# CHAPTER II

# LITERATURE REVIEW

Over the last decade, significant effort has been directed at improving quality of medical care. Quality of care has been defined by the Institute of Medicine as the "degree to which health services for individuals and populations increase the likelihood of desired health outcomes" [11]. Quality of care has become a central issue in the arena of healthcare policy debate. In part, this is caused by the concern that the increased emphasis on cost cutting will compromise quality. In addition, there is increasing recognition of the wide variation in healthcare practices and, more importantly, of the potential effect of this variance of healthcare delivery on outcomes [12]. Tools to measure quality are becoming widespread and focus either on the outcomes of care or the processes of care. For intensive care units (ICUs), the main focus on outcome measurement has been on the development of severity scoring systems for mortality prediction. These scoring systems have been used to perform risk adjustment on ICU mortality rates to compare outcomes across ICUs. Using these risk-adjustment tools, it becomes feasible to identify high-performance centers and to identify the processes of care that result in superior outcomes. Therein lays the opportunity to uncover and institute "best practices" to improve quality.

In this chapter, the available severity scoring systems for mortality prediction are presented, with emphasis on the most used today: APACHE II and III, SAPS II and MPM II. The steps needed for the development of such systems are outlined: selection of the patients, choice of outcome and predictor variables, data collection and assembly of the systems, the choice of methods and their application. The importance of the validation of the scoring systems is stressed. The recommendations for validation of ICU scoring systems and what to do when a scoring system calibrates poorly on an external data set are presented.

## Available severity scoring systems

The development of severity scoring systems started more than twenty years ago. The first general severity scoring system was the Acute Physiology and Chronic Health

Evaluation (APACHE) [13]. Developed in 1981 at the George Washington University Medical Center, the APACHE scoring system demonstrated to provide accurate and reliable measures of severity of illness in critically ill patients [14 - 16].

Two years later, Le Gall et al. published a simplified version of this scoring system, known as the Simplified Acute Physiology Score (SAPS) [17]. Another simplification of the original APACHE, the APACHE II, was published in 1985 by the developers of the original scoring system [18]. This scoring system introduced the possibility of mortality prediction, requiring for that purpose the selection of a primary reason for ICU admission from a list of 50 diagnoses. Additional contributions to severity in the intensive care setting were the Mortality Probability Model (MPM)[19], developed using multiple logistic regression techniques to choose and weigh the variables rather than by a consensus from a panel of experts.

The more recent developments in severity scoring systems comprise the third version of APACHE (APACHE III) and the second versions of the SAPS (SAPS II) and MPM (MPM II). All were built using logistic regression techniques to choose and weigh the variables and are able to provide predictions of hospital mortality. They have been shown to perform better than their previous versions. [20]

APACHE II

APACHE II was developed on the basis of data collected from 13 North American hospitals from 1979 to 1982[18] . A panel of experts using clinical judgment and documented physiologic relationships did the choice of variables and their weights. The scoring system uses the worst value during the first 24 hours in the ICU for 12 physiologic variables (weighted from 0 to 4 points), age, surgical status (emergency surgical, elective surgical or non-surgical) and previous health status. The selection of a primary reason for ICU admission is necessary to be included in a logistic regression scoring system that transforms scores in probabilities of mortality. The APACHE II score ranges from 0 to 71 points: up to 60 for physiologic variables, up to 6 for age and up to 5 for chronic health status. This system soon became the most worldwide utilized severity scoring system, and it is still in use today in many ICUs.

APACHE III

The APACHE III was developed on the basis of a large North American database of intensive care patients in 1988-89 [21]. The selection of the 40 participating hospitals was done in order to be representative of the continental North American hospitals with more than 200 acute-care beds. Excluded from its development were patients with a length of stay in the ICU lower than 4 hours, patients younger than 16 years, and patients with burn injuries, with chest pain admitted for ruling out myocardial infarction and those in the immediate postoperative phase of coronary artery bypass surgery.

This scoring system comprises APACHE III score based on acute physiological variables, age and chronic health, and the APACHE III predictive equation. The equation uses the APACHE III score and reference data on major disease-categories, the surgical status and the site of treatment immediately before ICU admission for estimating the risk of hospital mortality of ICU patients. The APACHE III score ranges from 0 to 299 points, including up to 252 points for the 18 physiology variables, up to 24 points for age and up to 23 points for chronic health. All physiologic variables are assessed as the worst values during the first 24 hours in the ICU. This strategy was chosen because it results in greater data availability (less proportion of missing values) and explanatory power [21].

