Severity of Illness Scoring Systems

Jean-Louis Vincent, MD, PhD, FCCM

Key words: calibration, discrimination

Disclosures: The author has not disclosed any potential conflicts of interest.

Severity of illness is a fairly nonspecific term referring to the seriousness or magnitude of a disease process and, by extension, the risk of complications and death; that is, the more severely ill a patient is, the more likely that patient is to succumb to the disease. Thus, the ability to quantify severity of illness could clearly be of use to clinicians and health care managers, and there has been considerable interest in developing techniques to measure severity of illness. We are all aware of individual patient characteristics that can affect mortality, such as age, certain comorbid diseases, serum lactate levels, and others. Severity of illness scores attempt to combine several such characteristics into more complex and complete scoring systems. Importantly, the term severity of illness will mean different things to different people, which will thus influence how it is measured and interpreted. For example, a general intensivist will be more interested in a global index in individual patients and how it may change in response to treatment; a nephrologist will be more interested in an organ-specific score assessing the severity of renal injury; a hospital manager will be more interested in how the severity of illness will affect resource use and costs; and a research team designing a clinical trial will be most interested in defining a homogeneous study population of patients or assessing the impact of an intervention on disease severity. Different scoring systems have thus been developed to meet these different needs. So-called generic scoring systems assess severity of illness in general ICU patients, whereas organ- or disease-specific scores have been developed for use in selected groups of patients or for
specific situations, such as the Glasgow Coma Scale, Injury Severity Score, and Model for End-Stage Liver Disease. Scoring systems have also been created that link severity of illness to resource allocation, such as the Therapeutic Intervention Scoring System.

Since their introduction some 30 years ago, severity of illness scores have become a part of routine practice in many ICUs. Those most widely used in the general ICU environment, which will be the focus of this chapter, can be broadly divided into 2 groups:

1. **Outcome prediction scores** that measure severity of illness once at ICU admission and use it to calculate a risk of death, for example, the Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), and Mortality Probability Model (MPM)

2. **Organ dysfunction scores** that can quantify the severity of organ dysfunction at admission and during the ICU stay, for example, the multiple organ dysfunction score (MODS) and the sequential organ failure assessment (SOFA)

### OUTCOME PREDICTION SCORES

There are 3 main outcome prediction scores used in adult critical care medicine: APACHE, SAPS, and MPM. All were originally developed 25 to 30 years ago but have been altered and adapted over the years to make them easier to use or better suited to changing patient demographics. Similar scores have been developed for use in children, such as the Pediatric Index of Mortality. Importantly, none of these scores is designed for individual outcome prediction.

Once a prognostic system has been developed, it needs to be validated in populations other than that in which it was developed to determine how well it performs, that is, how accurate it is at predicting prognosis. To this end, 2 objective measures are widely used:

1. **Calibration** is a measure of how well the estimated probability of mortality compares with the actual, observed mortality across the whole spectrum of possible outcomes. Calibration is generally measured using the Hosmer-Lemeshow goodness-of-fit test and reported as the chi-square value: a score with good calibration will have a P value >0.05.

2. **Discrimination** refers to how well the model predicts the correct outcome, that is, how well it discriminates between individuals who will live and those who will die. Discrimination is generally assessed by the area under the receiver operating characteristic (ROC) curve.

As it is impossible for any model to have perfect calibration and perfect discrimination at the same time, when selecting a score based on these 2 measures, it is important to consider the purpose for which the score is being used. For example, if the purpose is to compare quality of care between ICUs, better calibration is more important, whereas to predict outcome for individual patients, good discrimination is preferable.

