A Comparative Study of Clinical Severity Scoring Systems in ICUs in Taiwan

Chao-Hsiun Tang[1], Che-Ming Yang[1,2], Chi-Yuan Chuang[3], Ming-Lee Chang[4], Yu-Chwen Huang[4], Chin-Feng Huang[4]

Department of Healthcare Administration[1],Taipei Medical University, Taipei, Taiwan; Department of Nuclear Medicine[2], Chest Medicine[3], Nursing[4], Taipei Medical University-Municipal Wan Fang Hospital, Taipei, Taiwan

ABSTRACT

Objective: The objective of this study is to compare the performance of five commonly applied severity measures. Severity scoring systems have been developed in response to an increased emphasis on the evaluation and monitoring of health care services. The Acute Physiology and Chronic Health Evaluation (APACHE) II system is the only one that has been widely applied in Taiwan. This study is designed to ascertain the outcome prediction abilities of various severity measures in ICUs in Taiwan. **Materials and Methods:** All five severity instruments were applied to the same patient sample to assess the correlations and relative performance of all five systems. The data collection process was done in a 600-bed regional hospital with 101 valid cases recorded. **Results:** All five major severity scores, Multiple Organ Dysfunction Score (MODS), Therapeutic Intervention Scoring System (TISS), APACHE II, Mortality Probability Model 24 (MPM24), and Simplified Acute Physiology Score (SAPS), significantly correlated with each other. The accuracy of mortality prediction of each measure ranged from 0.71 to 0.88 as illustrated by the area under the receiver operating characteristics (ROC) curve. The predictive power of each severity measure against the total expenses for each admission was poor in light of the fact that the only significant coefficient of determination was as low as 0.064. APACHE II's performance was as good as all the other systems. MODS performed better in predicting costs of surgical cases with the coefficient of determination reaching 0.331. **Conclusions:** The application of APACHE II in Taiwan's ICUs as the only standard severity measure is justifiable based on our findings. MODS appears to have a better expense predictive power. However, the expense predictive power of TISS was not as good as expected. (***Tzu Chi Med J*** 2005; **17**:239-245)

Key words: severity of illness, ICU, APACHE II

INTRODUCTION

Severity scoring systems in the intensive care unit have been developed in response to an increased emphasis on the evaluation and monitoring of health care services [1]. According to Gregoire [2], there are four major purposes of severity-of-illness scoring systems. First, scoring systems are used in clinical trials for matching. Second, scoring systems are used to quantify severity of illness for administrative decisions such as resource allocation. Third, scoring systems assess ICU performance and compare the quality of care. Fourth, scoring systems are used to assess the prognosis of individual patients.

The most frequently used generic severity indices in ICUs according to the literature are Acute Physiology and Chronic Health Evaluation (APACHE) II, the Simplified Acute Physiology Score (SAPS), the Mortality Probability Model (MPM), the Multiple Organ Dysfunction Score (MODS), and Therapeutic Intervention Scoring System (TISS)[3,4]. Four of these five are

Received: January 24, 2005, Revised: March 14, 2005, Accepted: April 1, 2005
Address reprint requests and correspondence to: Dr. Che-Ming Yang, Department of Nuclear Medicine, Taipei Municipal Wan Fang Hospital, 111, Section 3, Hsing Long Road, Taipei, Taiwan

physiology-based; only TISS is service intensity based.

One of the most well-received generic severity measures based upon clinical data is the APACHE series, which calculates the probability of death independent of diagnosis. The APACHE score is calculated based on acute physiologic parameters and other clinical information. APACHE is actually less disease specific than other severity measurements in that it predicts the probability of dying independent of the disease. There are already three versions of this measure: APACHE I, II, and III. The latter two versions are more disease specific than the first version [5-7]. Numerous studies have been done in the past to explore APACHE's construct validity. For instance, Kruse et al[8] compared the predictions of physicians and critical care nurses with the APACHE II scores for 366 patients admitted to an ICU; they found no significant differences between the accuracy of clinical judgments and APACHE II scores. In a study done by Wong [9], the outcome predicted by the APACHE II score and the observed outcome had good correlation, and after controlling for severity of illness by using APACHE II scores, the hospital death rate was comparable between Canadian and US patients. Similar results have also been found in various disease-specific studies such as with acute myocardial infarction [10]. However, APACHE II has been criticized because it lacks validity in certain types of patients, such as burn and coronary artery bypass graft (CABG) patients. A recent study conducted in the UK to compare the performance of five severity-of-illness scoring systems used commonly in ICU patients also concluded that APACHE II is the most appropriate model for comparisons of mortality in different ICUs because of its superior calibration [4,7,11].

