Severity Scoring Systems in the Modern Intensive Care Unit

G Clermont,* MSc, MD, CM, D C Angus,** MB, ChB, MPH

Abstract

In recent years, several factors have led to increasing focus on the meaning of appropriateness of care and clinical performance in the intensive care unit (ICU). The emergence of new and expensive treatment modalities, a deeper reflection on what constitutes a desirable outcome, increasing financial pressure from cost containment efforts, and new attitudes regarding end-of-life decisions are reshaping the delivery of intensive care worldwide.

This quest for a measure of ICU performance has led to the development of severity adjustment systems that will allow standardised comparisons of outcome and resource use across ICUs. These systems, for many years used only in the research setting, have evolved to become sophisticated, computer-based decision-support tools, in some instances commercially developed, and capable of predicting a diverse set of outcomes. Their application has broadened to include ICU performance assessment, individualpatient decision-making, and pre- and post-hoc risk stratification in randomised trials.

In this paper, we review the popular scoring systems currently in use; design issues in the development and evaluation of new scoring systems; current applications of scoring systems; and future directions.

Ann Acad Med Singapore 1998; **27:397-403**

Key words: ICU performance, Mortality prediction, Severity model

Overview of Current Prediction Systems

The first major general severity adjustment system, the Acute Physiology and Chronic Health Evaluation (APACHE) system, was published in 1981.[1] Since then, APACHE, the Mortality Prediction Model (MPM), and the Simplified Acute Physiology Score (SAPS) systems have dominated the literature as general outcome prediction models. These systems share the construct that demographic and physiologic information collected at ICU admission and during the first 24 hours in the ICU is predictive of hospital mortality. As opposed to prediction models that are organ- or disease-specific, these general models were developed using large, heterogeneous populations of critically ill patients. APACHE III was constructed from a large cohort of US patients recruited towards the end of the last decade.* SAPS II and MPM II were both created on a large, combined North American and European, slightly more recent cohort.[3,4]

APACHE

The APACHE model consists of two distinct components: a chronic health score, reflective of the pre-morbid condition of the patient, and an acute physiology score, calculated over the first 24 hours of an ICU admission. The latest model, APACHE III, introduced in 1991, includes an expanded array of admission diagnoses and considers the location of the patient prior to ICU admission.* The latter modification was driven by the important observation that the implication of a given score differed depending on whether or not that score reflected prior therapy.[5,6] Failure to account for this fact may result in a lead-time bias (the capacity of APACHE III to eliminate lead-time bias, however, has not been evaluated independently). APACHE III is also designed to include a daily update of the probability of death[7] and, when coupled with additional elements pertaining to resource use, can be used to predict the need for continued ICU care.[8]

MPM

The Mortality Probability Model (MPM) is not a score but rather a direct prediction engine relying mainly on binary variables such as "heart rate greater than 150" or "presence of acute renal failure" as predictors.[9] MPM is the only model that provides a prediction of death at

* *Visiting Assistant Professor of Anesthesiology and Critical Care Medicine*
** *Associate Professor of Anesthesiology, Critical Care Medicine and Medicine*
Health Delivery and Systems Evaluation Team (HeDSET)
Department of Anesthesiology and Critical Care Medicine
University of Pittsburgh, USA
Address for Reprints: Dr Derek C Angus, 606B Scaife Hall, Critical Care Medicine, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA 15213, USA.
Email: angus@smtp.anes.upmc.edu

admission to the ICU (MPM_0). It was also developed to yield updated probabilities of hospital death at 24, 48 and 72 hours (MPM_{24}, MPM_{48}, MPM_{72}, respectively) into the ICU course, as early progress is often indicative of outcome. The system does not include a specific admission diagnosis[3] although cardiopulmonary resuscitation prior to admission, an ominous marker of adverse prognosis, is one of the dichotomous variables.

SAPS

The Simplified Acute Physiology Score (SAPS) was first developed as a simplified version of the acute physiology component of APACHE." As with APACHE III, SAPS II, the most recent version of SAPS, predicts hospital mortality from data collected during the first 24 hours of ICU stay though it does not include an admission diagnosis.[4]

Comparisons of Current Systems

SAPS and APACHE II have been compared for performance within disease categories such as myocardial infarction,[11] acute pancreatitis,[12] and liver transplantation.[13] Their performances were generally comparable. There is only one published report of the relative performance of the more recent instruments (MPM II, SAPS II, APACHE II and APACHE III) on a large prospective database of mixed ICU patients.[14] The newer generations appear to outperform their predecessors, but there is currently insufficient data to suggest recommending one prediction model over another.

