*Acta Anaesthesiol Scand 2008; **52**: 467–478*
*Printed in Singapore. All rights reserved*



ACTA ANAESTHESIOLOGICA SCANDINAVICA
doi: 10.1111/j.1399-6576.2008.01586.x

*Review Article*

# Severity scoring in the ICU: a review

K. STRAND[1] and H. FLAATTEN[1,2]
[1]*Department of Anaesthesia and Intensive Care, Haukeland University Hospital, Bergen, Norway and* [2]*Department of Surgical Sciences, Faculty of Medicine, University of Bergen, Bergen, Norway*

**Background:** Patients in the intensive care unit (ICU) require huge resources because of the dysfunction of several of their vital organs. The heterogeneity and complexity of the ICU patient have generated interest in systems able to measure severity of illness as a method of predicting outcome, comparing quality-of-care and stratification for clinical trials.
**Methods:** By searching Medline and EMBASE for publications describing scoring systems in the ICU, the most frequently used systems, defined as resulting in more than 50 references, are included in this review. Scoring systems belong to one of four classes prognostic, single-organ failure, trauma scores and organ dysfunction (OD). The different systems are described and discussed.
**Results:** Three different prognostic scoring systems, including several versions, four single OD scores and three OD scores, were included in this review.
**Conclusion:** Different forms of scoring systems are frequently used in the ICU. They have become a necessary tool to describe ICU populations and to explain differences in mortality. As there are several pitfalls related to the interpretation of the numbers supplied by the systems, they should not be used without knowledge on the science of severity scoring.

*Accepted for publication 14 November 2007*

**Key words:** Intensive care; case mix; multiple-organ failure; APACHE; prognosis; risk assessment; severity of illness index.



THE use of scoring systems to predict risk of mortality and evaluating outcome in critically ill patients is important in modern medicine. The first such system in widespread use was the APGAR score introduced in 1953 to asses the vitality of the newborn (1). The Glasgow Coma Scale (GCS) and Ranson score are other examples of systems that have gained widespread use. Within intensive care, a large number of scoring systems aimed either at the general intensive care unit (ICU) patient or defined subgroups have been developed during the last two decades. Prognostic or general severity scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE) (2) and Simplified Acute Physiology Score (SAPS) (3) estimate risk based on data available within the first 24 h of ICU stay. The standard mortality ratio (SMR), a key element in ICU benchmarking, can be calculated using these systems. Disease-specific scoring systems have been developed for several important subgroups treated in the ICU, such as pancreatitis, hepatic failure and adult respiratory distress syndrome. Because the ICU treats patients with one or more organ dysfunction (OD), several organ failure scoring systems have also been developed in the last 10 years. Scoring systems are also important in clinical trials and in the monitoring of quality-of-care. In this review, we present an overview of the most common severity scoring systems used in the ICU, explain the terms frequently encountered in the literature, and discuss the inherent possibilities and limitations of these systems.

## Methods

We searched Medline and EMBASE for relevant publications using the following search terms: 'severity score', 'mortality prediction', 'organ failure score', 'trauma score', 'scoring systems' and 'intensive care' or 'critical care'. Scoring systems were included in the review if the search for the specific system yielded more than 50 citations. Scoring systems primarily aimed to assess the use of resources in ICU patients are left out in this review.

ICU scoring systems can be divided into four major groups: general risk-prognostication scores

(severity of illness scores), disease-specific risk-prognostication scores, trauma scoring and OD (failure) scoring. Because trauma should be evaluated and scored at hospital admission, usually in the emergency department, these scoring systems, although of relevance to the ICU, are omitted from our discussion.

## General risk-prognostication systems

The basis for development of both the APACHE system (4) and the SAPS (5) in 1982 was the assumption that the severity of acute disease could be measured by quantifying the degree of abnormality of physiologic variables. These first versions were soon replaced by more sophisticated models using prospective sampled patient data and advanced logistic regression analysis. The characteristics of the general risk-prognostication systems included in this review are shown in Table 1. External validation data for these systems are displayed in Table 2.

### *APACHE II*

The APACHE II model (2), published in 1985, was developed due to the complexity of the original model and it has become the most frequently used general mortality prediction model (MPM). The original number of physiologic variables was reduced from 34 to 12 and some were re-weighted. Multivariate analysis was performed as part of the reduction process and variables were selected on the basis of clinical experience.

Patients under the age of 16 were not included. In addition to the acute physiology variables, age, operative status and the presence of severe chronic OD or immune suppression were incorporated. A list of coefficients corresponding to the admission category is supplied to increase the precision of the model and is incorporated into the final equation. The APACHE II has been extensively validated. Despite being the oldest system in common use, it still performs well in most validation studies. In a large study of more than 141,000 patients, the APACHE II had a slightly lower discrimination than APACHE III, SAPS II and MPM II (7). Calibration was found unsatisfactory in all systems.

### *APACHE III*

APACHE III was developed as a further refinement of APACHE II (34). Data from 17,440 patients in

*Table 1*

Characteristics of general risk-prognostication systems.

| | Year published | Development database | Origin of database | Age (years) | Collection of data | Simplicity of scoring | aROC* | GOF H-L C-test (*P*)† | External validation |
|---|---|---|---|---|---|---|---|---|---|
| APACHE II | 1985 | 5005 | USA | > 16 | First 24 h in ICU | + + | 0.86 | – | + + + |
| APACHE III | 1993 | 7848† | USA | > 16 | First 24 h in ICU | + | 0.90 | – | + + |
| APACHE IV | 2006 | 66,335† | USA | > 16 | First 24 h in ICU | + | 0.88 | 16.8 (0.08) | – |
| SAPS II | 1993 | 13,152 | Europe/North-America | > 18 | First 24 h in ICU | + + + | 0.86 | – | + + + |
| SAPS III | 2005 | 13,428‡ | All continents | > 16 | ICU admission ± 1 h | + + | 0.85 | 14.3 (0.16) | – |
| MPM II 0 | 1993 | 12,610 | Europe/North-America | > 18 | ICU admission | + + + | 0.82 | (0.327) | + + |
| MPM II 24 | 1993 | 10,357 | Europe/North-America | > 18 | At 24 h in ICU | + + + | 0.84 | (0.231) | + + |

*Discrimination (aROC, area under curve of receiver operating characteristic) in original publication.
†Calibration (goodness-of-fit Hosmer–Lewenshow *C*-statistic) in original publication.
‡Estimated by authors based on information in original publication.
APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, mortality prediction model; ICU, intensive care unit; ROC, receiver operating curves.

