Performance of the ICU

Moreno, Rui Paolo Jinó

Chapter 1

Introduction

A review of general outcome prediction models in intensive care

1. INTRODUCTION

The expansion of intensive care units (ICUs) in the last years has contributed to a better management of severe and complex diseases. After a history of more than one century, the ICU has become a central asset of the modern hospital. The technological developments in the last decades, together with a better understanding of the natural history and physiopathological mechanisms of the critically ill provided the possibility of the temporary replacement of the function of failing organs. However, this evolution was not made without costs.

The ICUs are today one of the largest consumers of hospital resources. Some reports estimate that the ICU consumes about 20 % of total hospital resources, and concern about the appropriate use appears justified [1]. Thus, the evaluation of the effectiveness became an important issue of ICU research in the 1990s.

The evaluation of the effectiveness of the ICU with strict scientific methods and criteria is not a simple task. According to standard methodological criteria, such an evaluation should preferably be based on formal randomised double-blinded studies, for instance comparing the effectiveness of care in the ICU with standard care in general wards [2]. However, such study proposals meet with opposition on the basis of ethical arguments in view of the general belief that ICU-treatment is appropriate and mandatory. As a matter of fact, even the appropriate evaluation of certain procedures of the daily practice in the ICU has been hampered by such objections [3,4]. Other approaches have been applied, like studies using historical controls. Some of these studies concluded positively in regard to the utility of intensive care [5-7] but others did not [8-10]. These studies can however be rightfully criticised on methodological grounds.

Apparently, we have to live with the general assumption that the development of intensive care, concerning its medical activities and their indications, followed the paths of multiple natural experiments. Consequently, we should compare the effectiveness of intensive care at the level of the individual ICU. In other words, the question to answer is whether the outcome of the patients treated in one ICU is in accordance with the expected outcome. With this approach, the actual outcome in the population under analysis is compared to the outcome in a reference population while controlling for case mix factors by using general outcome prediction models. The reference population is set as a gold standard if the model used for prediction is based on data from outstanding ICUs, or just as a reference population if the model is based on non-representative or non-selected ICUs [11]. This approach comparing actual and predicted outcomes is not new. It was for example used in the comparison of hospital mortality rates, using the difference between the observed mortality and that predicted by a model, controlling for some case mix factors [12,13]. Concerning the ICU, several investigators proposed the use of the ratio between observed and predicted

deaths (standardised mortality ratio, SMR) as an indicator of the effectiveness of care. The assumption is that although the ICUs may admit very heterogeneous groups of patients concerning relevant outcome markers such as large differences in age, previous health status or acute health status, the actually existent outcome prediction models can account for the most part of these characteristics [14].

The use of this methodology requires, first, that the outcome of interest is relevant, clearly defined and susceptible to accurate measurement, and second, that the outcome prediction model is able to control for important patient baseline characteristics which, in turn, are related to the outcome of interest. Thus, the full understanding of prediction models including their methodological limitations is essential for the clinicians and managers who use such models as a tool in the process of quality control in the ICU [15].

2. THE AVAILABLE MODELS

The development of general outcome prediction models started fifteen years ago (Table 1).

The first general outcome prediction model was the Acute Physiology and Chronic Health Evaluation (APACHE) [16]. Developed in 1981 at the George Washington University Medical Center, the APACHE scoring system demonstrated to provide accurate and reliable measures of severity of illness in critically ill patients [17-19].

Two years later, Le Gall et al. published a simplified version of this model, known as the Simplified Acute Physiology Score (SAPS) [20]. Another simplification of the original APACHE, the APACHE II, was published in 1985 by the developers of the original model [21]. This scoring system introduced the possibility of mortality prediction, requiring for that purpose the selection of a primary reason for ICU admission from a list of 50 diagnoses. Additional contributions to outcome prediction in the intensive care setting were the Mortality Probability Models (MPM) [22], developed using multiple logistic regression techniques to choose and weigh the variables rather than by a consensus from a panel of experts.

The most recent developments in outcome prediction models comprise the third version of APACHE (APACHE III) and the second versions of the SAPS (SAPS II) and MPM (MPM II). All were built using logistic regression techniques to choose and weigh the variables and are able to provide predictions of hospital mortality. They have been shown to perform better than their previous versions [23], and represent at this time the state of the art in this field. These models will be described in more detail in the next few sections together with APACHE II because of its still widespread use.

Table 1: General outcome prediction models.

Characteristics	APACHE	SAPS	APACHE II	MPM [a]	APACHE III	SAPS II	MPM II [b]
Year	1981	1984	1985	1988	1991	1993	1993
Participating countries	1	1	1	1	1	12	12
Participating ICUs	2	8	13	1	40	137	140
Number of patients	705	679	5815	2783	17440	12997	19124
Selection of variables and their weights	Panel of experts	Panel of experts	Panel of experts	Multiple logistic regression	Multiple logistic regression	Multiple logistic regression	Multiple logistic regression
Variables:							
Age	No	Yes	Yes	Yes	Yes	Yes	Yes
Patient origin	No	No	No	No	Yes	No	No
Surgical status	No	No	Yes	Yes	Yes	Yes	Yes
Chronic health status	Yes	No	Yes	Yes	Yes	Yes	Yes
Physiology	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Acute diagnosis	No	No	Yes [c]	No	Yes [d]	No	Yes [e]
Number of variables	34	14	17	11	26	17	15
Score	Yes	Yes	Yes	No	Yes	Yes	No
Equation to predict mortality	No	No	Yes	Yes	Yes	Yes	Yes

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Models.

[a]: These models were based on previous research from the same developers [24,25].

[b]: The numbers shown are for the admission component of the model (MPM II_0). The MPM II_{24}, was developed using data from 15925 patients in the same centres participating in the development of the admission model.

[c]: chosen from a list of 50 diagnoses.

[d]: chosen from a list of 78 diagnoses.

[e]: the MPM II_{24} model uses only 13 variables.

2.1 APACHE II

APACHE II was developed on the basis of data collected from 13 North American hospitals from 1979 to 1982 [21]. A panel of experts using clinical judgement and documented physiologic relationships did the choice of variables and their weights.

The model uses the worst value during the first 24 hours in the ICU for 12 physiologic variables (weighted from 0 to 4 points), age, surgical status (emergency surgical, elective surgical or non-surgical) and previous health status. The selection of a primary reason for ICU admission is necessary to be included in a logistic regression model that transforms scores in probabilities of mortality. The APACHE II score ranges from 0 to 71 points: up to 60 for physiologic variables, up to 6 for age and up to 5 for chronic health status. This system soon became the most worldwide utilised outcome prediction model, and it is still in use today in many ICUs.

2.2 APACHE III

The APACHE III was developed on the basis of a large North American database of intensive care patients in 1988-89 [26]. The selection of the 40 participating hospitals was done in order to be representative of the continental North American hospitals with more than 200 acute-care beds. Excluded from its development were patients with a length of stay in the ICU lower than 4 hours, patients younger than 16 years, and patients with burn injuries, with chest pain admitted for ruling out myocardial infarction and those in the immediate post-operative phase of coronary artery bypass surgery.

This model comprises a) APACHE III score based on acute physiological variables, age and chronic health, and b) the APACHE III predictive equation. The equation uses the APACHE III score and reference data on major disease-categories, the surgical status and the site of treatment immediately before ICU admission for estimating the risk of hospital mortality of ICU patients. The APACHE III score ranges from 0 to 299 points, including up to 252 points for the 18 physiology variables, up to 24 points for age and up to 23 points for chronic health. All physiologic variables are assessed as the worst values during the first 24 hours in the ICU. This strategy was chosen because it results in greater data availability (less proportion of missing values) and explanatory power [26].

The conversion of the score to a probability of mortality is accomplished with individual logistic regression equations for each of the 78 specific diagnostic categories and the nine patient origins. It is now a proprietary system; the equations are not in the public domain and must be purchased from APACHE Medical Systems, Washington, DC. This has limited its use, especially outside the United States, although it has been used recently in a cohort of Brazilian ICUs [27,28]. Moreover, the selection of a single admission diagnosis is sometimes difficult if not impossible in intensive care patients [29,30].

2.3 SAPS II

SAPS II was described in 1993 by Le Gall et al. [31], based on a European/North American multicenter study. It was developed and validated in a large cohort of patients from 110 hospitals in Europe and 27 in North America. Excluded from its development were patients aged less than 18 years, and burn patients, coronary care patients and cardiac surgery patients.

This model includes 17 variables to compute a score: 12 physiological variables, age, type of admission (medical and scheduled/unscheduled surgery), and three underlying diagnoses (acquired immunodeficiency syndrome, metastatic cancer and hematologic malignancy). SAPS II score ranges from 0 to 163 points (up to 116 for physiology, up to 17 for age and up to 30 for underlying diagnosis).

SAPS II uses the worst recorded values for the physiological variables during the first 24 hours in the ICU and do not need the selection of a single diagnosis for the computation of the probability of death in the hospital.

2.4 MPM II

MPM II was described in 1993 by Lemeshow et al. [32] based on the same database as SAPS II plus data collected in six ICUs of four teaching hospitals in the United States of America. The exclusion criteria used for its development were the same as those used for SAPS II. In this model the final result is only expressed as a probability of mortality (and not as a score). The actual MPM version incorporates models to predict mortality at admission (MPM II_0) and at 24 hours after admission to the ICU (MPM II_{24}). These were subsequently supplemented by models for the 48 (MPM II_{48}) and the 72 hours (MPM II_{72}) in the ICU, developed on the basis of a smaller database [33].

MPM II_0 contains 15 variables: age, three physiologic variables (coma or deep stupor, heart rate and systolic blood pressure), three chronic diseases (chronic renal failure, cirrhosis, and metastatic cancer), five acute diagnoses (acute renal failure, cardiac dysrhythmia, cerebrovascular accident, gastrointestinal bleeding, and intracranial mass effect), type of admission (medical or surgical unscheduled), mechanical ventilation and cardiopulmonary resuscitation prior to admission. All variables are evaluated on the basis of data collected within one hour at admission to the ICU.