SAPS II was developed by logistic regression analysis of data from a joint European-North American study in the 1990s. The SAPS II score is made up of 17 variables: age, type of admission (scheduled surgical, unscheduled surgical or medical), 12 physiological variables and three variables related to underlying disease; the cumulative scores of the 17 variables are converted into the probability of mortality for each patient [12].

The MPM II was developed from data from 139 ICUs in 12 countries in the 1990s. A major difference from APACHE II and SAPS II is that, while APACHE II and SAPS II are performed 24 hours after admission to the ICU, the MPM system contains models that can be performed both on admission (MPM_0) and at 24 hours (MPM_{24})[13].

MODS originated because of multiple organ dysfunction syndrome, which exits because intensive care prolongs survival. A systematic review of the literature by Marshall et al [14] showed that the respiratory, renal, hepatic, cardiovascular, gastrointestinal, hematological, and neurological systems were the most commonly evaluated systems in published papers; they identified acceptable representative variables for six of the seven systems (the gastrointestinal system was not included). The six variables are the PaO_2/FiO_2, serum creatinine, serum bilirubin, pressure-adjusted heart rate, platelet count and Glascow coma scale (GCS).

TISS was first introduced in 1974 and was developed as a proxy measure of the severity of illness for a patient by quantifying the type and amount of nursing care provided. TISS operates under the premise that regardless of the diagnosis, therapeutic support results from the severity of the illness. Data collection general guidelines are as follows: data should be collected at the same time each day, preferably by the same observer; a TISS item should be checked if it was performed at any time during the previous 24 hours; and when several related interventions are applied within the same time period only one set of points is awarded for the maximum intervention [15].

MATERIALS AND METHODS

If we look at the patient care flow from the perspective of a system paradigm, patient physiology represents the input, whereas service intensity represents the process. As the patient goes through the inpatient process, potential output or outcome measurements could be mortality or expenses. This reasoning framework is illustrated in Fig. 1. The logical deduction is that physiological indicators predict mortality better whereas service indicators perform better in predicting medical expenses.

Recording the APACHE II scores for ICU patients is currently required for purposes of accreditation in Taiwan. However, local research on the validity of APACHE II is sporadic. The purpose of this study is to assess the performance of APACHE II relative to other common severity scoring systems. Particularly, we are

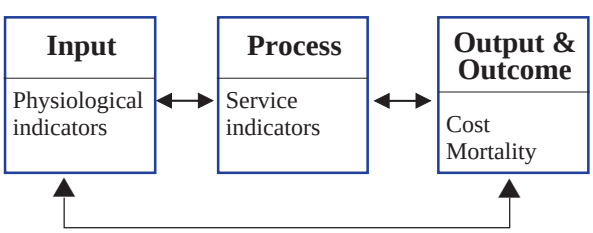

Fig. 1. Theoretical framework.

interested in knowing how APACHE II stands in terms of its performance in predicting mortality and costs compared to other scoring systems to determine whether it is reasonable to rely on APACHE II as the sole severity indicator in Taiwan's critical care settings. In that regard, three hypotheses were proposed:

1. All five severity scoring systems are significantly correlated.
2. The service indicator- based severity scoring system predicts costs better than physiological indicator-based severity scoring systems.
3. Physiological indicator- based severity scoring systems predict mortality better than the service indicator-based severity scoring system.

Since the objective of this study was to compare the five most commonly applied severity measures, the research method was to record all five severity instruments for the same patient sample to assess the correlations and relative performance of all five systems.