*Table 2*

Performance of general risk-prognostication systems.

| Author, year (Reference) | System | Data collection | No. of patients | ICU population | Country | Discrimination | Calibration (*P*) |
|---|---|---|---|---|---|---|---|
| Soares, 2006 (6) | SAPS III | 2003–2005 | 952 | Oncologic | Brazil | 0.87 | 13.6 (0.092) |
| | SAPS III CSA | | | | | 0.87 | 9.1 (0.331) |
| | SAPS II | | | | | 0.88 | 32.1 (0.001) |
| Harrison, 2006 (7) | APACHE II | 1995–2003 | 1,41,106 | General | UK | 0.80 | ($\ll$ 0.01) |
| | APACHE III | | | | | 0.83 | ($\ll$ 0.01) |
| | SAPS II | | | | | 0.82 | ($\ll$ 0.01) |
| | MPM II | | | | | 0.82 | ($\ll$ 0.01) |
| LeGall, 2005 (8) | SAPS II | 1998–1999 | 77,490 | General | France | 0.86 | 1162.9 (0.0001) |
| Lima, 2005 (9) | APACHE II | 2000–2001 | 324 | Renal failure | Brazil | 0.72 | (0.001) |
| | SAPS | | | | | 0.77 | ($<$0.001) |
| | LODS | | | | | 0.68 | ($<$0.001) |
| Aegerter, 2005 (10) | SAPS II | 1999–2000 | 13,739 | General | France | 0.87 | ($<$0.001) |
| Reiter, 2004 (11) | SAPS II | 1998–2001 | 35,637 | Trauma | Austria | 0.87 | 290 (0.001) |
| Ihnsook, 2003 (12) | APACHE III | 1999–2000 | 284 | General | Korea | 0.91 | 6.5 (6.54) |
| Beck, 2003 (13) | APACHE II | 1993–1996 | 16,646 | General | UK | 0.84 | 232.1 |
| | APACHE III | | | | | 0.87 | 443.3 |
| | SAPS II | | | | | 0.85 | 287.5 |
| Pettila, 2002 (14) | APACHE III | 1995 | 520 | General | Finland | 0.83 | 9.3 (0.041) |
| | LODS | | | | | 0.81 | 5.4 (0.8) |
| Cook, 2002 (15) | APACHE III | 1995–2000 | 5681 | General | Australia | 0.89 | 28.8 (0.001) |
| Timsit, 2001 (16) | SAPS II | | 893 | LOS $>$ 4 days | France | 0.74 | 37.4 (0.001) |
| | LODS | | | | | 0.68 | 21.2 (0.01) |
| | APACHE II | | | | | 0.71 | 19.0 (0.01) |
| Capuzzo, 2000 (17) | SAPS II | 1994–1997 | 1721 | General | France | 0.82 | 9.3 ($>$0.5) |
| | APACHE II | | | | | 0.81 | 5.1 ($>$0.9) |
| Rue, 2000 (18) | MPM II 0 | 1995 | 1441 | General | Spain | 0.80 | 67.9 (0.001) |
| | MPM II 24 | | | | | 0.84 | 28.3 (0.002) |
| | MPM II 48 | | | | | 0.82 | 31.2 (0.001) |
| | MPM II 72 | | | | | 0.81 | 35.4 (0.001) |
| Metnitz, 2000 (19) | SAPS II | 1997–1998 | 2901 | General | Austria | 0.83 | 69.1 (0.001) |
| Glance, 2000 (20) | APACHE II | | 6806 | General | USA | 0.74 | 233 (0.001) |
| Katsaragakis, 2000 (21) | APACHE II | 1992–1997 | | General | Greece | 0.84 | 18.1 (0.02) |
| | SAPS II | | | | | 0.87 | 60.5 (0.001) |
| Livingston, 2000 (22) | APACHE II | | 10,393 | General | Scotland | 0.76 | 67.4 (0.001) |
| | APACHE III | | | | | 0.80 | 366 (0.001) |
| | SAPS II | | | | | 0.78 | 142.0 (0.001) |
| | MPM II 0 | | | | | 0.74 | 451.9 (0.001) |
| | MPM II 24 | | | | | 0.79 | 100.8 (0.001) |
| Markgraf, 2000 (23) | APACHE II | | 3108 | General | Germany | 0.83 | 11.8 |
| | APACHE III | | | | | 0.85 | 48.4 |
| | SAPS II | | | | | 0.85 | 20.5 |
| Patel, 1999 (24) | APACHE II | 1996–1997 | 302 | General | USA | 0.70 | 14.3 (0.073) |
| | SAPS II | | | | | 0.67 | 22.6 (0.05) |
| | MPM II | | | | | 0.70 | 20.7 (0.05) |
| Vassar, 1999 (25) | APACHE II | 1988–1990 | 2414 | Trauma | USA | 0.85 | 92.6 |
| | APACHE III | | | | | 0.89 | 7.0 ($>$0.05) |
| Metnitz, 1999 (26) | SAPS II | 1997 | 1733 | General | Austria | 0.81 | 91.8 (0.0001) |
| Zimmermann, 1999 (27) | APACHE III | 1993–1996 | 37,668 | General | USA | 0.89 | 48.7 (0.0001) |
| Nouira, 1998 (28) | APACHE II | 1994 | 1325 | General | Tunisia | 0.82 | 26.0 (0.001) |
| | SAPS II | | | | | 0.84 | 73.8 (0.001) |
| | MPM II 0 | | | | | 0.85 | 36.7 (0.001) |
| | MPM II 24 | | | | | 0.88 | 29.6 (0.001) |
| Tan, 2000 (29) | APACHE II | | 1064 | General | Singapore | 0.88 | 44.0 (0.001) |
| | SAPS II | | | | | 0.87 | 49.1 (0.001) |
| Moreno, 1998 (30) | SAPS II | 1993–1994 | 10,027 | General | Europe | 0.82 | 208.4 (0.0001) |
| | MPM II | | | | | 0.79 | 368.2 (0.0001) |
| Moreno, 1997 (31) | APACHE II | | 982 | General | Portugal | 0.82 | ($<$0.001) |
| | SAPS II | | | | | 0.78 | ($<$0.001) |
| Apolone, 1996 (32) | SAPS II | 1994 | 1393 | General | Italy | 0.80 | 71.0 (0.001) |
| Bastos, 1996 (33) | APACHE III | 1990–1991 | 1734 | General | Brazil | 0.82 | 400.3 (0.0001) |

