CHAPTER V

DISCUSSION

Discussion

Benchmarking is the process of measuring and comparing outcomes with a standard. For intensive care units, the most frequently used benchmark has been hospital mortality rate; but to be useful, the mortality benchmark must be adjusted for variations in patient characteristics [1]. To do this, intensive care unit scoring systems use statistical techniques to predict an expected mortality rate that is adjusted for differences in diagnosis, physiologic abnormalities, and other important outcome determinants. The mortality benchmark is therefore based on patient outcomes at the hospitals where the system was developed and is adjusted for differences in patient characteristics. The accuracy of scoring systems is assessed by measuring how well the model distinguishes patients who die from patients who survive (discrimination) and the degree of correspondence between observed and predicted mortality (calibration) across the entire range of risk and within patient subgroups.

In the present study, the performance and validity of the third generation ICU scoring system, SAPS II and MPM_{24} II were evaluated in Thai adult ICUs. (The APACHE III were not included in the study because its equations are not in the public domain.) Both scoring systems showed excellent discrimination, although the discrimination was found to be slightly better for MPM_{24} II than for SAPS II; the same is true for the percentage of accuracy (overall correct classification). Good discrimination of both scoring systems has been reported in previous studies. The area under the ROC curves of both systems in the present study was higher than that in the other reports. Previously reported area under the ROC curves of SAPS II included 0.817 in Portugal [7], 0.840 in Tunisia [15], 0.87 in Hong Kong [16], 0.87 in Greece [17], 0.79 in Saudi Arabia [6], 0.843 in Scotland [5] and 0.88 in the original SAPS II [3]. In Thailand, Khwannimit et al. [10] also found good discrimination of SAPS II (area under the ROC curves of 0.888), however, Lertsithichi et al. found area under the ROC curve of SAPS II of 0.81 in surgical patients [9].

Reported area under the ROC curves of MPM_{24} II included 0.799 in Scotland [5], 0.882 in Tunisia [15], and 0.84 in Saudi Arabia [6]. In Thailand, there has been only one preliminary study on the area under the ROC curve of MPM_{24} II [9], which reported excellent discrimination and good calibration of the scoring system. A limitation of this study was the relatively small case mix populations.

It was found that calibration of SAPS II was inadequate. This lack of overall goodness-of-fit is similar to findings of other previous studies [6-7, 15-18]. Potential reasons for poor calibration in the present population might include the following: (1) difference in the case-mix of the study population compared with that in which the model was developed; (2) deterioration of the calibration of the scoring systems over time possibly influenced by medical progress and advancement in the science; (3) differences among ICUs in the quality of care.

In contrast, calibration was modest for MPM_{24} II with a non-significant Hosmer-Lemeshow statistic. The result is in agreement with other reports on the performance of the scoring system [4, 6, 18]. MPM II systems have been documented to have high reproducibility [19], which might explain the better calibration of MPM_{24} II in the present study.

When assessing the relative performance of severity of illness scoring systems it is important to appreciate that the most relevant assessment of their performance will depend on the proposed application. If models are to be used to assess quality of care by derivation of standardized mortality ratios, then calibration is the more significant measure of performance with discrimination being secondary to this. Castella et al. [20] stated that a model must have appropriate calibration before it can be applied to a population outside its original development population and only then can the system's discrimination be analyzed.

Moreover, from a practical standpoint, a simple system based on a limited number of general demographic and physiological variables may be a better option. Such a system is also more likely to be free from discrepancies associated with case-mix differences. Since, data collection and processing is easier and more practical for MPM_{24} II, it is one such scoring system. In evaluating the data collection burden, data abstraction times are longer for SAPS II [20].

The development of MPM (Lemeshow et al. 1985) was distinctly different from that of both APACHE and SAPS. The first difference was that MPM was entirely statistically derived. A large number of variables (137 at admission and 75 at 24 h) were collected on 755 consecutive general medical and surgical ICU patients. Various statistical techniques were used to determine the relative importance and weight of each variable. This process allowed the developers to retain only those variables that were commonly collected, non-ambiguous, and shown statistically to have a strong association with survival status. Finally, a multiple logistic regression model was developed. Using multiple logistic regression, the developers further reduced the number of significant variables and objectively derived weights for each of those remaining. The final multiple logistic regression models directly computed an estimate of the probability of the patient dying during the hospital stay rather than a point score. The results were very compact and simple models with only seven variables at admission and seven variables at 24 h, and very little time was required to collect and record the necessary information.