Data Collection Considerations

Wide use of general prediction models'may be associated with lower data quality. In a recent evaluation, well-trained data collectors, all of whom were nurses or residents, made clinically-significant errors measuring the acute physiology assessment in 18% of patients.[15] There is no reason to believe that the widespread application of scoring systems will not result in significant errors in mortality predictions in cases where the data collection is performed by less trained personnel or where the data is not audited regularly. At a minimum, it appears wise that data be collected by a small number of individuals and that regular assessment of inter-rater reliability be performed.[16]

APACHE III includes the diagnosis on admission in the regression equation of the predicted probability of death. As reasonable as this might appear to be, a single most appropriate admission diagnosis may be difficult to determine in a number of cases, independent of the clinical competence of the treating team. Accordingly, the use of a large number of diagnostic categories may actually create a misclassification bias and dilute the power of the prediction instrument. Furthermore, the inclusion of large numbers of disease categories reduces one's confidence in the prediction equations when the number of patients within some disease categories in the development cohort seems insufficient. Other factors that could lead to misclassification include variable measurement accuracy, the normalcy assumption (unmeasured variables are presumed to have values within the normal range), the categorical assignment of weights for APACHE and SAPS, and the subjectivity of the cut-offs for the binary MPM variables.

Design Considerations for Future Models

Given the abundance of data, and the limited resources available to collect them, the ICU of the future will have to concentrate efforts on collecting data directly relevant to ICU performance assessment and improvement. Therefore, prediction models that relate specific outcomes to clinical severity, as assessed by demographic, diagnostic and physiologic variables, and process of care prior to admission, will represent one of the building blocks of this assessment. Several considerations apply to the data that should be collected in the design of improved models.

Choosing the Outcome

The current prediction models were developed as predictors of hospital mortality. The advantage of this endpoint is its obvious clinical relevance and ease of measurement, avoiding the subjective element involved with "softer" endpoints, such as quality of life or functional status. However, it is becoming increasingly clear that hospital mortality is only one of the relevant outcomes of the ICU process, and possibly not the most pertinent in a number of situations.[17] ICU mortality, floor mortality after discharge from the ICU, and ICU readmission rates are all legitimate outcome measures which may be of greater interest in particular contexts. Similarly, resource use indices, such as ICU and hospital length of stay and costs, are also correlated with patient severity. And, of course, quality-adjusted long-term survival or functional outcome may be of much greater importance to both the patient and society.

It is also clear that the episode of critical illness might last beyond the end of the initial hospitalisation. Patients may be discharged prematurely or transferred to another hospital either for more advanced care or chronic ventilator management. Variations in discharge practices bias the interpretation of hospital discharge status, and it might therefore be more reasonable to generate predictions of fixed-time, such as 30- or 90-day, mortality. Too long an interval, however, may introduce additional bias if the prediction model is used to compare populations with fundamentally different baseline mortality rates.

Defining the Population

Models must be developed and validated on populations that reflect the intended target population

(the population on which the model will eventually be used). In particular, though designed to adjust for differences in case-mix, these models can only account for differences up to a point.[18] If a particular subgroup of patients behaves quite differently from the remaining patients but represents only a small fraction of the original cohort, the model may fail to capture the specific nuances of that subgroup's behaviour. If the model is then subsequently applied to a new cohort in which this subgroup represents a dominant fraction, the model may fail to predict accurately. In some instances, such as coronary bypass surgery where survival is excellent despite high acute physiology scores, patients are analysed separately while other groups, such as children, patients admitted to rule out myocardial infarction, and burn victims, are excluded altogether.

No disease- or organ-specific model has rivalled generalised models in the sophistication of the methods used for their development but, as the basic rules in model development and validation become popularised, we will likely witness the emergence of several sophisticated population-specific prediction models as methodological extensions of the generalised models.