Studies published since 1996 assessing performance of general risk-prognostication systems. Studies including less than 200 patients, done outside the ICU, published in non-English languages or lacking adequate measures of discrimination and calibration, are not shown. Internal validation studies are not included. Discrimination, aROC; calibration, H-L GOF C-statistic.

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, mortality prediction model; ICU, intensive care unit; ROC, receiver operating curves.

North American ICUs were prospectively collected in a database. Seventeen physiologic variables and seven chronic health items were selected following logistic regression analysis. The resulting APACHE III score comprises the three subscores age (0–24 points), acute physiology (0–252 points) and chronic health evaluation (0–23 points).

A second objective of the developers was to refine mortality prediction by correcting for risk in individually defined patient groups. Risk of death in 78 diagnostic categories was included in the final database. As a result of expected declining prognostic performance, the original APACHE III was recalibrated in 1998. The APACHE III equations are owned by a commercial company and are not in the public domain, which has limited its validation and use.

Compared with APACHE II there seems to be a trend for APACHE III to perform slightly better with regard to discrimination, but there is no consistent improvement in calibration when assessing studies published in the last 10 years (7, 38, 23).

### *APACHE IV*

Published in 2006, the APACHE IV system is based on 110,558 patients from 2003 and 2004 in 104 American intensive care or coronary care units (35). The score is made up of the acute physiology score (APS), age and admission circumstances, totalling 142 variables of which 115 are admission diagnoses. In contrast to SAPS III, the APS was found to be the most important factor (65.6% of predictive power), followed by disease group and age. As in earlier APACHE models, the APS was based on the most abnormal values registered during the first 24 h after ICU admission.

APACHE IV also includes a separate scoring system for coronary bypass patients.

### *MPM II*

When it was published in 1985, the MPM was the first general severity model to assess risk of death at ICU admission (36). In 1993, a revision of the MPM was published based on data from 10 North American hospitals in addition to the European/North American Study of Severity Systems (37). A total of 19,124 patients were included in the database. Patients under the age of 18, burn patients, coronary care patients or cardiac surgery patients were excluded. Prediction models for assessment at admission and 24 h were developed originally, but models for assessment at 48 and 72 h were published the following year.

MPM 0 includes a total of 15 variables collected at ICU admission; MPM 24 consists of eight variables collected at 24 h, as well as five variables obtained from the MPM 0.

As the models consist mainly of dichotomous variables, scoring is very simple. Although the development data are similar to the data used in SAPS II, the combined risk-based system differs from the strict physiology-based system of SAPS II.

To develop models based on information available further into the ICU stay, data from 2049 admissions for 48-h models and 975 patients for 72-h models were available. Using these data, the developers developed new logistic regression coefficients keeping the same variables as in the 24-h model, but collecting them at 48 and 72 h.

The strength of the MPM II models lies in their simplicity of scoring and the possibility of sequential assessment of mortality risk throughout the ICU stay. Recent validation studies have found MPM II to perform well (7, 18, 38), but it has not been as extensively validated as the SAPS and APACHE systems.

### *SAPS II*

This version was published in 1994 (3) based on a European/North American database, which included 13,152 patients. It used logistic regression analysis to select variables, weighting and conversion of the score to give the probability of hospital mortality for ICU patients over the age of 18. Coronary care patients, burn patients and cardiac surgery patients were excluded. The developers focused on maintaining a scoring system based mainly on physiological variables. Twelve physiologic variables were included in addition to age, admission type and the presence of metastatic or haematological cancer or AIDS. SAPS II has been extensively studied and validated. There seem to be quite convincing evidence of the ability to maintain good discrimination across different populations, but calibration is often poor as seen with the other general risk prediction models (7, 10).

### *SAPS III*

The SAPS III Outcomes Research Group published their new scoring system in 2005 (39). It was based on a prospective study of 16,784 patients aged 16 years or older from all continents (40). It was

realized that a mainly physiology-based scoring system (SAPS II) had serious shortcomings facing case-mix and lead-time bias. Three subscores, namely patient characteristics before admission (five variables), circumstances of admission (five variables) and acute physiology (10 variables) are summed up to produce the SAPS III score. These subscores provide 50%, 22.5% and 27.5% of the predictive power of the score. The patients' worst physiologic parameters at ICU admission (± 1 h) are recorded. Probability of mortality is calculated using the total SAPS III score in a general or customized equation based on the location of the hospital. Calibration and discrimination in the original data set were shown to vary widely across the world; the best predictive results were shown in North Europe (SMR 0.96) and the worst predictive results in South and Central America (SMR 1.30).

## Disease- and organ-specific prognostic scores

Scores to quantify single-organ failure or a specific disease are often used outside the ICU and knowledge of these scores may be valuable when communicating within the ICU. They have seldom been developed using large prospectively collected data and logistic regression analysis. Their use is often not validated for ICU patients with concomitant organ failure, but they continue to be used to guide treatment and prognostication.

