Severity of illness scoring systems in the intensive care unit

Mark T. Keegan, MB, MRCPI; Ognjen Gajic, MD; Bekele Afessa, MD

Objective: Adult intensive care unit prognostic models have been used for predicting patient outcome for three decades. The goal of this review is to describe the different versions of the main adult intensive care unit prognostic models and discuss their potential roles.

Data Source: PubMed search and review of the relevant medical literature.

Summary: The main prognostic models for assessing the overall severity of illness in critically ill adults are Acute Physiology and Chronic Health Evaluation, Simplified Acute Physiology Score, and Mortality Probability Model. Simplified Acute Physiology Score and Mortality Probability Model have been updated to their third versions and Acute Physiology and Chronic Health Evaluation to its fourth version. The development of prognostic models is usually followed by internal and external validation and performance assessment. Performance is assessed by area under the receiver operating characteristic curve for discrimination and Hosmer-Lemeshow statistic for calibration. The areas under the receiver operating characteristic curve of Simplified Acute Physiology Score 3, Acute Physiology and Chronic Health Evaluation IV, and Mortality Probability Model III were 0.85, 0.88, and 0.82, respectively, and all these three fourth-generation models had good calibration. The models have been extensively used for case-mix adjustment in clinical research and epidemiology, but their role in benchmarking, performance improvement, resource use, and clinical decision support has been less well studied.

Conclusions: The fourth-generation Acute Physiology and Chronic Health Evaluation, Simplified Acute Physiology Score 3, Acute Physiology and Chronic Health Evaluation IV, and Mortality Probability Model III adult prognostic models, perform well in predicting mortality. Future studies are needed to determine their roles for benchmarking, performance improvement, resource use, and clinical decision support.

Key Words: APACHE; benchmarking; critical care; intensive care unit; mortality; outcome assessment; statistical models

The severity of illness assessment scoring systems may be disease- and organ-specific, or global. The first intensive care unit (ICU) prognostic model used to assess patients’ overall disease severity was the Therapeutic Intervention Scoring System (1). The Therapeutic Intervention Scoring System was first described as a severity index based on treatment intensity. Its performance (discrimination and calibration) in predicting mortality was not well described. It is a direct measure of treatment intensity, not necessarily disease severity. As a result, its current application is limited to the assessment of workload and resource allocation in the ICU (2, 3). During the last three decades, several physiological-based ICU prognostic models have emerged. The main prognostic models for assessing the overall severity of illness in critically ill adults are Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), and Mortality Probability Model (MPM) (4–13). Pediatric Risk of Mortality and Pediatric Index of Mortality scores are used in critically ill pediatric patients (14–16). There are also scores primarily designed to describe the degree of organ dysfunction, not survival, in the critically ill (17–20). This review focuses on the global prognostic models used to assess the severity of illness in critically ill adults.

Model Development, Validation, Performance, and Customization. In a recent publication, our group reviewed the basics of model development (4). The commonly used predictor variables include age, comorbidities, physiological abnormalities, acute diagnoses, and lead-time bias. Lead-time bias refers to the inaccuracy in risk prediction that occurs when treatment and measurement occur at different times (21, 22). Lead-time bias has most effect in medical patients and emergency admissions (22). The main outcome measure is usually short-term mortality. The APACHE III and IV models have also included length of ICU and hospital stay and duration of mechanical ventilation (23–26). The relationships between the predictor and outcome variables of the development model need independent validation (27, 28). A mortality prognostic model must differentiate between survivors and nonsurvivors and be well calibrated and reliable (29). It also has to be periodically updated to reflect the change in medical practice and case mix over time (4, 30). The performance of the ICU prognostic models is assessed by the area under the receiver operating characteristic curve for discrimination and the Hosmer-Lemeshow statistic for calibration (4, 28, 31–33). The area under the receiver operating characteristic curve in critically ill adults is greater than 0.70 (34). The calibration is considered good if the Hosmer-Lemeshow statistic $p$ value is >.05.
Because of changes in case-mix and clinical practice, the performances of prognostic models deteriorate over time (35). To counterbalance the deterioration, models are often customized (36, 37) by adding new predictor variables at times (38, 39). The customized SAPS3 equations are already undergoing further customization for region-specific benchmarking purposes (40, 41). Recent data suggest that MPM0 III may not be significantly affected by case-mix (42). There are limited data to recommend how often the performance of such models should be assessed. We believe model performance should be assessed periodically and model upgrading may be needed every 4 yrs.