For the sake of precision and manageability, the plan was to collect at least 100 observations. All patients included in this study came from the same hospital within the study timeframe without further selection. The target hospital was a regional teaching hospital located in Taipei city. The target hospital was also a medical school affiliated general hospital with around 500 acute care beds and 150 special care beds, of which 50 are intensive care beds, separated into surgical and medical units. It has a staff of 1,200 including 100 physicians and 550 nurses. The target was chosen because of convenience and because of its medical school affiliation which guarantees certain levels of quality assurance that are essential for reliability of the data.

The data collection process was performed by the two head nurses of the medical and surgical intensive care units. Only two reviewers were chosen to insure competency and stability. Head nurses are supposed to be more knowledgeable and experienced than their staff. A limited number of reviewers entails less variation between reviewers. After proper training to ensure the reviewers' comprehension of the five instruments, the inter-rater reliability was assessed between the two head nurses. The kappa values for the five instruments: APACHE II, MPM_{24}, SAPS, MODS and TISS, were 1, 1, 1, 1, 0.91 respectively, which implies acceptable inter-rater reliability.

The correlations between the five systems were assessed by Pearson's correlation coefficient. Discrimination was assessed using the receiver operating characteristics (ROC) curve to evaluate how well the instrument distinguished between life and death. The cost prediction capability was assessed by linear regression.

RESULTS

Data collection was done between February and September, 2000. A total of 104 cases were recorded. Of the 104, 101 were valid. All patients were cared for by the two head nurses personally during their shifts. Since the head nurse was generally in charge of fewer patients, it took 8 months to gather sufficient samples. MPM_{24} distinguished between expected surgical cases and medical plus unexpected surgical cases. Since there were no unexpected surgical cases among our sample cases, this classification was used to separate surgical cases from medical cases.

Of the 101 patients, 60.4% were males and 80.2% were medical cases. The average age was 70.9 years and the survival rate was 62.4% (Table 1). The average APACHE II score of the sample population was 16.39;

Table 1. Descriptive Statistics of the Sample

Case number		101
Gender	Female	39.6%
	Male	60.4%
Age		70.9 y/o (±14.4918)
Type of admission	Medical	80.2%
	Surgical	19.8%
Survival rate		62.4%
Charges (NT$)		329,010 (±390,302.9968)

Table 2. Descriptive Statistics of Severity Scores

	Minimum	Maximum	Mean	Std. Deviation
MODS	0.00	16.00	4.35	3.6261
APACHE II	4.00	51.00	16.39	7.9975
SAPS	6.00	112.00	43.68	21.7334
TISS	8.00	63.00	27.05	12.5995
$PRMPM_{24}$	0.01	0.99	0.40	0.3321
PRSAPS	0.00	0.99	0.36	0.3176

the average MODS score was 4.35; the average SAPS II score was 43.68; the average TISS score was 27.05; the average probability of mortality predicted by MPM_{24} was 0.4; and the average probability of mortality predicted by SAPS II was 0.36 (Table 2).

Bivariate analysis of the five major severity scores showed they all significantly correlated with each other (Table 3). The predictive power of each severity measure against the total expenses for each admission was ascertained in a linear regression fashion. Since the distribution of the charges was highly skewed, a natural logarithm transformation was performed and the linear regressions were then executed against the natural logarithm of the charges. The only significant model appeared to have MODS as the independent variable. Its r^2 was still low at 0.064 when all cases were taken into account (Table 4). Since surgical cases tended to cost more due to the high expenditures for the operation and hence the distribution of the cases was distorted, subgroup analyses based on medical and surgical distinction were also conducted. After we separated medical from surgical cases, the r^2 jumped to 0.331 for surgical cases alone and decreased to 0.026 for medical cases in the MODS model (Table 4). However, even with this categorization, we still could not find a significant model for TISS and the predictive power did not change to a significant degree for any of the other measures either.