MPM II_{24} is based on 13 variables: age, six physiologic variables (coma or deep stupor, creatinine, confirmed infection, hypoxemia, prothrombin time and urinary output), three variables ascertained at admission (cirrhosis, intracranial mass effect, metastatic cancer), type of admission (medical or surgical unscheduled), mechanical ventilation and use of vasoactive drugs. The physiologic variables are evaluated based on worst values during the first 24 hours

in the ICU.

MPM II_{48} and MPM II_{72} use the same variables as MPM II_{24}, with different weights for the computation of the predicted risk of death. Both are based on the worst values during the preceding 24 hours.

3. THE DEVELOPMENT OF THE MODELS

All the outcome prediction models aim at predicting outcome on the basis of a given set of variables: they estimate what should be the outcome of a given patient, with a certain clinical condition (defined by the values of the given set of variables) as if this patient was treated in a hypothetical reference ICU used to develop the model. Several steps are necessary for the development of these models (Table 2).

Table 2. Steps in the development of a model

1. Patient selection
2. Outcome selection
3. Predictor variables selection and data collection
4. Assembly of the model
5. Validation of the model
6. Model updates and modifications

3.1 PATIENT SELECTION

Although named "general", none of the outcome prediction models presently available is applicable to all ICU patients. Burn patients, patients admitted with acute coronary disease (or to rule out myocardial infarction), young patients (less than 16 or 18 years of age), patients in the post-operative of coronary artery bypass surgery or with a very short length of stay in the ICU were explicitly excluded from the development of the models. This limitation becomes important when we study specialised ICUs with particular patient demographics, but it can also be of importance in general ICUs (with mixed medical and surgical patients).

For example, the EURICUS-I study [34] has shown that in general ICUs the utilisation of exclusion criteria for SAPS II based on diagnosis alone leads to the exclusion of 23.1 % of the patients from the subsequent application of the models. More important than that percentage is the clustering observed at ICU level, with values ranging from 0 to 62.2 %. When we apply all the exclusion criteria described by the developers the numbers are even

greater, with a volume of excluded patients reaching 67.5 % of all admitted patients at ICU level. This implies indeed that the application of the model to some ICUs (with subsequent performance evaluation) will be based on not more than roughly 1/3 of the admitted patients.

In summary, when applying a specific model to an ICU, attention should be given to the number and type of excluded patients. Further measurements and actions should be undertaken only if the evaluation is based on a representative number of the admitted patients.

3.2 OUTCOME SELECTION

Outcome can be seen as one or more events in the course of a disease process, such as morbidity, mortality, time to recovery from disease, or quality of life. Until now, all general outcome prediction models in intensive care focus exclusively on hospital mortality. This measure is considered the gold standard since it is easy to define and measure, and represents a clinically very relevant endpoint. Its validity is subjected to debate since it may be prone to bias (some even speak of manipulation). It has been demonstrated that some hospitals tend to change the location of deaths (*e.g.* by discharging patients to die) and other hospitals discharge patients very early in the course of their disease [35,36].

The use of hospital mortality for the evaluation of the performance of the ICU can also be questioned [37]. To evaluate ICU performance by using overall hospital mortality suggests that the ICU performance is the only determinant of the hospital performance, which obviously is not the case. Post-ICU mortality varies considerably from institution to institution and can be independent from ICU factors, such as quality of care in general wards.

Recently, some models have been published with alternative outcomes, *e.g.* survival at six months [38]. They have not been yet widely applied and they solve only part of the pitfalls mentioned to above.

More research is needed to identify the appropriate outcome measures for performance evaluation. As reported at a recent consensus conference, mortality as such is insufficient for the assessment of ICU outcome, and it should therefore be supplemented by for example measurements of quality of life [39].

3.3 PREDICTOR VARIABLES SELECTION AND DATA COLLECTION

The next step in the development of an outcome prediction model is the choice of a preliminary list of candidate variables. These, usually selected by experts in the field, can include demographic, clinical and laboratory variables. Each should be relevant for the outcome of interest, supported by the available literature and with documentation explaining how they relate to the outcome, precisely defined, routinely available, and occurring

frequently enough to be meaningful. In addition, they should not be confounded with the outcome of interest unless present before ICU admission [40]. Some authors have recommended that the chosen variables be biological in nature, rather than sociological or behavioural [41], and based on clinical rather than administrative data [42,43].

Some considerations regarding the choice of variables are especially important. The variables should have a high degree of intra-observer reliability (the same observer collecting data on the same variable at two different moments should arrive at the same result) and inter-observer reliability (two different observers collecting data from the same patient at the same moment should arrive at the same result). Measurements of reliability should always be computed and reported, together with the evaluation of the degree of data completeness and correctness.

Special attention should be given to the choice of variables that are less prone to deliberate manipulation. This practice, termed *gaming* in the United States of America seems to be increasing. For example, a recent paper by Green and Wintfeld [44] describes the reported incidence of risk factors according to the New York State evaluation of the performance of cardiac surgeons. The authors observed that with the implementation of the program the reported rate of congestive heart failure rose from 1.7 % in 1989 to 7.6 % in 1991.

The collection of data can be done retrospectively or prospectively. The second approach is preferred since it allows the collection of more accurate and complete data. Special attention should be given to missing data, especially in the case of non-random patterns of missing data. Although several techniques have been described to deal with this problem [45-48] none is perfect [49]. As a consequence, the amount and pattern of missing values can influence the choice of variables [26].

3.4 ASSEMBLY OF THE MODEL

The following step is data analysis and data reduction. The initial list of candidate variables, submitted to a combination of logistic regression techniques [50], smoothing methods [51,52], and clinical judgement is reduced to a smaller number of variables to be included in the model. This reduction is based on the general scientific principle of parsimony: theories with simpler explanations are considered more plausible than more complex ones. It ensures that the model to be developed will have a higher precision, and will lead to more interpretable models. Attention should be given at this stage to multi-collinearity and interactions [53].

After the final set of variables is chosen and their weights evaluated, the variables are usually combined in a single score. In each case, the score results from the sum of the weights assigned to the chosen variables, according to the magnitude of change from the accepted normal values. This approach was chosen by the developers of APACHE III [26] and SAPS

II [31] but not in the case of MPM II [32].

The next step is to relate the aggregated score (or the chosen set of variables, as in the case of MPM II) to the outcome of interest. All presently available general outcome prediction models use multiple logistic regression analysis for that purpose. In this technique, the dependent or outcome variable *y* is related to a set of independent or predictive variables by the equation:

$$Y = b_0 + b_1x_1 + b_2x_2 \ldots b_kx_k$$

where b_0 is the intercept of the model, x_1 to x_k represent the predictor variables and b_1 to b_k are the estimated regression coefficients.

Then, a logistic transformation is applied, with the probability of death being given by:

$$Pr = \frac{e^{Logit}}{(1 + e^{Logit})}$$

where Pr represents the probability of death and *logit* is *Y* as described before. Logistic transformation has the propriety to transform an S-shaped relationship between two variables into a linear one (on the logit scale).

Figure 1 plots the relationship between SAPS II score and probability of mortality. In the extremes of the score (very low or very high values) the associated changes in the probability of mortality are quite small. For intermediate values, however, even small changes in the score are associated with greater changes in the probability of death.

technique, the dependent or outcome variable y is related to a set of independent or predictive variables by the equation:

$$Y = b_0 + b_1x_1 + b_2x_2 \ldots b_kx_k$$

where b_0 is the intercept of the model, x_1 to x_k represent the predictor variables and b_1 to b_k are the estimated regression coefficients.

Then, a logistic transformation is applied, with the probability of death being given by:

$$\Pr = \frac{e^{Logit}}{(1+e^{Logit})}$$

where Pr represents the probability of death and *logit* is Y as described before. Logistic transformation has the propriety to transform an S-shaped relationship between two variables into a linear one (on the logit scale).

Figure 1 plots the relationship between SAPS II score and probability of mortality. In the extremes of the score (very low or very high values) the associated changes in the probability of mortality are quite small. For intermediate values, however, even small changes in the score are associated with greater changes in the probability of death.

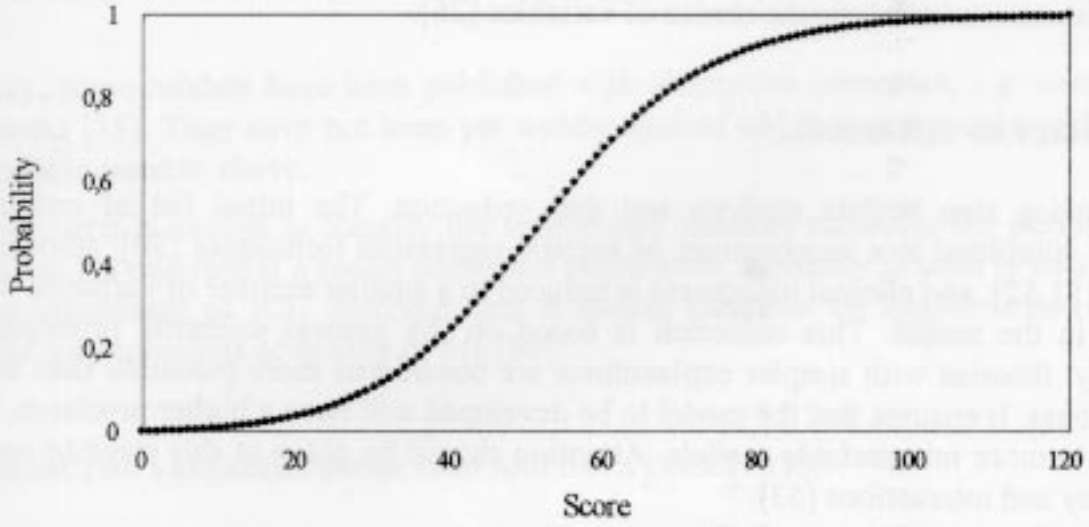

Figure 1. Relationship between the new Simplified Acute Physiology Score (SAPS II) and corresponding probability of hospital mortality, according to the logistic regression equation described by the developers of the model [31].