### *GCS*

The GCS was developed as a method for assessing depth and duration of impaired consciousness (41) and is one of the most widespread clinical scores in medicine. Motor response, verbal response and response to pain are noted, producing a total score from 3 to 15. A score of 14–15 indicates mild injury, 9–13 moderate injury and 3–8 severe injury. Its strengths lie in the ease of calculation and reproducibility. The GCS has become a standard method of assessing unconsciousness and coma, but its use outside the setting of trauma and traumatic brain injury is problematic. Its use is not encouraged in patients with other reasons for unconsciousness such as intoxication and epileptic activity. It has no place in assessing the depth of sedation in the ICU. The importance of the GCS in the ICU, with the exception of neuro-intensive care, is probably its inclusion in more complex scoring systems.

### *Ranson score*

The Ranson score was originally developed from a cohort of 100 patients with pancreatitis from a single centre (42). After univariate analysis, 11 variables were found to be associated with morbidity and mortality. Patients with severe pancreatitis are often admitted to the ICU and the Ranson score is still widely used despite the lack of formal validation and several complaints concerning the development of the score. In a study of critically ill patients with severe acute pancreatitis (43), the relatively simple Ranson score performed well, but not better than APACHE III.

### *Child–Pugh (CP)*

Child and Turcotte first proposed a classification system of liver failure in 1964 (44), later modified by Pugh in 1973 (45). The CP classification system grades the patients into three groups. When developing the CP score, empirical methods were used to select variables. Inclusion of two subjective variables (ascites and encephalopathy) may weaken inter-observer reliability and they are often altered by therapy. The CP is in common use and has been extensively validated outside the ICU. Within the ICU, performance has been moderate. Discrimination using receiver operating curves (ROC) was 0.72–0.75 in two studies of patients admitted to the ICU with liver failure (46, 47). These studies show the CP to perform worse than other prediction systems such as APACHE II and Sequential Organ Failure Assessment (SOFA). A modification of the CP, which includes serum creatinine as a variable, has been proposed, but has not yet been validated in an ICU setting.

### *Risk, injury, failure, loss and end-stage kidney (RIFLE) classification*

Acute kidney failure is a frequent and important predictor of mortality in the ICU population (48). To establish a uniform classification of acute kidney injury, the RIFLE classification was proposed by the acute dialysis initiative in 2004 (49). Three severity levels of acute kidney injury (risk, injury and failure) and two outcome classes (loss and end-stage) were proposed. Characterization of acute kidney injury is based on urine output and the elevation of serum creatinine compared with baseline. In a validation study of the risk, injury and failure criteria in the ICU setting, patients in the injury and failure groups were shown to have a significantly

increased risk of mortality even after the correction for non-renal organ failure and other confounding factors (50).

## OD scoring systems

Multiple OD syndrome is the leading cause of death for patients admitted to the ICU (51, 52). The general severity scoring systems, with the exception of MPM 48–72, do not consider OD that develops after the first 24 h of ICU stay. Definitions of multi-organ failure do not take into account the fact that the development and resolution of organ failure is a continuum of alterations and severity rather than a definite event. The characteristics of OD scoring systems included in this review are displayed in Table 3.

### *SOFA*

The Sepsis-Related Organ Failure Score (53), later renamed SOFA, was developed by a conference initiated by the European Society of Intensive Care Medicine in 1994. During development, there was focus on keeping the score objective and independent of therapy, making the collection of variables uncomplicated in most ICUs. The SOFA score uses routinely collected data for the calculation of a score of 0–4 for each organ, the higher number meaning more severe failure. SOFA comprises separate daily scores for respiratory, renal, cardiovascular, CNS, coagulation and hepatic failure. The scores can be used in several ways, as individual scores (each organ), as the sum of scores on one single ICU day or the sum of the worst scores during the ICU stay. There is a slight breach of the intent to keep the score independent of therapy, because evaluation of cardiovascular function is partially based on the choice of vasoactive drugs.

The developers validated the score retrospectively on 1643 septic patients showing increasing mortality with increasing SOFA score for each patient and good distribution of values between patients. In a later prospective, multi-centre study of 1449 patients, the developers confirmed these findings (54). Inter-observer reliability has been assessed by Arts et al. (55) who found good accuracy and precision.

### *Multiple-Organ Dysfunction Score (MODS)*

Published in 1995, the MODS (56) had similar goals as SOFA, in recognition of the need for a classification and prognosis system that could quantify the effect of multiple-organ failure on outcome. A number of test variables based on extensive literature reviews and previous experience were evaluated for their ability to predict mortality in a dose-dependent manner on 336 cases from a database of 692 surgical ICU patients. The variables were then evaluated on the grounds of simplicity and independence of therapy, and calibrated according to risk of mortality. Finally, validation of the score was performed on the remaining 356 cases in the database. The authors chose a scaling of each organ failure from 0 to 4, calibrating risk of death as a linear correlation between 0 ($<5\%$) and 4 ($>50\%$). Cardiovascular, respiratory, haematological, CNS, hepatic and renal failure were included in the score. In contrast to the Logistic Organ Dysfunction System (LODS) and SOFA, the worst scores of the whole ICU stay are recorded. The sum of these scores produce the final MODS score. As in the SOFA system, the simplicity of the scoring process is excellent, but the problems of evaluating circulatory failure prevail. MODS uses a composite variable, the pressure-adjusted heart rate ($HR \times MAP/CVP$) that includes the central venous pressure (CVP), not readily available in all ICU patients.

*Table 3*

Characteristics of organ dysfunction scoring systems.

| System | Year published | Development | Type | Collection of data | Scoring | Scales | Linear scale |
|---|---|---|---|---|---|---|---|
| SOFA | 1996 | Consensus | Organ dysfunction assessment | Worst during last 24 h | Daily | 1–4 in six organs | No |
| MODS | 1995 | Clinical/database | Organ dysfunction assessment | Morning | Daily | 0–4 in six organs | Yes |
| LODS | 1996 | Logistic regression | Risk prediction/organ dysfunction assessment | Worst during last 24 h | First day | 1–5 in six organs | Yes |

SOFA, Sequential Organ Failure Assessment; MODS, Multiple Organ Dysfunction Score; LODS, Logistic Organ Dysfunction System.