The accuracy of mortality prediction of each system was assessed using the area under the ROC curve. The area under the ROC curve was as follows: 0.714 for MODS; 0.790 for APACHE II; 0.782 for TISS; 0.847 for the mortality probability of MPM_{24}; and 0.884 for the probability of SAPS. (Fig. 2-6). The re-

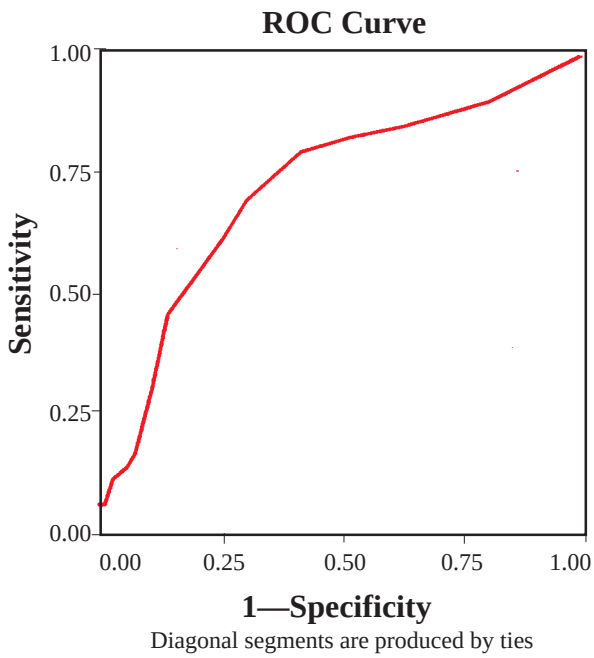

Fig. 2. The ROC curve for MODS.

Table 3. Bivariate Correlations of the Five Severity Measures

	$PRMPM_{24}$	PRSAPS II	MODS	SAPS II	TISS	APACHE II
$PRMPM_{24}$	1.000	0.726*	0.649*	0.724*	0.590*	0.622*
PRSAPS II		1.000	0.644*	0.969*	0.626*	0.669*
MODS			1.000	0.665*	0.585*	0.558*
SAPS II				1.000	0.622*	0.690*
TISS					1.000	0.541*
APACHE II						1.000

*: Pearson's correlation is significant at the 0.01 level (2-tailed)

Table 4. Linear Regression Analysis of Charges and Severity Measures

variable	Independent Total			Medical			Surgical		
	r	r^2	p	r	r^2	p	r	r^2	p
MODS	0.252	0.064	0.011	0.039	0.026	0.079	0.576	0.331	0.008
APACHE II	0.025	0.016	0.111	0.171	0.029	0.128	0.262	0.069	0.265
TISS	0.168	0.016	0.093	0.019	0.006	0.224	0.228	0.052	0.333
$PRMPM_{24}$	0.049	0.002	0.626	0.145	0.021	0.198	0.126	0.016	0.598
PRSAPS	0.043	0.002	0.666	0.027	0.001	0.812	0.145	0.021	0.543

*: The dependent variable is the natural logarithm of charges

sults were quite similar to other studies, such as Kruse et al [8].

In summary, for hypothesis 1, we did find that the five severity scoring systems were highly correlated. For hypothesis 2, the only service indicator based severity scoring system, TISS, did not predict costs better than the physiological indicator- based severity scoring systems. On the contrary, MODS performed better in surgical cases. For hypothesis 3, physiological indicator- based severity scoring systems did not predict mor-