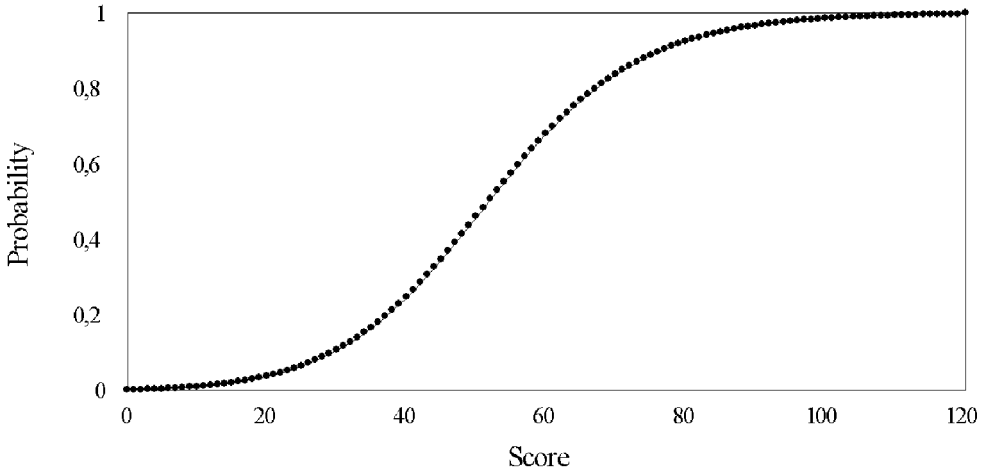

Figure 1. Relationship between the new Simplified Acute Physiology Score (SAPS II) and corresponding probability of hospital mortality, according to the logistic regression equation described by the developers of the model [31].

In order to fit multiple logistic regression models, we must keep in mind that they are based on certain assumptions, to be verified by the data analyst [54]. From these, the most important are:

Independence of observations: the patients in the data sample are statistically independent of one another.

Linearity assumption: on a certain scale of the dependent variable (*Y)*, each predictive variable (*x*) is linearly related to *Y*. In the case of logistic regression for binary responses, it is assumed that *x* is linearly related to the log odds of the response for patients sub-grouped by values of *x*. If this assumption is not verified then adequate transformations of the predictive variables should be found, *e.g.* restricted cubic spines [55].

Additive assumption: most regression models assume the additivity of effects of the predictors. If this assumption is not verified, then significant interactions should be tested and described.

Distributional assumptions: multiple logistic regression models assume that the distribution

of the residuals is binomial.

3.5 VALIDATION OF THE MODEL

All statistical models developed for prediction need validation, since it is necessary to know their predictive validity.

This validation process can be done in three different ways: cross-validation, external validation and temporal validation. In cross-validation, data are randomly divided in two portions, one for model development and the other for model validation, although other techniques can be used, *e.g.* bootstrapping or jackknife [56,57]. In external validation, data developed in one population are validated in an external, independent population. In temporal validation, a model developed at a specific point in time is later validated in the same setting at some future date.

All the general outcome prediction models in use today have been subjected to cross-validation. This step is necessary but not sufficient, since it does not assure *per se* the adequacy of the model when subjected to external validation. When the developers of the original models utilise randomisation to split the original database in two groups, development and validation samples, one may expect that all the variables (and non measured case mix factors) will be randomly distributed in the two sub-groups. Consequently, both sub-groups are expected to represent equal samples from the same underlying distribution and can not be considered true independent samples. The models analysed in this way are therefore expected to perform better on the validation sample than in an independent population. Independent validation in different populations is needed before general utilisation, since variations in case mix, local policies, quality of care and quality of data collection have been shown to affect the performance of the equations used to predict mortality [58-64].

The main question to be answered at this stage is the adequacy of outcome predictions when compared to the actual outcomes [65]. Three major issues must be evaluated [66]. The first is *calibration* or *reliability*, or how well the model predictions compare with the observed outcomes. The second is the model *discrimination*, *refinement* or *spread*, which evaluates how well the model can distinguish between observations with a positive or a negative outcome. A third source of model deviation can be the existence in the test set of *subsets of observations* in which the model does not perform well. Opposed to the first two, where a lot of research has been done and consensual techniques have emerged, there is little in the literature on how to identify these observations or what to do when the fit is unsatisfactory [66]. The evaluation of calibration and discrimination in the overall population has been named *overall goodness of fit*. The evaluation of the appropriateness of the predictions across sub-groups has been termed *uniformity of fit*. These issues will be reviewed in more detail in the following sections.

3.5.1 Overall goodness of fit

The evaluation of the overall goodness of fit comprises the evaluation of calibration and discrimination in the population under analysis.

Calibration evaluates the degree of correspondence between the estimated probabilities of mortality and the observed mortality in the sample under analysis. Four methods are usually employed in this analysis: overall observed/expected (O/E) mortality ratios, Flora's Z score [67], Hosmer-Lemeshow goodness-of-fit tests [50,68,69] and calibration curves.

Overall O/E mortality ratios are computed dividing the overall observed mortality rate (*i.e.* the actual number of deaths) by the predicted number of deaths (resulting from the sum of the individual probabilities of mortality assigned by the model); additional computations can be made to estimate the confidence interval for the ratio [70,71]. In a perfectly calibrated model this value should be one. This test is easy to perform but can be misleading. Consider the following example:

> A hypothetical model is used to predict the mortality of a population of 1000 patients with an actual mortality of 10 % (100 patients). Those patients included 500 less severe patients with an observed mortality rate of 5 % (25 patients) and 500 more severe patients with an observed mortality rate of 15 % (75 patients). The model, poorly calibrated, assigned the same probability of mortality to all patients (10 %), predicting in the overall population 100 deaths.
>
> Overall O/E ratio is one (100/100) but this ratio resulted from two errors: 25/50 = 0.5 (O/E ratio in less severe patients) and 75/50 = l.5 (O/E ratio in more severe patients).

Flora's Z score is based on a statistical technique that compares the number of survivors observed in the given data set with the number that would be predicted from the baseline survival curve. The difference is then standardised and compared to a table of the normal distribution [67]. The statistic used is:

$$Z = \frac{S - \sum_{i=1}^{n} Pi}{\sqrt{\sum_{i=1}^{n} PiQi}}$$

where *S* is the total number of survivors among the *n* patients, *Pi* is the probability of survival estimated by the model for the *i* patient and *Q* is 1-*Pi* or the probability of death estimated

by the model for the *i* patient. This approach suffers from the same drawbacks of the overall O/E ratios. For example, its application to the example given before will result in a perfect fit.

Hosmer-Lemeshow goodness-of-fit tests are two chi-square statistics proposed for the formal evaluation of the calibration of outcome prediction models [50,68,69]. In Hosmer-Lemeshow H test, patients are divided in 10 strata, according to their predicted probabilities: 0.0 to 0.1, 0.1 to 0.2 … 0.9 to 1.0. Then, a chi-square statistic is used to compare the actual and the expected number of deaths and the actual and expected number of survivors in each of the strata. The used formula is defined as:

$$C_g = H_g = \sum_{l=1}^{g} \frac{(o_l - e_l)^2}{e_l(1 - \bar{p}_l)}$$

where *g* is the number of groups (usually 10), o_l is the number of observed events occurred in the l-th group, e_l is the number of expected events in the same group and $\bar{p}_l$ is the average estimated probability, always in the l-th th group.

The resulting statistic is then compared with a chi-square table with 8 degrees of freedom (model development) or 10 degrees of freedom (model validation) in order to evaluate if the resulting discrepancies can be explained only by sampling variance. Large differences suggest that the model is not well calibrated. Hosmer-Lemeshow test C is similar, with the 10 strata being formed based on deciles of the predicted mortalities. The same authors demonstrated that the grouping method at the basis of the C statistic is preferable to the one based on fixed cut points, especially when many of the estimated probabilities are small [50]. These tests are currently regarded as an obligatory part of calibration evaluation [40], although they are also subjects of criticism [72]. It should be noted that the analysed sample must be large enough to have the statistical power needed for the detection of lack of fit [73].

Calibration curves are also used for reporting information on the calibration of a model. As Figure 2 shows, this type of charts compares the observed mortality rate with the one expected by the model. They can be misleading since the number of patients in each group tend to go down from left to right (*i.e.* when we move from low probabilities to high probabilities of death) and as a consequence even small non-significant differences in the higher severity groups appear visually more important than significant differences in the low probability groups.

They are not a formal statistical test of the adequacy of the model, being used only for information purposes. However, the publication of calibration curves has been recommended by a recent consensus conference as part of the standard assessment of the validation of a

Flora's Z score is based on a statistical technique that compares the number of survivors observed in the given data set with the number that would be predicted from the baseline survival curve. The difference is then standardised and compared to a table of the normal distribution [67]. The statistic used is:

$$Z = \frac{S - \sum_{i=1}^{n} Pi}{\sqrt{\sum_{i=1}^{n} PiQi}}$$

where S is the total number of survivors among the n patients, Pi is the probability of survival estimated by the model for the i patient and Q is 1-Pi or the probability of death estimated by the model for the i patient. This approach suffers from the same drawbacks of the overall O/E ratios. For example, its application to the example given before will result in a perfect fit.

Hosmer-Lemeshow goodness-of-fit tests are two chi-square statistics proposed for the formal evaluation of the calibration of outcome prediction models [50,68,69]. In Hosmer-Lemeshow H test, patients are divided in 10 strata, according to their predicted probabilities: 0.0 to 0.1, 0.1 to 0.2 ... 0.9 to 1.0. Then, a chi-square statistic is used to compare the actual and the expected number of deaths and the actual and expected number of survivors in each of the strata. The used formula is defined as:

$$\hat{C}_g = \hat{H}_g = \sum_{l=l}^{g} \frac{(o_l - e_l)^2}{e_l(1-\bar{\pi}_l)}$$

where g is the number of groups (usually 10), o_l is the number of observed events occurred in the l-th group, e_l is the number of expected events in the same group and $\bar{\pi}_l$ is the average estimated probability, always in the l-th th group.