*LODS*

The European/North American Study of Severity Systems provided data for the LODS in 1996 (57). It was the first OD score to be developed with the use of multivariate regression analysis of a large database. A total of 13,152 patients over the age of 18 were included, excluding burn patients, coronary care patients and cardiac surgery patients. Twelve variables for six organ systems (neurologic, cardiovascular, renal, pulmonary, haematologic and hepatic) were chosen to define OD. These variables were recorded as the worst value during the first 24 h in the ICU and do not include therapeutic interventions (except mechanical ventilation) or physiologic variables not readily available in all ICU patients. Four severity levels were identified assigning the scores 0, 1, 3 or 5 for each organ system according to the severity of failure.

LODS was developed for the evaluation of OD on the first day of ICU stay and not as a tool for monitoring disease progression, although there are modifications of LODS where scoring is performed on a daily basis.

In addition to the scoring of points, the LODS includes a logistic regression equation that provides an estimate of the severity of OD, using the probability of hospital mortality as a surrogate. The role of mortality prediction in the model separates LODS from merely descriptive models such as SOFA and MODS.

Using this equation, LODS was validated on the 2605 patients in the validation sample and was found to have excellent calibration and very good discrimination.

## Statistical terms

The *discrimination* of the system is the ability of the model to distinguish patients who survive from patients who die. The model should assign higher probabilities of death to non-survivors than to survivors. Traditionally, the terms sensitivity and specificity are used in this respect, but more often ROC are used to give a graphical expression (58). The ROC plot the sensitivity of a test against 1-specificity (Fig. 1). The area under the receiver operating characteristic (aROC) then represents the combined performance. The perfect test will have an aROC of 1, and is 0.5 when it is no better than chance alone. Prediction models should have an aROC larger than 0.7; values higher than 0.80 are considered good while values above 0.9 are excellent.

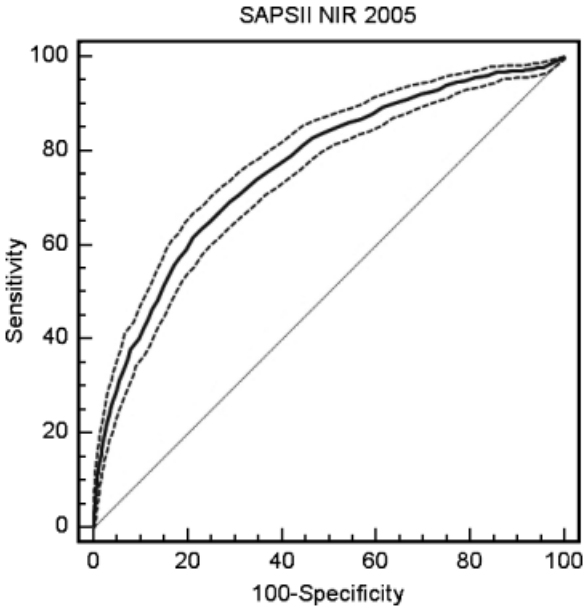


*Fig. 1. Discrimination: data taken from the Norwegian Registry of Intensive Care (2005). Area under the curve is 0.765 (graph also shows 95% CI).*

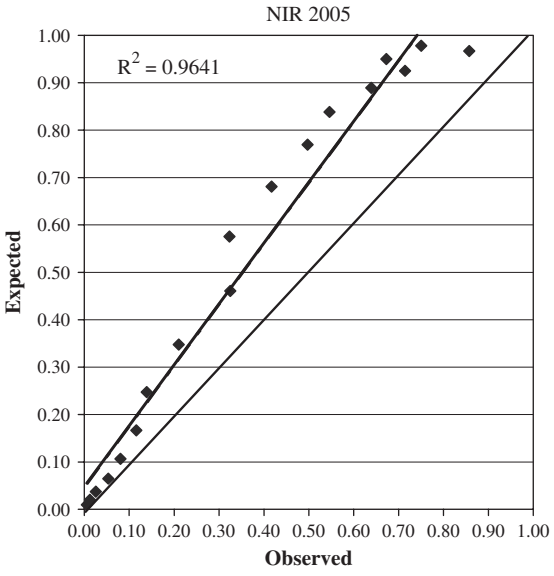


*Fig. 2. Calibration: example taken from the Norwegian Registry of Intensive Care (2005).*

The *calibration* of a system describes its prognostic accuracy at different levels of risk. When all patients are divided into 5% or 10% intervals according to mortality risk, the observed and expected deaths can be compared graphically (Fig. 2). There might be differences in the performance of a model at different risk levels. As seen in several

validation studies on independent populations, discrimination may be good although calibration is poor. Goodness-of-fit statistics examine the difference between the observed frequency and the expected frequency for groups of patients. The statistic can be used to determine whether the model provides a good fit for the data. If the *P*-value is large, then the model is well calibrated and fits the data well; if the *P*-value is small (smaller than $\alpha$), then the model is not well calibrated. One such statistic is the Hosmer–Lemeshow goodness-of-fit statistic (59).

A crude measurement of calibration is the SMR, which is calculated by dividing the observed number of deaths by the expected number. This ratio is often referred to as the O/E ratio. The expected number of deaths is the sum of all the individual mortality risks provided by the severity scoring system in use.

## Discussion

There are several issues regarding the performance of general outcome prediction models that warrant cautious interpretation and use of the results. The most important limiting factor is the lack of individual prognostic ability. These models were not designed for this purpose and should not be used in such a context. The problem of individual prognostication is still a matter of clinical judgement and not an issue of calculation (60).

According to the developers, the scores should be validated for the specific populations in which they are to be used. This is mainly attributed to differences between the population in the development database and the population at hand.