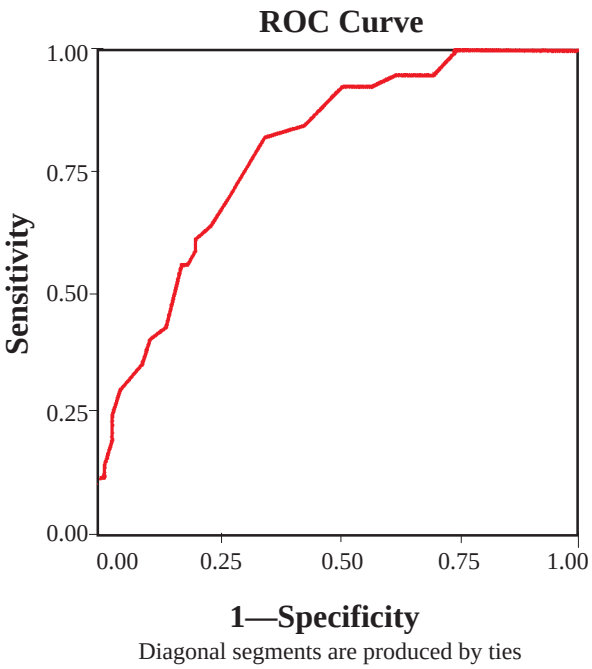

Fig. 3. The ROC curve for APACHE II.

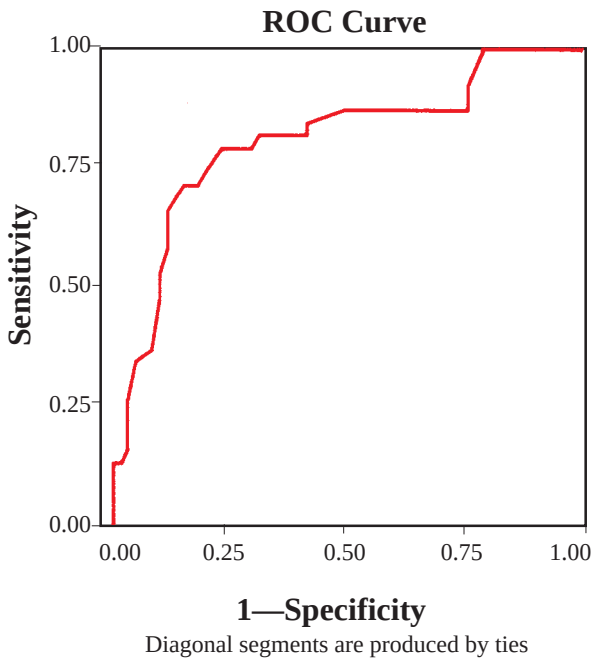

Fig. 4. The ROC curve for TISS.

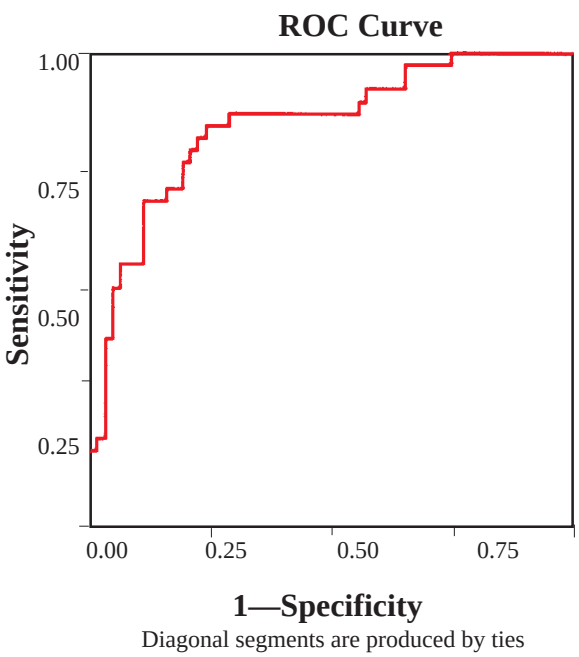

Fig. 5. The ROC curve for PRMPM24.

