Scoring systems for use in intensive care unit (ICU) patients have been introduced and developed over the last 30 years. They allow an assessment of the severity of disease and provide an estimate of in-hospital mortality. This estimate is achieved by collating routinely measured data specific to a patient (Table 1). A weighting is applied to each variable, and the sum of the weighted individual scores produces the severity score. Various factors have been shown to increase the risk of in-hospital mortality after admission to ICU, including increasing age and severity of acute illness, certain pre-existing medical conditions (e.g. malignancy, immunosuppression, and requirement for renal replacement therapy), and emergency admission to ICU. Before the 1980s, there were no scoring systems applicable to critical care populations which would allow outcomes from different critical care units to be compared. Since then, many scoring systems have been developed, though only a small minority are used. Several of these systems are known simply by their acronym (e.g. APACHE and MODS).

Physiology-based scoring systems are applied to critically ill patients and have a number of advantages over diagnosis-based systems that may be used in other patient groups. Any patient admitted to ICU can have single or multiple organ failure and therefore will not fit a clearly defined diagnostic group. Sometimes, no diagnosis can be made, either on admission or retrospectively. A diagnosis-based scoring system will therefore be inapplicable.

Scoring systems essentially consists of two parts: a severity score, which is a number (generally the higher this is the more severe the condition) and a calculated probability of mortality. Most commonly, this is the risk of in-hospital mortality, though other outcome measures (e.g. survival to 28 days post-hospital discharge) can also be modelled. In order to develop a scoring system, a database incorporating a large amount of patient data from many ICUs, and ideally from many different countries, is required. The applied variables can be grouped into five categories: age, co-morbidities, physiological abnormalities, acute diagnosis, and interventions.

### Classification of scoring systems

There is no agreed classification of the scoring systems that are used in critically ill patients. Scores can be applied either to a single set of data or repeated over time. The available methods include:

1. **Anatomical scoring.** These depend on the anatomical area involved. Anatomical scoring systems are mainly used for trauma patients [e.g. abbreviated injury score (AIS) and injury severity score (ISS)].

2. **Therapeutic weighted scores.** These are based on the assumption that very ill patients require a greater number of interventions and procedures that are more complex than patients who are less ill. Examples include the therapeutic intervention scoring system (TISS).

3. **Organ-specific scoring.** This is similar to therapeutic scoring; the underlying premise is the sicker a patient the more organ systems will be involved, ranging from organ dysfunction to failure [e.g. sepsis-related organ failure assessment (SOFA)].

4. **Physiological assessment.** It is based on the degree of derangement of routinely measured physiological variables [e.g. acute physiology and chronic health evaluation (APACHE) and simplified acute physiology score (SAPS)].

5. **Simple scales.** It is based on clinical judgement (e.g. survive or die).

6. **Disease specific.** [e.g. Ranson’s criteria for acute pancreatitis, subarachnoid haemorrhage assessment using the World Federation of Neurosurgeons score, and liver failure assessment using Child-Pugh or model for end-stage liver disease (MELD) scoring].

### Key points

Scoring systems are widely used in critical care medicine

They allow a quantification of the severity of illness and a probability of in-hospital mortality

Scoring systems must only be used with understanding of their limitations

No scoring system is ideal
The ideal scoring system

The ideal scoring system would have the following characteristics:

1. On the basis of easily/routinely recordable variables
2. Well calibrated
3. A high level of discrimination
4. Applicable to all patient populations
5. Can be used in different countries
6. The ability to predict functional status or quality of life after ICU discharge.

No scoring system currently incorporates all these features.

Types of scoring systems

Most critical care severity scores are calculated from the data obtained on the first day of ICU admission [e.g. the APACHE, the SAPS, and the mortality prediction model (MPM)]. Other scoring systems are repetitive and collect data sequentially throughout the duration of ICU stay or over the first few days (Table 2). Examples of repetitive systems are the SOFA and Multiple Organ Dysfunction Score (MODS). Both first day and sequential scoring systems can be further divided into subjective and objective scores. Subjective scores are produced by taking variables that have been agreed by a panel of experts, and then applying a numerical weighting to each variable to produce a subjective score. The weighting is usually determined by consensus opinion. Objective scores are developed from a large database of clinical data taken from many ICUs. A computer-based multipurpose probability model is then used to determine which variables to use and the weighting to be applied to each variable.

Assessment of scoring systems

Once a scoring system has been produced, its performance should be assessed and validated. This process refers to the ability of the score to predict mortality, and must be carried out on a different population to that used to assemble the score.2 This can occur by randomly splitting the original population into two groups: the first to produce the score and the other to validate the model, or by using a completely separate population.1 Model calibration and discrimination are then assessed.

Model calibration

Calibration assesses the degree of correspondence between the estimated probability of mortality and that actually observed. This can be tested using a goodness of fit test, most commonly the Hosmer–Lemeshow C statistic. Over the range of probabilities, the expected and observed mortality are compared and a P-value derived. Calibration is considered to be good if the predicted mortality is close to the observed mortality.3

If a scoring model predicts that a patient has a probability of in-hospital mortality of 0.25, it means that, in a sample population of 100 patients, 25 would be expected to die and 75 patients would survive. When the number of deaths in the actual population is near to that predicted by the scoring system, the model is considered well calibrated.