The resulting statistic is then compared with a chi-square table with 8 degrees of freedom (model development) or 10 degrees of freedom (model validation) in order to evaluate if the resulting discrepancies can be explained only by sampling variance. Large differences suggest that the model is not well calibrated. Hosmer-Lemeshow test C is similar, with the 10 strata being formed based on deciles of the predicted mortalities. The same authors demonstrated that the grouping method at the basis of the C statistic is preferable to the one based on fixed cut points, especially when many of the estimated probabilities are small [50]. These tests are currently regarded as an obligatory part of calibration evaluation [40], although they are also subjects of criticism [72]. It should be noted that the analysed sample must be large enough to have the statistical power needed for the detection of lack of fit [73].

Calibration curves are also used for reporting information on the calibration of a model. As Figure 2 shows, this type of charts compares the observed mortality rate with the one

model [40].

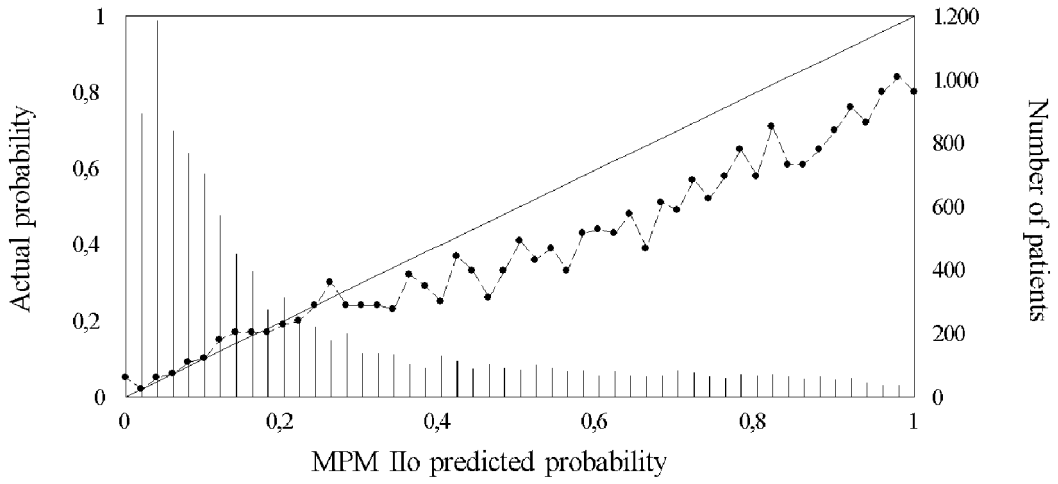

Figure 2. Calibration curve for the new admission Mortality Probability Model (MPM II_0) in EURICUS-I sample. The *solid line* represents perfect correspondence between actual and predicted risk of death and the *dotted line* the observed versus predicted risk of death. *Bars* provide the distribution of patients in the analysed groups.

Discrimination evaluates how well the model can distinguish between observations with a positive or a negative outcome. This assessment can be done by a non-parametric test like Harrell's c-index, using the rank of the magnitude of the assessment error [54]. This index measures the probability that for two randomly chosen patients, the one with the higher probabilistic prediction has the outcome of interest. It has been shown that this index relates directly to the area under a receiver operating characteristic (ROC) curve and can be obtained as the parameter from the Mann-Whitney-Wilcoxon rank sum test statistic [74].

The concept of the *area under the ROC curve* derives from psychophysics and has been applied in signal processing. In a ROC curve, a series of two-by-two tables is constructed with cut points that vary from the lowest possible value to the highest possible value of the score. For each table, the true positive rate (or sensitivity) and the false positive rate (or 1 minus specificity) is computed. The final plot of all pairs of true positive rates versus false positive rates is the visual representation of the ROC curve (Figure 3).

The interpretation of the area under the ROC curve is easy: a model with a perfect discrimination has an area of 1.0, a model which discrimination is no better than chance an area of 0.5. For third generation outcome prediction models this area is usually over 0.80 [26,31,32].

Methods for comparing the areas under ROC curves have been described [75-77], but can lead to misleading conclusions if the shape of the curves is different [78]. It has however been shown that this method is not as affected by the size of the sample as calibration measures [73].

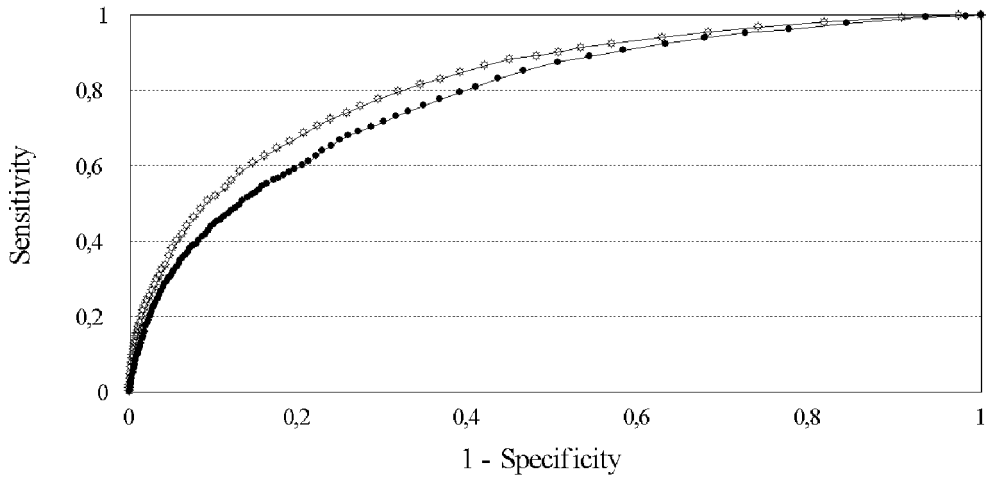

Figure 3. Receiver operating characteristic (ROC) curves for the new Simplified Acute Physiology Score (SAPS II) (*) and the new admission Mortality Probability Model (MPM II_0) (•) in EURICUS-I sample. For both models are presented the relationship between true positives (*sensitivity*) and false positives (*one minus specificity*).

Other measures have been used, based on *classification tables*, with authors reporting sensitivity, specificity, positive and negative predictive values and overall correct classification rates. They are, however, of limited utility in the validation of a model, because they have to use a fixed cut point (usually 10, 50 or 90 %). Take for example the following case:

> A hypothetical model is used to predict the mortality of a population of 1000 patients with a mortality rate of 12 %. Suppose that the application of the model would result

in all the patients being classified with a low risk of mortality (10 to 14 %).

So, in the aggregate, we should expect that a certain number of patients would die (10 to 14 %). However, a classification table with a 50 % cut point would classify those patients as *predicted survivors* and result in a very poor sensitivity when comparing them to *observed survivors*. However, in the present case, the model is predicting outcome in an accurate way.

Moreover, they can depend on the mortality rate of the sample, with models having generally low values on sensitivity especially when the analysed sample have a relatively high proportion of patients with low probabilities of mortality, since relatively few patients will have a probability of mortality greater than the chosen decision criteria [79].

The important difference between calibration and discrimination should be kept in mind. Lets look at the following two examples:

A hypothetical model is used to predict the mortality of a population of 1000 patients with an overall mortality rate of 20 %. 500 men (with a mortality rate of 40 %) and 500 women (with a mortality rate of 0 %) compose this population. Suppose that the application of the model would result in all patients (men and women) being classified as having a risk of mortality of 20 %. So, in the aggregate, the model would demonstrate very good calibration (observed 200 deaths, predicted 200 deaths) but a discrimination of no better than chance (area under ROC curve 0.50). In this case the model is very well calibrated but useless in the daily practice.

Another hypothetical model is used to predict the mortality in the same population. Suppose that the application of the model would result in all male patients who died being classified as having a risk of mortality of 60 % and all male who lived and all women classified as having a probability of 40 %. So, in the aggregate, the model would demonstrate a perfect discrimination (area under ROC curve 1.0) but a very poor calibration (observed 300 deaths, predicted 440 deaths).

The relative importance of calibration and discrimination varies with the utilisation that will be given to the model, with some authors arguing that for research or quality assurance purposes (group comparisons) calibration is especially important [37] and that for decisions about individual patients both descriptors are very relevant [80].

From a methodological point of view poor calibration of a model can be corrected. However, there is nothing that can adjust a model when it presents poor discrimination [48,66,81].

3.5.2 Uniformity of fit

The evaluation of the calibration and discrimination in the population under analysis is now

standard practice. More complex and less frequently performed is the identification of subsets of observations for which the model does not perform well. These observations are analogous to the influential observations in model building and its contribution to the overall deviation of the model can be exceptionally high [66].

The most important subsets are related to case mix, that is, the baseline characteristics of the population that are supposed to be related to the outcome of interest (Table 3).

Table 3. Principal components of case mix

1. Location in the hospital before ICU admission
2. Surgical status (non-operative, elective surgery, unscheduled surgery)
3. Degree of physiologic derangement
4. Physiologic reserve (age, chronic illnesses)
5. Diagnosis

The evaluation of the impact of case mix not directly related to severity of illness on the performance of outcome prediction models has been subject to less investigation. Although some authors such as Rowan and Goldhill in the United Kingdom [60,82] and Apolone and Sicignano in Italy [63,83] have suggested that the performance of a model could depend to a large extent on the composition of the population being studied, no consensus exists regarding how to define or evaluate the behaviour of a model in sub-populations.

Recently, we proposed the evaluation of these issues through a formal test of performance, comprising discrimination, calibration and O/E mortality ratios within clinically relevant sub-groups, in order to identify whether performance is identical in all groups (uniformity of fit) [84].

An example of such potential bias is the effect of previous therapy on the degree of physiologic derangement present at admission to the ICU (lead-time bias) [85,86]. If the patient has been stabilised prior to ICU admission (*e.g.* in the operative theatre or in the emergency room), the degree of physiologic derangement present at ICU admission will not reflect the physiologic severity of the underlying cause for admission. This problem can be less severe in the case of MPM II, since the variables included are not easily affected by acute therapy and are independent from the acute diagnosis. APACHE III accounts for this problem in the equations used for mortality prediction. For all the other systems the only way to estimate its influence is to analyse the performance of the score in subgroups defined by location in the hospital prior to ICU admission.