The conversion of the score to a probability of mortality is accomplished with individual logistic regression equations for each of the 78 specific diagnostic categories and the nine patient origins. It is now a proprietary system; the equations are not in the public domain and must be purchased from APACHE Medical Systems, Washington, DC. This has limited its use, especially outside the United States.

SAPS II

SAPS II was described in 1993 by Le Gall et al. [3], based on a European/North American multi-center study. It was developed and validated in a large cohort of patients from 110 h ospitals in Europe and 27 hospitals in North America. Excluded from its development were patients aged less than 18 years, and burn patients, coronary care patients and cardiac surgery patients.

This scoring system includes 17 variables to compute a score: 12 physiological variables, age, type of admission (medical and scheduled/unscheduled surgery), and three underlying diagnoses (acquired immunodeficiency syndrome, metastasis cancer

and hematological malignancy). SAPS II score ranges from 0 to 163 points (up to 116 for physiology, up to 17 for age and up to 30 for underlying diagnosis).

SAPS II uses the worst recorded values for the physiological variables during the first 24 hours in the ICU and do not need the selection of a single diagnosis for the computation of the probability of death in the hospital.

MPM II

MPM II was described in 1993 by Lemeshow et al. [4] based on the same database as SAPS II plus data collected in six ICUs of four teaching hospitals in the United States of America. The exclusion criteria used for its development were the same as those used for SAPS II. In this scoring system the final result is only expressed as a probability of mortality (and not as a score). The actual MPM version incorporates scoring systems to predict mortality at admission ($MPM_0$ II) and at 24 hours after admission to the ICU ($MPM_{24}$ II). These were subsequently supplemented by scoring systems for the 48 ($MPM_{48}$ II) and the 72 hours ($MPM_{72}$ II) in the ICU, developed on the basis of a smaller database [22].

$MPM_0$ II contains 15 variables: age, three physiologic variables (coma or deep stupor, heart rate and systolic blood pressure), three chronic diseases (chronic renal failure, cirrhosis, and metastasis cancer), five acute diagnoses (acute renal failure, cardiac dysrhythmia, cerebrovascular accident, gastrointestinal bleeding, and intracranial mass effect), type of admission (medical or surgical unscheduled), mechanical ventilation and cardiopulmonary resuscitation prior to admission. All variables are evaluated on the basis of data collected within one hour at admission to the ICU.

$MPM_{24}$ II is based on 13 variables: age, six physiologic variables (coma or deep stupor, creatinine, confirmed infection, hypoxemia, prothrombin time and urinary output), three variables ascertained at admission (cirrhosis, intracranial mass effect, and metastasis cancer), and type of admission (medical or surgical unscheduled), mechanical ventilation and use of vasoactive drugs. The physiologic variables are evaluated based on worst values during the first 24 hours in the ICU.

$MPM_{48}$ II and $MPM_{72}$ II use the same variables as $MPM_{24}$ II, with different weights for the computation of the predicted risk of death. Both are based on the worst values during the preceding 24 hours.

The available severity scoring systems is summarized in table 2.1.

Table 2.1: The available severity scoring systems

| Characteristics | APACHE | SAPS | APACHE II | MPM [a] | APACHE III | SAPS II | MPM II [b] | APACHE IV |
|---|---|---|---|---|---|---|---|---|
| Year | 1981 | 1984 | 1985 | 1985 | 1991 | 1993 | 1993 | 2006 |
| Participating countries | 1 | 1 | 1 | 1 | 1 | 12 | 12 | 45 |
| Participating ICUs | 2 | 8 | 13 | 1 | 40 | 139 | 140 | 104 |
| Number of patients | 705 | 679 | 5815 | 2783 | 17440 | 12997 | 19124 | 110588 |
| Selection of variables and their weights | Panel of experts | Panel of experts | Panel of experts | Multiple logistic regression | Multiple logistic regression | Multiple logistic regression | Multiple logistic regression | Multiple logistic regression |
| Variables | | | | | | | | |
| Age | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Patient origin | No | No | No | No | Yes | No | No | Yes |
| Surgical status | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Chronic health status | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Physiology | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Acute diagnosis | No | No | Yes [c] | No | Yes [d] | No | Yes | Yes |
| Number of variables | 34 | 14 | 17 | 11 | 26 | 17 | 15 [e] | 29 |
| Score | Yes | Yes | Yes | No | Yes | Yes | No | Yes |
| Equation to predict mortality | No | No | Yes | Yes | Yes | Yes | Yes | Yes |

[a]: These scoring systems were based on previous research from the same developers.

[b]: The numbers shown are for the admission component of the scoring system ($MPM_0$ II). The $MPM_{24}$ II, was developed using data from 15925 patients in the same centers participating in the development of the admission scoring system.

