Comparing ICU Populations: Background and Current Methods

J.E. Zimmerman, E.A. Draper, and D.P. Wagner

Key Messages

- Intensive care unit (ICU) scoring systems provide case mix adjusted benchmarks which can be used to compare mortality rates among hospitals
- Differences between observed and predicted hospital mortality are often due to factors other than the quality of ICU therapy
- Patient characteristics that have a significant influence on hospital mortality are not fully accounted for in current ICU scoring systems
- The prognostic impact of identical patient risk factors varies within the healthcare systems of different regions and countries
- Future ICU prognostic systems will be more complex and more accurate; they will require automated data collection and periodic adjustment for changes in therapy over time

Several years ago an international course was developed to assist in the planning and optimal use of intensive care. The course was organized in response to rising costs and increasing demands for ICU services in the host country, and was attended by ICU physicians and government officials representing national and regional health funding agencies. The participants described and compared patient demographics, clinical characteristics, and ICU resource use for 7,609 to 16,662 ICU admissions; and then compared observed and predicted mortality rates for the three countries. ICU admissions from the host country were significantly more often nonoperative, had more comorbid conditions, particularly metastatic cancer, and a higher severity of illness. The host country's ICUs provided less technologic monitoring and similar amounts of therapy on the first ICU day, but ICU stay was twice as long as in the two other countries. Observed hospital mortality was similar to predicted in two countries but significantly higher (21.2%) than predicted (19.6%) in the host country.

Many ICU physicians from the host country questioned the possibility of comparing patients and practices from three different countries; and the accuracy of the equation used to predict mortality. They emphasized that differences in ICU length of stay and in observed vs. predicted mortality might be accounted for by differences in duration of prior therapy, interhospital transfer practices, diagnoses, comorbidities, ICU discharge practices, and the infrequent use of do not resuscitate (DNR) orders. Many expressed concern that the government officials in attendance would not recognize these limitations.

W. J. Sibbald et al. (eds.), *Evaluating Critical Care*

Introduction

To evaluate critical care services, outcome data must be collected and then compared to a performance benchmark or standard. The outcomes that are examined include mortality, complication rates, hospital and ICU length of stay, staffing level, or the use of treatment resources. A simple comparison of these outcomes, however, is frequently unsatisfactory because the characteristics of patients treated in different ICUs are not the same. In addition, the ICUs that are compared will often differ because of variations in hospital referral patterns, teaching status, and location. To meaningfully evaluate ICU performance, therefore, data comparisons must be adjusted for variations in both patient and hospital characteristics.

When comparing ICU outcomes, three approaches have been used to adjust for patient and institutional characteristics. First, the outcomes of a single ICU can be examined over time. If there is no significant change in patient characteristics, differences in performance can be meaningfully compared. Examples of change over time comparisons include assessments of mortality and resource use before and after changes in ICU organizational structure [1, 2]. Second, an ICU's outcome data can be compared with that of units with similar characteristics. Useful comparisons are possible when the hospitals, ICUs, and patients are similar. For example, comparing length of stay and the process of care after cardiac surgery has been possible because the hospitals, ICUs, and operative procedures are reasonably similar [3,4]. Third, an ICU scoring system can be used to compare observed outcomes to a case-mix adjusted standard. ICU scoring systems use statistical techniques to predict outcomes that are prospectively adjusted to reflect differences in hospital and patient characteristics such as diagnosis, severity of illness, and other known outcome determinants. The patient data on which these case-mix adjusted outcome predictions are based provide a standard or benchmark which is then compared to the observed measure of ICU performance. In addition to comparing ICU outcomes to an average standard, benchmarking can also identify ICUs with the best outcomes and provide insights about the clinical practices associated with their superior performance.

Background

The 1970s: You Can't Predict Mortality

Variations in severity of illness are a major reason why comparison of outcome data across ICUs requires adjustment for patient differences. In the early 1970s, methods for adjusting clinical outcomes for patient differences were limited to a few specific disorders. For trauma patients severity was defined by the type and extent of injury using the Injury Severity Scoring System [5]; for burn injuries by the extent of third degree burn area using the Burn Index [6]; and for head injury patients by the impact of the injury on neurological function using the Glasgow Coma Score (GCS) [7]. Each of these severity measurements correlated with hospital mortality rates, but each was limited in explanatory power and patient applicability.

In the late 1970s the use of physiological measures was a new approach to defining severity. Using physiological measurements was helpful because abnormalities are common to many acute diseases and the extent of derangement represented an objective and reproducible way to measure severity. Two major approaches were used to choose the physiological measures and decide on the importance or weight for each one. The first approach was to collect physiological information during treatment. The physiological patterns of survivors and non-survivors were contrasted and the measures weighted based on how often specific physiological values were associated with survival versus death. These methods were used by Siegal et al. [8] in septic shock, by Shoemaker et al. [9] for postoperative patients, and by Teres et al. [10] in developing the Mortality Prediction Model (MPM I). The second approach was to select and weight the physiological measures before treatment. Selection and weighting was based on prior studies and expert opinion. This was the method used in developing the Acute Physiology, and Chronic Health Evaluation (APACHE I) system [11].