3.6 MODEL UPDATES AND MODIFICATIONS

All outcome prediction models require periodic updating. Changes in the population baseline characteristics, improvements in the outcome of major diseases or the introduction of new diagnostic tests with improved accuracy, all imply a modification in the models used. Moreover, the utilisation of a model outside its development population can also require modifications (customisation). Two recent examples in the literature exemplify this problem.

Sirio et al. in Japan [59], utilising APACHE II, demonstrated lower risk adjusted mortality for Japanese ICUs than for ICUs in the United States of America. They attributed the results to a different case mix, namely the frequency in Japan of patients with elective oesophageal and gastric cancer surgery. APACHE II, which was developed in 1985 in 13 American hospitals, is today not calibrated to be of use in Japan. Its utilisation will require recalibration, that is, the computation of new equations relating severity of illness to mortality.

In Italy, Apolone et al. [63] demonstrated very clearly that SAPS II, developed 5 years ago, was not able to adequately predict mortality in a sample of 99 Italian ICUs. After recalibration, the performance of the model has improved but not all the problems were solved.

More recently, our group faced this problem when using SAPS II in a large international study (EURICUS-I). We have shown that recalibration of a model is feasible, preferably by re-computing all the coefficients and not only by changing the relation between the final score and outcome. However, some problems met with when the model was applied to this population remained to be solved [87,88].

These modifications can also be necessary for the application of a general outcome prediction model to a specific group of patients, *e.g.* sepsis patients. This technique has been applied in the past with success [89-91] but it is not certain that it will be successfully applicable in all instances.

4. MODEL APPLICABILITY AND UTILITY

Once a model has been developed and validated its applicability and utility should be accessed. Can it be used outside specific research conditions? Is it useful?

The first question has yet to be answered. All the developers of general outcome prediction models have advocated its general applicability in the intensive care setting. However, the literature contains sufficient examples of models developed in large populations that failed later when applied in other settings [27,58-64]. The question can only be answered by testing

the overall goodness of fit and the uniformity of fit of a model in the population in which this specific model is to be applied.

Another important problem is the frequent lack of adherence to the application rules of the model used. Different definitions of the variables, time frames for data collection, frequency of assessment, exclusion criteria and handling of data prior to analysis all can potentially affect the application of a model to a different population [92,93]. On the other hand, a strict adherence to the original rules is often not possible since these are quite frequently not available in detail. A recent review of these problems has been published, raising many doubts about the applicability of the models outside a research setting [94].

The potential applications of these models, and thus their usefulness, can be analysed separately for individual patients and for groups.

4.1 INDIVIDUAL PATIENTS

Some evidence suggests that statistical models can perform as well or better than clinicians in predicting outcomes [95-102]. Statistical models may help clinicians in the complex process of decision-making [103-105]. This view is however not universally supported [106-108], and can be specially controversial in the process of withholding or withdrawing treatment [109]. Moreover, the application of different models to the same patient frequently results in very different predictions [110]. An example is given in Figure 4, in which the predicted probabilities of death in the Portuguese study using two different models are plotted. Many problems must be solved to find a method applicable to individual patients and current recommendations preclude the use of statistical models for the prediction of outcome of individual patients [39].

The first problem is the probabilistic nature of the predictions compared with the binary nature of the decisions that have to be taken. All the existing models give us probabilistic predictions. At their very best, an accurate model can only say that there are 46 % of chances that a specific patient will die. This prediction can be entirely true but useless in the process of decision making. The clinician can not know if that specific patient will fall in the 46 of each 100 that will die or in the 54 of each 100 that will survive. Consequently, to withdraw therapy in an individual patients based in the information given by such a model, is impossible. Attempts to use daily evaluation and adjustments to increase the certainty of the predictions have been made [111-114] but have failed to fulfil previous expectations when applied in other settings [115]. They can however be of use in discussions with relatives if they are properly calibrated for that specific ICU.

Another problem is the time frame needed for prediction. For a model to be of use in the process of decision making, the prediction must be given when the decision is to be taken. This precludes its use in the decision to admit a patient to intensive care (since most models

need at least 24 hours in the ICU before a prediction can be made). Even at later moments in time, when the model reaches a conclusion, only a small subset of patients can benefit from it. As an example, the utilisation of one of such systems (Riyadh Intensive Care Program) using computerised trend analysis of daily APACHE II scores corrected for organ system failure [111] in two ICUs resulted in a sensitivity of 14 % in one study [115] and of 23 % in another [116].

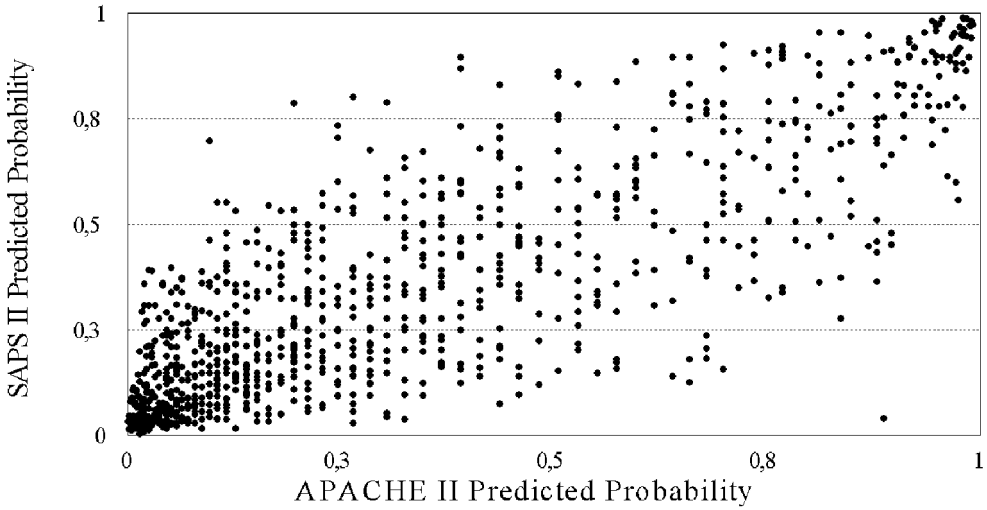

Figure 4. Plot of the new Simplified Acute Physiology Score (SAPS II) versus the Acute Physiology and Chronic Health Evaluation (APACHE II) Score in the Portuguese Multicentric study. SAPS II and APACHE II are highly related but there are a large number of outliers. The predictions seem more related in the extremes of risk than in the middle range.

It should be noted that the use of outcome prediction models has been advocated for decisions such as the use of total parenteral nutrition [117] or the identification of futility in intensive care, allowing the early withdrawal of therapy [116]. Some authors have inclusively demonstrated that the provision of the attendant clinician with objective estimates of patient outcome did not affect unfavourably the quality of care while reducing costs and increasing the availability of beds [118]. Maybe that in the future, with the application of other techniques [119-122], the remaining problems can be solved and the statistical methods will gain a place in the process of decision making in intensive care.

A field where the outcome prediction models have gained acceptance is clinical trials. They have been used in the aggregated level to assure the comparability of groups [123-125] and also for the identification of eligible patients and as a criteria for randomisation [126].

4.2 GROUP EVALUATION

At the aggregated level, outcome prediction models have been proposed for two objectives: allocation of resources and evaluation of performance.

A lot of research has been done on the identification and characterisation of low risk patients [99,127-131]. This type of patients, that received only monitoring and floor care services during the ICU stay could be discharged earlier to less intensive areas in the hospital [105,132]. One may argue that these patients have a low risk both of need of active-life support and of mortality, exactly because they were treated in an ICU. In other words, their treatment in less intensive areas of the hospital could lead to a more frequent deterioration of their clinical situation [133].

Moreover, the cost of a patient in intensive care depends mainly on the amount of nursing workload required. Patient-related characteristics (diagnosis, degree of physiologic derangement) are not the only determinants of this consumption; this depends also on standing practices and policies of care in each ICU. To focus only on patient characteristics at admission or during the first hours in the ICU is to forget the subsequent process of care. It seems therefore more logical to focus on the amount of nursing manpower available versus the amount of nursing manpower required during the stay of the patient in the ICU to evaluate the match between provision and consumption of resources, and the appropriateness of their use. This strategy has been followed in the past with success [134-137] and seems preferable to strategies more based on the condition of the patient or on the ratio between observed and predicted length of stay [138-140].

Inversely, outcome prediction models have been used to identify patients expected to use more resources [141]. Unfortunately, such patients can rarely be identified at admission since their degree of physiologic derangement during the first day in the ICU is quite variable but usually not very high [142-144]. But, on the other hand, should such identification be possible, what would we do with that knowledge: withhold treatment in potentially expensive patients?

Another area in which outcome prediction models were used during the last decade is performance evaluation. Several researchers proposed the use of outcome prediction models to evaluate the performance of the ICU. This was done under the assumption that existent models can take into account the most important determinants of mortality (age, previous health status, admission diagnosis, and level of physiologic derangement) [14]. It has been proposed to use the standardised mortality ratio (SMR) to evaluate performance. This ratio is obtained by dividing the actual number of deaths by the predicted number of deaths (resulting from the sum of the individual probabilities computed by the model) in the sample under analysis. The interpretation of the SMR is easy: a ratio less than one implies a performance better than the reference group and a ratio greater than one implies a performance lower than the reference group. This measure has been used for purposes as

diverse as international comparisons of intensive care [18,59,60,145-147], hospital comparisons [14,17,21,27,64,138,139,148,149], audits of intensive care [150-154], and evaluation of managerial effectiveness [149,155,156].

Figure 5 gives an example of this methodology, in which ICU A shows a performance lower than the reference (its SMR is higher than one) while ICU B shows a performance higher than the reference (its SMR is lower than one).