A score's performance may vary from that reported in the developmental and initial validation studies for several reasons. First, data collection for score calculation needs to be performed in the same manner and following the same rules as were used during the development of the model in question. In addition, specific populations (eg, patients aged <16 years, cardiac surgery patients, patients with burns, patients with a very short length of ICU stay) were often not included in the original development databases and these scores may not be accurate if such patients are included in validation populations. Finally, changes in patient demographics, ICU admission policies, therapeutic measures, and the use of withdrawing and withholding decisions will influence the performance of outcome prediction scores.
Acute Physiology and Chronic Health Evaluation

The APACHE system was the first of the outcome prediction scores to be developed and is the most widely used of these scores, particularly in the United States. Developed by Knaus et al. in 1981 using a nominal group process, APACHE was designed as a “conceptual model of the key elements that influenced a patient’s outcome from severe illness.” In the original APACHE, the score was separated into 2 parts:

1. A score related to the health status of the patient prior to admission, including functional status, medical needs, and presence of chronic diseases over the 6 months prior to admission. Patients were allocated to 1 of 4 categories from A (previous good health) to D (severe restriction of activity because of disease).

2. A score related to the actual physiological status of the patient on admission. Thirty-four physiological variables were assessed over the first 32 hours of admission and awarded a weighted score so that sicker patients had higher scores.

A patient’s APACHE classification, therefore, consisted of a number (the physiology score) and a letter (the chronic health status), for example, 33-B. Original and subsequent validation studies demonstrated that the APACHE score was consistently accurate in predicting ICU and hospital mortality rates in different hospitals and across different countries.

In 1985, Knaus and colleagues published the second version of APACHE, APACHE II (Table 1 on page 878), which is currently the world’s most widely used outcome prediction score for critically ill patients. APACHE II is simpler than the original system in that there are just 12 physiological variables, and the effects of chronic health status are incorporated directly into the model so that there is just 1 overall score, ranging from 0 to 71. The effect of age is also included. The recorded value for each physiological variable in APACHE II is based on the worst value recorded during the first 24 hours of a patient’s admission to the ICU. A score greater than 34 is associated with a mortality rate of more than 80%.

APACHE III was developed in 1991 and further updated in 1999. APACHE III never gained wide support, partly because it is more complex but also because it was marketed commercially so the software was not freely available. APACHE III was remodeled in 2006, using more refined statistical methods to create APACHE IV, which uses the same physiological variables and weights but different predictor variables. In the validation set of more than 100,000 patients, the area under the ROC curve was 0.88 and the Hosmer-Lemeshow chi-square was 16.9 ($P = 0.8$). APACHE IV also includes ICU length of stay prediction equations, which may facilitate the assessment and comparison of ICU efficiency and resource use.

Simplified Acute Physiology Score

The SAPS, developed and validated in France in 1984, used 13 weighted physiological variables and age to provide an indication of the risk of death of ICU patients. In 1993, SAPS II was developed from a much larger international database of 13,152 patients using logistic regression techniques. SAPS II includes 12 physiological variables, age, type of admission (scheduled surgical, unscheduled surgical, or medical), and 3 variables related to underlying disease (AIDS, metastatic cancer, and hematologic malignancy) (Table 1 on page 878). As for APACHE II, for the physiological variables, the worst value over the first 24 hours of ICU admission is taken into account. The calibration (Hosmer-Lemeshow test) of SAPS II on the validation set of 4,783 patients was 3.7 ($P = 0.883$) and the discrimination (area under the ROC curve, or AUC) 0.86.

In 2005, SAPS 3, a completely new outcome prediction model, was created using complex statistical techniques to select and weight variables. A database of 16,784 patients from 303 ICUs in 35 countries was used to develop the score with the aim of increasing representativeness to all countries. SAPS 3 is divided into 3 subscores:

1. Data related to patient characteristics prior to admission (age, comorbidities, use of vasoactive drugs before ICU
Table 1.
Key Components of the Three Main Outcome Prediction Scores