A second major difference was that, while APACHE II and SAPS were performed 24 h after admission to the ICU and used the 'worst' value during the first day for each of its variables, the MPM system contained models that could be performed both immediately upon admission and at 24 h. Third, the variables were more condition based rather than the physiologically based variables predominating in APACHE II and SAPS. Fourth, the variables collected were generally 'yes' or 'no' answers.

Subsequently, MPM II (Lemeshow et al. 1993) was developed with data collected in two separate studies, again using consecutive general medical and surgical ICU admissions. One of these datasets consisted of approximately 6000 cases collected at six United States hospitals. For the second dataset, the developers of MPM and SAPS (Le Gall et al. 1993) joined forces to collect data on over 14 000 patients in 137 hospitals in Europe and North America during 1990 and 1991. Together, the 19,124 cases ultimately included came from a diverse sample of ICUs consisting of 41% university hospitals, 27% university-affiliated hospitals, 16% community teaching hospitals, and 16% community non-teaching hospitals. The data consisted of a large set of variables including all those used in APACHE II, the original MPM, and SAPS, together with others.

For the admission model (MPM_0), the 19,124 patients were randomly separated into two groups: 12,610 cases (65%) were used for model development, and 6,514 (35%) were used later for validation purposes. The association of each possible independent variable with hospital mortality was assessed. This list was further analyzed for frequency of inclusion, ease of interpretation, and strength of association with mortality. Through multiple logistic regression techniques, variables were further reduced and those remaining were each assigned a weight. Next, variables were assessed for removal if their elimination improved calibration while not harming discrimination. The resulting model contains 15 variables and directly computes a probability of mortality during the entire hospital stay.

A similar process was performed using data available for the 10,357 patients in the development sample still alive at 24 h. The resulting model contains 13 variables: five values already collected in the admission model, two re-evaluated at 24 h, and six new variables. This model (MPM_{24} II) was tested and demonstrated excellent discrimination and good calibration.

Because of the ease of collection, the acceptance and use of MPM_0 II and MPM_{24} II have been significant. Unlike APACHE III, the MPM II probability of mortality does not depend on a single diagnosis because individual diagnoses, other than those included as part of the model itself, did not seem be significantly related to mortality. This is a significant simplifying factor since the necessity of designating a single overriding diagnosis within the first 24 h of an ICU admission is fraught with obvious difficulties. It is estimated that the collection time for any of the MPM II models is less than 2 min per patient (with some sites reporting less than 1 min) when the data are collected concurrently.

As with any ICU or similar model, the results from an MPM II calculation should be regarded as an estimate and a probability. Any multiple logistic regression probability model is developed from a specific population and can only purport to perform in that population. All the ICU scoring systems were developed from a diverse population of general medical and surgical patients in ICUs who were receiving intensive care and from data collected at specific times. Therefore their accuracy should only be assumed in similar settings. In an effort to deal with this, MPM II was designed to allow for

customization to account for specific differences. For example, while a model might perform very well across the entire range of hospitals, there might be a benefit in customizing for a certain subgroup such as hospitals in a specific country. In addition, as the MPM II database grows it will be possible and necessary to customize the models to reflect general changes in mortality resulting from changes in treatment and improvements in the quality of care.

