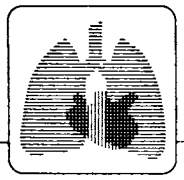

critical care

Predicting Outcome after ICU Admission*

The Art and Science of Assessing Risk

Daniel P. Schuster, M.D.†

(Chest 1992; 102:1861-70)

APACHE = Acute Physiology and Chronic Health Evaluation; MPM = Mortality Prediction Model; ROC = receiver operating characteristic; SAPS = Simplified Acute Physiology Score; TISS = Therapeutic Intervention Scoring System

Many clinicians have concluded that effective and efficient intensive care means restricting such care to only those who need it and will benefit from it.[1,2] More recently, third-party payors and government and other regulatory agencies have also sought to determine whether the quality of intensive care meets acceptable professional standards, in either absolute or relative terms (*eg*, relative to that in other intensive care units [ICUs]).[3,4]

To address these different issues, intensivists and others interested in ICU outcome need a "predictive instrument" or "tool" against which they can judge the performance of their own ICU. This review will emphasize features common to several such instruments, while identifying areas of continuing concern. Two systems in particular, the Acute Physiology and Chronic Health Evaluation (APACHE) system and the Mortality Prediction Model (MPM), will be used as paradigms for discussion. Both systems have evolved through several versions. The latest version of APACHE, for instance, is termed APACHE-III.[5] Both systems use a set of "indicators" or "predictors" to estimate the expected outcome of different patient groups—as a result of or despite intensive care. The expected outcome is calculated and is then compared against actual outcome, with inferences drawn about the appropriateness of an admission or the quality of care. Unfortunately, debate still continues about the accuracy of both these and other available instruments.

Creating a Predictive Instrument

Most studies of ICU outcome, including those involving APACHE or MPM, attempt to identify "risk factors" or other "predictors" of outcome (Table 1). Usually a specific subset of ICU patients is defined, and potential predictors are evaluated for their association, if any, with a particular outcome. Studies involving the adult respiratory distress syndrome (ARDS), septic shock or multiorgan system failure, nontraumatic coma, cardiopulmonary resuscitation, cancer requiring intensive care support, or mechanical ventilation in general are examples in which outcome has been evaluated in this manner. Several recent reviews have summarized many of these reports.[6-10] Since these studies arrive at different sets of specific risk factors with differing degrees of predictive power or accuracy, it is difficult to know which set, if any, has real clinical value.

In general, most ICU predictors can be classified into at least one of five categories (Table 2).[11] For instance, Fowler et al[12] evaluated 47 demographic, clinical, and physiologic variables for their relationship to mortality in ARDS. They found that only low numbers of band forms, low pH, and low HCO_3 level were independently predictive. In essence, however, all three are suggestive of the patient's physiologic response to acute lung injury. Likewise, Menzies et

Table 1—*Elements Important in Creating a Useful Predictive Instrument*

Patient selection
Outcome selection
Variable (predictor) selection
Data collection
Relating predictors to outcome
Validation
Impact evaluation
Updates

Table 2—*Categories of ICU Risk Factors*

Associated illness (chronic health, comorbidities)
Underlying cause and severity of indication for ICU admission
Physiologic derangements (especially if related to underlying cause)
Response to therapy
Complications (especially if unanticipated

*From the Respiratory and Critical Care Division, Department of Internal Medicine, Washington University Medical School, St Louis.
†Established Investigator, American Heart Association and Career Investigator, American Lung Association.
Reprint requests: Dr. Schuster, Pulmonary Division, Campus Box 8052, 660 South Euclid, St. Louis 63110

al[13] found that the preadmission life-style score, a history of cor pulmonale or left ventricular failure, serum albumin concentration (all indicators of chronic health), and FEV_1 (indicating the severity of the underlying disease) predicted outcome in patients with chronic obstructive pulmonary disease requiring mechanical ventilation. Similar groupings can be found in the findings of other studies.

Why should a particular set of predictors be significantly associated with outcome and not another? Will the proposed set of predictors continue to work well when applied to a new group of similar patients? If not, were the two groups really comparable? These and similar questions repeatedly plague the literature of ICU outcomes research. Because of the lack of answers to these questions, relatively few sets of predictors ever get reapplied prospectively, for either research or clinical purposes.

Patient and Outcome Selection

Since all studies of ICU outcome focus on just a subset of the total ICU patient population, there is an implicit assumption that the most powerful set of predictors will be specific for a particular type of patient or problem. Interestingly, more general systems like APACHE and MPM challenge this notion, which remains one of the most controversial aspects of their development. Even these systems, however, have excluded some patient groups. For instance, burn patients, patients under 16 years of age, patients admitted specifically to "rule out" myocardial infarction because of chest pain, and patients admitted to the ICU after coronary artery bypass surgery were all excluded from the development phase of the APACHE-III scoring system (although coronary bypass patients apparently will be analyzed and reported on in the future).[5] Lemeshow et al[14,15] excluded a similar set of patients when they originally developed the MPM. Depending on the population demographics in a particular ICU, these exclusions could render either instrument inappropriate for evaluation or management.

Ideally, both predictor and outcome variables should be unambiguously defined. Wasson et al[16] have recommended that the variables be biological in nature, rather than sociological or behavioral, which will presumably result in less measurement bias. For example, APACHE and MPM both use hospital mortality because it is easily measured and is ultimately, perhaps, the most important outcome. However, ICU mortality might be more directly relevant to a quality assurance evaluation of proper ICU management. Whether systems that use hospital mortality as their main outcome variable can be used to effectively and appropriately evaluate other "secondary" outcomes (*eg*, length of stay or duration of mechanical ventilatory support) is still unknown.