Model discrimination

Model discrimination reviews the ability of the scoring model to discriminate between patients who die from those who survive, based on the predicted mortalities. Methods include calculation of the area under the receiver operating characteristic (ROC) curve or by using a classification matrix. The two most important parts of the classification matrix are the specificity and sensitivity. In MPMs, these are not absolute levels, and a huge grey area exists between those who die and those who survive. Therefore, a number of classification matrices are constructed with sensitivity and specificity values across the range.3

A pair of sensitivity–specificity values produces the ROC curve across the range of mortality prediction scores. The area under the resultant curve (AUC) represents the number of patients who died. The curve is analysed using complex computerized statistical processes to assess the discrimination.3 Clearly, if this AUC is around 0.50, the performance of the scoring system is no better than a coin toss. Typically, model developers require an AUC of the ROC curve to be >0.70.1
Issues related to model assessment

Despite the methods of validating a scoring system, there remain a number of issues related to the design and assessment of the models that could affect their reliability. The populations on which the model is developed and validated are split randomly or chosen at random, thereby reducing any bias. However, given the significant length of time it can take to obtain the data required to develop and validate a scoring system, it is possible that many factors can have changed during this period. Thus, if poor goodness-of-fit is obtained during validation, it may be difficult to state for certain if this due to sample or model problems. Sample size also has a major influence on the validity of the scoring system: too small a population lends towards the risk of the score being unable to distinguish and assess reliably between different patient groups. Clearly, a large population is required, but just how large is not known. In addition, a scoring system must be modelled and validated against a real cohort of ICU patients, but it is difficult to be sure how representative this cohort is of the wider population of critically ill patients. Indeed, does a representative ICU population exist? In practice, these questions are hard to determine and so we assume that by using a large cohort to produce and validate a particular model it is more likely to reflect a typical ICU patient population.

Commonly used scoring systems

Acute physiology and chronic health evaluation

The Acute Physiology and Chronic Health Evaluation (APACHE) score is probably the best-known and most widely used score. The original APACHE score was first used in 1981 and scores for three patient factors that influence acute illness outcome (pre-existing disease, patient reserve, and severity of acute illness). These included 34 individual variables, a chronic health evaluation, and the two combined to produce the severity score.

The APACHE II scoring system was released in 1985 and incorporated a number of changes from the original APACHE. These included a reduction in the number of variables to 12 by eliminating infrequently measured variables such as lactate and osmolality. The weighting of other variables were altered; most notably, the weightings for Glasgow Coma Score and acute renal failure were increased. In addition, weightings were added for end-organ dysfunction and points given for emergency or non-operative admissions. Each variable is weighted from 0 to 4, with higher scores denoting an increasing deviation from normal. The APACHE II is measured during the first 24 h of ICU admission; the maximum score is 71. A score of 25 represents a predicted mortality of 50% and a score of over 35 represents a predicted mortality of 80%. The APACHE II severity score has shown a good calibration and discriminatory value across a range of disease processes, and remains the most commonly used international severity scoring system worldwide.

APACHE III, released in 1991, was developed with the objectives of improved statistical power, ability to predict individual patient outcome, and identify the factors in ICU care that influence outcome variations. The weightings are far more complex than the two previous scoring systems, but notably are the addition of HIV and haematological malignancy (as well as disseminated malignancy and liver disease) to the chronic health points. The performance of the APACHE III severity score is slightly better than that of APACHE II, but the former has not achieved widespread acceptance perhaps because the statistical analysis used to score it is under copyright control.

Simplified acute physiology score

The SAPS was first released in 1984 as an alternative to APACHE scoring. The original score is obtained in the first 24 h of ICU admission by assessment of 14 physiological variables and their degree of deviation from normal, but no input of pre-existing disease was included. It has been superseded by the SAPS II and SAPS III, both of which assess the 12 physiological variables in the first 24 h of ICU admission and include weightings for pre-admission health status and age.

Mortality prediction model

The MPM is based on two models and allows a probability of in-hospital death to be calculated, rather than a severity score that needs to be converted. Assessment of chronic health status, acute diagnosis, and weightings for physiological variables allows a prediction of death to be made. Data at admission and 24 h after ICU admission are included. The newer MPM II is based on multiple regression analysis from a large population and includes weightings for physiology, acute and chronic illness, age, and therapeutic interventions. Sequential calculations can be made at 0, 24, 48, and 72 h from ICU admission.

Sepsis-related organ failure assessment

The SOFA was produced by a group from the European Society of Intensive Care Medicine to describe the degree of organ dysfunction associated with sepsis. However, it has since been validated to describe the degree of organ dysfunction in patient groups with organ dysfunctions not due to sepsis. Six organ systems—respiratory, cardiovascular, central nervous systems, renal, coagulation, and liver—are weighted (each 1–4) to give a final score [6–24 (maximum)].