<table>
<thead>
<tr>
<th>APACHE II(^{14})</th>
<th>SAPS II(^{6})</th>
<th>MPM II(^{20})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physiological variables</td>
<td>Physiological variables</td>
<td>Physiological variables</td>
</tr>
<tr>
<td>Temperature</td>
<td>Temperature</td>
<td>Heart rate</td>
</tr>
<tr>
<td>Mean arterial pressure</td>
<td>Systolic blood pressure</td>
<td>Systolic blood pressure</td>
</tr>
<tr>
<td>Heart rate</td>
<td>Heart rate</td>
<td>Coma (GCS 3-5)</td>
</tr>
<tr>
<td>Respiratory rate</td>
<td>Serum bicarbonate level</td>
<td>Acute diagnoses</td>
</tr>
<tr>
<td>Oxygenation (Pao(_2) or a-aDO(_2))</td>
<td>Pao(_2)/FiO(_2)</td>
<td>Acute renal failure</td>
</tr>
<tr>
<td>Arterial pH</td>
<td>Serum bilirubin</td>
<td>Cardiac dysrhythmia</td>
</tr>
<tr>
<td>Serum sodium</td>
<td>Serum sodium</td>
<td>Cerebrovascular incident</td>
</tr>
<tr>
<td>Serum potassium</td>
<td>Serum potassium</td>
<td>Gastrointestinal bleeding</td>
</tr>
<tr>
<td>Serum creatinine</td>
<td>Serum urea or urea nitrogen</td>
<td>Intracranial mass effect</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>Urine output</td>
<td>Chronic diagnoses</td>
</tr>
<tr>
<td>White blood cell count</td>
<td>White blood cell count</td>
<td>Chronic renal failure</td>
</tr>
<tr>
<td>Glasgow Coma Scale score</td>
<td>Glasgow Coma Scale score</td>
<td>Cirrhosis</td>
</tr>
<tr>
<td><strong>Age points</strong></td>
<td><strong>Age</strong></td>
<td><strong>Metastatic cancer</strong></td>
</tr>
<tr>
<td>Chronic health points</td>
<td>Type of admission</td>
<td>CPR prior to admission</td>
</tr>
<tr>
<td>AIDS</td>
<td>Mechanical ventilation</td>
<td></td>
</tr>
<tr>
<td>Hematological malignancy</td>
<td>Nonelective surgery</td>
<td></td>
</tr>
<tr>
<td>Metastatic cancer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{14}\) a-aDO\(_2\), alveolar-arterial oxygen pressure difference; APACHE, Acute Physiology and Chronic Health Evaluation; CPR, cardiopulmonary resuscitation; GCS, Glasgow Coma Scale; MPM, Mortality Probability Model; SAPS, Simplified Acute Physiology Score.

admission, intrahospital location before ICU admission, and length of stay in the hospital before ICU admission)

2. Data related to the circumstances of the admission (reasons for ICU admission, planned/unplanned ICU admission, surgical status at ICU admission, anatomical site of surgery, and presence of infection at ICU admission)

3. Data related to the degree of physiological derangement within 1 hour (in contrast to the 24-hour time window in the SAPS II model) before or after ICU admission (lowest estimated Glasgow Coma Scale score, highest heart rate, lowest systolic blood pressure, highest bilirubin, highest body temperature, highest creatinine, highest leukocyte count, lowest platelet count, lowest pH, and ventilatory support and oxygenation)

There are thus 20 variables in total and the total score can range from 0 to 217. Unlike the APACHE and MPM scores, SAPS 3 includes customized equations for prediction of hospital mortality in 7 geographical regions (Australasia; Central, South America; Central, Western Europe; Eastern Europe; North Europe; Southern Europe, Mediterranean; and North America). The SAPS 3 score has been shown to exhibit good discrimination, calibration, and goodness of fit.\(^{19}\)