Variable Selection and Data Collection

Although hundreds of variables can be measured in the ICU, only a subset can be evaluated in any one study. Usually a combination of demographic, clinical, and laboratory variables (*eg*, age, sex, primary diagnosis, physical signs, blood gas values, and electrolyte concentrations) are included. In effect, the choice of variables reflects a (sometimes unstated) hypothesis that some subset will be related to outcome. For instance, the developers of the APACHE system[17,18] stated that their original variable list was generated by a "team of experts." This team eventually agreed upon 34 physiologic variables representing the seven major organ systems, which by training and experience they believed were associated with hospital mortality. In many other studies of ICU outcome, a given set of variables may bear no intuitive relationship to the underlying disease, even though the variables are "significantly" associated with outcome, because the initial variable list was not chosen with an explicit hypothesis in mind.

Just as few studies describe the process used to choose the initial variable set, many also fail to fully describe how quality data collection was ensured. For instance, it is common for there to be data missing when measurements are collected retrospectively, since some information might not have been collected at the appropriate times in all patients. Even prospectively collected data often require judgment (*eg*, diagnostic labels that may not meet well-defined criteria). When these methodologic issues are not adequately addressed, confidence in the results is often undermined.

Data Analysis

To reduce the number of variables finally evaluated for their association with a particular outcome, APACHE,[5,17,18] the Simplified Acute Physiology Score (SAPS),[19] and the Therapeutic Intervention Scoring System (TISS)[20,21] all summarize a large set of variables into a single "score." In each case, the score represents the sum of values ("weights") assigned to the chosen predictors. For continuous (*ie*, physiologic) variables, selected ranges are defined. The weights assigned to each range are related to the magnitude of departure from accepted normal values. For instance, in APACHE-III,[5] heart rates between 50 and 99 beats per minute are given a value of 0. As the value for heart rate becomes more and more abnormal (either high or low), weights assigned for scoring purposes increase.

In APACHE-II, SAPS, and TISS, values for the weights vary between 0 and 4, corresponding to clinically intuitive distinctions such as "normal" and

mildly, moderately, and severely abnormal. As with the choice of predictors themselves, systems that employ such semiquantitative estimates for the weights or the range of values over which they apply are potentially biased. Indeed, this was an early criticism of the APACHE system.[14,22] Recently, in APACHE-III, more sophisticated (but not necessarily better) statistical techniques have been used to derive weights for the physiologic variables.[5,23] It is not yet clear from data published so far that the new weighting system is more powerful than that used with the more familiar APACHE-II.

An alternative to producing a score is to use statistical techniques such as univariate or linear discriminant function analysis to reduce the initial variable list, leaving only those variables that are individually associated with outcome for further analysis.[24] This approach was used in developing the MPM. The variable list is then reduced further by keeping only those variables that remain statistically associated with outcome after all variables from the reduced variable list are considered simultaneously.

Although the assumptions and theoretical pitfalls of "scoring" versus statistical modeling have been discussed,[9,23-25] no report has evaluated and compared both approaches when applied to the same data base. Thus, the "best" approach remains controversial.

Relating Predictors to Outcome

Regression techniques are often used to relate predictors to outcome. For multiple-variable (multivariate) analyses, an equation (*ie*, a "model") such as

$$y = b_0 + b_1x_1 + b_2x_2 \ldots b_ix_i \quad (1)$$

describes a linear relationship where y is the dependent, "response," or "outcome" variable; x_1 through x_i are the individual predictors; b_1 through b_i are coefficients (analogous to a slope in a simple linear regression model for one independent variable); and b_0 is a constant (the y intercept in a simple linear regression model). (I have taken some liberties, for the purposes of clarity and simplicity, with the notation used in formal analyses by some statisticians.)

Regression analysis is used to predict outcome for given values of the predictor set. For a univariate model, the correlation between x and y is expressed by calculating Pearson's product moment correlation coefficient, r. The square of r (r^2) is called the "coefficient of determination." When expressed as a percentage, r^2 is usually interpreted to indicate the percentage of variation in y (outcome) "explained" by the variation in x (the predictor). Thus, with a correlation coefficient of 0.8, r^2 would be 0.64, which is interpreted to mean that the observed variation in the predictor explains 64 percent of the variation in outcome. The value for r^2 can also be calculated for multivariate regression models, with a similar interpretation for the overall result.

Table 3—*Predictors and Coefficients for Regression Equations Used in APACHE-II and MPM**

Predictor	Coefficient
APACHE-II[64]	
APACHE-II score†	0.146
Diagnosis responsible for ICU admission	Variable
Admission after emergency surgery?‡	0.603
Intercept	−3.517
MPM[60]	
LOC?‡	2.8902
Emergency admission?‡	1.2671
CPR prior to ICU admission?‡	1.0137
Cancer?‡	0.94131
CRF at time of ICU admission?‡	0.64049
Infection?‡	0.55592
Age	0.047789
Previous ICU admission?‡	0.43946
Heart rate	0.00736
Surgical service?‡	−0.37987
Systolic BP	−0.04591
(Systolic BP)^2	0.000116
Intercept	−2.9678

*In both systems, the outcome variable is the probability of death. LOC = level of consciousness (*ie*, presence or absence of coma or stupor); CPR—cardiopulmonary resuscitation prior to ICU admission; CRF = history of chronic renal failure? BP = blood pressure. The values for each predictor and their coefficients are used to calculate the logit value used in equation 2.

†Sum of physiology, age, and chronic health points (weights).

‡Categorical variable (the listed value if true; a value of 0 if not true).