[c]: chosen from a list of 50 diagnoses.

[d]: chosen from a list of 78 diagnoses.

[e]: the $MPM_{24}$ II scoring system uses only 13 variables.

Development of the severity scoring systems

All the severity scoring systems aim at predicting outcome on the basis of a given set of variables: they estimate what should be the outcome of a given patient, with a certain clinical condition (defined by the values of the given set of variables) as if this patient was treated in a hypothetical reference ICU used to develop the scoring system. Several steps are necessary for the development of these scoring systems (Table 2.2).

Table 2: Steps in the development of a scoring system

| Steps in the development of a scoring system |
|---|
| 1. Patient selection |
| 2. Outcome selection |
| 3. Predictor variables selection and data collection |
| 4. Assembly of the scoring system |
| 5. Validation of the scoring system |
| 6. Scoring system updates and modifications |

Patient selection

None of the severity scoring systems presently available is applicable to all ICU patients. Burn patients, patients admitted with acute coronary disease (or to rule out myocardial infarction), young patients (less than 16 or 18 years of age), patients in the post-operative of coronary artery bypass surgery or with a very short length of stay in the ICU were explicitly excluded from the development of the scoring systems. This limitation becomes important when studying specialized ICUs with particular patient demographics. In other words, when applying a specific scoring system to an ICU, attention should be given to the number and type of excluded patients. Further measurements and actions should be undertaken only if the evaluation is based on a representative number of the admitted patients.

Outcome selection

Outcome can be seen as one or more events in the course of a disease process, such as morbidity, mortality, time to recovery from disease, or quality of life. Until now,

all general sever ity sc oring systems in intensive care focus exclusively on hospital mortality. This measure is considered the gold standard since it is easy to define and measure, and represents a clinically very relevant endpoint.

**Predictor variables selection and data collection**

The next step in the development of a severity scoring system is the choice of a preliminary list of candidate variables. These, usually selected by experts in the field, can include demographic, clinical and laboratory variables. Each should be relevant for the outcome of interest, supported by the available literature and with documentation explaining how they relate to the outcome, precisely defined, routinely available, and occurring frequently enough to be meaningful. In addition, they should not be confounded with the outcome of interest unless present before ICU admission [23].

The collection of data can be done retrospectively or prospectively. The second approach is preferred since it allows the collection of more accurate and complete data.

**Assembly of the scoring system**

The following step is data analysis and data reduction. The initial list of candidate variables, submitted to a combination of logistic regression techniques [24], smoothing methods, and clinical judgment is reduced to a smaller number of variables to be included in the scoring system. This reduction is based on the general scientific principle of parsimony: theories with simpler explanations are considered more plausible than more complex ones. It ensures that the scoring system to be developed will have a higher precision, and will lead to more interpretable scoring systems. Attention should be given at this stage to multi-colinearity and interactions.

After the final set of variables is chosen and their weights evaluated, the variables are usually combined in a single score. In each case, the score results from the sum of the weights assigned to the chosen variables, according to the magnitude of change from the accepted normal values. This approach was chosen by the developers of APACHE III [21] and SAPS II [3] but not in the case of MPM II [4].

The next step is to relate the aggregated score (or the chosen set of variables, as in the case of MPM II) to the outcome of interest. All presently available general severity scoring systems use multiple logistic regression analysis for that purpose. In

this technique, the dependent or outcome variable *y* is related to a set of independent or predictive variables by the equation:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots \beta_k x_k$$

where $\beta_0$ is the intercept of the scoring system, $x_1$ to $x_k$ represent the predictor variables and $\beta_1$ to $\beta_k$ are the estimated regression coefficients. Then, a logistic transformation is applied, with the probability of death being given by:

$$\Pr = \frac{e^{\text{logit}}}{\left(1 + e^{\text{logit}}\right)}$$

where Pr represents the probability of death and logit is *Y* as described before.

Validation of the scoring system

All statistical scoring systems developed for prediction need validation, since it is necessary to know their predictive validity.

This validation process can be done in three different ways: cross-validation, external validation and temporal validation. In cross-validation, data are randomly divided in two portions, one for scoring system development and the other for scoring system validation. In external validation, data developed in one population are validated in an external, independent population. In temporal validation, a scoring system developed at a specific point in time is later validated in the same setting at some future date.

All the general severity scoring systems in use today have been subjected to cross-validation. This step is necessary but not sufficient, since it does not assure *per se* the adequacy of the scoring system when subjected to external validation. When the developers of the original scoring systems utilize randomization to split the original database in two groups, development and validation samples, one may expect that all the variables (and non measured case-mix factors) will be randomly distributed in the two sub-groups. Consequently, both subgroups are expected to represent equal samples from the same underlying distribution and can not be considered true independent samples. The scoring systems analyzed in this way are therefore expected

to perform better on the validation sample than in an independent population. Independent validation in different populations is needed before general utilization, since variations in case-mix, local policies, quality of care and quality of data collection have been shown to affect the performance of the equations used to predict mortality [7, 25-28].