**Mortality Probability Model**

The first MPM was developed using data from patients in just 1 ICU and published in 1985.\(^{7}\) Unlike the APACHE and SAPS systems, MPM consisted of 2 models: an admission model that included 7 admission variables, none of which was treatment related, and a 24-hour model that included 7 variables reflecting treatments and the patient’s condition in the ICU, which was designed for patients who
stayed in the ICU for more than 24 hours. A revised model, MPM II, was developed in 1993 using logistic regression techniques on a database of 12,610 ICU patients from 12 countries. MPM II also consists of 2 scores: an admission model, MPM\textsubscript{IV}, which contains 15 variables (Table 1), and a 24-hour model, MPM\textsubscript{III}, which contains 5 of the admission variables and 8 additional variables. Unlike the APACHE and SAPS systems, in which variables are weighted, in MPM II each variable (except age, which is entered as the actual age in years) is designated as present or absent and given a score of 1 or 0 accordingly.

A third version of MPM\textsubscript{IV} was developed in 2007 using a database of 124,885 patients from 135 ICUs in 98 hospitals (all in America). MPM\textsubscript{IV}-III consists of 16 variables, all obtained within 1 hour of ICU admission, of which 3 are physiological parameters. These are used to estimate the probability of mortality at hospital discharge.

Comparing Score Performance

As stated earlier, performance of the 3 scores will differ according to how much the population in which they are being used differs from that of the developmental cohort. As such it is difficult to determine which score, if any, is better globally. In a retrospective study of 11,300 patients from 35 hospitals in California, Kuzniewicz et al\textsuperscript{27} reported that discrimination and calibration were adequate for APACHE IV, MPM\textsubscript{IV}-III, and SAPS II, with discrimination of APACHE IV slightly better than that of the other 2 scores (AUC 0.892 for APACHE IV, 0.873 for SAPS II, and 0.809 for MPM\textsubscript{IV} III, \( P < 0.001 \)). Sakr et al\textsuperscript{31} reported that SAPS II had poor calibration and discrimination in a surgical ICU population that included a large proportion of cardiac surgery patients. Customization of scores to the local patient case-mix is one technique that can help improve the calibration in individual countries or regions. In a retrospective analysis of prospectively collected data from a German surgical ICU, Sakr et al\textsuperscript{24} reported that the discriminative abilities of SAPS 3, APACHE II, and SAPS II were similar (AUC 0.80 for APACHE II, 0.83 for SAPS II, and 0.84 for SAPS 3) and that all 3 scores had poor calibration which improved after customization to the local population. Similarly, Brinkman et al\textsuperscript{27} noted that the calibration of APACHE IV improved after customization to their Dutch population ( Hosmer-Lemeshow C statistic = 823 before and 147 after customization) without altering the good discrimination.

In Austria, Metnitz et al\textsuperscript{32} reported poor calibration of the general SAPS 3 score (\( C = 90.29, P < 0.001 \)) and of the SAPS 3 equation for Central and Western Europe (\( C = 45.61, P < 0.001 \)); after customization, however, calibration was excellent (\( C = 5.61, P = 0.847 \)). Discrimination of all 3 SAPS 3 equations was good (AUC 0.82).\textsuperscript{25}

ORGAN DYSFUNCTION SCORES

As the general management of intensive care patients has improved and, with it, survival rates, there has been an increasing interest in the ability to assess outcomes other than mortality. Many patients who die in the ICU now do so as a result of multiple organ failure, and the presence of organ dysfunction is associated with prolonged ICU stays and increased use of resources. Organ failure scores, by describing the degree of organ dysfunction in individual ICU patients over time, thus provide an important ongoing measure of morbidity. There have been many attempts to develop organ dysfunction scores over the last few decades, but we will focus on 2 of the scores most commonly used in general ICU patients: MODS\textsuperscript{28} and its successor, SOFA.\textsuperscript{29}