### *Case-mix*

One of the main reasons for using these scores is the difference in the type of patients admitted. The mix of patients admitted to the ICU varies widely within national health systems, as well as between countries with different traditions in intensive care. University and referral hospitals often have a case-mix with sicker patients. It is obvious that comparing outcomes between ICUs on different levels in the referral system without using an MPM will give results without any real value. Simulation studies have shown differences in risk to alter the predictive performance of the systems (61). Although recognized as a key factor in these models, the ability to adjust for differences in case-mix may still be the weakest point when it comes to comparing the quality of ICUs using SMRs (62). Studies of these models in different groups within ICU populations show that they do not necessarily perform well in subgroups, meaning they will be vulnerable to overrepresentation of certain diagnostic categories. Including diagnostic categories in a model will always be a trade-off for the number of patients in each category. Because the number of patients in each disease category is relatively small, the risk-based models have not been shown to have superior performance over a more physiology-based system such as SAPS II. Newer models such as APACHE IV and SAPS III have addressed this problem by increasing the size of the developmental database and utilizing more advanced statistical modelling.

### *Lead-time bias*

In systems providing advanced medical therapy and the presence of emergency physicians even before hospital admission, sedation and the stabilization of respiratory and circulatory insufficiency before and during transport will influence variables such as GCS and blood pressure collected in the ICU. In the same manner, a patient stabilized in the emergency room before ICU admission will have less deviant physiology values resulting in a lower predicted risk of death in physiology-based systems. This influence of pre-ICU care on mortality prediction is often referred to as 'lead-time bias'. Lead-time bias has been shown to be a factor in the calculation of risk prediction. In a study of 76 patients in the United Kingdom, Tunnell et al. (63) showed that the inclusion of variables collected before ICU admission resulted in an increased severity of illness scores, although not reaching statistical significance. SAPS III takes account of both the pre-ICU hospital LOS and some therapeutic actions before ICU admission, in the prognostic model.

### *Boyd and Grounds effect*

Concern regarding the timing of data collection is not limited only to the problem of lead-time bias. In order to obtain as much data as possible, it has been common to gather data over the first 24-h period in the ICU. Poor care in the first 24-h will result in patients with higher predicted probability of death due to the more extreme physiological values allowed to develop when adequate

stabilization is not performed. This effect is referred to as the Boyd and Grounds effect (40,64). Scoring systems such as MPM II 0 and the newly developed SAPS III avoid this by collecting data at the time of admission.

*Inter-observer reliability*

A pre-requisite for comparing outcomes is confidence in uniformity of scoring. Models providing excellent results in internal validation samples may not be clinically useful if the variables chosen are not strictly defined or counterintuitive, resulting in variation of interpretation among physicians. Polderman et al. (65) found an inter-observer variability resulting in a score variability of 15% when using APACHE II. When testing MPM II for inter-observer reliability, the authors found high agreement for risk prediction, but significant variation among the variables in the model (66). Automated data collection systems are becoming widespread in the ICU. While reducing workload and the risk of human error, these systems tend to capture more extreme values and this may also affect the performance of the scoring system (67).

The selection of hospital mortality as an outcome can be criticized since it is difficult to keep track of patients when they are moved between hospitals. SMRs should thus be interpreted with caution when the rate of inter-hospital transfers is high. Comparisons between units should be done between hospitals at the same referral level and confidence intervals should always be supplied in order to give the estimate of normal variation (68).

The use of organ- or disease-specific scoring systems is a tempting method to avoid the problem with case-mix in the ICU. For scores such as Ranson and CP, validation studies are few and inconclusive. There is little evidence that such scores perform better than general severity scores in the critically ill patient. GCS continues to be an important parameter in the evaluation of patients with reduced consciousness and is an important part of most severity scores at admission and throughout the ICU stay.

The available risk prediction systems tend to lose predictive performance in patients with prolonged ICU stay (69). Both treatment and disease state change during the ICU stay, and hence it is of interest to gather data throughout the period. The APS of the APACHE system was shown to have greater prognostic ability on the current day than on the day of admission (70). When OD scores are collected on a daily basis, they can supply several derived scores, such as the total, maximum or delta score that may be of prognostic value (71). Studies comparing the prognostic abilities of three OD scoring systems when using daily scoring have not revealed clinically important differences (72). LODS, the only model with a risk prediction formula, did not perform very well when studied in Austrian ICUs (73). Calibration increased after customization of the formula. This study provides evidence that prediction based on organ failure alone is not a good alternative to the general scoring systems, at least when OD scoring is done only once during the ICU stay.

Although the predictive ability of these systems has been studied extensively, we must not forget that these OD scores were not primarily developed as severity scoring systems, but to describe organ failure. Scoring on a daily basis is a valuable clinical tool in the assessment of severity only when the system used really reflects changes in OD (74). In this regard, all three OD scores seem to perform as intended. When used as a daily routine, it is important that the scoring is not too complicated or time consuming. The ideal system will not incorporate variables that are not measured routinely or include interventions that may vary with local treatment policies.

Evaluation of OD may be of great value in clinical trials in ICU patients. Using mortality alone as an end-point has given disappointing results in many promising trials. To provide outcomes more sensitive to treatment effects, there is a need for scoring systems that are able to describe changes in organ function. For their use in such trials, it is a pre-requisite that organ failure is measured using variables that are objective, reliable and reflect the degree of failure.

## Conclusion

Severity scoring is necessary in both quality control and management of the ICU. Individual risk assessment is not available, but existing severity scores have been shown to give us valuable information when used on ICU cohorts. Recently developed scoring systems promise even better performance, but validation studies are not yet available. As there are several pitfalls related to the interpretation of the numbers supplied by the systems, they should not be used for clinical or

administrative purposes without in-depth knowledge of the science of severity scoring.