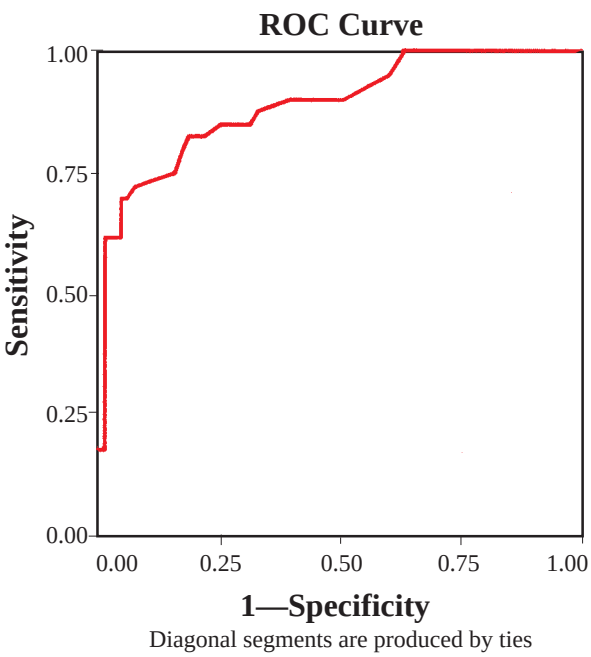

Fig. 6. The ROC curve for PRSAPS.

tality better than the service indicator based severity scoring system.

DISCUSSION

Local researchers and hospitals in Taiwan have utilized different severity scoring systems to various degrees and conducted related studies, which usually corresponded to the findings of other studies. Mackay Memorial Hospital compared the predicting ability of APACHE II and the Organ System Failure (OSF) system [16]. The mean APACHE II score for the alive group was lower than that for the dead group. The number of organ system failures was higher in the dead group than in the alive group. For each seven point increment in the APACHE II score, there was a significant increase in the death rate. The sensitivity was 55% and specificity was 94% when using an APACHE II score ≥ 28 as the cut-off point. If a risk of hospital death ≥ 60% was taken as the cutoff point, the sensitivity was 53% and specificity was 97%. The sensitivity of the OSF measures was 58% and the specificity was 87% if more than 2 organ system failures was taken as the cut-off point.

In Lin et al's study [16] in predicting deaths of MICU patents, the area under the ROC curve of the APACHE II score appeared to be greater than that of the OSF measures. This implied that APACHE II scores had a better predictive power of hospital death than OSF measures. The risk of death calculated from APACHE II scores had an even greater predictive power than the APACHE II score alone, which was evidenced by a greater area under the curve.

Lin et al's results [16] showed that both APACHE II scores and OSF measurements have good predictive powers for hospital deaths in MICU patients. APACHE II scores and risk of death methods were better than OSF measures in predicting hospital deaths. The predictability of the APACHE II risk of death method applied to their patients was comparable to the results of Knaus et al's study [6]. When a 70% risk of death was used as the cut-off point to predict the mortality of MICU patients, a positive predictive value of 96% was achieved.

In 1995, another local study was done by Chao et al [17] to select the prognosis index for chest patients in the intensive care unit. They applied the APACHE II, the 1983 TISS, the condition index score (CIS), the discrete condition and demographic variables (DCDV), and the intensive care nursing requirements (ICNR) to their patients. They studied 117 patients admitted to the adult medical and surgical ICUs of Taiwan Provincial Taoyuan General Hospital in 1995 with a primary diagnosis of chest disease. APACHE II scores for medical patients averaged 18.25 ± 7.791 and for surgical patients 9.48 ± 6.459. TISS scores for medical patients averaged 26.45 (6.085, and for surgical patients 27.19 ± 2.587. They found that the DCDV-24hr score was the best survival predictor for chest patients in ICUs, that APACHE II also provided valid predictions, and that TISS does not predict survival [17].

The use of the APACHE II system as the severity assessment tool in Taiwan's ICUs has been taken for granted. With the rapid development in the severity scoring field, further local research is imperative to justify the continuing use of this system. The findings of our study indicate that APACHE II's mortality prediction ability is better than TISS and MODS, and slightly inferior to SAPS and MPM. This is similar to Chao et al's [17] and Lin et al's findings [16] that APACHE II is better than TISS and the organ system failure approach in predicting mortality. The APACHE II system is quite easy to use and local professionals have become familiar with its application in intensive care. Our findings do support the argument that the application of APACHE II in Taiwan's ICUs is still valid despite the developement of other new severity of illness measurement modalities.