The main question to be answered at this stage is the adequacy of outcome predictions when compared to the actual outcomes. Three major issues must be evaluated [29]. The first is *calibration*, or how well the scoring system predictions compare with the observed outcomes. The second is the scoring system *discrimination*, which evaluates how well the scoring system can distinguish between observations with a positive or a negative outcome. A third source of scoring system deviation can be the existence in the test set of *subsets of observations* in which the scoring system does not perform well. Opposed to the first two, where a lot of research has been done and consensual techniques have emerged, there is little in the literature on how to identify these observations or what to do when the fit is unsatisfactory [29]. The evaluation of calibration and discrimination in the overall population has been named *overall goodness of fit*. The evaluation of the appropriateness of the predictions across sub-groups has been termed *uniformity of fit* [29].

The evaluation of the overall goodness of fit comprises the evaluation of calibration and discrimination in the population under analysis. Calibration evaluates the degree of correspondence between the estimated probabilities of mortality and the observed mortality in the sample under analysis. Three methods are usually employed in this analysis: overall observed/expected (O/E) mortality ratios, Hosmer-Lemeshow goodness-of-fit tests [24] and calibration curves.

*Overall O/E mortality ratios* are computed dividing the overall observed mortality rate (*i.e* .the actual number of deaths) by the predicted number of deaths (resulting from the sum of the individual probabilities of mortality assigned by the scoring system); additional computations can be made to estimate the confidence interval for the ratio [30]. In a perfectly calibrated scoring system this value should be one.

Overall O/E mortality ratios in the ICU setting are known as standardized mortality ratio. This is an overall summary statistic providing information about how the

overall mortality rate agrees with the mortality prediction for the sample. It compares the observed number of deaths to the predicted number of deaths. Large departures will indicate failure of the scoring systems to predict the probability of patient death in that context. In other words, it is the ratio of observed deaths to predicted deaths or observed mortality rate to predicted mortality rate, and can be calculated as:

$$SMR = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \pi_i}$$

There are n patients indexed by i, $\pi$ is the estimate of the probability of death provided by the scoring system. For a patient who dies, the outcome, $Y_i$ =1 or $Y_i$ = 0 if the patient survives.

The estimate of the 95% CI of the SMR is:

$$SMR \pm 1.96 \frac{\sqrt{\sum_{i=1}^{n} \hat{\pi}_i \cdot (1 - \hat{\pi}_i)}}{\sum_{i=1}^{n} \hat{\pi}_i}$$

*Hosmer-Lemeshow goodness-of-fit tests* are two chi-square statistics proposed for the formal evaluation of the calibration of severity scoring systems [24]. In Hosmer-Lemeshow H test, patients are divided in 10 strata, according to their predicted probabilities: 0.0 to 0.1, 0.1 to 0.2 ... 0.9 to 1.0. Then, a chi-square statistic is used to compare the actual and the expected number of deaths and the actual and expected number of survivors in each of the strata. The used formula is defined as:

$$G_{HL}^2 = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi_8^2$$

where

$n_j$ = Number of observations in the $j^{\text{th}}$ group

$O_j = \sum_i y_{ij}$ = Observed number of cases in the $j^{\text{th}}$ group

$E_j = \sum_i \hat{p}_{ij}$ = Expected number of cases in the $j^{\text{th}}$ group

The resulting statistic is then compared with a chi-square table with 8 degrees of freedom in order to evaluate if the resulting discrepancies can be explained only by sampling variance. Large differences suggest that the scoring system is not well calibrated. Hosmer-Lemeshow test C is similar, with the 10 strata being formed based on deciles of the predicted mortalities. The same authors demonstrated that the grouping method at the basis of the C statistic is preferable to the one based on fixed cut points, especially when many of the estimated probabilities are small. These tests are currently regarded as an obligatory part of calibration evaluation.

*Calibration curves* are also used for reporting information on the calibration of a scoring system. This type of charts compares the observed mortality rate with the one expected by the scoring system. They are not a formal statistical test of the adequacy of the scoring system, being used only for information purposes. However, the publication of calibration curves has been recommended by a recent consensus conference as part of the standard assessment of the validation of a scoring system.