Predictors and their coefficients for APACHE-II and MPM are shown in Table 3. Comparable equations have been developed for APACHE-III but are proprietary information. The APACHE-II system[17,18] uses three predictors: the APACHE-II score, the reason (diagnosis) for ICU admission, and whether the admission occurred after emergency surgery. The APACHE-II score, in turn, represents the sum of weights assigned to 12 physiologic variables, to age, and to a value for chronic health. (Analogously, the APACHE-III equations also include three predictors: the APACHE-III score, diagnosis, and treatment location prior to ICU admission. The APACHE-III score, in turn, represents the sum of weights assigned to 17 physiologic variables, to age, and to chronic health.) In contrast, MPM uses 12 predictors, culled from an original list of 377 variables; this list was reduced by a process of linear discriminant function and stepwise regression analyses.[14,15]

When the predictors are not linearly related to the outcome variable, transformations can be performed, so that models like that described in equation 1 can be used. A commonly used method is logit transformation. A logit is the natural logarithm of the ratio of two probabilities. For instance, the probability of dying (Pr) could be defined as:

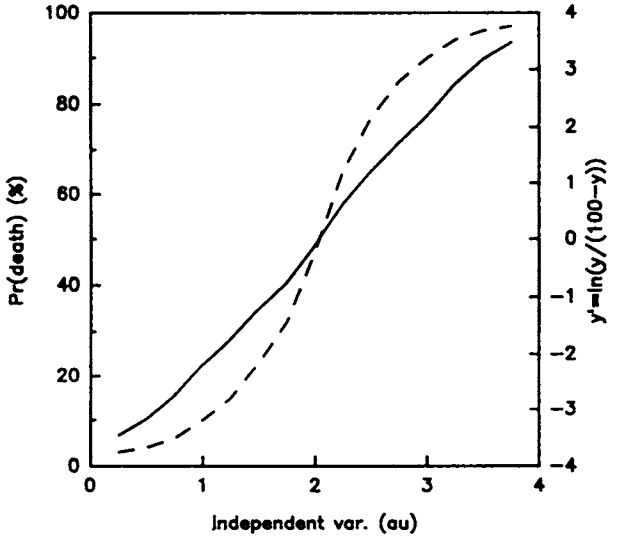

FIGURE 1. Example of how logit transformation can change a sigmoidal relationship between two variables into a linear one. In this case, y = probability (Pr) of death. Values for the independent variable (var) might represent weights assigned to specific ranges for the actual values for that variable.

$$\Pr = e^{\text{logit}}/(1 + e^{\text{logit}}), \quad (2)$$

where logit = y in equation 1.

Logistic transformation has the property of transforming S-shaped relationships between two variables into linear ones (Fig 1). Intuitively, an S-shaped relationship is appropriate for many outcome prediction models because a range of x values might exist for which the outcome (*eg*, the risk of death) is near zero (the normal range), another range of values for which the risk is maximal and unchanging (severely abnormal values), and a final range for which the risk changes as the value changes. Multiple regression analyses that use logit transformations are referred to as multiple logistic regression analyses.

Regression equations for APACHE-II and MPM (Table 3) in effect reveal the implicit hypothesis for each model. For APACHE, the hypothesis is that ICU outcome must be related to the severity of the acute illness for which the patient was admitted, with "severity" quantified as the "acute physiology score," as modified by age, the patient's chronic health before developing the acute illness, and the patient's underlying diagnosis. In contrast, even though the final predictor list in the MPM was generated strictly by statistical techniques, it is interesting to note that the list includes many conditions which are known, either intuitively (*eg*, emergency vs nonemergency admission, readmission) or from other studies (cardiopulmonary resuscitation, diagnosis of cancer, or chronic renal failure), to affect ICU prognosis. Perhaps, then, it is not surprising that despite the differences in the two sets of predictors, each system can predict ICU outcome (see next section).

VALIDATING THE INSTRUMENT

Predictive models can be validated by comparing the model's predictions, derived from a "training" data base, against the actual observed outcome in a test set. The test group can be the group from which the original model was derived (compared by so-called cross-validation techniques), a portion of the group from which the data were originally collected but from which data were not used to develop the model (the split-sample method), or a completely new sample.[24] The last approach, although preferred because it minimizes potential bias, is expensive and often impractical. When a validation is performed by a different set of investigators on a new group of patients, the results are rarely as good as in the original report because it is difficult to exactly duplicate the methods of the original developers.[26] Furthermore, the new predictions include natural variations (noise) not included in the original development.

A variety of statistics are used to compare predictions with actual outcomes, including summary statistics (such as the coefficient of determination), classification rates, area under receiver operating characteristic (ROC) curves, and goodness-of-fit statistics. The coefficient of determination, as noted previously, is interpreted to indicate the percentage of variation in the outcome variable described by the model as a whole (the set of independent variables). This property is sometimes referred to as "explanatory power." A high value for r^2 does not mean that the predictive model is free of significant bias. Likewise, a low value for r^2 does not mean that the model's predictions are inaccurate—only that uncertainty around any given prediction is great.

A model can be evaluated for its ability to discriminate between patients who are likely to either die or not die by classifying them into two groups: "predicted to die" and "predicted to not die." Since the probability of death has continuous values from 0 percent to 100 percent, a cut-point must be chosen to classify the patient into one category or the other. For instance, if a cut-point of 50 percent is chosen, then any patient with a predicted chance of dying greater than 50 percent would be classified as "predicted to die." After classifying a group of patients, the result can be compared against actual outcomes in a 2×2 table, allowing familiar descriptive statistics to be calculated (Table 4). Interestingly, when a 50 percent cut-point is used, virtually all proposed ICU predictive instruments have a false classification rate of approximately 10 percent to 15 percent.[22,27,28] Thus, it has been said,[27] all systems are approximately equivalent, and none is good enough for individual prediction.

A problem with this interpretation is revealed by the following example. Assume that for 100 patients,

Table 4—*Hypothetical 2 × 2 Classification Table Using a Cut-point of 50%**

	Predicted	
	Died	Lived
Observed		
Died	10	0
Lived	10	80

*Sensitivity = 10/10 = 100%; specificity = 80/90 = 89%; False classification rate = 10/100 = 10%; predictive value of a positive test = 10/20 = 50%.