Adult ICU Prognostic Models. The main adult ICU prognostic models include APACHE, SAPS, and MPM (4–13, 43, 44). Recent reviews have addressed the main components of these models (4, 45). In addition to predicting mortality, APACHE III and IV provide predictions for ICU and hospital length of stay, duration of mechanical ventilation, risk of needing an active treatment during the ICU stay, and potential transfer from the ICU (4, 23, 24, 26). A recent multicentered study from California developed ICU length of stay prediction models based on SAPS II, MPM0 III, and recalibrated APACHE IV length of stay models (46). The study showed that APACHE IV and MPM0 III were more accurate than SAPS II for predicting ICU length of stay and APACHE IV was the most accurate and best calibrated model.

The history of the current adult ICU prognostic models goes back to development of the original APACHE three decades ago (7). Its second generation, APACHE II, is an ICU prognostic scoring system that is the most widely used in the medical literature (5). APACHE III was narrowly disseminated because of its proprietary nature. SAPS I was developed on data from eight ICUs in France (9) and SAPS II from 137 ICUs in 12 countries (8). The MPM I model was created from a small number of easily available variables from a single medical center (10). Fifteen variables were used in the MPM0 II (43).

Fourth-Generation Adult ICU Prognostic Models. Studies evaluating the performance of the older generation adult prognostic models showed performance degradation over time manifested by worsening discrimination and calibration of the model (47). This led to the development of fourth-generation adult ICU prognostic models. There are differences among these models. Data for SAPS3 were collected as part of a research project. The data for APACHE IV and MPM III were obtained from ICUs that had bought the APACHE or Project Impact Critical Care systems (both owned by Cerner Corporation, Kansas City, MO). Because institutions that participated in the development of either of these models were not randomly selected, the findings may not apply to other ICUs. Because several medical centers continue to participate in Cerner-owned APACHE and Project Impact activities, upgrades and newer versions are likely to be developed when APACHE IV and MPM III show performance degradation (48).

The calibration of the initial fourth-generation models was good as well as the discrimination. The areas under the receiver operating characteristic curve of SAPS3, APACHE IV, and MPM0 III were 0.85, 0.88, and 0.82, respectively (11–13, 49). APACHE IV and MPM0 III were validated in a multicentered study of 11,300 ICU patients from California (50). APACHE IV had better discrimination and longer data extraction time than MPM0 III (50). MPM0 III was recently validated on 55,459 patients from 103 ICUs, 25 of which did not participate in the original development (48). The fact that all three fourth-generation models are free from charge may help their use for research, healthcare delivery, and performance measure. APACHE IV is the most complex and may require software support. MPM0 III is the least complex.

A recent focus of SAPS3 researchers has been highlighting the need for institutional or regional customization (40, 41, 51–56). All patients included in the development of APACHE IV and MPM III were from the United States. In contrast, patients from five continents were included in the development of SAPS3. With its customized models, SAPS3 appears to be a good candidate for an international benchmark. However, the number of patients included from some of the countries is very small and the results may not be generalizable. There are external SAPS3 validation data from different countries. However, most of these data are limited to small sample size and a narrow patient case-mix (11–13, 49, 51, 52, 54–59). A recent SAPS3 external validation study of 28,357 patients from 147 Italian ICUs showed good discrimination but poor calibration (53). Similar findings were noted in an Austrian multicentered study (41).

Because MPM0 III and SAPS3 are based on data obtained within 1 hr of ICU admission, they can be used to assess severity of illness before ICU interventions take place. The prognostic models assume missing data as normal, which may adversely affect the performance of the severity scores (60). Because of the multiplicity of data to be collected, missing data may have the highest impact on the performance of APACHE IV (Table 1). The performances of prognostic models in predicting outcome are likely to be compromised by the lack of uniformity in data acquisition (61). Several ICUs use computer interfaces with their laboratory and bedside monitor systems to extract data. Others still enter data manually. SAPS3 was calibrated for manual data acquisition.

All patients included in the development of APACHE IV and MPM III were from the United States. In contrast, patients from five continents were included in the development of SAPS3. However, the number of patients included from some of the countries was very small and the results may not be generalizable.