*Discrimination* evaluates how well the scoring system can distinguish between observations with a positive or a negative outcome. This index measures the probability that for two randomly chosen patients, the one with the higher probabilistic prediction has the outcome of interest. It has been shown that this index relates directly to the area under a receiver operating characteristic (ROC) curve.

The concept of the *area under the ROC curve* derives from psychophysics and has been applied in signal processing. In a ROC curve, a series of two-by-two tables is constructed with cut points that vary from the lowest possible value to the highest possible value of the score. For each table, the true positive rate (or sensitivity) and the false positive rate (or 1 minus specificity) is computed. The final plot of all pairs of true positive rates versus false positive rates is the visual representation of the ROC curve. The interpretation of the area under the ROC curve is easy: a scoring system with a perfect discrimination has an area of 1.0, a scoring system which discrimination is no better than chance has an area of 0.5. For third generation severity scoring systems this area is usually over 0.80 [26-28].

Methods for comparing the areas under ROC curves have been described by Hanley and McNeil in Radiology 1983; 148(3): 839-43 [31]. The general approach to

assessing whether the difference in the areas under two ROC curves derived from the same set of patients is random or real is to calculate a critical ratio z, defined as

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}$$

$$SE(A\hat{r}a_1 - A\hat{r}a_2) = \sqrt{SE^2(A\hat{r}a_1) + SE^2(A\hat{r}a_2) - 2rSE(A\hat{r}a_1)\,SE(A\hat{r}a_2)}$$

where ı and *SE1* refer to the observed area and estimated standard error of the ROC area associated with modality 1; where *A2* and *SE2* refer to corresponding quantities for modality 2; and where *r* represents the estimated correlation between ı and *A2.2* This quantity z is then referred to tables of the normal distribution and values of z above some cutoff, *e.g.*, $z \geq 1.96$, are taken as evidence that the "true" ROC areas are different.

Two intermediate correlation coefficients are required, which are then converted into a correlation between 1 and *A2 via* a table that we supply below. The first is *rN*, the correlation coefficient for the ratings given to images from non-diseased patients by the two modalities. The second is *rA* □ the correlation coefficient for the ratings of diseased patients imaged by the two modalities. Each of these can be calculated in traditional ways using either the Pearson product-moment correlation method or the Kendall tau. The former approach is usually used for results derived from an interval scale whereas the latter is more appropriate for results obtained from an ordinal scale. ROC curves in radiology are derived from ordinal scale data and therefore we have used the Kendall tau for calculating *rN* and *rA*. Standard statistical packages *(e.g.*, SPSS, SAS) provide tau; when the number of rating categories is small, however, say four or less, the calculation can also be performed manually.

Once the correlations between the ratings *(rN* among the normals, *rA* among the abnormals) are obtained, it is necessary to calculate the correlation that they induce between the two areas 1 and *A*, for ease of notation we have called this *r* (without any subscript). This is the coefficient present in Equations 2 and 3. Tabulation of *r*

(Appendex D) is the fundamental contribution of this paper. It has also been shown that this method is not as affected by the size of the sample as calibration measures [31].

Other measures have been used, based on *classification tables*, with authors reporting sensitivity, specificity, positive and negative predictive values and overall correct classification rates. They are, however, of limited utility in the validation of a scoring system, because they have to use a fixed cut point (usually 10, 50 or 90 %). Moreover, they can depend on the mortality rate of the sample, with scoring systems having generally low values on sensitivity especially when the analyzed sample have a relatively high proportion of patients with low probabilities of mortality, since relatively few patients will have a probability of mortality greater than the chosen decision criteria.

The relative importance of calibration and discrimination varies with the utilization that will be given to the scoring system, with some authors arguing that for research or quality assurance purposes (group comparisons) calibration is especially important [5] and that for decisions about individual patients both descriptors are very relevant. From a methodological point of view poor calibration of a scoring system can be corrected. However, there is nothing that can adjust a scoring system when it presents poor discrimination.

**Scoring system update and modifications**

All severity scoring systems require periodic updating. Changes in the population baseline characteristics, improvements in the outcome of major diseases or the introduction of new diagnostic tests with improved accuracy, all imply a modification in the scoring systems used. Moreover, the utilization of a scoring system outside its development population can also require modifications (customization). Two examples in the literature exemplify this problem.

Sirio et al. in Japan [26], utilizing APACHE II, demonstrated lower risk adjusted mortality for Japanese ICUs than for ICUs in the United States of America. They attributed the results to a different case-mix, namely the frequency in Japan of patients with elective esophageal and gastric cancer surgery. APACHE II, which was developed in 1985 in 13 American hospitals, is today not calibrated to be of use in Japan. Its utilization will require recalibration, that is, the computation of new equations relating severity of illness to mortality.