20 are found with a 50 percent chance of dying. With a 50 percent cut-point, all such patients would be classified as "predicted to die." (Table 4). Assume further that only ten actually die. The result would be a false classification rate of 10 percent. However, in terms of the actual prediction, the model was exactly correct (*ie*, 50 percent of the patients were predicted to die, and 50 percent died). This example reveals an important distinction between discrimination (*ie*, how well does the model discriminate between patients who will die or live?) and calibration (*ie*, how closely do predictions correlate with actual outcome across the entire range of risk [from 0 percent to 100 percent]?). Furthermore, true and false classification rates are dependent on the mortality rate of the sample.[24] Thus, conclusions about the accuracy of predictive models based on true or false classification rates are potentially misleading.

The relative importance of discrimination versus calibration depends upon how the predictive instrument is to be used. For research or quality assurance purposes (group comparisons), calibration is especially important. For decisions about individual patients, both descriptors are relevant. When faced with a specific patient, we want to know, as accurately as possible, exactly what are the chances of dying (calibration). Having determined the probability, we then need to decide whether to act upon that decision (*ie*, to treat or not to treat). In essence, we have chosen a cut-point and have classified the patient into one of two possible categories (discrimination).

Of course, the choice of 50 percent as a cut-point for classification is arbitrary. To reduce the number of false-positive (*ie*, patients classified as predicted to die who actually live), a higher (more specific) cut-point should be used, say 90 percent. In so doing, new figures for sensitivity and specificity would be calculated. Indeed, any number of cut-points could be chosen, each time calculating new sensitivity and specificity figures. By plotting the pairs of true-positive (sensitivity) and true-negative (1 − specificity) rates at each cut-point, a ROC curve is developed (Fig 2).[29,30] This curve demonstrates the continuous trade-off between sensitivity and specificity. As the model

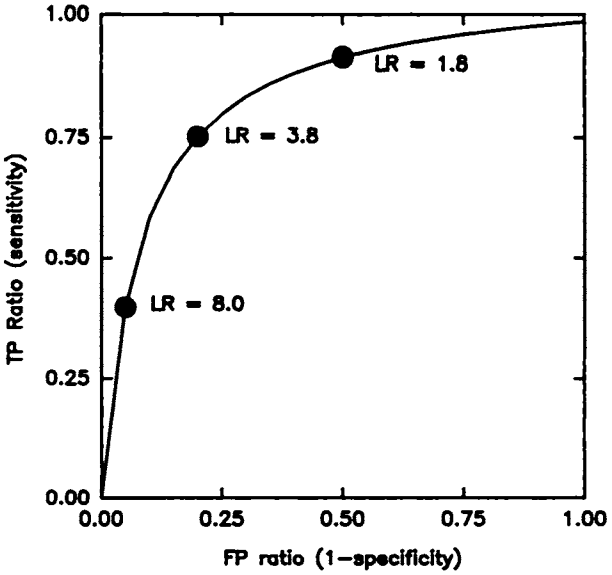

FIGURE 2. Hypothetical receiver operating characteristic curve showing three different cut-points. Of the three points shown, the one with the highest specificity and likelihood ratio (LR) has the lowest sensitivity. The choice of the appropriate cut-point should include the cost of making decisions based on classifying patients according to that cut-point. TP = true-positive; FP = false-positive.

becomes more "perfect" (*ie*, able to achieve 100 percent sensitivity and specificity regardless of cut-point), the area under the ROC curve trends toward 1.0; as the performance of the model becomes more random, the area under the curve trends toward 0.5.

McNeil et al[29] have discussed the factors that should be considered when choosing a cut-point. When the costs (in all senses of this word) of acting upon a positive prediction are high, the chance of a false positive should be minimized (*ie*, high specificity should be required). An applicable circumstance in the ICU would be if therapy were to be withdrawn based upon a prediction of death. In contrast, if the costs are trivial, one should maximize true positives by increasing sensitivity. When the relative costs are unknown, an intermediate cut-point, such as 50 percent, is often used. Unfortunately, the rationale for choosing a given cut-point is rarely given in studies of ICU outcome that report these statistics.

At all points along the ROC curve, the slope of the curve is the ratio of true positives and false positives, also known as the likelihood ratio. The likelihood ratio can be a useful calculation, because Bayes' theorem indicates that the odds of an event occurring are equal to the prior probability multiplied by the likelihood ratio.[31] This potential application of Bayes' theorem to ICU decision-making is discussed in the next section.

Model calibration can best be described by comparing the predicted risk to the actual outcome over a stratified range of potential risks. A graph of these comparisons results in a calibration curve (Fig 3). Where discrepancies between observed and actual

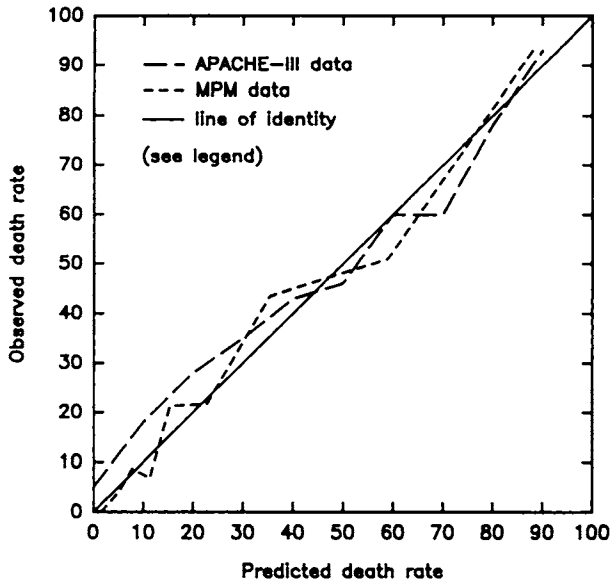

FIGURE 3. Estimated calibration curves of observed versus predicted death rates for APACHE-III and MPM. The APACHE-III data were estimated from the validation data set shown in Figure 4 of reference 5. Although the predictions for these data did not use the non-physiology score coefficients for the entire APACHE-III data base, the calibration curve shown in Figure 7 of reference 5 (as an alternative) includes data from both the development and the validation halves of the data base. The MPM data are derived from Table 7 of reference 14. These curves are shown for illustrative purposes only and are not meant to indicate the accuracy of either instrument in other or later versions.

outcomes occur, simple inspection of the curve does not reveal whether the discrepancy is "important" or clinically "significant." Although Lemeshow and Hosmer[32] have developed a "goodness-of-fit" statistic to describe how faithfully the predicted and observed results compare overall, the clinical impact of discrepancies must still be evaluated separately. Here indeed it would be useful to understand not only how often patients were misclassified but also why. Such information is difficult to glean from most published studies of ICU outcome.