In addition to measuring severity of illness, APACHE I was also used to predict group mortality [12]. To do this a multivariate regression equation was developed using information on age, gender, chronic health status, a 34 item acute physiology score, and the organ system dysfunction responsible for ICU admission. Regression coefficients were obtained using data for 613 ICU patients at the George Washington University Hospital and used to predict the number of deaths among 795 ICU patients at five university hospitals [12]. There was close agreement between observed and predicted death rates at the five hospitals. These results were later confirmed by similar analyses for 1,260 emergency ICU admissions at five United States and seven French tertiary care hospitals [13], and at 14 hospitals in the United States, France, Spain and Finland [14].

The 1980s: We Must Simplify

Knowledge gained in the 1970s produced reliable methods for severity measurement and risk stratification. These scoring systems, however, required multi-institutional validation; and APACHE I was too complex for use in clinical trials or for evaluating performance in individual ICUs. APACHE II was introduced in 1985 and incorporated major changes to the original APACHE system [15]. The number of physiological variables was reduced from 34 to 12 and higher scores were assigned to renal and neurological variables; scoring for emergency operative status was added, and chronic health evaluation was changed to reflect the impact of aging, and chronic cardiac, pulmonary, renal, or liver disease. The APACHE II score ranged from 0 to 71 with an increasing score reflecting an increased severity of disease and a higher risk of hospital death. In addition, an equation to predict risk of death was developed and coefficients published to reflect the prognostic impact of the APACHE II score, emergency surgery, and 49 disease categories.

The Simplified Acute Physiology Score (SAPS) was introduced in 1984 and measured severity using weights for physiologic variables similar to those used in

APACHE II. SAPS, however, emphasized simplifying severity scoring rather than predicting hospital mortality [16]. The 14 variable MPM 24 hour model and an 11 variable MPM 48 hour model also emphasized simplicity, but focused on predicting risk of death rather than severity scoring [17]. MPM also used a different analytic approach. Instead of using an expert panel to select and weight predictor variables, MPM used objective statistical reduction techniques to identify a smaller subset of the strongest outcome predictors.

During the later part of the 1980s and early 1990s APACHE II, MPM, and SAPS were used to describe ICU populations, to predict mortality for ICU patient groups, and to compare severity in clinical trials. There were growing concerns, however, about errors in prediction caused by differences in patient selection [18,19] and lead time bias [20]. There were also concerns about the size and representativeness of the databases used to develop the three systems, and about poor calibration within patient subgroups [21, 22] and across geographical locations [23].

Table 1. Variables used by the APACHE III, MPM II admission, MPM II 24, 48, 72 hour, and SAPS II systems for predicting hospital mortality

Prior Health Status	
APACHE III	7 comorbidities plus age
MPM II admission	3 comorbidities plus age
MPM II 24, 48, 72 hr.	2 comorbidities plus age
SAPS II	3 comorbidities plus age
Physiological Measures	
APACHE III	6 vital signs, 11 laboratory tests
MPM II admission	3 vital signs
MPM II 24, 48, 72 hr.	2 vital signs, 3 laboratory tests
SAPS II	5 vital signs, 7 laboratory tests
Timing and Selection for ICU	
APACHE III	7 locations; length of stay before ICU
MPM II admission	CPR, ventilator at or before ICU
MPM 24, 48, 72 hr.	not used
SAPS II	not used
ICU Admission Diagnosis	
APACHE III	78 diagnoses or disease categories
MPM II admission	5 acute diagnoses
MPM II 24, 48, 72 hr.	2 acute diagnoses
SAPS II	not used
Other Information	
APACHE III	4 hospital characteristics medical, elective, or emergency surgery
MPM II admission	medical or unscheduled surgery
MPM II 24, 48, 72 hr.	medical or unscheduled surgery constants for 24, 48, and 72 hours
SAPS II	medical, scheduled, or unscheduled surgery

APACHE III, Acute Physiology and Chronic Health Evaluation; MPM II, Mortality Probability Models; SAPS II, Simplified Acute Physiology Score; CPR, cardiopulmonary resuscitation.

The 1990s: The Limits of ICU Comparisons

Between 1991 and 1993 the three major scoring systems were refined and updated using the knowledge and experience gained during the 1980s. As a result of these refinements each system became more complex. The increase in complexity is shown in Table 1 which displays the type and number of predictor variables used in APACHE III, MPM II, and SAPS II [24–26]. The reference databases on which predictions are based reflect more contemporary (1988–1992) treatment results, include more patients (12,997–19,124), and more ICUs (42–140). Although each system predicts hospital mortality rate, the capabilities of APACHE III were expanded to include outcomes such as ICU and hospital length of stay, TISS score, risk for active therapy, use of pulmonary artery catheters, laboratory studies, and duration of mechanical ventilation. Each system was internally validated using a development and validation set, and within each database the systems demonstrated excellent discrimination (ability to identify patients who live or die), and calibration (correlation between predicted and observed mortality).

During the late 1990s studies using APACHE III [27–32], SAPS II [32, 33–37], and MPM II [36,38] frequently revealed an observed mortality that was different from expected. Each system predicted a mortality rate that was adjusted for differences in the variables included in each system. The predicted mortality provided a benchmark that was based on the effectiveness of therapy at the time and places where the systems were developed. These predictions, however, were not adjusted for unmeasured variables, for the passage of time, or for differences in the process, amount, or timing of treatment. Most of these studies detected differences between observed and predicted mortality which were explained using one or more of the following approaches:

1) Inadequate predictive equations. If the observed and predicted mortality were not uniform across all ranges of risk the system was poorly calibrated and it was concluded that the system is inaccurate.
2) Predicted mortality was treated as a 'gold standard'. When observed mortality exceeded predicted it was concluded that care must be suboptimal.
3) Observed outcome was treated as a 'gold standard'. When multiple prognostic systems were tested, the 'best' system had a predicted mortality that was closest to observed. Unfortunately APACHE II, a benchmark based on 1979 to 1981 treatment standards, was often selected.
4) Factors that might account for differences between observed and predicted mortality were analyzed and, if possible, identified.