## References


1. Apgar V. A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* 1953; **32**: 260–72.
2. Knaus WA, Draper EA, Wagner DP et al. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; **13**: 818–29.
3. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; **270**: 2957–63.
4. Knaus WA, Draper EA, Wagner DP. Toward quality review in intensive care: the APACHE system. *Qual Rev Bull* 1983; **9**: 196–204.
5. Le Gall JR, Loirat P, Alperovitch A et al. A simplified acute physiology score for ICU patients. *Crit Care Med* 1984; **12**: 975–7.
6. Soares M, Salluh JI. Validation of the SAPS 3 admission prognostic model in patients with cancer in need of intensive care. *Intensive Care Med* 2006; **32**: 1839–44.
7. Harrison DA, Brady AR, Parry GJ et al. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med* 2006; **34**: 1378–88.
8. Le Gall JR, Neumann A, Hemery F et al. Mortality prediction using SAPS II: an update for French intensive care units. *Crit Care* 2005; **9**: R645–52.
9. Lima EQ, Dirce MT, Castro I et al. Mortality risk factors and validation of severity scoring systems in critically ill patients with acute renal failure. *Ren Fail* 2005; **27**: 547–56.
10. Aegerter P, Boumendil A, Retbi A et al. SAPS II revisited. *Intensive Care Med* 2005; **31**: 416–23.
11. Reiter A, Mauritz W, Jordan B et al. Improving risk adjustment in critically ill trauma patients: the TRISS–SAPS Score. *J Trauma* 2004; **57**: 375–80.
12. Ihnsook J, Myunghee K, Jungsoon K. Predictive accuracy of severity scoring system: a prospective cohort study using APACHE III in a Korean intensive care unit. *Int J Nurs Stud* 2003; **40**: 219–26.
13. Beck DH, Smith GB, Pappachan JV et al. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Med* 2003; **29**: 249–56.
14. Pettila V, Pettila M, Sarna S et al. Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill. *Crit Care Med* 2002; **30**: 1705–11.
15. Cook DA, Joyce CJ, Barnett RJ et al. Prospective independent validation of APACHE III models in an Australian tertiary adult intensive care unit. *Anaesth Intensive Care* 2002; **30**: 308–15.
16. Timsit JF, Fosse JP, Troche G et al. Accuracy of a composite score using daily SAPS II and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 h. *Intensive Care Med* 2001; **27**: 1012–21.
17. Capuzzo M, Valpondi V, Sgarbi A et al. Validation of severity scoring systems SAPS II and APACHE II in a single-center population. *Intensive Care Med* 2000; **26**: 1779–85.
18. Rue M, Artigas A, Alvarez M et al. Performance of the Mortality Probability Models in assessing severity of illness during the first week in the intensive care unit. *Crit Care Med* 2000; **28**: 2819–24.
19. Metnitz PG, Lang T, Vesely H et al. Ratios of observed to expected mortality are affected by differences in case mix and quality of care. *Intensive Care Med* 2000; **26**: 1466–72.
20. Glance LG, Osler TM, Papadakos P. Effect of mortality rate on the performance of the Acute Physiology and Chronic Health Evaluation II: a simulation study. *Crit Care Med* 2000; **28**: 3424–8.
21. Katsaragakis S, Papadimitropoulos K, Antonakis P et al. Comparison of Acute Physiology and Chronic Health Evaluation II (APACHE II) and Simplified Acute Physiology Score II (SAPS II) scoring systems in a single Greek intensive care unit. *Crit Care Med* 2000; **28**: 426–32.
22. Livingston BM, MacKirdy FN, Howie JC et al. Assessment of the performance of five intensive care scoring models within a large Scottish database. *Crit Care Med* 2000; **28**: 1820–7.
23. Markgraf R, Deutschinoff G, Pientka L et al. Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. *Crit Care Med* 2000; **28**: 26–33.
24. Patel PA, Grant BJ. Application of mortality prediction systems to individual intensive care units. *Intensive Care Med* 1999; **25**: 977–82.
25. Vassar MJ, Lewis Jr FR, Chambers JA et al. Prediction of outcome in intensive care unit trauma patients: a multicenter study of Acute Physiology and Chronic Health Evaluation (APACHE), Trauma and Injury Severity Score (TRISS), and a 24-hour intensive care unit (ICU) point system. *J Trauma* 1999; **47**: 324–9.
26. Metnitz PG, Valentin A, Vesely H et al. Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Simplified Acute Physiology Score. *Intensive Care Med* 1999; **25**: 192–7.
27. Zimmerman JE, Wagner DP, Draper EA et al. Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. *Crit Care Med* 1998; **26**: 1317–26.
28. Nouira S, Belghith M, Elatrous S et al. Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units. *Crit Care Med* 1998; **26**: 852–9.
29. Tan IK. APACHE II and SAPS II are poorly calibrated in a Hong Kong intensive care unit. *Ann Acad Med Singapore* 1998; **27**: 318–22.
30. Moreno R, Miranda DR, Fidler V et al. Evaluation of two outcome prediction models on an independent database. *Crit Care Med* 1998; **26**: 50–61.
31. Moreno R, Morais P. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. *Intensive Care Med* 1997; **23**: 640–4.
32. Apolone G, Bertolini G, D'Amico R et al. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva. *Intensive Care Med* 1996; **22**: 1368–78.
33. Bastos PG, Sun X, Wagner DP et al. Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. *Intensive Care Med* 1996; **22**: 564–70.
34. Knaus WA, Wagner DP, Draper EA et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; **100**: 1619–36.

