Projet de recherche et développement PRD 15-I-2012).

### Availability of data and materials

The datasets generated and/or analyzed during the present study are not publicly available, owing to currently ongoing research studies.

### Authors' contributions

AP participated in the conception of the statistical plan analysis, analyzed data, participated in the interpretation of results, and helped in drafting the manuscript. TVP conceived of this ancillary study, participated in data analysis and the interpretation of results, and revised the manuscript critically. PM participated in the conception of this ancillary study, participated in the interpretation of results, and revised the manuscript critically. MC supervised data collection, conducted the control of data quality, participated in the interpretation of results, and revised the manuscript critically. CC participated in the conception of the design of this ancillary study, conceived of the statistical analysis plan, participated in data analysis and interpretation of the results, and drafted the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have competing interests.

### Consent for publication

Not applicable.

### Ethics approval and consent to participate

The ELOISE study was granted approval by local ethics committees when needed; in some countries, the approval was not required, owing to the nature of the study being noninterventional. This is explained in a previously published article [15]. This article presents the results of an ancillary study of the ELOISE study. The list of ethical bodies is provided in Additional file 7.

## Publisher's Note



### Author details

[1]Clinical Research Center, Faculty of Medicine, University of Geneva, Geneva 4, 1211 Geneva, Switzerland. [2]Division of clinical epidemiology, Department of health and community medicine, University Hospitals of Geneva, Rue Gabrielle Perret-Gentil 4, 1211 Geneva, Switzerland. [3]Department of Anesthesiology, Intensive Care and Pharmacology, University Hospitals of Geneva, Rue Gabrielle Perret-Gentil 4, 1211 Geneva, Switzerland. [4]Intensive Care Unit, Lugano Regional Hospital, Via Tesserete 46, 6900 Lugano, Switzerland. [5]Department of Morphology, Surgery and Experimental Medicine, Section of Anesthesia and Intensive Care, Sant'Anna Hospital, University of Ferrara, Via Aldo Moro 8, Cona, 44124 Ferrara, Italy.

Received: 21 November 2016 Accepted: 17 March 2017
Published online: 04 April 2017

### References

1. de Vos M, Graafmans W, Keesman E, Westert G, van der Voort PH. Quality measurement at intensive care units: which indicators should we use? J Crit Care. 2007;22:267–74.
2. Glance LG, Osler TM, Dick A. Rating the quality of intensive care units: is it a function of the intensive care unit scoring system? Crit Care Med. 2002;30:1976–82.
3. Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. Crit Care. 2010;14:207.
4. Vosylius S, Sipylaite J, Ivaskevicius J. Evaluation of intensive care unit performance in Lithuania using the SAPS II system. Eur J Anaesthesiol. 2004;21:619–24.
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. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med. 2005;31:1345–55.

7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38.
8. Metnitz PGH, 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. Intensive Care Med. 1999;25:192–7.
9. Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G. 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.
10. 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.
11. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G, et al. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Intensive Care Med. 1996;22:1368–78.
12. Villers D, Fulgencio JP, Gouzes C, Hémery F, Blériot JP, Garrigues B, et al. ICU performance: results of a French study involving 80,000 ICU stays [in French]. Ann Fr Anesth Reanim. 2006;25:1111–8.
13. Capuzzo M, Moreno RP, Le Gall JR. Outcome prediction in critical care: the Simplified Acute Physiology Score models. Curr Opin Crit Care. 2008;14:485–90.
14. Sprung CL, Cohen SL, Sjokvist P, Baras M, Bulow HH, Hovilehto S, et al. End-of-life practices in European intensive care units: the Ethicus Study. JAMA. 2003;290:790–7.
15. Capuzzo M, Volta C, Tassinati T, Moreno R, Valentin A, Guidet B, et al. Hospital mortality of adults admitted to intensive care units in hospitals with and without intermediate care units: a multicentre European cohort study. Crit Care. 2014;18:551.
16. Copas JB. Plotting *p* against *x*. J R Stat Soc: Ser C: Appl Stat. 1983;32:25–31.
17. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78:1–3.
18. Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G. Calibration belt for quality-of-care assessment based on dichotomous outcomes. PLoS One. 2011;6:16110.
19. Capuzzo M, Scaramuzza A, Vaccarini B, Gilli G, Zannoli S, Farabegoli L, et al. Validation of SAPS 3 admission score and comparison with SAPS II. Acta Anaesthesiol Scand. 2009;53:589–94.
20. Desa K, Peric M, Husedzinovic I, Sustic A, Korusic A, Karadza V, et al. Prognostic performance of the Simplified Acute Physiology Score II in major Croatian hospitals: a prospective multicenter study. Croat Med J. 2012;53:442–9.
21. Haaland OA, Lindemark F, Flaatten H, Kvale R, Johansson KA. A calibration study of SAPS II with Norwegian intensive care registry data. Acta Anaesthesiol Scand. 2014;58:701–8.
22. Nassar AP, Malbouisson LM, Moreno R. Evaluation of Simplified Acute Physiology Score 3 performance: a systematic review of external validation studies. Crit Care. 2014;18:117.
23. Suistomaa M, Kari A, Ruokonen E, Takala J. Sampling rate causes bias in APACHE II and SAPS II scores. Intensive Care Med. 2000;26:1773–8.
24. Green SM. Cheerio, laddie! Bidding farewell to the Glasgow Coma Scale. Ann Emerg Med. 2011;58:427–30.