The multi-institutional nature of the original databases upon which the models were built favours the minimisation of biases related to inter-institutional variation. However, recent work from the United Kingdom has elegantly demonstrated that, even if the patient cohort appears similar by disease-mix, the country in which the patients are treated can significantly affect the calibration of the model.[19]

Selecting the Predictor Variables

The choice of predictor variables and their relative weights is central to the performance of a prediction model. The more recent prediction models were built using past experience to limit the number of variables before using statistical techniques such as discriminant analysis, factor analysis and stepwise logistic regression to reduce a potentially large number of candidate predictors to a manageable set.

The weights attributed to each of the acute physiologic variables in the derivation of the global score were determined subjectively in APACHE II and SAPS. SAPS II and APACHE III used advanced statistical techniques to derive the weights of each variable.[4,20] The authors of APACHE III report that using continuous weights did not improve the predictive power of the model. Since MPM does not use a score, the issue of attributing weights to variables is moot. Thresholds of positivity had to be defined for many of the physiologic variables and, although this determination carried an element of subjectivity, classification was consistent across data collectors.[3]

Validating the Model

Validation is a complex, yet important issue covering all phases of model development from data acquisition and monitoring to the evaluation of model performance. The particular steps that are taken during the later stages of model development to verify the capacity of the model to adequately describe the target population deserve a brief description. Criteria for the validation of predictive models have been published recently[16,21-23] and are centred on the concepts of independent validation samples, calibration, and discrimination.

An important step in model validation is to determine whether the model performs well on a population with a case-mix similar to the one that was used to develop the model. This usually involves dividing the patient database into development and validation sets. The model is then constructed on the development set and tested on the validation set. This split-sample validation technique, used in the development of all three general models, is only one of many approaches that may be used. The proportion of patients assigned to the validation set is not fixed, but typically ranges from 10% to 50% of the patient database. This assignment can also be either random or clustered by ICU or hospital. Although not used for the predictive tools under discussion, clustered segregation offers the potential advantage of uncovering systematic differences that are cluster-specific.[24] Accordingly, if the performance of a model is comparable from the development set to the validation set, this model can be interpreted as more robust and applied to target populations with more confidence.

Calibration refers to the capacity of the model to predict mortality with accuracy over the entire range of risk. This is an estimate of the "fit" of the model. In binary outcome models (e.g. alive or dead), the familiar R^2 parameter of linear regression does not carry a simple interpretation, and its use as a measure of goodness-of-fit is therefore limited. For such models, Hosmer and Lemeshow have described the statistic (the Hosmer and Lemeshow 'C' statistic) that has been commonly used to estimate goodness-of-fit of the general models.[25] The validation population is divided into groups (usually deciles) of increasing predicted risk, and calibration is estimated by determining the correspondence between predicted and actual outcome within each risk group.

Calibration is especially important if the instrument is to be used for research or quality assurance purposes, where it is desirable that a model predicts reasonably well the approximate number of deaths over the entire range of risk. Good calibration, however, does not imply that the model has good discrimination-the capacity to correctly predict the outcome of individual patients. This property, akin to the well described concepts of sensitivity and specificity, is commonly expressed as a receiver operating characteristic (ROC) curve.[26,27]

The area under the ROC curve (AUC) is a prevalence-independent measure of discrimination that carries the following intuitive meaning: given that the AUC = n (between 0 and 1), a randomly selected non-survivor will have a higher score than a randomly selected survivor 100 x n% of the time.[26] Non-discriminatory models will have an AUC of 0.5 (equivalent to chance prediction) while models that achieve perfect discrimination between survivors and non-survivors will have an AUC of 1 .0. ROC curves also possess the following properties:

(1) they are independent of mortality in a given sample, and are thus an intrinsic property of the predictor (as opposed to sensitivity, specificity, and accuracy, which all depend on the prevalence of the outcome of interest);
(2) different models constructed on the same database can be compared visually;
(3) there are formal analytic techniques to compare ROC curves to each other and to random chance; and
(4) sensitivity and specificity are readily obtained at various thresholds of the predictor. Varying the threshold may be relevant in mortality models, as identification of low- and high-risk populations with a high degree of accuracy is desirable in practical applications.

All future models should report on calibration and discrimination. Reporting on discrimination alone is insufficient.[28] Since risk might be particularly easy or difficult to assess in given populations, it is recommended that the performance of a known system on the particular population used in the study be reported alongside the new predictive model to provide some insight into comparative performance of the proposed system with some standard.