Multiple Organ Dysfunction Score

The MODS was developed by Marshall et al\textsuperscript{28} and published in 1995. Seven organ systems (respiratory, cardiovascular, renal, hepatic, hematological, central nervous system, and gastrointestinal) were selected for further consideration based on a literature review of 30 manuscripts that had described organ dysfunction, and potential variables that could be used to describe the organs of each system were selected (Table 2 on page 880). No reliable descriptor of gastrointestinal function could be identified, so this system was not included in the final model. For the cardiovascular system, a composite variable, the pressure-adjusted heart rate (heart rate x central venous pressure/mean blood pressure) was created. In patients without a central line, this variable is scored as
normal. A score of 0 (normal) to 4 (most dysfunction) is given for each organ system, based on the first results for each parameter that day, giving a total maximum score of 24. The score was developed in 336 patients admitted to 1 surgical ICU and validated in 356 patients admitted to the same ICU. It was later validated in many other groups of patients. Not surprisingly, increasing MODS scores correlated well with ICU outcome, with no deaths occurring in patients with MODS scores of 0 and with an ICU mortality rate of 100% for patients with scores of more than 20 in the initial validation study.

**Sequential Organ Failure Assessment**

Following shortly after MODS, a consensus conference was held in Paris in 1994 that led to the development of SOFA (Table 3 on page 881). Six organ systems (respiratory, cardiovascular, renal, hepatic, central nervous, and coagulation systems) were again selected for inclusion.

In SOFA the worst value on each day is recorded, unlike MODS in which the first value of each day is used. The function of each organ system is scored from 0 (normal function) to 4 (most abnormal), giving a possible score of 0 to 24. Table 2 on page 880 shows the key similarities and differences between these 2 systems.

The SOFA participants believed that the cardiovascular component of MODS, the pressure-adjusted heart rate, was too complex for daily practice and not sufficiently accurate because it could be normal even in patients with very severe shock. With no other appropriate independent variable available to evaluate cardiovascular function, it was decided to use a treatment-related variable (dose of vasopressor agents). This approach may be considered as a limitation because it is somewhat subjective, depending on local treatment policies and even individual physician preferences. Furthermore, the score may need to be adapted as new therapies are developed and incorporated.

### Table 2.

**Key Characteristics of the Sequential Organ Failure Assessment (SOFA) and the Multiple Organ Dysfunction Score (MODS)**

<table>
<thead>
<tr>
<th>Score</th>
<th>Authors (Reference)</th>
<th>Year of Publication</th>
<th>Development</th>
<th>Organ Dysfunction Measures</th>
</tr>
</thead>
</table>
| MODS  | Marshall et al²⁸     | 1995                | Systematic review of the literature | First value of each day  
Respiratory: $P_{ac}/Fio_2$  
Coagulation: platelet count  
Hepatic: serum bilirubin  
Cardiovascular: pressure-adjusted heart rate (HR x MAP/CVP)  
Neurological: GCS  
Renal: serum creatinine |
| SOFA  | Vincent et al²⁹     | 1996                | Consensus  | Worst value over 24-h period  
Respiratory: $P_{ac}/Fio_2$  
Coagulation: platelet count  
Hepatic: serum bilirubin  
Cardiovascular: blood pressure, need for vasoactive drugs  
Neurological: GCS  
Renal: serum creatinine and urine output |

CVP, central venous pressure; GCS, Glasgow Coma Scale; HR, heart rate; MAP, mean arterial pressure.
into clinical practice. For example, vasopressin is now being used more widely as an adjunctive vasopressor agent than was the case when SOFA was developed. In recent studies, SOFA has been adjusted in different ways to account for this. In one study, additional points were prospectively assigned to patients who received vasopressin but no norepinephrine (2 points for vasopressin <0.04 U/min and 3 points if >0.04 U/min). In another study, subjects receiving vasopressin as a single vasoactive agent were given a cardiovascular component score of 3, whereas those on vasopressin plus additional agents received a score of 4. Interestingly, although Peres Bota et al reported similar performance in terms of mortality prediction between SOFA and MODS, when using just the cardiovascular components, outcome prediction was better for SOFA.