Fortunately, most studies used the later approach and as a result have provided information that should lead to improvements in comparing ICU patients and in prognostic accuracy. Based on available information, the ability to accurately predict hospital mortality is based on the following factors: First, data must be available and reliably collected. Second, data collection must be accurate and reproducible. Third, there must be adequate adjustment for variables known to influence mortality, such as patient selection, lead time bias, diagnosis, physiologic reserve, physiologic abnormalities, and environmental factors such as geograph-

ic location, hospital characteristics and practices. Fourth, the reference database must be broadly representative and the predictive equation accurate with regard to discrimination, calibration, and ability to account for mortality differences among subgroups. Fifth, the patient sample must be large enough to avoid randomness and have the power to detect clinically significant differences between observed and predicted mortality. Sixth, average treatment results must be similar to those for patients within the reference database. It is unlikely, however, that treatment results will be identical because quality of therapy differs among hospitals, before and after ICU admission, and over time. Nonetheless, when a prognostic system satisfies the first five criteria it can be used to compare effectiveness with the benchmark established by the reference database.

Which Variables Should Be Compared?

Based on the knowledge and experience gained from the studies using APACHE III [27–32], SAPS II [32, 33–37] and MPM II [36, 38], it is possible to identify prognostic variables that may require modification or should be investigated in future studies (Tables 2–4). Some of these variables have been shown to influence hospital mortality, but are not included in each of the current models. Others have not yet been tested in multi-institutional studies, but on the basis of recent studies appear to be strong candidates for future testing. Some variables reflect patient characteristics, others are environmental, but none should directly reflect the type, process, or amount of treatment. This is because comparison of observed and predicted mortality is intended to reflect the outcome from treating a single episode of critical illness.

Patient Characteristics That Influence Outcome

Chronic Health Status

APACHE III, SAPS II, and MPM II each account for the adverse impact of increasing age on hospital mortality. Compared to physiologic abnormalities and other prognostic factors the weighting and relative explanatory power of age is relatively small, a finding that has been demonstrated in multiple studies of survival from critical illness. Each prognostic system also considers comorbid conditions (Table 1). Metastatic cancer is common to each system, and cirrhosis and hematologic malignancy (leukemia/multiple myeloma) are included in both MPM II and APACHE III. During model development, 34 comorbid conditions were tested for inclusion in APACHE III and 12 for inclusion in SAPS II. Conditions such as diabetes, severe impairment of activities of daily living; and chronic cardiovascular pulmonary and renal diseases were tested for their independent impact on hospital mortality, but did not meet statistical requirements for inclusion. Although each of these variables are known to influence hospital mortality, it is likely that their significance was diminished by the impact of physiological variables or diagnostic information that either directly or indirectly reflect these

Table 2. Studies using APACHE III mortality predictions and proposed reasons for differences between observed and predicted hospital mortality.

Study Location	Year	Patients	ICUs	SMR	ROC	H–L X^2 Statistic	Proposed reasons for calibration differences
United States (Cleveland) [27]	1991–1995	116,340	38	0.90	0.90	C^2=2407	Early discharge to skilled nursing facilities, patient selection (more beds, more low severity admissions), unmeasured variables, earlier treatment withdrawal
United States [28]	1993–1996	37,668	285	1.01	0.89	C^2=48.7	Inaccurate disease labeling, faulty assessment of GCS, changes in treatment outcomes over time
United Kingdom [29]	1993–1995	12,793	17	1.25	0.89	C^2 = 333	Differences in case mix, patient selection (fewer beds more high severity admissions), lead time bias (due to delayed ICU admission), earlier ICU discharge, manual vs. automated data collection
United Kingdom [30]	1993–1996	1,144	1	1.35	0.85	H^2=130	Differences in patient selection, comorbidity, admission criteria, disease labeling, less pre–hospital care & trauma experience, lead time bias
Brazil [31]	1990–1991	1,734	10	1.67	0.82	C^2=400	Lack of ICU equipment, beds, nurse preparation, poor ward care, selection bias, lead time bias
Germany [32]	1991–1994	2,661	1	1.13	0.85	C^2=48.4	More medical and emergency surgery patients, GCS not accurate for 43% of patients, selection bias(frequent interhospital transfer) lead time bias, longer hospital stay

SMR=standardized mortality ratio; ROC=reciever operating curve area; GCS=Glasgow coma score; H–L X^2 statistic= Hosmer Lemeshow statistic; critical value for X^2 is 16.9 for p=.05, with 9 degrees of freedom. For a fixed difference between observed and predicted X^2 will increase with sample size.

comorbidities. It should be noted that the comorbidities included in each prognostic system have an impact on immunologic status and their prognostic importance probably reflects the association of infection with ICU and hospital mortality. Based on past findings it seems unlikely that future scoring systems will be substantially improved by adding items reflecting chronic health status.