Current Uses of General Scoring Systems

General prediction models are now widely used as a selection and stratification tool at the onset of therapeutic trials in ICU patients. They are also proposed as a tool to evaluate ICU performance by facilitating the comparability of ICU populations across institutions. Finally, they are proposed for use in complementing physicians' evaluations of the prognosis of individual patients, including the need for admission to an ICU and the appropriateness of aggressive therapy (although the extent to which scoring systems are being used in this particular fashion is unclear).

Stratification Tools for Research and Clinical Trials

The proliferation of large clinical trials in intensive care contributed to the need for reliable and valid stratification of critically ill patients to avoid selection bias at randomisation and to evaluate the impact of severity on the effectiveness of therapies. This becomes especially relevant as sufficiently powered trials must enroll patients from ICUs across several countries. There is a concern, however, that stratification and randomisation based on general prediction models may not accomplish this goal in trials of therapies targeted at specific subgroups of patients where more disease-specific predictors may have better discrimination. Consequently, general models are frequently complemented by disease-specific criteria. Furthermore, APACHE II and SAPS II were specifically constructed to provide predictions based on data from the first 24 hours following ICU admission, while the validity of APACHE III was evaluated up to one week into the ICU course. The application of scoring systems to extrapolate risk prediction to the moment of entry of a patient in a clinical trial, which may occur days or weeks after admission to the ICU, has not been explored prospectively.

Evaluation of ICU Performance and Quality Assessment

The concept of ICU performance is complicated and its definition controversial. Performance, once equated with effectiveness of care, now integrates the concepts of appropriateness of care, optimisation of care, quality of care and patient, user and family satisfaction (Fig. 1). The role of prediction models in the assessment of ICU performance is two-fold. First, they may help compare effectiveness of care across ICUs and longitudinally within an ICU or an institution as a quality improvement tool. Second, they may assist in generating resource use comparisons across institutions, appropriately weighted for severity.

Knaus et al[29] used APACHE II to obtain ratios of expected mortality to observed mortality across 13 US hospitals. This study helped originate the concept of using standardised mortality ratios (SMRs), as the ratio of observed over predicted mortality, to compare performance across ICUs. In the APACHE III database, the SMR was found to be significantly different (>2 SD) from 1 in 10 of the 42 ICUs evaluated.[30] Since, by random chance alone, only 2 ICUs should have been outside 2SD,

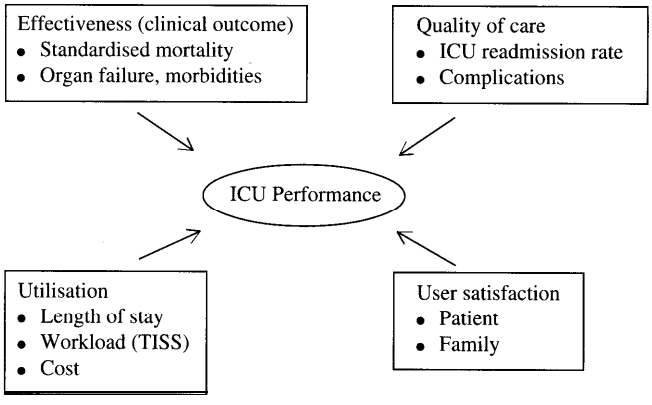

Fig. 1. ICU performance is a complicated concept, with many different perspectives, some of which may occasionally conflict

this study certainly suggests an excessive number of ICUs which were not "average".

However, the use of SMRs as an indicator of performance has been challenged,[31] especially as it bears no relation to ICU performance as assessed by an experienced team that evaluated multiple dimensions of ICU care.[8] It seems prudent to interpret an unfavourable SMR mainly as a guide to examine the particular situation of an ICU with respect to organisation, leadership, staffing, level of technology and communication, and local Do-Not-Resuscitate policies.[32] LeGall et al[4] also reported similar ratios as they related to overall ICU performance in 12 countries but omitted drawing inferences on the comparative quality of care provided in these countries.

It is apparent that SMRs may provide some estimate of effectiveness with respect to some standard, which may be, for example, the mean of all the ICUs being compared. However, it is apparent that good or bad SMRs may originate from either random variation or from a genuinely deviant effectiveness. Time trends of SMRs will assist differentiation between these two fundamentally very different causes of deviation from some standard. Clinicians, administrators, policymakers and mass media will hopefully avoid indiscriminate use of SMRs.