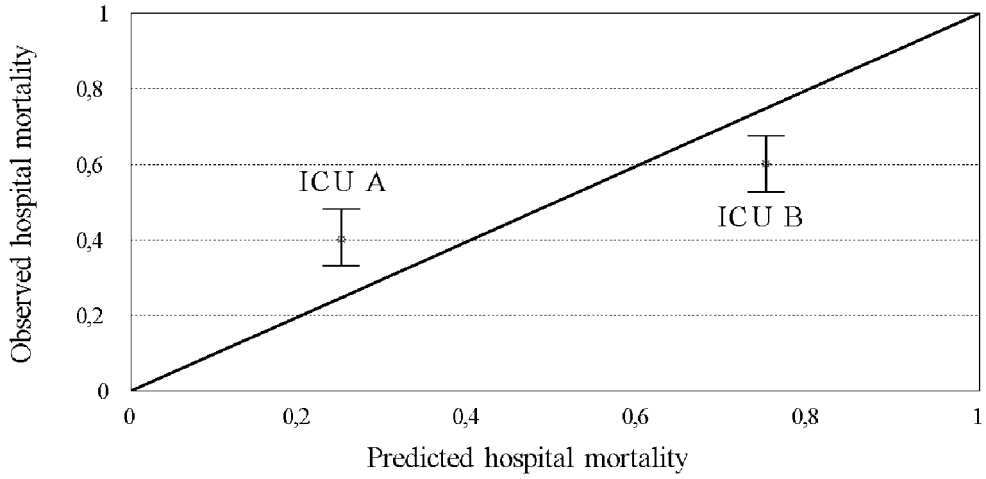

Figure 5. Use of the standardised mortality ratio (SMR) to evaluate the performance of the ICU. In this hypothetical example is plotted the SMR for two hypothetical ICUs: ICU A, with a ratio higher than 1 (performing *worst* than the reference) and ICU B with a ratio lower than 1 (performing *better* than the reference).

Prior to embracing this mode of utilisation, five questions should be answered.

Can the data needed for the computation of the models be collected in a standard and reliable way?

Can the models be used in the majority of the patients admitted to general intensive care units?

Are the presently available models able to take into account the differences in baseline patient characteristics known to influence mortality ?

Is the reference population adequately chosen and are the models well calibrated on that population?

Is the dimension of the sample under analysis large enough to yield the power for detecting significant differences?

Each of these assumptions has been challenged in the last years [60,92,93,96,157-161] and more research is required before this methodology can be universally applied.

The problem of sample size deserves a speciall word. As shown in Figure 6, the number of analysed patients is very important in the comparison of the ICU under analysis with the reference population. In small or quite heterogeneous populations we will not have the power needed to show significant differences, even if they exist (type II error).

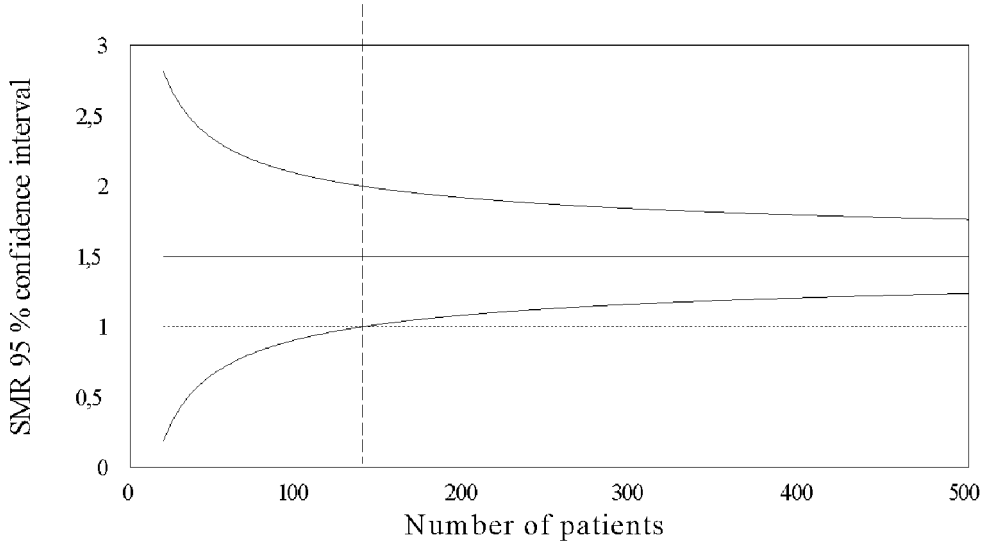

Figure 6. Relationship between the number of analysed patients and the 95 % confidence interval for the standardised mortality ratio (SMR). This hypothetical example describes an ICU with an overall SMR of 1.5 (observed mortality 15 %, predicted mortality 10 %). As shown, only with more than 130 patients analysed we can demonstrate that the SMR is significantly different from 1.

Moreover, the comparison between observed and predicted mortality, even if possible, should be more meaningful if performed in low-, medium-, and high-risk patients. Performance can vary inside ICUs according to the degree of severity of the admitted patients. This approach

has been advocated in the past on theoretical grounds [11,30,162] but it has been used in only one published study containing a number of patients in the middle and high-risk groups too small to allow definitive analysis [163]. This approach has also been used recently in the analysis of performance in the EURICUS-I study [34].

5. AIM OF THIS THESIS

In the preceding sections we have reviewed the methodological steps needed for the development and application of general outcome prediction models in intensive care. Our main focus of interest was the evaluation of performance of ICUs. Can it be measured? And if so: in which conditions?

The objective of the research described in this thesis is threefold. First, to evaluate the results of the application of a standard methodology for the measurement of performance in the intensive care setting. Subsequently, a more detailed analysis is undertaken addressing the reasons that could explain the incapability of the models to correctly adjust to patient baseline characteristics when applied to independent populations. Finally, an appraisal of the techniques that can be used to improve the predictive capability of the models is made.

The path of research followed was:

- to test the predictive accuracy of two general severity scores, SAPS II and APACHE II, when applied prospectively to a multicentric Portuguese database (**Chapter 2**);
- to test the generalizability of the findings when evaluated in a large multinational database, using SAPS II and MPM II_0 (**Chapter 3**);
- to study the main determinants of the length of stay in the ICU (LOS) and the way it is influenced by patient and hospital characteristics other than solely by the severity of illness of the patients; to address in more detail the effects of the LOS in the predictive performance of SAPS II (**Chapter 4**);
- to study the predictive ability (performance) of the models across relevant sub-groups, and to test the hypothesis that unequal performance is one of the main determinants of their frequent inability to accurately predict mortality when applied to independent populations (**Chapter 5**);
- to explore two different strategies to improve the predictive accuracy of MPM II_0, testing formally its effects on the overall performance and the performance across sub-groups of the model (**Chapter 6**).

6. REFERENCES

Reis Miranda D, Gyldmark M. Evaluating and understanding of costs in the intensive care unit. In: Ryan DW, ed. Current practice in critical illness. London: Chapman & Hall, 1996:129-49.

Kalb PE, Miller DH. Utilization strategies for intensive care units. JAMA 1989;261:2389-95.

Connors AF, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA 1996;276:889-97.

Dalen JE, Bone RC. Is it time to pull the pulmonary artery catheter ? JAMA 1996;11:916-8.

Rogers RM, Weiler C, Ruppenthal B. Impact of the respiratory intensive care unit in survival of patients with acute respiratory failure. Chest 1972;62:94-7.

Skidmore FD. A review of 460 patients admitted to the intensive therapy unit in a general hospital. Br J Surg 1973;60:1-16.

Feller I, Tholen D, Cornell RG. Improvements in burn care. JAMA 1980;244:2074-8.

Griner PF. Treatment of acute pulmonary edema: conventional or intensive care? Ann Intern Med 1972;77:501-6.

Piper KW, Griner PF. Suicide attempts with drug overdose: oucomes of intensive care vs. conventional floor care. Arch Intern Med 1974;134:703-6.

Hook EW, Horton CA, Schaberg DR. Failure of intensive care unit support to influence mortality from pneumococcal bacteremia. JAMA 1983;249:1055-7.

Teres D, Lemeshow S. Using severity measures to describe high performance intensive care units. Crit Care Clin 1993;9:543-54.

Jencks SF, Daley J, Draper D, Thomas N, Lenhart G, Walker J. Interpreting hospital data: the role of clinical risk adjustment. JAMA 1988;260:3611-6.

Daley J, Jenks S, Draper D, Lenhart G, Thomas N, Walker J. Predicting hospital mortality for Medicare patients: a method for patients with stroke, pneumonia, myocardial infarction, and congestive hearth failure. JAMA 1988;260:3617-24.

Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An evaluation of outcome from intensive care in major medical centers. Ann Intern Med 1986;104:410-8.

Berwick D. Continuous improvement as an ideal in health care. N Engl J Med 1989;320:53-6.

Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE - Acute Physiology and Chronic Health Evaluation: a physiologically based classification system. Crit Care Med 1981;9:591-7.

Knaus WA, Draper EA, Wagner DP, et al. Evaluating outcome from intensive care: A preliminary multihospital comparison. Crit Care Med 1982;10:491-6.

Knaus WA, Le Gall JR, Wagner DP, et al. A comparison of intensive care in the U.S.A. and France. Lancet 1982;642-6.

Wagner DP, Draper EA, Abizanda Campos R, et al. Initial International use of APACHE: an acute severity of disease measure. Med Decis Making 1984;4:297.

Le Gall JR, Loirat P, Alperovitch A, et al. A Simplified Acute Physiologic Score for ICU patients. Crit Care Med 1984;12:975-7.

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

2. Lemeshow S, Teres D, Avrunin J, et al. Refining intensive care unit outcome by using changing probabilities of mortality. Crit Care Med 1988;16:470-7.

3. Castella X, Artigas A, Bion J, Kari A, The European / North American Severity Study Group. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. Crit Care Med 1995;23:1327-35.

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

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

6. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619-36.

7. Bastos PG, Sun X, Wagner DP, Knaus WA, Zimmerman JE, The Brazil APACHE III Study Group. Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. Intensive Care Med 1996;22:564-70.