Physiological Measures

In contrast to age and comorbidities, physiological measures account for the largest proportion of explanatory power of APACHE III and SAPS II for hospital mortality. In aggregate, APACHE III, SAPS II, and MPM II use a total of 21 physiological measurements, but only heart rate, blood pressure, and a modified GCS are common to each system. The MPM II models are less reliant on physiological measures and in aggregate use only three vital signs and three laboratory tests. APACHE III uses 17 physiological measures compared to 12 measures for SAPS II. Variables that are unique to each system include hematocrit, albumin, glucose, and creatinine in APACHE III; and potassium and prothrombin time in SAPS II. It is unlikely, however, that simply adding measures that are unique to another system will improve accuracy. This is because, except for hematocrit and prothrombin time, the remaining unique variables were tested and did not meet statistical criteria for inclusion during model development. Future studies, therefore, should test hematocrit and prothrombin time as potential predictor variables. In addition, measures such as platelet count [39] and pupillary reactions [40] have been shown to have an independent impact on mortality and also deserve future testing.

Improving the weighting and accuracy of measuring the GCS would greatly enhance prognostic accuracy for each system. Recent evidence suggests that the GCS requires additional weighting for head trauma patients [28]. In addition, several studies have demonstrated scoring difficulties in measuring GCS for as many as 43% of patients [28, 31, 32]. The GCS measurement problem appears to be caused by the wide variation across ICUs in the use of deep sedation and paralysis, treatments that can make accurate scoring impossible. Because no severity weighting is applied when GCS cannot be assessed, the effect is to underpredict hospital mortality for patients whose GCS might otherwise reflect a substantial neurological deficit. Such problems are particularly common among trauma, neurological, and respiratory patients [28, 32]. Because there is no better measure of neurological function, data collectors should use the following approach to minimize errors in recording GCS:

1) The GCS should be obtained for as many patients as possible. Frequent inability to record GCS due to sedation is often a sign of poor data reliability.
2) Carefully adhere to definitions provided in the original description of the GCS. Avoid scoring until hypotension or hypoxemia have been stabilized, and use an ocular score of 1 for patients with severe periorbital swelling.
3) If the patient is sedated or paralyzed but an accurate GCS was previously recorded use that score. This approach was specified for SAPS II and should now be used for each system [26].

Table 3. Studies using SAPS II mortality predictions and proposed reasons for differences between observed and predicted hospital mortality.

Study Location	Year	Patients	ICUs	SMR	ROC	H-L X^2 Statistic	Proposed reasons for calibration differences
Italy [33]	1994	1,393	99	1.14	0.80	C^2=71	Insufficient diagnostic data, imprecise variable definitions, lead time bias, international and regional differences in case-mix/therapy
Portugal [34]	1994–1995	982	19	1.02	0.82	C^2=32.7 H^2=49.7	Insufficient diagnostic data, imprecise variable definitions, international and regional differences,
Austria [35]	1997	1,733	9	0.85	0.81	C^2=91.8 H^2=89.1	Unmeasured case mix differences, unreliable GCS assessment
12 European Countries [36]	1994–1995	10,027	89	1.15	0.82	C^2=218.2	Insufficient diagnostic data, unmeasured clinical and nonclinical variables, changes in effectiveness of therapy over time
Tunisia [37]	1994–1995	1,325	3	1.27	0.84	H^2=73.8	Differences in patient selection (fewer ICU beds), resource limitations, international differences, unmeasured case mix differences
Germany [32]	1991–1994	2,661	1	1.16	0.85	C^2=20.5	Unmeasured case mix differences, unreliable GCS assessment, lead time bias, longer hospital length of stay.

SMR=standardized mortality ratio; GCS = Glasgow coma score; ROC = reciever operating curve area; H–L X^2 statistic = Hosmer Lemeshow statistic; critical value for X^2 is 16.9 for p=.05, with 9 degrees of freedom. For a fixed difference between observed and predicted X^2 will increase with sample size.

4) If there is a suspicion that neurological status has changed and sedation or paralysis can be safely reduced, repeat GCS assessment after therapy is reduced.
5) If direct measurement is impossible, record a GCS of 15. Although risk will be underestimated, it will cause a systematic error that might be corrected by recalibration.

Diagnosis

Tables 2–4 demonstrate that insufficient diagnostic data or inaccurate disease labeling were frequently proposed as reasons for differences between observed and predicted mortality [28, 30, 33–37]. This reflects a widely held belief that a patient's ICU admission diagnosis provides important prognostic information. For example, a recent study of 37,668 ICU admissions reported an observed hospital mortality rate of 3.3% for asthma compared to 37.5% for noncardiac pulmonary edema, and 18.2% for gastrointestinal bleeding due to varices compared to 8.4% for bleeding due to diverticulitis or angiodysplasia [28]. For each of the above diagnoses, specific disease labeling provided more accurate prognostic information than could be achieved by aggregating patients into medical, respiratory, or gastrointestinal subgroups. The same study also demonstrated improved prognostic accuracy when mortality rates for combined diagnoses such as unstable angina and acute myocardial infarction or bacterial and viral pneumonia were predicted separately, and when residual organ system categories were disaggregated into specific diagnostic categories [28]. Unfortunately, developing coefficients for specific diseases requires a very large reference database. In addition, imprecise diagnostic labeling and difficulty choosing a single diagnosis are also a source of error in predicting mortality risk. These issues will be discussed further in the section on data reliability.