Optimisation of care, the concept of minimising resource use without adversely affecting effectiveness, will also use prediction models to standardise utilisation across ICUs treating populations with discrepancies in disease severity. A preliminary indicator of ICU performance should perhaps be based on some balance between standardised effectiveness and standardised utilisation. An elegant first step towards this goal was reported by Rapoport et al,[33] who describes a standardised utilisation ratio (SUR), akin to an SMR, that benchmarked ICUs with respect to utilisation compared to the average of all ICUs included in the report (Fig. 2).

Another drawback to measuring ICU performance is that there are no gold standards of ICU care, both in terms of effectiveness and utilisation. It is difficult to know what represents the best achievable outcome for an ICU under the best of circumstances. Similarly, it is unfortunately very difficult to know how much resources can be reduced before a decrease in effectiveness can be documented. Although it would seem desirable to redefine SMRs and SURs in terms of absolute best effectiveness and leanest cost not compromising care, this is elusive. Comparison to a "good" ICU, one which has shown consistency in low SMRs or SURs, is more feasible.

Individual Patient Decision-making

Although there is abundant literature on the inaccuracy of clinical judgment in a variety of clinical situations,[34] clinicians appear to be adept at determining the probability of hospital death in critically ill patients.[35,36] In one investigation, clinical judgment performed as well as APACHE III, with clinicians generally having better discrimination but lower calibration: outcome predictions were biased toward over-estimation of the chances of dying.[37] Scoring systems have thus been suggested as adjuncts to clinical judgment in the critical care arena.[38] This suggestion seems intuitive in low-risk patients. However, the predictive accuracy of general models in higher risk patients several days into their ICU courses has received little attention. Accordingly, their value in assisting clinicians to judge whether to limit or withdraw care in these patients deserves further investigation.

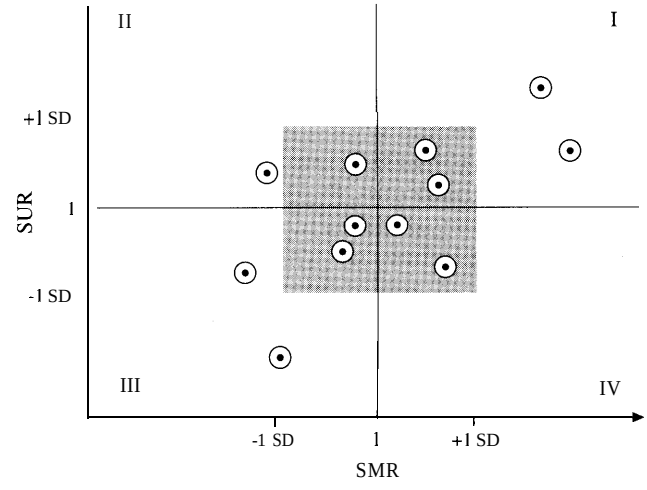

Fig. 2. ICUs (•) can be evaluated with respect to their SMR and SUR when compared to some standard. ICUs which have SMRs or SURs (see text) > 1 SD (outside the shaded box) above the standard are less effective or consume more resources than expected. ICUs > 1 SD below the norm are deemed more effective or more economical. Quadrant III corresponds to ICUs that are both more effective and more resource conscious, an ideal situation. Quadrant I represents the worst combination of a less effective, more expensive ICU. Consistency over time of the position of a particular ICU on this graph would carry more useful information (adaptedfrom Rapoport et al[33]).

Prediction models can theoretically be used prospectively to classify individuals at different degrees of risk. Thus, the scores could be used to augment clinical decision-making regarding ICU triage and withdrawal of support. APACHE II has been suggested as an attempt to rationalise resource use by directing low-risk surgical patients (predicted risk of mortality <10%) away from the ICU.[39] The mortality of such low-risk patients, had they not been admitted to the ICU, however, is unknown.

It would therefore seem more appropriate to use scoring systems for decisions such as appropriate timing of discharge from the ICU,[40] ordering of expensive tests, or institution of resource-intensive therapies. These issues must be formally investigated before recommendations can be formulated, even at the low spectrum of expected mortality. The discriminatory power of the predictive instruments is, however, insufficient to draw inferences

from existing data for such applications. Scoring systems should, at best, complement the judgment of the clinician, not supplant it.