Predictor sets from most studies of ICU outcome have not been validated by an independent set of investigators. Without question, the most carefully studied system is APACHE-II, and the results have been mixed: sometimes the original predictions of the APACHE developers have been supported, and sometimes they have not.[33-41] The negative studies obviously have raised concerns about accepting the APACHE-II results as a standard for ICU evaluation.[27,42]

It is not always clear why APACHE-II has at times failed to achieve the same level of performance as originally reported. Certainly one possibility is that APACHE-II is not in fact an accurate predictor of ICU outcome, at least for certain groups of patients. Another reason, as noted earlier, is that strict attention may not have been paid to duplicating the methods of the original studies.[26] Knaus et al[5] have also pointed out that some investigators have simply but incorrectly reported only acute physiology scores, not predicted probabilities of death, to compare patient populations with disparate diseases (see next section).

On the other hand, these "negative" studies have revealed several potentially important sources of bias in using any predictive instrument. "Lead-time bias" occurs when a model is applied after certain assumptions about the model are no longer true. In the case of APACHE, lead-time bias results when patients are partially treated prior to ICU admission (*eg*, in the emergency room, the operating room, another ICU, or another hospital).[34,36] Doing so violates one of the APACHE premises, namely, that the score reflects the physiologic severity of the underlying cause for ICU admission, independent of any treatment that would be given subsequently. The new APACHE-III system includes coefficients for treatment location, which are meant to adjust for this problem, but how well they perform in this regard has not been reported yet.

Another source of error is selection bias (*ie*, patients are selected for evaluation by criteria that were not in

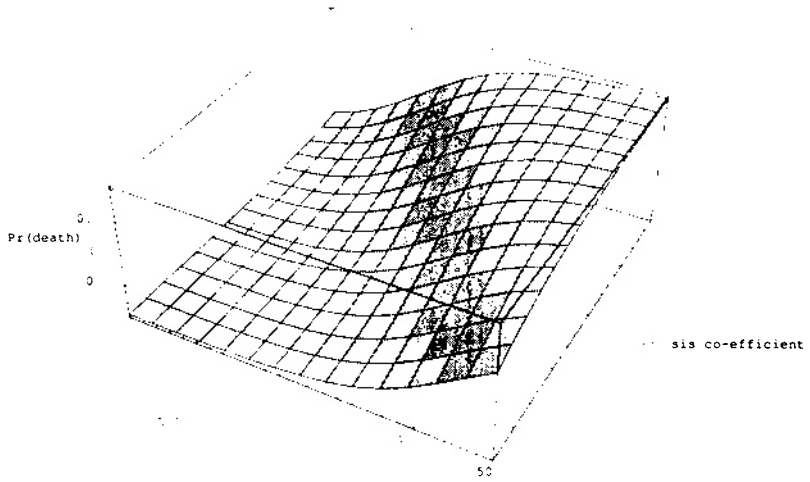

FIGURE 4. Surface contour plot relating probability of death, APACHE-II score, and underlying diagnosis (as quantified by the coefficients given in reference 18). No emergency surgery was assumed in developing this plot. Note that for low to intermediate values of the APACHE-II score, most of the change in estimating the probability (Pr) of death is related to the underlying diagnosis.

fact used to develop the original model). Again, using APACHE as an example, the predictive equation for the study was developed using the one chief reason for ICU admission as one of the independent variables. Estimating a patient's risk of death by using one of the other diagnostic labels, even if relevant to that patient, is an incorrect step and can lead to erroneous predictions. Since different users may choose different diagnoses, they may arrive at different predicted risks of dying. Indeed, this was one of the original motivations for SAPS, the simplification of APACHE.[19] As an example, consider a bone marrow transplantation patient with a stuporous mental state, neutropenia, diffuse pulmonary infiltrates, and hypotension. Is the primary reason for ICU admission noncardiogenic pulmonary edema, aspiration, or sepsis (all diagnostic possibilities in APACHE-II)? Since each is associated with a different partial regression coefficient for the diagnosis predictor, the calculated risk of death will be different (ranging from 44 percent to 53 percent, assuming a physiology score of 24 points) (Fig 4). On the other hand, any one group of experts is likely to consistently apply the same set of (unstated) diagnostic criteria. This may be one reason why validation studies by split-sample techniques have been favorable, unlike studies by other investigators in a new group of patients.

The MPM model may be less susceptible to lead-time bias and selection because the variables included in its predictive equation are not as easily affected by acute therapy and are independent of the underlying diagnosis. On the other hand, MPM may still be dependent on case mix, since the predictive equation represents the patient population of a specific center.[43]

Several predictive instruments have been compared to one another and to "clinical judgment."[19,22,44-48] In general, despite some small but admittedly statistically significant differences, the instruments have fared comparably.[49] In the most recent reports, APACHE-III shows a statistically significant improvement over its predecessor, APACHE-II. Whether this difference is of clinical importance is not yet clear. When the predictive models are compared with clinical judgment, physicians seem to do a somewhat better job at discrimination, while the models are in general better calibrated over the range of stratified risk, especially in the intermediate ranges, where there is usually the greatest clinical uncertainty.[44]

Impact of the Instrument

Predictive instruments can be used for research purposes to show that the groups being studied (*eg*, a test group and a placebo group) are similar with respect to baseline severity of disease. Presumably, similar predicted risks would imply similar severity of disease at baseline. However, a common error in using APACHE (and probably other scoring systems as well) in clinical research studies has been to simply report the raw scores for the different study groups, when those groups include different reasons for ICU admission. Although the scores may be statistically similar, they only imply similar severity of illness within a very specific diagnostic group. When multiple diagnoses are involved, it is essential to compare the groups by comparing the predicted risks of death.