Selection for ICU Care

When a prognostic system is used to predict mortality in a new population it is important that the reference database from which the estimated patient risks are derived contains only patients chosen by similar selection criteria. The APACHE III, SAPS II, and MPM II databases were all created using consecutive ICU admissions to a diverse group of hospitals in the US, Canada, and Europe. Unfortunately, the decision to admit a patient to ICU is not uniform among hospitals, and admission source, particularly the transfer of a patient from another hospital to an ICU, is associated with a higher mortality rate [41, 42].

Several studies have suggested that variations in selection criteria for ICU admission may have caused differences between observed and predicted mortality (Tables 2–4). One study suggested a selection bias due to frequent admission of low severity patients because of the ready availability of beds [27], but

Table 4. Studies using MPM II mortality predictions and proposed reasons for differences between observed and predicted hospital mortality.

Study Location	Year	Patients	ICUs	SMR	ROC	H-L X^2 Statistic	Proposed reasons for calibration differences
Tunisia [38]							
MPM II_0	1994–1995	1,325	3	0.91	0.85	C^2=36.7 H^2=19.9	Unmeasured case mix differences, treatment differences including lack of full-time intensivist lack of nursing education skill and motivation
MPM II_{24}	1994–1995	1,325	3	1.14	0.88	C^2=38.0 H^2=29.6	
12 European Countries [36]							
MPM II_0	1994–1995	10,027	89	0.85	0.78	C^2=437.1 H^2=368.2	Inadequate diagnostic data, failure to adjust for prior location and other unmeasured clinical and nonclinical variables, changes in quality of treatment over time

SMR=standardized mortality ratio; ROC=reciever operating curve area; H-L X^2 statistic= Hosmer Lemeshow statistic; critical value for X^2 is 16.9 for p=.05, with 9 degrees of freedom. For a fixed difference between observed and predicted X^2 will increase with sample size.

others suggested that selection bias was caused by too few beds and consequent delays in ICU admission [29–31, 37]. Selection bias has also been attributed to differences in the frequency of interhospital transfer [31, 36, 37] and in the reasons for transfer [32]. In an independent US database, APACHE III adjusted well for the prognostic implications of the selection differences reflected by patient location before ICU admission [28]. Studies from Europe and developing countries, however, suggest that the selection variable of APACHE III does not adequately adjust for international differences in selection for intensive care.

Lead Time Bias

When data for all ICU admissions are not obtained at approximately the same time in the course of an acute illness or within a similar time period after major surgery, prognostic estimates will be inaccurate because physiological measures reflect different phases of critical illness [18, 19]. Adjustment for differences in patient location before ICU admission (selection) allows for some of these differences, but cannot account for delays in interhospital transfer, or for extensive amounts of intensive care therapy on hospital wards before the patient physically arrives in ICU [20, 43]. Based on this knowledge, adjustments for prior location and for the length of hospital stay before ICU admission were included as prognostic variables in APACHE III [41]. These adjustments work well in the US [28], but international differences in lead time have been proposed as a potential explanation for differences between observed and APACHE III predicted mortality in Europe and Brazil [29–32]. Failure to account for selection and lead time bias were also proposed as reasons for miscalibration in studies using SAPS II and MPM II [33, 36, 38]. Based on this knowledge, prognostic systems should account for location before ICU admission and the length of hospital stay before ICU admission. It seems likely that adjustment for these variables will be required at the national level. At present, however, no method has been proposed to adjust for differences between observed and predicted mortality that might be related to the quality or process of care before ICU admission [30, 44] or after ICU discharge [45]. Accounting for such treatment differences, however, should not be necessary since the outcome of interest is mortality from a single episode of critical illness.

Environmental Factors that Influence Outcome

Hospital Practices

Hospital practices that are not directly related to the quality of ICU and hospital care also influence hospital mortality. One example is the discharge of ICU patients directly to skilled nursing or long term acute care facilities. These ICU patients are typically stable, but at discharge still require mechanical ventilation or dialysis and skilled nursing care. The number of long term acute care facili-

ties in the US has increased substantially during the last 10 years, but their availability and the transfer practices of individual hospitals vary widely. Direct transfer of these patients from ICU to long term acute care facilities can markedly reduce observed hospital mortality [27]. This is because these 'chronically critically ill patients' are at high risk for death after transfer [46]. Frequent transfer of such patients, therefore, can result in substantial overprediction of hospital mortality.

Variations in the frequency of limiting or withdrawing therapy must also be considered when comparing observed and predicted hospital mortality. These practices vary among hospitals [47], and have also increased over time [47, 48]. Interhospital differences in treatment withdrawal practices can have a marked, but variable influence on observed mortality. Among otherwise identical patients, observed mortality will be higher at hospitals where lifesupport withdrawal is frequent compared to hospitals where withdrawal is infrequent. In two recent analyses, differences in the frequency of treatment withdrawal, and in withdrawal among high versus low risk patients were thought to have a marked and varied impact on observed versus predicted mortality [28, 49]. Studies in the US using APACHE III have also shown a small but significant difference in hospital mortality that is associated with hospital size and teaching status [28, 41]. The exact reasons for these differences are uncertain, but should be examined in future studies.