Model Use. Knowledge about the probability of clinical outcome may provide help to healthcare policymakers, hospital administrators, clinicians, patients, and their families in selecting treatment options. Rising costs of health care and concerns about quality of care have led to efforts aimed at determining outcome associated with medical services. Quality of care can be measured by comparing observed and predicted outcomes. Discordance between the predicted and observed outcomes is considered to indicate better or worse than average quality of care. The ICU adult prognostic models are attractive to measure predicted outcomes in the critically ill. However, the predictor variables should be resistant to manipulation and subjectivity and the models should be reliable and valid before they are applied to assess quality of care (28). There are several factors unrelated to quality of care such as patients' preferences for life support and response to disease, the surrounding environment, and effect of treatment that influence outcome and may have not been included in the prognostic models (47). Most of the prognostic models do not include patients' preferences for life support as a predictor variable. However, the MPM investigators have persistently shown that
Table 1. Variables included in the fourth-generation prognostic models

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>SAPS 3 (11, 12)</th>
<th>APACHE IV (13)</th>
<th>MPM III (49)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Length of hospital stay before ICU admission</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ICU admission source</td>
<td>3</td>
<td>8</td>
<td>No</td>
</tr>
<tr>
<td>Type of ICU admission</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Chronic comorbidities</td>
<td>6</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>CPR before ICU admission</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Resuscitation status</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Surgical status at ICU admission</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Anatomic site of surgery</td>
<td>5</td>
<td>116</td>
<td>5</td>
</tr>
<tr>
<td>Reasons for ICU admission/acute diagnosis</td>
<td>10</td>
<td>116</td>
<td>5</td>
</tr>
<tr>
<td>Acute infection at ICU admission</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Mechanical ventilation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Vasovagal drug therapy before ICU admission</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Clinical physiological variables</td>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Laboratory physiological variables</td>
<td>6</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

SAPS, Simplified Acute Physiology Score; APACHE, Acute Physiology and Chronic Health Evaluation; MPM, Mortality Probability Model; ICU, intensive care unit; CPR, cardiopulmonary resuscitation.

Modified with permission from Afessa et al (4).

Table 2. Potential uses of adult intensive care unit prognostic models

<table>
<thead>
<tr>
<th>Level</th>
<th>Potential Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>National level</td>
<td>Benchmarking</td>
</tr>
<tr>
<td>Institution level</td>
<td>Internal use of quality improvement efforts such as comparing performances of hospitals against the national average.</td>
</tr>
<tr>
<td></td>
<td>Institutional self-monitoring for competitive or contractual reasons.</td>
</tr>
<tr>
<td></td>
<td>Monitoring by regulatory agencies and payors.</td>
</tr>
<tr>
<td></td>
<td>Adjustment of outcomes in clinical trials.</td>
</tr>
<tr>
<td></td>
<td>Helping patients select among different hospitals.</td>
</tr>
<tr>
<td>Physician level</td>
<td>Quality improvement for individual physician.</td>
</tr>
<tr>
<td></td>
<td>Institution’s use of outcome information on individual physicians.</td>
</tr>
<tr>
<td></td>
<td>Help patients select physician.</td>
</tr>
<tr>
<td></td>
<td>Use of persistent poor performance for sanctions or license withdrawal.</td>
</tr>
<tr>
<td>Patient level</td>
<td>Help patients in the decisionmaking process.</td>
</tr>
<tr>
<td></td>
<td>Resource allocation</td>
</tr>
</tbody>
</table>

the Do-Not-Resuscitate status of a patient influences outcome (10, 44, 49). We have indicated previously that lead-time bias influences patient outcome (21, 22). However, most of the prognostic models do not include any measure of lead-time bias. The specific ICU admission reason is one of the main predictors of outcome. APACHE IV includes 116 ICU admission reasons compared with ten and five for SAPS3 and MPM III, respectively (11–13, 49). Despite their limitations, the predictive models have potential uses at the national, hospital, physician, and patient levels (Table 2) (28).

**Benchmarking.** Clinicians and investigators need to know why some ICUs save more lives than others (62). Transparency and severity-adjusted data analysis linked to the process of care are likely to lead to this path. The US News and World Report has been ranking thousands of hospitals in the United States for several years (63). Thompson Reuters has started reporting the 100 top US hospitals annually (64). Several Internet web sites claim to provide physician ratings. An organization had published ranking of ICUs based on their performances (65). Most of these rankings are based on administrative data. The Centers for Medicare and Medicaid Services is planning to transition from fee-for-service payment systems to value-based purchasing (66, 67). Compared with the severity models derived from administrative data, the ICU adult prognostic models are better tools for risk-adjusted quality assessment. Administrative data do not distinguish between medical conditions present at hospital admission and complications that occur after admission (68, 69). Including complications as pre-existing conditions weakens the performance of a model in risk stratification (68). Compared with models based on administrative data, the ICU severity score models are based on predictor variables available at ICU admission and their performances have been well described. Although not yet implemented, the Joint Commission has a plan requiring hospitals to publicly report their risk-adjusted mortality and length of stay as part of the ICU quality core measures (70). Some states have already started participating in ICU prognostic model severity-adjusted benchmarking (46, 50). Standardized mortality ratio is widely used to evaluate performance. The standardized mortality ratio should be reported with its 95% confidence intervals (71). Similar measures can be established for other outcomes such as length of stay (72). Benchmarking helps to identify variations in clinical outcome and changes in practice patterns over time (73). The appropriate application of benchmarking at the national and community levels may provide reliable information to regulatory agencies, payers, healthcare providers, and patients. However, it requires buy-in by governments, payers, hospitals, healthcare providers, and the public. Over two decades ago, hospitals, physicians, and employers from the Cleveland metropolitan area formed the Cleveland Health Quality Choice to implement a standardized measurement system to evaluate patient outcome in the ICU (73). They used the APACHE III prognostic system for severity adjustment. Although the overall study was compromised by the use of a nonrecalibrated model and changing hospital discharge practices, it highlighted the feasibility of community-based benchmarking. Hospital discharge practices influence hospital mortality. In a multicentered study from California, Vailevikis et al described the association between acute care hospital transfers and early postdischarge mortality and recommended using the standardized mortality ratio based on 30-day, instead of hospital, mortality (74). Benchmarking provides opportunities to improve performance based on findings from good and bad performers (72, 75–77). Internal benchmarking can also be used to highlight weaknesses and strengths within the same institution (78).