For quality assurance purposes, not only must the system be well calibrated over the range of risk (especially in the intermediate range, where quality might actually have an impact on outcome), but it must also be appropriate for the patient population being evaluated, (*ie*, the case mix of the ICU must be well represented in the study used to generate the predictive model). Obviously, if a system is to be used for general purposes, the patient data base must be very large and broad. With a data base of more than 17,000 patients, APACHE-III is particularly impressive in this regard.[5] However, if APACHE is going to be used for quality assurance, then strict adherence to rules for deciding the ICU admitting diagnosis will be essential, since APACHE seems to be especially sensitive to the classification of admitting ICU diagnosis (Fig 4). Because of this problem, a model like MPM, which is independent of diagnosis, is attractive. However, until it can be shown that MPM is independent of case mix, it would be premature to accept MPM or to reject APACHE for this reason alone.

Even if a particular instrument could predict outcome with great accuracy, it would still be important to show that the information could be used to identify problems within a particular ICU, that behavior could be changed, and that subsequent patient outcome could be favorably affected. For instance, Knaus et al[50] suggested that differences in ICU management structure partially explained the variation in mortality among 13 ICUs. However, their study did not determine whether changes in structure would actually rectify these differences. Systems that predict mortality do not evaluate other, perhaps equally important, aspects of "quality" care. Whether mortality rates track these other aspects remains to be shown. The new APACHE-III system does estimate some other parameters (expected TISS points as a measure of "efficiency," expected length of stay in the ICU, expected use of pulmonary artery catheterization). These new aspects offer great promise, but data about them from the APACHE-III data base have yet to be reported.

The predicted mortality, if it becomes the standard against which any one ICU will be judged, represents the average mortality from the ICUs participating in the study. This "average" standard might actually be lower than what is expected by our profession or our

society. It is important that performance also be evaluated in terms of some absolute standard, as difficult as this might be to determine.

The most controversial potential use for predictive models is to use them to affect individual case management, especially for issues of withholding or withdrawing treatment.[51,52] These twin problems apply to the extremes of predicted risk. On average, at the extremes, clinical judgment seems to be equal to, if not better than, predictive models.[44,53] However, variability in clinical estimates is great,[54,56] and the tendency to overestimate risk clinically can be significant.[57] As a result, Knaus et al[5] have suggested that the information provided by predictive models could still be useful, because they should be more reliable (*ie*, risk estimates for a given type of patient would be reproducible, regardless of the clinician's expertise or experience) and more credible (because they are derived from a data base that is much larger than any one clinician's experience).

Even so, most clinicians are loath to withhold a therapeutic trial from a patient once he or she has been admitted to the ICU, regardless of their own predictive accuracy or that of any "objective" predictive instrument. This time-honored practice recognizes the inherent uncertainty in all predictions.[49] By definition, no system will ever be able to predict an unpredictable outcome, either favorable or unfavorable, but it is precisely the unpredictable that so heavily influences outcome and cost.[58] Therapeutic trials seek to bolster the original prediction by showing that the patient fails to improve despite the best of care. Implicity, both MPM[15] and APACHE-III[5] have recognized this facet of clinical practice and have incorporated different predictive equations into their systems depending on the duration of ICU care. There have been no comparisons of these systems with these adjustments, and the actual impact of time on the predictions is as yet still inadequately described.

It is possible that predictive instruments could be used to guide individual patient management by adding to clinical judgment, rather than replacing it. For instance, both a clinical prediction and an "objective" prediction could be entered into yet another logistic regression equation to predict mortality.[44] Alternatively, clinical judgment could be used to set a "pretest probability" of dying. Then, using Bayes' theorem in a fashion analogous to its use for the interpretation of ventilation-perfusion scanning in the diagnosis of pulmonary emboli,[59] the prediction from one of the instruments could be used to adjust the pretest probability, yielding a final predicted probability of dying.[53,60] Because of its general familiarity, this process has great appeal, but its application to ICU decision making has not been evaluated in any way.

Even if it were indeed possible to correctly identify very-low- and very-high-risk patients, and even if therapy to such patients were withheld or withdrawn, it is not clear that these maneuvers would have a great impact on ICU operation or hospital economics.[27] The numbers of patients at the truly high end of predicted risk is generally small.[5,14,15] For instance, in the APACHE-III study, fewer than 10 percent of the patients studied had a predicted risk greater than 90 percent. To what extent any system could be used to minimize expenditures in this group remains to be shown, since the relationship between severity of illness and resource use is nonlinear[27,61] (*ie*, expenditures actually are lower for some high-risk patients because physicians have already limited resource use). Conversely, although the number of low-risk patients in most ICUs is high, these patients consume relatively few resources.[62] Furthermore, it remains to be shown that despite a lack of intervention, these patients would still do well if denied ICU admission.

Updates

Many sets of predictors fail to perform well in subsequent studies.[44] In addition to the issues already discussed, changes in therapy may have altered the nature of the ICU population (*eg*, patients being admitted with new complications not previously encountered in that group) or may have altered the prognosis of either the acute problem or the underlying disease. This problem can be addressed only by periodic updates of the data base, either verifying previous results or modifying the predictors and/or their weights accordingly. To date, only the developers of APACHE-III[5] seem to have the infrastructure necessary to make such a formidable undertaking possible.

Conclusions

It is quite clear that clinicians, administrators, and regulators would like an accurate predictive instrument against which to judge and evaluate clinical effectiveness, efficiency, and quality. Numerous isolated studies that report predictors of ICU outcome for selected patient groups have been of little value because they have not been appropriately validated. Nevertheless, *in toto*, they demonstrate the importance of chronic health, the nature and severity of the acute illness, the response to therapy, and the impact of unexpected complications on outcome.