8. Bastos PG, Knaus WA, Zimmerman JE, Magalhães Jr A, Wagner DP, The Brazil APACHE III Study Group. The importance of technology for achieving superior outcomes from intensive care. Intensive Care Med 1996;22:664-9.

9. Holt AW, Bury LK, Bersten AD, et al. Prospective evaluation of residents and nurses as severity score data collectors. Crit Care Med 1992;20:1688-91.

10. Teres D, Lemeshow S. Why severity models should be used with caution. Crit Care Clin 1994;10:93-110.

11. 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;270:2957-63.

12. 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;270:2478-86.

13. Lemeshow S, Klar J, Teres D, et al. Mortality Probability Models for patients in the intensive care unit for 48 or 72 hours: a prospective, multicenter study. Crit Care Med 1994;22:1351-8.

14. Organization and management of Intensive Care: a prospective study in 12 European countries. Reis Miranda D, Ryan DW, Schaufeli WB, Fidler V, eds. Berlin Heidelberg: Springer-Verlag, 1997.

15. Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data. The

role of length of stay and comorbidities. JAMA 1988;260:2240-6.

16. McKee M, Hunter D. What can comparisons of hospital death rates tell us about the quality of care? In: Delamothe T, ed. Outcomes in clinical practice. London: British Medical Journal Press, 1994:108-15.

17. Schuster DP. Predicting outcome after ICU admission. The art and science of assessing risk. Chest 1992;102:1861-70.

18. Knaus WA, Harrell FE, Lynn J, et al. The SUPPORT prognostic model. Objective estimates for seriously ill hospitalized adults. Study to understand prognoses and preferences for outcomes and risks of treatments. Ann Intern Med 1995;122:191-203.

19. Suter P, Armagandis A, Beaufils F, et al. Predicting outcome in ICU patients: consensus conference organized by the ESICM and the SRLF. Intensive Care Med 1994;20:390-7.

20. Hadorn DC, Keeler EB, Rogers WH, Brook RH. Assessing the performance of mortality prediction models. Santa Monica, CA, RAND/UCLA/Harvard Center for Health Care Financing Policy Research, 1993.

21. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;313:793-9.

22. Green J, Wintfeld N, Sharkey P, et al. The importance of severity of illness in assessing hospital mortality. JAMA 1990;263:241-6.

23. Smith DW, Pine M, Baily RC, et al. Using clinical variables to estimate the risk of patient mortality. Med Care 1991;29:1108-29.

24. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. N Engl J Med 1995;332:1229-32.

25. Buck SF. A method of estimation of missing data values in multivariate data suitable for use with an electronic computer. J R Stat Soc B 1960;22:302-7.

26. Timm NH. The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika 1970;35:417-37.

27. Donner A. The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. Am Stat 1982;36:378-81.

28. Harrell Jr. FE, Lee KL, Mark DB. Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-87.

29. Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995;48:209-19.

30. Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: John Wiley & Sons, Inc., 1989

31. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 1979;74:829-36.

32. Cleveland WS. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Stat 1981;35:54.

33. Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariate methods. Boston: PWS-KENT Publishing Company, 1987

34. Harrell Jr. FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247:2543-6.

35. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989;8:551-61.

36. Miller RG. The jackknife - a review. Biometrica 1974;61:1-15.

37. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 1983;37:36-48.

38. Castella X, Gilabert J, Torner F, Torres C. Mortality prediction models in intensive care: Acute Physiology and Chronic Health Evaluation II and Mortality Prediction Model compared. Crit Care Med 1991;19:191-7.

39. Sirio CA, Tajimi K, Tase C, et al. An initial comparison of intensive care in Japan and United States. Crit Care Med 1992;20:1207-15.

40. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II 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 1993;307:977-81.

41. Abizanda Campos R, Rodriguez MT, Ferrandiz A, et al. Evaluation of SAPS II mortality prediction capability. Comparison with SAPS I and APACHE II [Abstract]. Intensive Care Med 1994;20:S1.

42. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: A prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. Crit Care Med 1994;22:1392-401.

43. Apolone G, D'Amico R, Bertolini G, et al. The performance of SAPS II in a cohort of patients admitted in 99 Italian ICUs: results from the GiViTI. Intensive Care Med 1996;22:1368-78.

44. Moreno R, Morais P. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. Intensive Care Med 1997;23:177-86.

45. Hadorn DC, Draper D, Rogers WH, Keeler EB, Brook RH. Cross-validation performance of mortality prediction models. Stat Med 1992;11:475-89.

46. Miller ME, Hui SL. Validation techniques for logistic regression models. Stat Med 1991;10:1213-26.

47. Flora JD. A method for comparing survival of burn patients to a standard survival curve. J Trauma 1978;18:701-5.

48. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. Comm Stat 1980;A10:1043-69.

49. Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982;115:92-106.

50. Gardner MJ, Altman DG. Statistics with confidence. London: British Medical Journal Press, 1989

51. Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression estimates. Stat Med 1995;14:2161-72.

52. Champion HR, Copes WS, Sacco WJ, et al. Improved predictions from a severity characterization of trauma (ASCOT) over trauma and injury severity score (TRISS): results of an independent evaluation. J Trauma 1996;40:42-9.

53. Zhu B-P, Lemeshow S, Hosmer DW, Klarm J, Avrunin J, Teres D. Factors affecting the performance of the models in the Mortality Probability Model and strategies of customization: a simulation study. Crit Care Med 1996;24:57-63.

54. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

55. Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.

56. McClish DK. Comparing the areas under more than two independent ROC curves. Med Decis Making 1987;7:149-55.

57. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.

58. Hilden J. The area under the ROC curve and its competitors. Med Decis Making 1991;11:95-101.

59. Lemeshow S, Le Gall J-R. Modeling the severity of illness of ICU patients. A systems Update. JAMA 1994;272:1049-55.

60. Kollef MH, Schuster DP. Predicting intensive care unit outcome with scoring systems. Underlying concepts and principles. Crit Care Clin 1994;10:1-18.

61. van Houwelingen JC, le Cessie S. Predictive value of statistical models. Stat Med 1990;8:1303-25.

62. Goldhill DR, Withington PS. The effects of casemix adjustment on mortality as predicted by APACHE II. Intensive Care Med 1996;22:415-9.

63. Sicignano A, Carozzi C, Giudici D, et al. The influence of length of stay in the ICU on power of discrimination of a multipurpose severity score (SAPS). Intensive Care Med 1996;22:1048-51.

64. Moreno R, Apolone G, Fidler V, Reis Miranda D. Evaluation of the uniformity of fit of SAPS II and MPM II_0 on an independent database [abstract]. Intensive Care Med 1996;22:S267.

65. Dragsted L, Jorgensen J, Jensen NH, al. e. Interhospital comparisons of patient outcome from intensive care: importance of lead-time bias. Crit Care Med 1989;17:418-22.

66. Escarce JJ, Kelley MA. Admission source to the medical intensive care unit predict hospital death independent of APACHE II score. JAMA 1990;264:2389-93.

67. Moreno R, Apolone G, Fidler V, Reis Miranda D. The impact of first level customization on the

uniformity of fit of SAPS II [abstract]. Intensive Care Med 1996;22:S267.

68. Moreno R, Apolone G, Fidler V, Reis Miranda D. The effects of first versus second level customization on the goodness of fit of MPM II_0 [abstract]. Intensive Care Med 1996;22:S267.

69. Knaus WA, Harrell FE, Fisher CJ, et al. The clinical evaluation of new drugs for sepsis. A prospective study design based on survival analysis. JAMA 1993;270:1233-41.

70. Le Gall J-R, Lemeshow S, Leleu G, et al. Customized probability models for early severe sepsis in adult intensive care patients. JAMA 1995;273:644-50.

71. Knaus WA, Harrell FE, LaBrecque JF, et al. Use of predicted risk of mortality to evaluate the efficacy of anticytokine therapy in sepsis. Crit Care Med 1996;24:46-56.

72. Fery-Lemmonier E, Landais P, Loirat P, et al. Evaluation of severity scoring systems in ICUs - Translation, conversion, and definition ambiguities as a source of inter-observer variability in APACHE II, SAPS and OSF. Intensive Care Med 1995;21:356-360.

73. Abizanda R, Balerdi B, Lopez J, et al. Fallos de prediccion de resultados mediante APACHE II. Analisis de los errores de prediction de mortalidad en pacientes criticos. Med Clin Barc 1994;102:527-31.

74. Rowan K. The reliability of case mix measurements in intensive care. Curr Opin Crit Care 1996;2:209-13.

75. Perkins HS, Jonsen AR, Epstein WV. Providers as predictors: using outcome predictions in intensive care. Crit Care Med 1986;14:105-10.

76. Silverstein MD. Predicting instruments and clinical judgement in critical care. JAMA 1988;260:1758-9.

77. Dawes RM, Faust D, Mechl PE. Clinical versus actuarial judgement. Sci Med Man 1989;243:1674-88.

78. Kleinmuntz B. Why we still use our heads instead of formulas: toward an integrative approach. Psychol Bull 1990;107:296-310.

79. McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? Med Decis Making 1989;9:125-32.

80. Poses RM, Bekes C, Copare FJ, et al. The answer to "what are my chances, doctor?" depends on whom is asked: prognostic disagreement and inaccuracy for critically ill patients. Crit Care Med 1989;17:827-33.

81. Poses RM, Bekes C, Winkler RL, Scott WE, Copare FJ. Are two (inexperienced) heads better than one (experienced) head? Averaging house officers prognostic judgement for critically ill patients. Arch Intern Med 1990;150:1874-8.

82. Winkler RL, Poses RM. Evaluating and combining physicians' probabilities of survival in an intensive care unit. Management science 1993;39:1526-43.

83. Chang RWS, Lee B, Jacobs S, Lee B. Accuracy of decisions to withdraw therapy in critically ill patients: clinical judgement versus a computer model. Crit Care Med 1989;17:1091-7.

84. Knaus WA, Rauss A, Alperovitch A, et al. Do objective estimates of chances for survival influence decisions to withhold or withdraw treatment? Med Decis Making 1990;10:163-71.