Hospital Length of Stay

An ICU patient's risk of dying in the hospital is influenced by how long that patient remains in the hospital. APACHE III mortality predictions are adjusted for this influence using coefficients derived from a regression analysis of hospital length of stay among survivors [41]. This analysis incorporated all patient specific predictor variables, forecast a predicted hospital length of stay, and then calculated the mean difference between observed and predicted hospital length of stay for each ICU. Compared to hospitals where mean length of stay for survivors was within 1.6 days of predicted stay, shorter stays (1.6 days < predicted) decreased the multivariate odds ratio of death to 0.7; and longer stays (1.6 days > predicted) increased the odds ratio of death to 1.2.

Because this adjustment is unique to US hospitals it has not been possible to adjust for the impact of hospital length of stay on mortality in international studies using APACHE III. The potential importance of this adjustment is emphasized in Figure 1 which displays as much as a four fold difference in mean hospital length of stay after acute myocardial infarction in six countries. That similar variations exist within large ICU databases is suggested by reports of a 11.6 to 12.0 day mean hospital length of stay in the US [27, 28], 17.1 days in Brazil [31], 25.6 days in Germany [32], and 14.8 to 22.8 days in 10 European countries [26]. This information suggests that future studies of mortality among ICU patients should report hospital length of stay, and also investigate the impact of differences in hospital length of stay on mortality.

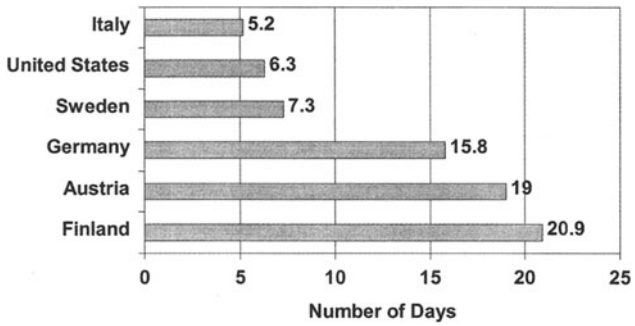

Fig. 1. Mean hospital length of stay in 1996 for patients with acute myocardial infarction (from the Organization for Economic Cooperation and Development)

Geographical Location

Many of the studies shown in Tables 2–4 either explicitly or implicitly suggest that APACHE III, SAPS II, and MPM II do not adequately adjust for national differences in intensive care. In addition, studies in the US [28,41] and Italy [33] have demonstrated differences in hospital mortality across geographic regions despite adjustment for patient differences. It is uncertain whether these mortality differences reflect national and regional variations in critical care practices, socioeconomic differences, variations in disease labeling, patient selection and lead time bias, frequency of treatment withdrawal, or duration of hospital stay. For example, in the US differences between observed and predicted hospital mortality evaporated after adjustment for hospital length of stay for survivors. It seems doubtful that each of the above differences could be accounted for by simply customizing current systems using national databases [36, 50, 51]. Instead, national and regional differences in each prognostic variable should be reported; examined for a significant independent impact on hospital mortality, and if significant added to the predictive model.

Changes Over Time

Reductions in mortality from critical illness are usually related to the introduction of new drugs (e.g., thrombolytic agents), new technologies (e.g., non-invasive positive pressure ventilation [NIPPV]) or new techniques (e.g., low tidal volume ventilation in acute respiratory distress syndrome [ARDS]). We all believe in medical progress, but attributing a lower observed versus predicted mortality to changes in treatment effectiveness over time is overly simplistic. Reductions in mortality due to improvements in treatment effectiveness are usually disease specific, e.g., thrombolytic therapy for acute myocardial infarction. For some diseases, however, it may be difficult to attribute improved mortality to therapy alone. For example, should the improved outcome for parasitic pneumonia [28] be attributed to better ICU therapy for *Pneumocystis carinii* pneumonia, to

improved therapy for human immunodeficiency virus (HIV) infection, or to recent changes in the definition of acquired immunodeficiency syndrome (AIDS)? Although medical progress usually improves survival, some changes over time tend to increase observed mortality, e.g., the recent increase in frequency of withdrawal of life support. In the United States we now adjust for the prognostic impact of changes in practices and therapy every one to two years by examining observed and APACHE III predicted mortality for each ICU admission diagnosis and adjusting coefficients as needed.

Data Accuracy and Reliability

If a prognostic scoring system cannot be applied reproducibly in a different population it will not perform accurately, irrespective of how well the system examines the variables that influence outcome [52]. The accuracy of outcome prediction is therefore in part determined by inter (between) observer reliability, i.e., the difference in quantifying data when different individuals score the same patient. A high degree of interobserver reliability was reported by the developers of APACHE III [24], SAPS II [26], and MPM II [25]. Interobserver reliability was excellent for discrete measures such as age and other demographic information; and for physiological measures that require little or no judgement on the part of data collectors. In contrast, interobserver reliability was lower for variables that require choosing the most deranged physiological measurement (e.g., temperature), or calculation (e.g.. PaO_2/FiO_2); and even lower for agreement on GCS parameters. Independent studies that have used APACHE III, SAPS II, and MPM II [27, 34, 36] have also reported good interobserver reliability during data collection, but have confirmed that interobserver reliability deteriorates for measures such as GCS [32, 53], ICU admission diagnosis, and physiological variables that require calculation [53, 54].