Although hospital mortality rates is usually used to judge hospital performance, it has several weaknesses, including the limitation of the risk adjustment to factors, that are identifiable and measurable (79). In addition to mortality, there are other important outcome measures that can be used for benchmarking. The APACHE prognostic system has models for predicting ICU length of stay (23, 26) and duration of mechanical ventilation (24). The MPM researchers introduced and subsequently revised a two-dimensional graphic tool (Rapoport-Teres graph) for benchmarking performance and resource use (80, 81). The Rapoport-Teres graph is constructed by plotting the normalized differences between actual
and predicted survival rates (standardized clinical performance index) on the x-axis and normalized differences expected and actual weighted hospital days values (standardized resource use index) on the y-axis (80, 81) (Fig. 1). A recent publication reported the standardized hospital length of stay, based on the adult ICU prognostic models, of several hospitals from California (46). The APACHE III database also provides accessories to track low-risk monitor admissions and readmissions (25, 72, 78). Previous ICU benchmark studies based on severity-adjusted outcome have identified policies and practices associated with ICUs that perform well and with good patient outcome: the existence of an alternative to ICU care, a mechanism for improving selection for ICU care, a mechanism to facilitate patient throughput, a mechanism to facilitate ICU discharge planning, reducing excess capacity, matching staffing to workload, process-related guidelines or protocols, care guidelines or protocols for high-volume diagnoses and care processes, performance monitoring and review, and empowering the medical director of the ICU to play an active role (72, 75). Implementation of such policies and practices is likely to improve patient outcome.

**Performance Improvement.** A well-performing prognostic model helps to make meaningful comparisons of a hospital’s current performance with its past. This will allow hospitals to identify their weaknesses and initiate interventions aimed at quality improvement and allow patients and third party payers to choose healthcare providers based on performance. However, changes in case-mix, practice patterns, and other secular trends may also influence differences in outcomes. The ICU prognostic models may facilitate the accreditation process by external organizations. The ICU severity models may also serve as tools for evaluation of the impact of new therapies as well as organizational and process of care changes (72, 75, 78).

The APACHE Critical Care Series and Project Impact have advanced the prognostic models by adding accessories to track readmission, sentinel event, reimbursement, and resource consumption (82). They regularly provide standardized and customized reports of outcome. Based on data from the APACHE III database, Zimmerman et al (72) have highlighted the policies and practices of ICUs with low mortality rate and efficient resource use. They have described the structural characteristics and process of care in ICUs with good performance.

**Resource Use.** Accurate estimation of severity of illness has the potential to help in the appropriate allocation of scarce ICU resources. With the scarcity of ICU beds in many hospitals, avoiding unnecessary ICU admission and transferring patients who do not need ICU are important. MPMIII and SAPS3 have the potential to be used as decision support for ICU admission triage because most of their predictor variables are available at admission (11, 12, 49). The Critical Care Series of the APACHE III clinical support system provides estimates for the risk of requiring specific critical care interventions and potential transfer from the ICU, including providing care in an intermediate unit with reduced cost (25, 82, 83). Using APACHE III data, Seneff et al (24) reported an accurate prediction of the average duration of mechanical ventilation for groups of ICU patients. The MPM researchers developed and updated models for predicting patients’ weighted hospital days (80, 81). Such predictions may be useful for resource allocation.