Unfortunately, even for well-studied systems like APACHE and MPM, we cannot yet answer definitely whether any one instrument accurately predicts ICU outcome, whether any one system is better than another, or whether any one system is better than clinical judgment. Based on limited information, the differences in performance between APACHE-II and MPM and between APACHE-II and APACHE-III

appear to be relatively small. However, the extremely large data base, the apparent attention to statistical detail, and the intent to update the data base on a regular basis suggest that predictions from APACHE-III should be significantly more reliable and credible than those from earlier versions or routine clinical judgment. Additional insights about the performance of APACHE-III, especially with respect to impact on ICU management, await additional reports.

Obviously, any system can be misused or misinterpreted. Comparing groups by comparing scores is appropriate only when the groups are homogenous with respect to admitting ICU diagnosis. In all other cases, groups must be compared by reporting the average predicted mortality for all patients within the group. To arrive at these estimates, careful attention must be paid to following the rules of each instrument and to identifying sources of bias, especially lead-time and selection biases.

Even if any one instrument meets all expected standards of accuracy, it remains to be shown that these tools can be used to identify problems, that the problems can be corrected, and that the corrections will have a favorable impact on ICU or patient management. It is not premature to want to use one of these instruments, but the user should have a clear idea of what it is that he wishes to use it for. As stated so eloquently in an anonymous editorial about these tools:[63] ". . . probability is only one factor to be taken into account when making a clinical decision. Statistics should be used as the drunken man uses the lamp post—for support rather than illumination."

ACKNOWLEDGMENTS: The author wishes to gratefully acknowledge the helpful comments of Drs Clay Dunagan and Michael Kahn, and the statistical review provided by Dr Michael Province.

A more complete list of references, especially pertaining to the outcome of specific ICU patient populations, is available from the author.

REFERENCES

1 Thibault GE, Mulley AG, Barnett GO, Goldstein RL, Reder VA, Sherman EL, et al. Medical intensive care: indications, interventions, and outcomes. N Engl J Med 1980; 302:938-42

2 Consensus Development Conference. National Institutes of Health, Bethesda, Md, March 7-9, 1983

3 Hoyt JW, Leisifer DJ, Rafkin HS. Critical care units. In: Wenzel RP, ed. Assessing quality health care: perspectives for clinicians. Baltimore: Williams & Wilkins, 1992:267-96

4 Iezzoni LI. Severity standardization and hospital quality assessment. In: Couch JB, ed. Health care quality management for the 21st century. Tampa, Fla: Hillsboro Printing Co, 1991; 177-234

5 Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991; 100:1619-36

6 Knaus WA. Prognosis with mechanical ventilation: the influence of disease, severity of disease, age, and chronic health status on survival from an acute illness. Am Rev Respir Dis 1989; 140:S8-13

7 Zimmerman JE, Knaus WA. Outcome prediction in adult intensive care. In: Shoemaker WC, Ayres S, Grenvik A, Holbrook P, Thompson WL, eds. Textbook of critical care. Philadelphia: WB Saunders, 1989:1147-65

8 Raffin TA. Intensive care unit survival of patients with systemic illness. Am Rev Respir Dis 1989; 140:S28-35

9 Teres D, Lemeshow S. Evaluating the severity of illness of critically ill patients. In: Shoemaker WC, ed. Diagnostic methods in critical care. New York: Marcel Dekker, 1987; 1-17

10 Civetta J. Prediction and definition of outcome in a cost-sensitive era. In: Civetta JM, Taylor RW, Kirby RR, eds. Intensive and critical Care. Philadelphia: J B Lippincott, 1988:1677-98

11 Hudson LD. Survival data in patients with acute and chronic lung disease requiring mechanical ventilation. Am Rev Respir Dis 1989; 140:S19-24

12 Fowler AA, Hamman RF, Zerbe GO, Benson KN, Hyers TM. Adult respiratory distress syndrome: prognosis after onset. Am Rev Respir Dis 1985; 132:472-78

13 Menzies R, Gibbons W, Goldberg P. Determinants of weaning and survival among patients with COPD who require mechanical ventilation for acute respiratory failure. Chest 1989; 95:398-405

14 Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit Care Med 1985; 13:519-25

15 Lemeshow S, Teres D, Avrunin JS, Gage RW. Refining intensive care unit outcome prediction by using changing probabilities of mortality. Crit Care Med 1988; 16:470-77

16 Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodological standards. N Engl J Med 1986; 313:793-99

17 Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE—acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981; 9:591-97

18 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985; 13:818-29

19 Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. Crit Care Med 1984; 12:975-77

20 Cullen DJ, Civetta JM, Briggs BA, Ferrara LC. Therapeutic intervention scoring system: a method for quantitative comparison of patient care. Crit Care Med 1974; 2:57-60

21 Keene AR, Cullen DJ. Therapeutic intervention scoring system: update 1983. Crit Care Med 1983; 11:1-3

22 Lemeshow S, Teres D, Avrunin JS, Pastides H. A comparison of methods to predict mortality of intensive care unit patients. Crit Care Med 1987; 15:715-22

23 Wagner D, Knaus W, Bergen M. Statistical methods (APACHE-IIII study design). Crit Care Med 1989; 17(suppl 2):S194-98

24 Ruttimann UE. Severity of illness indices: development and evaluation. In: Shoemaker WC, Ayres S, Grenvik A, Holbrook P, Thompson WL, eds. Textbook of critical care. Philadelphia: WB Saunders, 1442-46

25 Teres D, Avrunin JS, Lemeshow S. Severity-of-illness modeling. In: Intensive care medicine. Rippe JM, Irwin RS, Alpert JS, Fink MP, eds. Boston: Little, Brown Co, 1989:1953-59

26 Charlson ME, Ales KL, Simon R, MacKenzie CR. Why predictive indexes perform less well in validation studies. Arch Intern Med 1987; 147:2155-61