In Italy, Apolone et al. [28] demonstrated very clearly that SAPS II, developed in 1993, was not able to adequately predict mortality in a sample of 99 Italian ICUs. After recalibration, the performance of the scoring system has improved.

Scoring system applicability and utility

Once a scoring system has been developed and validated its applicability and utility should be accessed. All the developers of general severity scoring systems have advocated its general applicability in the intensive care setting. However, the literature contains sufficient examples of scoring systems developed in large populations that failed later when applied in other settings [7, 25 - 28]. The question can only be answered by testing the overall goodness of fit of a scoring system in the population in which this specific scoring system is to be applied. Another important problem is the frequent lack of adherence to the application rules of the scoring system used. Different definitions of the variables, time frames for data collection, frequency of assessment, exclusion criteria and handling of data prior to analysis all can potentially affect the application of a scoring system to a different population

Severity scoring systems in critical care medicine may have many applications in health care management and research [32]. Severity scoring is ideally suited to the ICU for the following reasons:

- The ICU population is well defined, and the provision of care is well circumscribed.
- There is substantial evidence that the degree of illness in the ICU is the major determinant of hospital mortality.
- These scoring systems have considerable power to risk-adjust ICU populations and to permit the evaluation of ICU effectiveness and efficiency [33].

As a result, severity scoring systems can provide a foundation for research and for the assessment of the quality of critical care. As examples:

- Severity scoring instruments play a major role in assessing the quality of ICU care. As an example, a number of studies have used APACHE scores to allow comparison of open and closed ICUs with respect to patient outcomes [34, 35]. In an open ICU, the primary admitting physician directs patient care with input from

critical care specialists via consultation; in a closed ICU, critical care specialists oversee patient care directly. A systematic review found that patients admitted to ICUs staffed by trained intensivists (closed ICUs) had lower mortality rates (relative risk 0.71, 95% CI 0.62-0.82) and shorter hospitalizations (RR 0.61, 95% CI 0.5-0.75) than patients in open ICUs [34].

- Severity scoring systems can be used for inter-hospital comparisons of ICUs [36-38]. One study using APACHE III suggests that teaching hospitals have higher ICU severity and somewhat better risk-adapted patient outcomes [37]. Other studies suggest that APACHE, SAPS, and MPM have a similar ability to identify ICUs that are outliers with respect to quality of care [40-41].
- Severity scores have been used to assess the outcome of critically ill patients transferred from other institutions, compared to those admitted directly from the emergency department or ward. As an example, one study controlled for severity of illness based on a predictive system, and found that transferred patients were independently at increased risk of mortality (RR 2.2, 95% CI 1.7-2.8) [42]. A similar approach has been used to identify other risk factors for adverse outcomes, including admission on a weekend and transfer out of the ICU at night [43-44].
- Severity scores have been utilized to manage some hospital resources. Assigning severity scores to ICU patients may identify those patients who can be placed in less expensive settings [45].
- ICU severity scoring systems have been used to compare critically ill patients in many therapeutic trials, particularly patients with sepsis and ARDS [46-49].

**Comparison of systems**

The three major ICU predictive scoring systems (APACHE, SAPS, MPM) all have been developed for more than 20 years. The current iterations of these scoring systems are more accurate than their predecessors. From literature review, the performances of these severity scoring systems are summarized in table 2.3.

Until now, there were only a few studies about the performance of these scoring systems in Thai ICU patients. Lertsithichai et al. [9] have found that SAPS II exhibited

good discrimination and calibration, in contrast to the study of Khwannimit et al. [10] which reported that APACHE II provided better discrimination than SAPS II but both systems showed poor calibration in over predicting mortality and suggested that customized or new severity scoring systems should be developed for critically ill patients in Thailand.

It is important to realize, however, that there are several limitations and methodological differences among these tools [40-41].

1. Data collection

Both APACHE and SAPS use the worst physiologic values measured within 24 hours of admission to the ICU. In contrast, $MPM_0$ data are collected immediately on ICU admission, and modified following 24 hours of hospitalization ($MPM_{24}$). Substitution of the worst 24-hour physiological variables with admission physiological variables to calculate the APACHE II score has been studied [51]. The retrospective evaluation of 11,107 ICU admissions found no difference in the discrimination ability, suggesting that an admission APACHE II score is an acceptable alternative to a worst 24-hour APACHE II score.

APACHE III requires precise physiologic measurements; by comparison, SAPS and MPM use broader physiologic categories, which may facilitate data recording.