In this study, although SAPS II provided excellent discrimination but showed poor calibration, while MPM_{24} II provided better discrimination and showed good calibration. The accuracy or overall correct classification rate was higher for MPM_{24} II than for SAPS II at the different cutoff points. From the ROC curve, the cutoff value that is closest to the upper left hand corner is 0.3. A cutoff point or decision criterion of 0.3 means that every patient with a risk greater than 0.30 is predicted to die. The overall correct classification rate for MPM_{24} II was 81.8%, with a sensitivity of 83.5%, a specificity of 81%, and positive likelihood ratio of 4.4. At this decision criterion, MPM_{24} II has both high sensitivity and high specificity. It is specific for minimizing the prediction of a positive outcome (survival) when it actually does not occur, and well sensitive to predict the outcome (survival) when it actually occurs. The likelihood ratio express that a patient with a risk greater than 0.30 is about 4.4 times more likely to be dead than a patient with a risk lower than 0.30. Furthermore, there was no significant difference between the SMR for MPM_{24} II and 1, as evident from the confidence intervals; this indicate that the scoring system gave overall accurate mortalities estimates. Because of its accuracy for Thai ICU patients, MPM_{24} II can be used to benchmark ICU performance using aggregate SMRs to assess quality of care.

When deciding which prognostic method is the most suitable for daily routine use, the specific conditions and requirements of the individual ICU setting have to betaken into consideration. The results of this study illustrate that prognostic methods which originally developed in another country should be validated for the population to which they will subsequently be applied. The better discriminative ability and the good calibration of MPM_{24} II may be advantageous for Thai ICU patients. Its simplicity makes it easy and quick to compute, as MPM_{24} II consists mainly of dichotomous variables. Although the development data are similar to the data used in SAPS II, the combined

risk-based system differs from the strict physiology-based system of SAPS II. Moreover, the widespread availability of computers and clinical databases facilitates calculation of predicted mortalities. MPM_{24} II was published with formulas in well known scientific journals. Software to compute MPM_{24} II has been made available via Internet without any charge.

Our study provides an external validation of the SAPS II and MPM_{24} II scoring systems in ICU patients; however, several limitations need to be addressed. First, as a single center study there may be bias concerning quality of ICU care, and ICU policy. Secondly, the case-mix in our study may differ from that in other ICUs as a high proportion of our patients were admitted after emergency surgery, limiting the extrapolation of our results to other populations. However, the study gave some insight into this issue, at least from a tertiary care perspective. Finally, postoperative patients are not homogeneous; however, the small sample size in the various subgroups in our study hinders exploration of the uniformity of fit among subgroups.

Conclusion

In conclusion, the findings from the present study confirmed that, in this group of ICU patients, the performance of the MPM_{24} II scoring system was superior over that of SAPS II. MPM_{24} II provided better discrimination and showed good calibration. Furthermore, data collection and processing is easier and more practical for MPM_{24} II. Hence, it could be successfully applied in Thai ICUs.

Suggestion for further studies

1. The present study was a single-center study, the results might be biased towards a certain case-mix, quality of ICU care, and ICU policy. In addition, the relatively small sample size was a relevant limiting factor in performed stratified analysis of calibration of both scoring systems. A multi-center study would have the benefit of fewer concerns of case-mix and a better sample size.

2. An explanation for the poor performance of the severity scoring systems is the presence of other factors, not measured by the present severity scores that can have a huge influence on the performance of the ICUs. It should be noted that those factors are

not randomly distributed between patients but clustered into ICUs; their effects on the performance of actual systems should be one of the main priorities of further research in this field. In other words, next generation severity scores should take into account not only the patient's variability (that is, baseline characteristics, severity of disease) but also the variability among ICUs (clinical and non-clinical factors that can influence outcome).

3. The application of a severity scoring system to a different population other the original database can only be done once the system has been tested and validated on that population, since variations in case-mix not accounted by the systems can have a significant impact on its performance. This problem is obviously more important when highly specialized ICUs with unique patient characteristics we study, but even its use in general ICUs can be dependent on the characteristics of the underlying populations. This stressed the necessity of periodic evaluation and or modifications of the systems, and can be especially important when the systems do not take into account diagnostic information, as is the case in SAPS II.

4. Changes in structure, organization and financing of the health care systems are having profound effects on the provision of hospital and intensive care and may further degrade the validity of standard severity systems. The earlier discharge of patients from acute care hospitals has become common practice in many countries and can affect the performance of the scoring systems that rely on vital status at hospital discharge as the principal outcome measure. In-hospital mortality may no longer represent an adequate end point and alternative outcome measurements are needed. For the present, a possible alternative is to focus more on patient stratification and description than on prediction, in other words to replace mortality with morbidity.

Conflict of interests

None declared.