Contrary to Chao et al's findings [17], TISS's mortality predictive ability was comparable to the other measures. MODS predicted costs better than TISS in surgical cases. Since TISS is a service based scoring system, it should be more related to costs than the other measures. However, since the service items included are mostly nursing services and expensive major procedures might not be included, TISS scores are likely to underestimate costs. Surprisingly, MODS does well in surgical case cost predictions, although it did not perform any better when medical and surgical cases were included together. Medical patients tend to have multiple system failures when they need intensive care. Conversely, most surgical cases do not have organ failure postoperatively. Therefore, we can reasonably infer that when multiple organs fail in surgical cases, costs are driven up proportionally. This explains why costs in surgical cases are more correlated with MODS scores than with the other scores.

We conclude that these five severity scoring systems are approximately equal in ascertaining an individual patient's condition. Their mortality predictive abilities are close, even though APACHE II is not the best predictor of mortality. However, they are all poor predictors of medical expenses except for MODS, which appears to do well in surgical cases. TISS is also not a good predictor of hospital expenses. Therefore, there is

no absolutely perfect substitute for APACHE II. Sample size constitutes a major limitation of our research. Since there were only 20 surgical cases, it is difficult to compare with confidence how the performances of these scoring systems differ among different admission types. Further multiple center studies will be helpful.

REFERENCES

1. Rowan K: Scoring Systems in Intensive Care. In: Webb AR, Shapiro M, eds. Oxford Textbook of Critical Care. UK: Oxford, 1999, pp 1048-1081.
2. Gregoire G, Russell JA: Assessment of severity of illness. In: Hall JB, Schmidt GA, eds. Principle of Critical Care. New York: McGraw-Hill Publication, 1998, pp 57-69.
3. Rue M, Artigas A, Alvarez M, Quintana S, Valero C: Performance of the Mortality Probability Models in assessing severity of illness during the first week in the intensive care unit. Crit Care Med 2000; **28**:2819-2824.
4. Livingston BM, MacKirdy FN, Howie JC, Jones R, Norrie JD: Assessment of the performance of five intensive care scoring models within a large Scottish database. Crit Care Med 2000; **28**:1820-1827.
5. Seneff MG, Zimmerman JE, Knaus WA: Using the APACHE III system to improve the quality of intensive care. International Journal of Intensive Care 1997; **7**: 237-243.
6. Knaus WA, Draper EA, Wagner DP, Zimmerman JE: APACHE II: A severity of disease classification system. Crit Care Med 1985; **13**:818-829.
7. Ridley S: Severity of illness scoring systems and performance appraisal. Anaesthesia 1998; **53**:1185-1194.
8. 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**:1739-1742.
9. Wong DT, Crofts SL, Gomez M, McGuire GP, Byrick RJ: Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients. Crit Care Med 1995; **23**:1177-1183.
10. Ludwigs U, Hulting J: Acute Physiology and Chronic Health Evaluation II scoring system in acute myocardial infarction: A prospective validation study. Crit Care Med 1995; **23**:854-859.
11. Angus DC: Scoring system fatigue and the search for a way forward. Crit Care Med 2000; **28**:2145-2146.
12. Le Gall JR, Lemeshow S, Saulnier F: A new Simplified Acute Physiology Score (SAPS II) based on a European/North American Multicenter Study. JAMA 1993; **270**:2957-2963.
13. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J: Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993; **270**:2478-2486.
14. Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ: Multiple Organ Dysfunction score: A reliable descriptor of a complex clinical outcome. Crit Care Med 1995; **23**:1638-1652.
15. Keene AR, Cullen DJ: Therapeutic intervention scoring system: Update 1983. Crit Care Med 1983; **11**:1-3.
16. Lin FJ, Lue YD, Kuo HT, Huang WC, Lin CC, Lee CM: Predicting outcomes in medical ICU patients, APACHE II scores or OSF measures? Journal of Emergency and Critical Care Medicine (Taipei) 1992; **3**:3-13.
17. Chao CL, Lee SW, Lai MC: Selection of prognosis index for chest patients in intensive care unit. Ann Inter Med (Taipei) 1997; **8**:46-51.