85. Zimmerman JE, Wagner DP, Draper EA, Knaus WA. Improving intensive care unit discharge decisions: supplementary physician judgement with predictions of next day risk for life support. Crit Care Med 1994;22:1373-84.

86. Branner AL, Godfrey LJ, Goetter WE. Prediction of outcome from critical illness: a comparison of clinical judgement with a prediction rule. Arch Intern Med 1989;149:1083-6.

87. Kruse JA, Thill-Baharozin MC, Carlson RW. Comparison of clinical assessment with APACHE II for predicting mortality risk in patients admitted to a medical intensive care unit. JAMA 1988;260:1739-42.

88. Marks RJ, Simons RS, Blizzard RA, et al. Predicting outcome in intensive therapy units - a comparison of APACHE II with subjective assessments. Intensive Care Med 1991;17:159-63.

89. Knaus WA, Wagner DP, Lynn J. Short-term mortality predictions for critically ill hospitalized adults: science and ethics. Sci Med Man 1991;254:389-94.

90. Lemeshow S, Klar J, Teres D. Outcome prediction for individual intensive care patients: useful, misused, or abused ? Intensive Care Med 1995;21:770-6.

91. Chang RWS, Jacobs S, Lee B. Predicting outcome among intensive care unit patients using computerised trend analysis of daily APACHE II scores corrected for organ system failure. Intensive Care Med 1988;14:558-66.

92. Chang RWS. Individual outcome prediction models for intensive care units. Lancet 1989;i:143-6.

93. Rogers J, Fuller HD. Use of daily acute physiology and chronic health evaluation (APACHE) II scores to predict individual patient survival rate. Crit Care Med 1994;22:1402-5.

94. Wagner DP, Knaus WA, Harrel Jr. FE, Zimmerman JE, Watts C. Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis. Crit Care Med 1994;22:1359-72.

95. Jacobs S, Arnold A, Clyburn PA, Willis BA. The Riyadh intensive care program applied to a mortality analysis of a teaching hospital intensive care unit. Anaesthesia 1992;47:775-80.

96. Atkinson S, Bihari D, Smithies M, Daly K, Mason R, McColl I. Identification of futility in intensive care. Lancet 1994;344:1203-6.

97. Chang RW, Jacobs S, Lee B. Use of APACHE II severity of disease classification to identify intensive-care-unit patients who would not benefit from total parenteral nutrition. Lancet 1986;1483-6.

98. Murray LS, Teasdale GM, Murray GD, et al. Does prediction of outcome alter patient management ? Lancet 1993;341:1487-91.

99. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991;115:843-8.

100. Schizas CN, Pattichis CS, Bonsett CA. Medical diagnostic systems: a case for neural networks. Technol Health Care 1994;2:1-18.

101. Baxt WG. Application of artificial neural networks to clinical medicine. Lancet 1995;346:1135-8.

102. Dybowski R, Weller P, Chang R. Prediction of outcome in critically ill patients using artificial neural network, synthesised by genetic algorithm. Lancet 1996;347:1146-50.

103. Greenman RL, Schein RMH, Martin MA, et al. A Controlled Clinical Trial of E5 Murine Monoclonal IgM Antibody to Endotoxin in the Treatment of Gram-Negative Sepsis. JAMA 1991;266:1097-102.

104. Ziegler EJ, Fisher CJ, Sprung CL, et al. Treatment of gram-negative bacteremia and septic shock with ha-1a human monoclonal antibody against endotoxin. A randomized, double-blind, placebo-controlled trial. N Engl J Med 1991;324:429-36.

105. Cohen J, Carlet J. INTERSEPT: an international, multicenter, placebo-controlled trial of monoclonal antibody to human tumor necrosis factor-alpha in patients with sepsis. International Sepsis Trial Study Group. Crit Care Med 1996;24:1431-40.

106. Gattinoni L, Brazzi L, Pelosi P, et al. A trial of goal orientated hemodynamic therapy in critically ill patients. N Engl J Med 1995;333:1025-32.

107. Henning RJ, McClish D, Daly B, et al. Clinical characteristics and resource utilization of ICU patients: implementation for organization of intensive care. Crit Care Med 1987;15:264-9.

108. Wagner DP, Knaus WA, Draper EA. Identification of low-risk monitor admissions to medical-surgical ICUs. Chest 1987;92:423-8.

109. Wagner DP, Knaus WA, Draper EA, et al. Identification of low-risk monitor patients within a medical-surgical ICU. Med Care 1983;21:425-33.

110. Zimmerman JE, Wagner DP, Knaus WA, Williams JF, Kolakowski D, Draper EA. The use of risk predictors to identify candidates for intermediate care units. Implications for intensive care unit utilization. Chest 1995;108:490-9.

111. Zimmerman JE, Wagner DP, Sun X, Knaus WA, Draper EA. Planning patient services for intermediate care units: insights based on care for intensive care unit low-risk monitor admissions. Crit Care Med 1996;24:1626-32.

112. Strauss MJ, LoGerfo JP, Yeltatzie JA, Temkin N, Hudson LD. Rationing of intensive care unit services. An everyday occurrence. JAMA 1986;255:1143-6.

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

114. Management of Intensive Care. Guidelines for better use of resources. Reis Miranda D, Williams A, Loirat P, eds. Dordrecht/Boston/London: Kluwer Academish Publishers, 1990.

115. Quality, efficiency, and organization of intensive care units in The Netherlands: an interdisciplinary study on medical and business aspects (Dutch language). Reis Miranda D, Spangenberg JFA, eds. Groningen: Van Denderen, 1992.

116. Reis Miranda D, Gimbrere J. The Netherlands. New Horiz 1994;2:357-63.

117. Moreno R, Reis Miranda D, Iapichino G. The efficiency of nursing manpower use in Europe [abstract]. Intensive Care Med 1996;22:S304.

118. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in Intensive Care Units. Ann Intern Med 1993;118:753-61.

119. Zimmerman JE, Shortell SM, Knaus WA, et al. Value and cost of teaching hospitals: a prospective, multicenter, inception cohort study. Crit Care Med 1993;21:1432-42.

120. Rapoport J, Teres D, Lemeshow S, Gehlbach S. A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. Crit Care Med 1994;22:1385-91.

121. Teres D, Rapoport J. Identifying patients with high risk of high cost. Chest 1991;99:530-1.

122. Cerra FB, Negro F, Abrams J. APACHE II score does not predict multiple organ failure or mortality in post-operative surgical patients. Arch Surg 1990;125:519-22.

123. Rapoport J, Teres D, Lemeshow S, Avrunin JS, Haber R. Explaining variability of cost using a severity of illness measure for ICU patients. Med Care 1990;28:338-48.

124. Oye RK, Bellamy PF. Patterns of resource consumption in medical intensive care. Chest 1991;99:695-89.

125. Zimmerman JE, Knaus WA, Judson JA, et al. Patient selection for intensive care: a comparison of New Zealand and United States Hospitals. Crit Care Med 1988;16:318-25.

126. Rapoport J, Teres D, Barnett R, et al. A comparison of intensive care unit utilization in Alberta and western Massachusetts. Crit Care Med 1995;23:1336-46.

127. 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;23:1177-83.

128. Le Gall JR, Loirat P, Nicolas F, et al. Utilisation d'un indice de gravité dans huit services de réanimation multidisciplinaire. Presse Médicale 1983;12:1757-61.

129. Zimmerman JE, Rousseau DM, Duffy J, et al. Intensive care at two teaching hospitals: an organizational case study. Am J Crit Care 1994;3:129-38.

130. Chisakuta AM, Alexander JP. Audit in Intensive Care. The APACHE II classification of severity of disease. Ulster Med J 1990;59:161-7.

131. Marsh HM, Krishan I, Naessens JM, et al. Assessment of prediction of mortality by using the APACHE II scoring system in intensive care units. Mayo Clin Proc 1990;65:1549-57.

132. Turner JS, Mudaliar YM, Chang RW, Morgan CJ. Acute Physiology and Chronic Health Evaluation (APACHE II) scoring in a cardiothoracic intensive care unit. Crit Care Med 1991;19:1266-9.

133. Oh TE, Hutchinson R, Short S, Buckley T, Lin E, Leung D. Verification of the Acute Physiology and Chronic Health Evaluation Scoring system in a Hong Kong intensive care unit. Crit Care Med 1993;21:698-705.

134. Moreno RP, Estrada H, Pereira E, Massa L. Movimento assistencial da Unidade de Cuidados Intensivos Polivalente do Hospital de Santo António dos Capuchos. Acta Med Port 1994;7:13-20.

135. Zimmerman JE, Shortell SM, Rousseau DM, et al. Improving intensive care: observations based on organizational case studies in nine intensive care units: a prospective, multicenter study. Crit Care Med 1993;21:1443-51.

136. Shortell SM, Zimmerman JE, Rousseau DM, et al. The performance of intensive care units: does good management make a difference ? Med Care 1994;32:508-25.

137. Park RE, Brook RH, Kosecoff J, et al. Explaining variations in hospital death rates: ramdomness, severity of illness, quality of care. JAMA 1990;264:484-90.

138. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland - I: Variations in case mix of adult admissions to general intensive care units and impact on outcome. Br Med J 1993;307:972-7.

139. Best WR, Comper DC. The ratio of observed-to-expected mortality as a quality of care indicator in non-surgical VA patients. Med Care 1994;32:390-400.

140. Fisher M. Intensive care: do intensivists matter ? Intensive Care World 1995;12:71-2.

141. Fisher M, Herkes RG. Intensive care: speciality without frontiers. In: Parker M, Shapiro MJ, Porembka DT, eds. Critical Care State of the Art. California: Society of Critical Care Medicine, 1995:15:9-27.

142. Teres D, Lieberman S. Are we ready to regionalize pediatric intensive care? Crit Care Med 1991;19:139-40.

143. Pollack MM, Alexander SR, Clarke N, et al. Improved outcomes from tertiary center pediatric intensive care: a statewide comparison of tertiary and nontertiary care facilities. Crit Care Med 1990;19:150-9.