The APACHE III instrument can recalculate estimated mortality on a daily basis [18]. This method may have greater predictive power than a single projection based on the first 24 hours of ICU admission [52]. However, the Study to Understand Prognosis and Preferences for Outcomes and Risks of Treatment (SUPPORT) examined how this updated prognosis affected clinical decisions, and found that communication between physicians, patients, and their families needed substantial improvement before predictive technology would enhance clinical decision-making.

2. Mortality calculation

APACHE III and APACHE IV use proprietary computer software to calculate predicted mortality. SAPS and MPM use published equations in which the severity score is entered into equations whose solutions provide the predicted mortalities.

3. Cost

APACHE III and APACHE IV require proprietary computer technology and substantial data collection. APACHE II score calculators are available to the public. MPM and SAPS require somewhat less data and no additional computer investment.

**Customization of the severity scoring systems**

When a predictive severity scoring system calibrates poorly on an external data set, one may try to improve its performance by customizing the system to the data. Several strategies exist to do so [65, 66]. For instance, one may choose to re-estimate a system's coefficients on the new data, and to add or remove terms from the system. A simpler customization strategy is to re-estimate the intercept and slope of the linear predictor, by fitting a new logistic regression equation with observed outcome as the dependent variable and the logit-transformed original predictions as the independent variable, this has been termed "logistic recalibration" [67].

This strategy will only be effective when the scoring system shows structural errors on the external data set, that is, when the average predicted risk is either too high or too low, or when the predicted probabilities either vary not enough or too much, and not when the calibration problems are more subtle. This type of customization will not change the order of the predictions, thus leaving AUC values unaffected and it was earlier shown to be effective in the context of ICU prognosis [68].

Table 2.3 The performances of various severity scoring systems

| Authors | n = death/total (% death) | APACHE II | APACHE III | SAPS II | $MPM_0$ II | $MPM_{24}$ II |
|---|---|---|---|---|---|---|
| Le Gall, et al.[3] (1993) | n=2867/13,152 (21.8%) | | | ROC AUC =0.86<br>GOF $\chi2$ =3.7<br>P value =0.883 | | |
| Poses R, et al. [62] (1996) | n = 38/201 (18.9%) | ROC AUC =0.87 | | | | |
| Moreno, et al.[7] (1998) | n = 2005/10,027 (19.99%) | | | ROC AUC =0.822<br>GOF $\chi2$ =208<br>$p < 0.001$ | ROC AUC =0.785<br>GOF $\chi2$ =368<br>$p < 0.001$ | |
| Markgraf R, et al.[56] (2000) | n = 491/2661 (18.45%) | ROC AUC =0.83<br>GOF $\chi2$ =11.8<br>$p > 0.1$ | ROC AUC =0.846<br>GOF $\chi2$ =48.4<br>$p < 0.001$ | ROC AUC =0.846<br>GOF $\chi2$ =20.5<br>$p < 0.001$ | | |
| Cook, et al.[61] (2006) | n = 338/3455 (9.9 %) | | ROC AUC =0.92<br>GOF $\chi2$ =7.4<br>$p = 0.01$ | | | |
| Livingston, et al.[5] (2000) | n = 3055/10393 (29.39 %) | ROC AUC =0.805<br>GOF $\chi2$=36.4<br>$p = < 0.001$ | ROC AUC =0.845<br>GOF $\chi2$ =331<br>$p = < 0.001$ | ROC AUC =0.843<br>GOF $\chi2$ =57.75<br>$p = < 0.001$ | ROC AUC =0.785<br>GOF $\chi2$ =307<br>$p = < 0.001$ | ROC AUC =0.799<br>GOF $\chi2$ =101<br>$p = < 0.001$ |
| Arabi Y, et al.[6] (2002) | n = 310/969 (32 %) | ROC AUC =0.83<br>GOF $\chi2$=21.9<br>$p = 0.005$ | | ROC AUC =0.79<br>GOF $\chi2$=43.4<br>$p = < 0.001$ | ROC AUC =0.85<br>GOF $\chi2$=26.6<br>$p = < 0.001$ | ROC AUC =0.84<br>GOF $\chi2$= 14.7<br>$p = 0.06$ |