SOFA was initially validated in a mixed medical–surgical ICU population and has since been used and validated in many different patient groups. Because it is so easy to use and variables are readily and routinely measured in most ICUs, SOFA is now the most widely used of the organ dysfunction scores and has been incorporated as an outcome measure in many clinical trials (eg, Wernerman et al and De Backer et al). Increasing SOFA scores on admission or during the ICU stay (maximum SOFA score) correlate well with mortality. Moreover, changes in SOFA score over time are useful in predicting outcome, with Lopes Ferreira et al reporting that regardless of

### Table 3.
The Sequential Organ Failure Assessment (SOFA) Score

<table>
<thead>
<tr>
<th>SOFA Score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Respiration</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaO/FiO, mm Hg</td>
<td>&gt;400</td>
<td>≤400</td>
<td>≤300</td>
<td>≤200</td>
<td>≤100</td>
</tr>
<tr>
<td>Coagulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platelets ×10³/mm³</td>
<td>&gt;150</td>
<td>≤150</td>
<td>≤100</td>
<td>≤50</td>
<td>≤20</td>
</tr>
<tr>
<td>Liver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilirubin, mg/dL</td>
<td>&lt;1.2</td>
<td>1.2-1.9</td>
<td>2.0-5.9</td>
<td>6.0-11.9</td>
<td>&gt;12.0</td>
</tr>
<tr>
<td>(μmol/L)</td>
<td>(&lt;20)</td>
<td>(20-32)</td>
<td>(33-101)</td>
<td>(102-204)</td>
<td>(&gt;204)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypotension</td>
<td>No hypotension</td>
<td>MAP &lt;70 mm Hg</td>
<td>Dopamine ≤5 or Dobutamine (any dose)²</td>
<td>Dopamine &gt;5 or Epinephrine ≤0.1 or Norepinephrine ≤0.1³</td>
<td>Dopamine &gt;15 or Epinephrine &gt;0.1 or Norepinephrine &gt;0.1³</td>
</tr>
<tr>
<td>Central nervous system</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCS score</td>
<td>15</td>
<td>13-14</td>
<td>10-12</td>
<td>6-9</td>
<td>&lt;6</td>
</tr>
<tr>
<td>Renal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creatinine, mg/dL</td>
<td>&lt;1.2</td>
<td>1.2-1.9</td>
<td>2.0-3.4</td>
<td>3.5-4.9</td>
<td>&gt;5.0</td>
</tr>
<tr>
<td>(μmol/L)</td>
<td>(&lt;110)</td>
<td>(110-170)</td>
<td>(171-299)</td>
<td>(300-440)</td>
<td>(&gt;440)</td>
</tr>
<tr>
<td>Or urine output</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;500 mL/d</td>
<td>&lt;200 mL/d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GCS, Glasgow Coma Scale; MAP, mean arterial pressure.

* Adrenergic agents administered for at least 1 hour (doses given are in μg/kg/min).
the initial SOFA score, an increase in SOFA score over the first 48 hours of the ICU stay was associated with a mortality rate of 53%, no change in score was associated with a mortality rate of 31%, and a decrease in the score was associated with a mortality rate of 23% ($P < 0.01$).

As well as providing a global assessment of total organ dysfunction, the individual organ scores can be considered separately, thus providing a description of patterns of organ dysfunction at any point in time. Increasing SOFA scores for each organ are associated with increased mortality rates.

### SUMMARY

Severity of illness scores are now used to some extent by all ICUs, and with increasingly automated data collection, calculation has never been easier. Such scores can be useful to guide in prognostication, to assess ongoing disease development and organ (dys)function, to compare ICU performance over time and across units, and to compare clinical trial populations and outcomes. The various types of scores have been developed for different purposes and should not be considered as mutually exclusive but rather complementary, as each provides different and additional information. The limitations of each scoring system should be remembered when interpreting the results. All of the main outcome prediction scores have recently been updated to adapt to changing ICU demographics and treatments, and further changes will be needed periodically.

### REFERENCES