We believe that the reliability of data collection can and has been improved [27, 28, 55]. Because current prognostic systems have become more complex it is no longer possible to replicate methods directly from journals [56]. It is our experience that a comprehensive instruction manual is needed to precisely describe methods and definitions. This instruction manual is supplemented by on-site data collector training, an instructional video, training exercises, direct data collection supervision, and ongoing telephone assistance. After data for 100 to 200 patients is collected, formal interobserver reliability testing is performed for a 10% random sample. In addition, emphasis is placed on careful design of data collection forms, the use of software based error checking for manually collected data, and automated data collection. A recent study of prognostic scoring using an automated ICU information system versus manual data collection showed that automation improved the detection of physiological abnormalities [57]. This resulted in increased severity scores, which in turn increased predicted mortality. Automation improves the reliability of collecting laboratory data; and of recording comorbidity and ICU admission diagnosis through the use of computerized pick lists. Computer software also calculates mean arterial pressure and PaO_2/FiO_2 and oxygen delivery values and assists in identifying the most abnormal physiological values.

Conclusion

During the past three decades there has been significant progress in predicting group mortality for ICU patient groups. Although simplification of prognostic scoring is a desirable goal, current information suggests that improvements in prognostic accuracy will require the addition of more prognostic variables. As in the past, independent studies have suggested a substantial number of variables that require further testing. We believe that each prognostic model should adjust for variables that significantly influence outcomes, and that simple re-calibration is insufficient to adjust for the absence of proven outcome predictors. Unfortunately, as the number of predictor variables increase, data reliability and ease of use decreases. We believe that technology, not simplification, is the answer to this challenge. Accurate predictive models, however, can only provide a benchmark not a true 'gold standard'.

The Rest of the Story

The course participants believed that a common data set made it possible to compare patients' resource use and outcomes in the three countries. Because the cultures, ICU policies, and health care systems varied markedly across the different countries, the participants believed that country-specific efforts were essential to provide national standards, which could then be compared internationally. Shortly after the course, leaders of the host country's national and regional governments met with the course's ICU physician organizers and representatives of the national intensive care society. Subsequent to this meeting the government and intensive care society supported collection of a nationally representative ICU data set and the development of a customized mortality predictive equation. In 1999, the host country's ICU society made data collection instruments and automated calculation of predicted mortality available over the Internet. ICUs now compare their observed and predicted mortality rates and use this information to assess individual ICU performance.

References

1. Carson SS, Stocking C, Podsadecki T, et al (1996) Effects of organizational change in the medical intensive care unit of a teaching hospital: a comparison of "open" and "closed" formats. JAMA 276:322–328
2. Manthous CA, Amoateng YA, Al-Kharrat T, et al (1997) Effects of a medical intensivist on patient care in a community teaching hospital. Mayo Clin Proc 72:391–399
3. Engleman RM (1996) Mechanisms to reduce hospital stays. Ann Thorac Surg 61: S26–S29
4. Cheng DCH, Karske J, Peniston C, et al (1996) Early tracheal extubation after coronary artery bypass graft surgery reduces costs and improves resource use. Anesthesiology 85:1300–1310
5. Baker SP, O'Neil B, Haddun W, Long WB (1974) The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. J Trauma 14:187–196
6. Feller I, Tholen D, Cornell RG (1980) Improvements in burn care, 1965 to 1979. JAMA 244:2074–2078

7. Teasdale G, Jennett B (1974) Assessment of coma and impaired consciousness: a practical scale. Lancet ii: 81–84
8. Siegel JH, Goldwyn RM, Friedman HP (1971) Patterns and processes in the evaluation of human septic shock. Surgery 70:232–240
9. Shoemaker WP, Chang P, Czer L (1979) Cardiovascular monitoring in post-operative patients. Crit Care Med 7: 237–241
10. Teres D, Brown RB, Lemeshow S. (1982) Predicting mortality of intensive care unit patients: the importance of coma. Crit Care Med 10: 86–94
11. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE (1981) APACHE – acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9: 591–597
12. Knaus WA, Draper EA, Wagner DP, et al (1982) Evaluating outcome from intensive care: a preliminary multihospital comparison. Crit Care Med 10: 491–496
13. Knaus WA, LeGall JR, Wagner DP, et al (1982) A comparison of intensive care in the U.S.A. and France. Lancet ii: 642–646
14. Wagner DP, Draper EA, Abizanda-Campos R, et al (1984) Initial international use of APACHE an acute severity of disease measure. Med Decis Making 4: 297–313
15. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II – a severity of disease classification system. Crit Care Med 13: 818–829
16. Le Gall JR, Loirat P, Alperovitch A, et al (1984) A simplified acute physiology score for ICU patients. Crit Care Med 12: 975–977
17. Lemeshow S, Teres D, Pastides H, Avrunin JS, Gage RW (1985) A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit Care Med 13: 519–525
18. Escarce JJ, Kelly MA (1990) Admission source to the medical intensive care unit predicts hospital death independent of APACHE II score. JAMA 264:2389–2394
19. Borlase BC, Baxter JK, Kenney PR, Forse RA, Benotti PN, Blackburn GL (1991) Elective intrahospital admissions versus acute interhospital transfers to a surgical intensive care unit: cost and outcome prediction. J Trauma. 31: 915–919
20. Dragsted L, Jorgenson J, Jensen NH, et al (1989) Interhospital comparisons of patient outcome from intensive care: importance of lead time bias. Crit Care Med 17: 418–422
21. Rowan KM, Kerr JH, Major E, McPherson K, Vessey MD (1993) Intensive Care Society's study in Britain and Ireland – II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. Br Med J 307:977–981
22. Jacobs S, Chang RWS, Lee B, Lee B (1988) Audit of intensive care: a 30 month experience using the APACHE II severity of disease classification system. Intensive Care Med 14:567–574
23. Sirio Ca, Tajimi K, Tase C, et al (1992) An initial comparison of intensive care in Japan and the United States. Crit Care Med 20: 1207–1215
24. Knaus WA, Wagner DP, Draper EA, et al (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100:1619–1636.
25. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rappoport J (1993) Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270: 2478–2486
26. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270:2957–2963
27. Sirio CA, Shepardson LB, Rotondi AJ, et al (1999) Community-wide assessment of intensive care outcomes using a physiologically based prognostic measure. Chest 115:793–801
28. Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA (1998) Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Crit Care Med 26:1317–1326
29. Pappachan JV, Millar B, Bennett D, Smith GB (1999) Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. Chest 115:802–810