Future Directions

The last decade has seen impressive sophistication in statistical methodologies. Further improvements in regression modeling are likely to include implementation of better fitting techniques. Validation techniques will be further refined and standardised. General models will be complemented, and not substituted, by disease- or process-specific predictors. The breadth of outcomes modeled will continue to expand into morbidity and quality-of-life related variables.

Entirely different classes of models are likely to emerge. Artificial neural networks (ANN) have shown increased predicting power when compared to the current generation of logistic models in special applications,[41,42] and reports have recently been published on the comparative performance of ANNs and general models in predicting ICU and hospital mortality.[43,44]

The concept of ICU performance is likely to draw increasing attention, and standards will probably emerge. Quality assessment and improvement are already major foci of attention at health institutions worldwide. Prediction systems will play an important role in defining quality and describing different aspects of ICU performance. ICU administrators should therefore acquire a more than superficial knowledge of prediction models and implement appropriate methods of data collection and data quality monitoring.

Outcome prediction tools have not yet achieved the goal of accurately predicting outcome in individual patients, but have otherwise reached widespread acceptance as useful instruments for risk stratification and comparison of ICU populations. It is likely that specific contexts will guide the particular instrument to be used.

REFERENCES

1. Knaus W A, Zimmerman J E, Wagner D P, Draper E A, Lawrence D E. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981; 9:591-7.
2. Knaus W A, Wagner D I', Draper E A, Zimmerman J E, Bergner M, Bastos P G, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991; 100:1619-36.
3. Lemeshow S, Teres D, Klar J, Avrunin J S, Gehlbach S H, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993; 270:2478-86.
4. Le Gall J R, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European / North American multicenter study. JAMA 1993; 270:2957-63.
5. Dragsted L, Jorgensen J, Jensen N H, Bonsing E, Jacobsen E, Knaus W A, et al. Interhospital comparisons of patient outcome from intensive care: importance of lead-time bias. Crit Care Med 1989; 17:418-22.
6. Escarce J J, Kelley M A. Admission source to the medical intensive care unit predicts hospital death independent of APACHE II score. JAMA 1990; 264:2389-94.
7. Wagner D I', Knaus W A, Harrell F E Jr, Zimmerman J E, Watts C. Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis. Crit Care Med 1994; 22:1359-72.
8. Zimmerman J E, Shortell S M, Rousseau D M, Duffy J, Gillies R R, Knaus W A, et al. Improving intensive care: observations based on organizational case studies in nine intensive care units: a prospective, multicenter study. Crit Care Med 1993; 21:1443-51.
9. Lemeshow S, Teres D, Pastides H, Avrunin J S, Steingrub J S. A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit Care Med 1985; 13:519-25.
10. Le Gall J R, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. Crit Care Med 1984; 12:975-7.
11. Moreau R, Soupison T, Vauquelin P, Derrida S, Beaucour H, Sicot C. Comparison of two simplified severity scores (SAPS and APACHE II) for patients with acute myocardial infarction. Crit Care Med 1989; 17:409-13.
12. Dominguez-Munoz J E, Carballo F, Garcia M J, de Diego J M, Campos R A, Yanguela J, et al. Evaluation of the clinical usefulness of APACHE II and SAPS systems in the initial prognostic classification of acute pancreatitis: a multicenter study. Pancreas 1993; 8:682-6.
13. Angus DC, Pretto E A, Abrams J A, Safar P. Life supporting first aid training of the lay public for disaster preparedness. Prehosp Disaster Med 1991; 6:257.
14. Castella X, Artigas A, Bion J, Kari A. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. The European / North American Severity Study Group. Crit Care Med 1995; 23:1327-35.
15. Holt A W, Bury L K, Bersten A D, Skowronski G A, Vedig A E. Prospective evaluation of residents and nurses as severity score data collectors. Crit Care Med 1992; 20:1688-91.
16. Hadorn D C, Keeler E B, Rogers W H, Brooks R H. Assessing the performance of mortality prediction models. Santa Monica, CA: RAND, 1993.
17. Petros A J, Marshall J C, van Saene H K. Should morbidity replace mortality as an endpoint for clinical trials in intensive care? Lancet 1995; 345:369-71.
18. Murphy-Filkins R, Teres D, Lemeshow S, Hosmer D W. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit. Crit Care Med 1996; 24:1968-73.
19. Rowan KM, Kerr J H, Major E, McPherson K, Short A, Vessey M P. Intensive Care Society's APACHE II study in Britain and Ireland-I: Variations in case mix of adult admissions to general intensive care units and impact on outcome. BMJ 1993; 307:972-7.
20. Wagner D, Draper E, Knaus W A. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Development of APACHE III. Crit Care Med 1989; 17:S199-203.
21. Angus D C, Pinsky M R. Risk prediction: Judging the judges. Intensive Care Med 1997; 23:363-5.
22. Hadorn D C, Draper D, Rogers W H, Keeler E B, Brook R H. Cross-validation performance of mortality prediction models. Stat Med 1992; 11:475-89.
23. Concato J, Feinstein A R, Holford T R. The risk of determining risk with multivariable models. Ann Intern Med 1993; 118:201-10.
24. Lemeshow S, Hosmer D W Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982; 115:92-106.
25. Hosmer D W Jr, Lemeshow S. Applied Logistic Regression. New York: John Wiley and Sons, 1989.
26. Hanley J A, McNeil B J. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology 1982; 143:29-36.
27. Zweig M H, Campbell G. Receiver Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39:561-77.
28. Linde-Zwirble W T, Angus D C, Chen J, Clermont G, Newbold R C III, Pinsky M R. How should we measure the performance of mortality models? Chest 1996; 110:30S.
29. Knaus W A, Draper E A, Wagner D P, Zimmerman J E. An evaluation of