27 Civetta JM, Hudson-Civetta JA, Nelson JD. Evaluation of APACHE II for cost containment and quality assurance. Ann Surg 1990; 212:266-76

28 Kirby RR, Civetta JM. Critical care. In: Brown D, ed. Risk and outcome in anesthesia. Philadelphia: JB Lippincott, 1988; 184-212

29 McNeil BJ, Keeler E, Adelstein SJ. Primer of certain elements of medical decision making. N Engl J Med 1975; 293:211-15
30 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36
31 Patterson RE, Horowitz SP. Importance of epidemiology and biostatistics in deciding clinical strategies for using diagnostic tests: a simplified approach using examples from coronary artery disease. J Am Coll Cardiol 1989; 13:1653-65
32 Lemeshow S, Howmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982; 115:92-106
33 Fedullo AJ, Swinburne AJ, Wahl GW, Bixby KR. APACHE II scoring and mortality in respiratory failure due to cardiogenic pulmonary edema. Crit Care Med 1988; 16:1218-21
34 Dragsted L, Jorgensen J, Jensen NH, Bonsing E, Jacobsen E, Knaus WA, et al. Interhospital comparisons of patient outcome from intensive care: importance of lead-time bias. Crit Care Med 1989; 17:418-22
35 Cerra FB, Negro F, Abrams J. APACHE II score does not predict multiple organ failure of mortality in postoperative surgical patients. Arch Surg 1990; 125:519-22
36 Escarce JJ, Kelley MA. Admission source to the medical intensive care unit predicts hospital death independent of APACHE II score. JAMA 1990; 264:2389-93
37 Teskey RJ, Calvin JE, McPhail I. Disease severity in the coronary care unit. Chest 1991; 100:1637-42
38 Schein M, Gecelter G. APACHE II score in massive upper gastrointestinal hemorrhage from peptic ulcer: prognostic value and potential clinical applications. Br J Surg 1989; 76:733-36
39 Maher ER, Robinson KN, Scoble JE, Farrimond JG, Browne RG, Sweny P, et al. Prognosis of critically-ill patients with acute renal failure: APACHE II score and other predictive factors. Q J Med 1989; 269:857-66
40 Dobkin JE, Cutler RE. Use of APACHE II classification to evaluate outcome of patients receiving hemodialysis in an intensive care unit. West J Med 1988; 149:547-50
41 Chang RWS, Jacobs S, Ffarcs J, Lee B, Pace N. Predicting deaths among intensive care unit patients. Crit Care Med 1988; 16:34-42
42 Civetta JM. Scoring systems: do we need a different approach? Crit Care Med 1991; 19:1460-61
43 Teres D, Lemeshow S, Avrunin JS, Gage RW. Multi-center validation of the admission mortality prediction model [abstract]. Crit Care Med 1988; 16:412
44 McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? Med Decis Making 1989; 9:125-32
45 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; 17:409-12
46 Schafer JH, Maurer A, Jochimsen F, Emde C, Wegscheider K, Arntz HR, et al. Outcome prediction models on admission in a medical intensive care unit: do they predict individual outcome? Crit Care Med 1990; 18:1111-17
47 Kruse JA, Thill-Baharozian MC, Carlson RW. Comparison of clinical assessment with APACHE II for predicting mortality risk in patients admitted to a medical intensive care unit. JAMA 1988; 260:1739-42
48 Chang RWS, Lee B, Jacobs S, Ffarcs J, Lee B. Accuracy of decision to withdraw therapy in critically ill patients: clinical judgment versus a computer model. Crit Care Med 1989; 17:1091-97
49 Sneff M, Knaus WA. Predicting patient outcome from intensive care: a guide to APACHE, MPM, SAPS, PRISM, and other prognostic scoring systems. J Intensive Care Med 1990; 5:33-52
50 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An evaluation of outcome from intensive care in major medical centers. Ann Intern Med 1986; 104:410-18
51 Knaus WA, Rauss A, Alperovitch A, Le Gall JR, Loirat P, Patios E, et al. Do objective estimates of chances for survival influence decision to withhold or withdraw treatment? Med Decis Making 1990; 10:163-71
52 Knaus WA, Wagner DP, Lynn J. Short-term mortality predictions for critically ill hospitalized adults: science and ethics. Science 1991; 13:17-35
53 Brannen AL, Godfrey LJ, Goetter WE. Prediction of outcome from critical illness: a comparison of clinical judgment with a prediction rule. Arch Intern Med 1989; 149:1083-86
54 Perkins HS, Jonsen AR, Epstein WV. Providers as predictors: using outcome predictions in intensive care. Crit Care Med 1986; 14:105-110
55 Poses RM, Bekes C, Copare FJ, Scott WE. The answer to "what are my chances, doctor?" depends on whom is asking: prognostic disagreement and inaccuracy for critically ill patients. Crit Care Med 1989; 17:827-33
56 Pearlman RA. Variability in physician estimates of survival for acute respiratory failure in chronic obstructive pulmonary disease. Chest 1987; 91:515-21
57 Detsky AS, Stricker SC, Mulley AG, Thibault GE. Prognosis, survival and the expenditure of hospital resources for patients in an intensive-care unit. N Engl J Med 1981; 305:667-72
58 Zook CJ, Moore FD. High-cost users of medical care. N Engl J Med 1980; 302:996-1002
59 Kelley MA, Carson JL, Palevsky HI, Schwartz JS. Diagnosing pulmonary embolism: new facts and strategies. Ann Intern Med 1991; 114:300-06
60 Chang RWS. Individual outcome prediction models for intensive care units. Lancet 1989; 143-47
61 Rapoport J, Teres D, Lemeshow S, Avrunin JS, Haber R. Explaining variability of cost using a severity-of-illness measure for ICU patients. Med Care 1990; 28:338-48
62 Oye RK, Bellamy PE. Patterns of resource consumption in medical intensive care. Chest 1991; 99:685-89
63 TPN and APACHE [editorial]. Lancet 1986; 1:1478