Table 2.3 The performances of various severity scoring systems (continued)

| Authors | n = death/total (% death) | APACHE II | APACHE III | SAPS II | $MPM_0$ II | $MPM_{24}$ II |
|---|---|---|---|---|---|---|
| Arabi Y, et al. [58] (2003) | n = 152/250 (60.8 %) | ROC AUC =0.782 GOF $\chi 2$=34.9 p = < 0.001 | | ROC AUC =0.797 GOF $\chi 2$=23.6 p = 0.003 | ROC AUC =0.806 GOF $\chi 2$=29.8 p = < 0.001 | ROC AUC =0.823 GOF $\chi 2$=24.8 p = 0.002 |
| Ihnsook, et al. [63] (2003) | n = 84/284 (29.57 %) | | ROC AUC =0.90 GOF $\chi 2$=6.54 p = 0.59 | | | |
| Vosglius, et al (2004) | n = 524/2067 (25.35 %) | | | ROC AUC =0.883 GOF $\chi 2$=56.98 p = < 0.001 | | |
| Desa K,et al [59] (2005) | n = 87/395 (22.03 %) | | | ROC AUC =0.827 GOF $\chi 2$=22.96 p = 0.003 | | |
| Le gall [55] (2005) | n = 16,645/77,490 (21.48 %) | | | ROC AUC =0.858 GOF $\chi 2$=116 p = < 0.001 | | |
| Ratanarat, et al. [57] (2005) | n = 178/482 (36.9 %) | ROC AUC =0.788 | | ROC AUC =0.746 | | |
| Zimmerman, et al. [53] (2006) | n = 15,035/110,558 (13.6 %) | | ROC AUC =0.86 GOF $\chi 2$=635 p = < 0.001 | ROC AUC =0.88 GOF $\chi 2$=16.9 p = 0.08 | | |
| Rivera-Fernandez, [54] et al. (2007) | n = 3,625/17,598 (20.6 %) | | ROC AUC =0.84 | ROC AUC =0.80 | | |
| Higgins, et al. [60] (2007) | n = 17,262/125,085 (13.8 %) | | | | ROC AUC =0.823 SMR 1.018 (.99-1.04) | |

Recommendations for a validation of ICU scoring system [69]

An ICU scoring system that accurately estimates the probability of in-hospital mortality can potentially be used as a risk adjustment tool to analyse mortality outcomes in an ICU. The scoring system's performance must however be thoroughly evaluated to determine whether the estimates of probability of death are accurate.

A publication of guidelines for standard reporting of the accuracy of diagnostic test by the Standards for Reporting of Diagnostic Accuracy (STARD) provide a useful example of an explicit and rigorous check list to serve as a guide to methodology and documentation. Reports on the performance of ICU death share similar characteristics to reports on the accuracy of a diagnostic or screening test. Therefore, the systematic approach used by STARD provides a framework for scoring system assessment. From this approach, the following recommendations are made for presentation of the performance of scoring systems that estimate the probability of death of ICU patients.

1. The Title, Abstract and Keywords should identify the report as that of the assessment of the performance of an ICU mortality prediction scoring system.

2. The Introduction should clearly state the aim of the scoring system, to validate an existing scoring system, to compare several scoring systems, or to adjust an existing scoring system. Relevant knowledge or previous assessments should be presented supporting the value of the current analysis.

3. The Methods section should describe the context of the analysis and dates of data collection. Single or multiple ICUs, and the type of hospital and ICU should be described. The study population and the method of patient eligibility and exclusion should be described. This will allow the reader to assess independence of the validation process, conditions affecting scoring system performance, and the applicability of the scoring system to a simulation of interest.

4. The scoring system application and the assessment process must be reproducible. An account of rules of data collection and a description of the scoring system development are required in a full, or referenced if it has already been described. If a new scoring system is presented, or existing scoring system modified, a description of the assumptions and the statistical methods used to develop the scoring system and to assess its performance should be given. The mortality and survival endpoints must be

defined. The methods by which patient data are divided into sets for scoring system development and testing must be given.

5. The accuracy of ICU outcome scoring systems should be assessed in terms of discrimination and calibration.

- For discrimination, the area under the ROC curve should be calculated, with standard error or confidence intervals for precision. For two scoring systems on the same sample, pair-wise comparison of scoring systems should be done.
- Calibration should be assessed by an overall assessment statistic and an assessment of fit across risk intervals. The most commonly used global indication is the SMR with confidence intervals. A graphical approach with a calibration curve incorporating either confidence intervals or a frequency of patients across intervals, should be presented. A numerical evaluation of goodness-of-fit using the H-LC statistic should be used. If more than one scoring system is being evaluated on the dataset, then a statistical comparison is suggested.

6. The results section should contain a descriptive analysis of the sample including characteristics of age, gender ad major diagnostic categories, severity of illness measurements and mortality rate. If this is an independent evaluation of a scoring system, it is useful to compare the patient characteristics of the validation data set with those for the sample on which the scoring system was developed. All missing or incomplete records must be accounted for. It can be useful to include a flow diagram of all admissions identifying readmissions and excluded patients as well as missing and incomplete records.

7. The discussion should evaluate the strengths and weaknesses of methodology and the scoring system in the context of the results of the analysis and contemporary relevant knowledge.