outcome from intensive care in major medical centers. Ann Intern Med 1986; 104:410-8.

30. Knaus W A, Wagner D P, Zimmerman J E, Draper E A. Variations in mortality and length of stay in intensive care units. Ann Intern Med 1993; 118:753-61.
31. Boyd 0, Grounds M. Can standardized mortality ratio be used to compare quality of intensive care unit performance? Crit Care Med 1994; 22:1706-9.
32. Lemeshow S, Le Gall J R. Modeling the severity of illness of ICU patients. A systems update. JAMA 1994; 272:1049-55.
33. Rapoport J, Teres D, Lemeshow S, Gehlbach S H. A method for assessing the clinical performance and cost- effectiveness of intensive care units: a multicenter inception cohort study. Crit Care Med 1994; 22:1385-91.
34. Dawes R M, Faust D, Meehl P E. Clinical versus actuarial judgment. Science 1989; 243:1668-74.
35. Christensen C, Cottrell J J, Murakami J, Mackesy M E, Fetzer A S, Elstein A S. Forecasting survival in the medical intensive care unit: a comparison of clinical prognoses with formal estimates. Methods Inf Med 1993; 32:302-8.
36. Meyer A A, Messick W J, Young P, Baker C C, Fakhry S M, Muakkassa F, et al. Prospective comparison of clinical judgment and APACHE II score in predicting the outcome in critically ill surgical patients. J Trauma 1992; 32:747-53.
37. McClish D K, Powell S H. How well can physicians estimate mortality in a medical intensive care unit. Med Decis Making 1989; 9:125-32.
38. Muller J M, Herzig S, Halber M, Stelzner M, Thul P. The acute physiology score as a stratification and prognostic criterion in patients in a surgical intensive care ward. Chirurg 1987; 58:334-40.
39. Wagner D P, Knaus W A, Draper E A. Identification of low-risk monitor admissions to medical- surgical ICUs. Chest 1987; 92:423-8.
40. Zimmerman J E, Wagner D P, Draper E A, Knaus W A. Improving intensive care unit discharge decisions: supplementing physician judgment with predictions of next day risk for life support. Crit Care Med 1994; 22:1373-84.
41. Buchman T G, Kubos K L, Seidler A J, Siegforth M J. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit. Crit Care Med 1994; 22:750-62.
42. Doyle H R, Dvorchik I, Mitchell S, Marino I R, Ebert F H, McMichael J, et al. Predicting outcomes after liver transplantation. A connectionist approach. Ann Surg 1994; 219:408-15.
43. Dybowski R, Weller P, Chand R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 1996; 347:1146-50.
44. DiRusso S, Clermont G, Linde-Zwirble W T, Angus D C. The use of an artificial neural network to predict hospital and intensive care mortality. Crit Care Med 1997; 25:A113.