30. Beck DH, Taylor BL, Millar B, Smith GB (1997) Prediction of outcome from intensive care: a prospective cohort study comparing acute physiology and chronic health evaluation II and III prognostic systems in a United Kingdom intensive care unit. Crit Care Med 25:9–15
31. Bastos PG, Sun X, Wagner DP, Knaus WA, Zimmerman JE, The Brazil APACHE III Study Group (1996) Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. Intensive Care Med 22:564–570
32. Markgraf R, Deutschinoff G, Pientka L, Scholten T (2000) Comparison of acute physiology and chronic health evaluation (APACHE II and III) and simplified acute physiology score (SAPS II): a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. Crit Care Med 28: 26–33
33. Apolone G, Bertolini G, D'Amico R, et al (1996) The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Intensive Care Med 22:1368–1378
34. Moreno R, Morais P (1997) Outcome prediction in intensive care: results of a prospective, multicenter, Portuguese study. Intensive Care Med 23:177–186
35. Metnitz PGH, Valentin A, Vesely H, et al (1999) Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Intensive Care Med 25:192–197
36. Moreno R, Miranda DR, Fidler V, Schilfgaarde RV (1998) Evaluation of two outcome prediction models on an independent database. Crit Care Med 26:50–61
37. Nouira S, Roupie E, El Atrouss S, et al (1998) Intensive care use in a developing country: a comparison between a Tunisian and a French unit. Intensive Care Med 24:1144–1151
38. Nouira S, Belghith M, Elatrous S, et al (1998) Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units. Crit Care Med 26:852–859
39. Vanderschuern S, Weerdt A, Malbrain M, et al (2000) Thrombocytopenia and prognosis in intensive care. Crit Care Med 28:1871–1876
40. Hamel MB, Goldman L, Teno J, et al (1995) Identification of comatose patients at high risk for death or severe disability. JAMA 273:1842–1848
41. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. (1993) Variations in mortality and length of stay in intensive care units. Ann Intern Med 118:753–761
42. Borlase BC, Baxter JK, Benotti PN, et al (1991) Surgical intensive care unit resource use in a specialty referral hospital: I. Predictors of early death and cost implications. Surgery 109:687–693
43. Porath A, Eldar N, Harman-Bohem I, Gurman G (1994) Evaluation of the APACHE II scoring system in an Israeli intensive care unit. Isr J Med Sci 30:514–520
44. Boyd O, Grounds RM (1993) Physiological scoring systems and audit. Lancet 341: 1573–1574
45. Goldhill DR Sumner A (1998) Outcome of intensive care patients in a group of British intensive care units. Crit Care Med 26:1337–1345
46. Seneff MG, Wagner DP, Thompson D, Honeycutt C, Silver M (2000) The impact of long term actue care facilities on the outcome and cost of care for patients undergoing prolonged mechanical ventilation. Crit Care Med 28: 342–350
47. Jayes RL, Zimmerman JE, Wagner DP, Knaus WA (1996) Variations in the use of do-not resuscitate orders in ICUs. Chest 110:1332–1339
48. Pendergrast TJ, Luce JM (1997) Increasing incidence of witholding and withdrawal of life support for the critically ill. Am J Respir Crit Care 155:15–20
49. Egol A, Willmitch B (1998) Witholding and withdrawing care affects severity adjusted outcome data. Crit Care Med 26:A25 (Abst)
50. Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D (1996) Factors affecting the performance of the models in the mortality probability model II system and strategies of customization: a simulation study. Crit Care Med 24:57–63
51. Teres D, Lemeshow S (1999) When to customize a severity model. Intensive Care Med 25:140–142
52. Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. Ann Intern Med 130:515–524
53. Chen LM, Martin CM, Morrison TL, Sibbald WJ (1999) Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. Crit Care Med 27:1999–2004

54. Holt AW, Bury LK, Bersten AD, Skowronski GA, Vedig AE (1992) Prospective evaluation of residents and nurses as severity score data collectors. Crit Care Med 20:16898–1691
55. Rosenthal GE, Harper DL (1994) Cleveland health quality choice: a model for collaborative community based outcomes assessment. J Comm J Qual Improv 20:425–442
56. Langham J, Goldfrad C, Rowan K (1998) Can case mix adjustment methods be replicated directly from the publication. Intensive Care Med 24 (Suppl 1): S23 (Abst)
57. Bosman RJ, Oudemans van Straaten HM, Zandstra DF (1998) The use of intensive care information system alters outcome prediction. Intensive Care Med 24:953–958