**Clinical Decision Support.** Probabilities of hospital mortality provide meaningful information to physicians when discussing patient prognosis. However, probabilities should not be used for making treatment decisions at the individual patient level (84). Patient and caregiver preferences as well as their spiritual and cultural beliefs have to be taken into account during the decision-making process by patients and family members. The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments showed that survival estimates combining an objective prognosis with a physician’s clinical estimate had better ability to identify patients with high probabilities of survival or death (85). This can be attributed to the fact that physician estimates of low ICU survival may lead to subsequent life support limitation (86). Currently, most patients and their families rely on prognostic information given to them by the physicians to make decisions. However, because of the biases of subjective estimates, physicians’ ability to correctly predict mortality is highly variable (87, 88). Assessment of futility is another important potential application for the use of severity of illness systems. Trends in the severity of illness provide important prognostic information (89).

In patients with high risk of death at ICU admission, lack of improvement in severity score indicates poor prognosis (89–91). Awaiting studies addressing their role in improving the clinicians’ estimates, the probabilities derived from the prognostic models should be used as “the drunken man uses the lamppost, for support rather than illumination” in making a clinical decision (92). Prognostic models will need to be subjected to the same scrutiny as drugs before they are used in decisions that impact on healthcare delivery and individual patient care.

Although there is scarcity of data supporting the use of severity scores for individual patient care, APACHE II-based administration of activated protein C for severe sepsis/septic shock has become an accepted practice. The initial randomized clinical trial showed that activated protein C reduces the mortality of patients with severe sepsis/septic shock (93). However, post hoc analysis suggested this benefit may not extend to patients with lower severity scores. A subsequent clinical trial showed no mortality benefit in patients with APACHE score <25 (94).

**Limitations.** There are several limitations inherent in the ICU prognostic models (95). Errors in data collection and entry and flaws in model development and validation weaken the performance of prognostic models. All adult ICU severity score models, including the three fourth-generation ones, were developed in nonrandomly selected ICUs, compromising the generalizability of their findings (11–13, 26, 49). Application of prognostic models requires unambiguous definitions of predictor and outcome variables and reproducible measurements easily available in clinical practice (96). Predictor variables may not be easily measured and certain laboratory values may not be rou-
tinely obtained. Lack of standardization in obtaining predictor values leads to missing data, compromising the performance of a model. A few years usually pass between performing a study and publishing its result. By the time the prognostic model studies are published, their prognostic accuracy may have degraded (96). Several factors, including lead-time bias, pre-ICU location, acute diagnosis, physiological reserve, and patients’ preferences for life support, influence mortality. Most of these prognostically important variables are not included in some of the latest prognostic models. Although the ICU models perform reasonably well in the general ICU population, they are far from perfect in identifying which individual patient will live or die. Most importantly, long-term survival and quality of life are not forecast by the prediction models.

Future Directions. Prognostic research has received limited attention compared with etiologic, diagnostic, and therapeutic research (97). Data addressing the impact of adult ICU prognostic models on healthcare providers’ behavior and patient outcome are scarce. Currently, existing hospital and healthcare provider ranking systems are based on administrative databases and are greatly influenced by the public relation policies of the individual hospital or healthcare provider. The development and application of robust prognostic models are prerequisites for meaningful ranking. The level of clinical detail, ICU-specific diagnoses, and variables make the current adult prognostic models attractive for use in epidemiologic and critical care outcomes research (98).

Future studies are needed to determine the role of the ICU severity scores in clinical practice. The APACHE II, III, and IV models provide risk stratification based on the worst value of the first 24 hrs (5–7, 13), MPM2 III and SAPS3 provide risk stratification based on data available within 1 hr of ICU admission (11, 12, 49). If these models can be modified to include values available before ICU admission, they can be incorporated in the ICU admission criteria. Risk stratification based on severity scores can identify patients who are at low risk of mortality and in whom ICU-level life-sustaining interventions are unlikely to be required (25). The identification of such patients can lead to their treatment in the non-ICU setting at reduced cost. The APACHE III and IV prognostic models provide daily risk stratification (6, 13). The potential use of the trend in risk stratification in patient disposition from the ICU needs further studies (90, 91). Internal and external ICU benchmark studies have shown that differences in policies and practices may partly explain the variations in performance among ICUs (72, 75, 76, 78). The impact of implementing the best policies and practices requires future studies. There is scarcity of data advocating the use of ICU severity scores in selecting treatment for individual patients. The lack of benefit of activated protein C in patients with severe sepsis and APACHE II score <25 highlights the potential role of risk stratification based on severity scores in selecting treatment (94). Before their full potential is realized, future prognostic models will have to include not only the improvement in the assessment of baseline severity of illness, but also the meaningful patient outcome beyond hospital mortality.

REFERENCES