Multiple organ dysfunction score

The MODS scores six organ systems: respiratory ($PO_2$/Fi$O_2$ ratio in arterial blood); renal (measurement of serum creatinine); hepatic (serum bilirubin concentration); cardiovascular (pressure-adjusted heart rate); haematological (platelet count); and central nervous system (Glasgow Coma Score) with weighted scores (0–4) awarded for increasing abnormality of each organ systems. Scoring is performed on a daily basis and so allows a day-by-day prediction for patients.
Comparison of performance

With so many scoring systems available, it would be ideal to be able to assess each system side by side both at the developmental stage and with the validation samples. Unfortunately, these data are not available and few comparisons between scores are available. The severity scores all show very good discriminatory values with AUC under the ROC curve ranging between 0.80 and 0.90, achieving good to excellent calibration assessments. The organ failure scores also show discriminatory values in the region of 0.90 and achieve good calibration.

Uses and abuses of scoring systems

Severity scoring systems allow generation of a score that reflects the severity of the condition resulting in ICU admission. The scores allow the factors that influence outcome and that differ between patients to be taken into account and can be standardized to allow comparison between patients. Inferences can be made regarding patient response to therapies and interventions if sequential scoring systems (e.g. SOFA, MODS) are monitored for several days after ICU admission. In addition, the APACHE III score has been shown to be of use for individual patients in triage.

Another important use for scoring systems in ICU is an audit tool. They can help individual ICUs to compare their performance over time. However, this type of comparison should be interpreted carefully and, in particular, comparisons between different units are susceptible to misinterpretation. If estimated probabilities of hospital death against actual mortality were calculated for a number of different ICUs, there would be a spread of results ranging from those with mortality below that expected to those above that expected. This does not mean that one ICU is performing better or worse than another because several factors other than simple clinical skills are involved. These include case mix (an ICU admitting less sick patients would be expected to have a lower mortality), access to technical and therapeutic modalities, and administrative and staffing variations (e.g. nursing staff per bed). Only once these factors have been taken into account can meaningful comparisons be made. The only way of accounting for these differences is to use the standardized mortality ratio. This method assumes that the scoring systems are accurate and are always correct at predicting mortality.

Apart from one or two exceptions (notably the Glasgow Coma Score, which is not a critical care scoring system), a higher score denotes more severe illness. However, certain disease states or conditions may generate very high severity scores, even though they do not generally result in high mortality. These are usually conditions associated with a high degree of physiological derangement but which are either self-limiting or can be managed to return towards normal relatively quickly. Classically, this arises with diabetic ketoacidosis but might also occur in patients admitted to ICU after surgery while still under the effects of general anaesthesia. In both cases, a high severity score would be obtained which might be potentially misleading.

A further potential problem is that scoring systems do not have a linear scale: a score of 20 does not mean a patient is twice as sick as another patient with a score of 10, and likewise does not have twice the risk of dying.

Severity scoring systems are also often used to stratify critically ill patients for possible inclusion in clinical trials. It is important to realize that the scores have been validated for a set time period (most commonly the first 24 h of ICU admission) or, in the case of repetitive scores, at set times. If the scoring system is used outside of these pre-validated limits, then reliability cannot be assumed, and some sort of stratification may be required before inferences can be made.

It is apparent that the use of physiological variables in scoring systems may give rise to potential bias and lead to the calculation of an inaccurate severity score. Values of the variables included can alter spontaneously or as a result of resuscitative therapy before admission of the patient to the ICU (occurring before transfer from a ward, emergency department or other ICU, or out-of-hospital care performed by ambulance personnel). This is termed lead time bias and can render the scoring system inaccurate. As a result, scoring performed on ICU admission can suggest a better severity and predicted mortality than is actually the case. In one study, six variables accounted for the most lead time bias: heart rate, blood pressure, respiratory rate, oxygenation, pH, and blood glucose.

However, the most important potential limitation of scoring systems is the inappropriate interpretation of the score. Clinicians must be aware that the probability of in-hospital mortality based on a particular score relates to a similar group of patients and not to an individual. This is important to understand before attempting to use scoring systems in clinical practice. So, although it can be useful to know the predicted mortality of a group of patients with a similar score, we cannot be sure which patients will die and which will survive. Consequently, scoring systems should not be used to make predictions for individual cases. Conversely, scoring systems can appropriately be used to assist the clinical decision-making as they do allow an objective assessment of a patient’s severity of illness, and therefore reflect the likelihood of mortality in a similar cohort of patients. Overall, they should be considered as a facet to assist the clinician.

Finally, all the scoring systems assess the severity of illness and the likelihood of in-hospital mortality. Of arguably more importance is the ability to predict outcome or morbidity after discharge from ICU; at present, no such scoring system exists. Such a system would provide potential invaluable information, particularly if it were combined with the currently available ICU scoring systems.

References


Please see multiple choice questions 27–31