35. Zimmerman JE, Kramer AA, McNair DS et al. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; **34**: 1297–310.
36. Teres D, Lemeshow S, Avrunin JS et al. Validation of the mortality prediction model for ICU patients. *Crit Care Med* 1987; **15**: 208–13.
37. Lemeshow S, Teres D, Klar J et al. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993; **270**: 2478–86.
38. Livingston BM, MacKirdy FN, Howie JC et al. Assessment of the performance of five intensive care scoring models within a large Scottish database. *Crit Care Med* 2000; **28**: 1820–7.
39. Moreno RP, Metnitz PG, Almeida E et al. SAPS 3 Investigators. 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.
40. Metnitz PG, Moreno RP, Almeida E et al. SAPS 3 Investigators. SAPS 3 – from evaluation of the patient to evaluation of the intensive care unit. Part 1: objectives, methods and cohort description. *Intensive Care Med* 2005; **31**: 1336–44.
41. Jennett B, Bond M. Assessment of outcome after severe brain damage: a practical scale. *Lancet* 1975; **1**: 480.
42. Ranson JHC, Rifkind KM, Roses DF et al. Prognostic signs and the role of operative management in acute pancreatitis. *Surg Gynecol Obstet* 1974; **139**: 69–81.
43. Eachempati SR, Hydo LJ, Barie PS. Severity scoring for prognostication in patients with severe acute pancreatitis: comparative analysis if the Ranson score and the APACHE III score. *Arch Surg* 2002; **137**: 730–6.
44. Child CG, Turcotte JG. Surgery and portal hypertension. *Major Probl Clin Surg* 1964; **1**: 1–85.
45. Pugh RNH, Murray-Lyon M, Dawson JL et al. Transection of the oesophagus for bleeding oesophageal varices. *Br J Surg* 1973; **60**: 646–9.
46. Ho YP, Chen YC, Yang C et al. Outcome prediction for critically ill cirrhotic patients: a comparison of APACHE II and Child–Pugh scoring systems. *J Intensive Care Med* 2004; **19**: 105–10.
47. Cholongitas E, Senzolo M, Patch D et al. Risk factors, sequential organ failure assessment and model for end-stage liver disease scores for predicting short term mortality in cirrhotic patients admitted to intensive care unit. *Aliment Pharmacol Ther* 2006; **33**: 883–93.
48. Ala-Kokko T, Ohtonen P, Laurila J et al. Development of renal failure during the initial 24 h of intensive care unit stay correlates with hospital mortality in trauma patients. *Acta Anaesthesiol Scand* 2006; **50**: 828–32.
49. Bellomo R, Ronco Cthe ADQI Workgroup et al. Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care* 2004; **8**: R204–12.
50. Hoste EA, Clermont G, Kersten A et al. RIFLE criteria for acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis. *Crit Care* 2006; **10**: R73.
51. Carrico CJ, Meakins JL, Marshall JC et al. Multiple-organ failure syndrome. *Arch Surg* 1986; **121**: 196–208.
52. Tran DD, Groeneveld ABJ, van der Meulen J et al. Age, chronic disease, sepsis, organ system failure, and mortality in a medical intensive care unit. *Crit Care Med* 1990; **18**: 474–9.
53. Vincent JL, Moreno R, Takala J et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996; **22**: 707–1.
54. Vincent JL, de Mendonca A, Cantraine F et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problems" of the European Society of Intensive Care Medicine. *Crit Care Med* 1998; **26**: 1793–800.
55. Arts DG, de Keizer NF, Vroom MB et al. Reliability and accuracy of Sequential Organ Failure Assessment (SOFA) scoring. *Crit Care Med* 2005; **33**: 1988–93.
56. Marshall JC, Cook DJ, Christou NV et al. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. *Crit Care Med* 1995; **23**: 1638–52.
57. Le Gall JR, Klar J, Lemeshow S et al. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA* 1996; **276**: 802–10.
58. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.
59. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Commun Statist* 1980; **10**: 1043–69.
60. Flaatten H. Prognostic scoring systems in the ICU. *Acta Anaesthesiol Scand* 2006; **50**: 1175–6.
61. Zhu BP, Lemeshow S, Hosmer DW et al. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study. *Crit Care Med* 1996; **24**: 57–63.
62. Metnitz PG, Lang T, Vesely H et al. Ratios of observed to expected mortality are affected by differences in case mix and quality of care. *Intensive Care Med* 2000; **26**: 1466–72.
63. Tunnell RD, Millar BW, Smith GB. The effect of lead time bias on severity of illness scoring, mortality prediction and standardised mortality ratio in intensive care–a pilot study. *Anaesthesia* 1998; **53**: 1045–53.
64. Boyd O, Grounds M. Can standardized mortality ratio be used to compare quality of intensive care unit performance? *Crit Care Med* 1994; **22**: 1706–9.
65. Polderman KH, Christiaans HM, Wester JP et al. Intra-observer variability in APACHE II scoring. *Intensive Care Med* 2001; **27**: 1550–2.
66. Rue M, Valero C, Quintana S et al. Interobserver variability of the measurement of the mortality probability models (MPM II) in the assessment of severity of illness. *Intensive Care Med* 2000; **26**: 286–91.
67. Bosman RJ, Oudemane van Straaten HM, Zandstra DF. The use of intensive care information systems alters outcome prediction. *Intensive Care Med* 1998; **24**: 953–8.
68. Teres D. The value and limits of severity adjusted mortality for ICU patients. *J Crit Care* 2004; **19**: 257–63.
69. Sicignano A, Carozzi CARCHIDIA et al. The influence of length of stay in ICU on power of discrimination of a multipurpose severity score (SAPS). *Intensive Care Med* 1996; **22**: 1048–51.
70. Wagner DP, Knaus WA, Harrell FE et al. Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis. *Crit Care Med* 1994; **22**: 1359–72.

71. Ulvik A, Wentzel-Larsen T, Flaatten H. Trauma patients in the intensive care unit: short- and long-term survival and predictors of 30-day mortality. *Acta Anaesthesiol Scand* 2007; **51**: 171–7.
72. Pettila V, Pettila M, Sarna S et al. Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill. *Crit Care Med* 2002; **30**: 1705–11.
73. Metnitz PG, Lang T, Valentin A et al. Evaluation of the logistic organ dysfunction system for the assessment of organ dysfunction and mortality in critically ill patients. *Intensive Care Med* 2001; **27**: 992–8.
74. Laurila J, Laurila PA, Saarnio J et al. Organ system dysfunction following open cholecystectomy for acute acalculous cholecystitis in critically ill patients. *Acta Anaesthesiol Scand* 2006; **5**: 173–9.

Address:
*Kristian Strand*
Department of Anaesthesia and Intensive Care
Stavanger University Hospital
PO Box 8100, 4068
Stavanger
Norway
e-mail: stkr@sus.no