Chapter 8

Summary
This thesis addresses the measurement of the performance of the ICU. The focus is on the clinimetric tools that have been proposed to this evaluation.

We reviewed the underlying theoretical principles used in the construction and application of general outcome prediction models. The models were then used on independent databases in order to evaluate if they were able to adequately control for patient baseline characteristics, namely, the severity of disease. Subsequently, a more detailed analysis was undertaken addressing the reasons why the models could not adjust correctly to patient characteristics when applied to other populations. Finally, an appraisal of the techniques that can be used to improve the predictive capability of the models was made.

On chapter 1, the available outcome prediction models are presented, with emphasis on the most used today: APACHE II and III, SAPS II and MPM II. The steps needed for the development of such models are outlined: selection of the patients, choice of outcome and predictor variables, data collection and assembly of the models with special attention to the methodological problems involved.

The importance of the validation of the models is stressed. The choice of methods and their application are presented. Finally, we address the utility, applicability and potential applications of the models and discuss the necessity of updates and modifications.

In chapter 2 we use formal statistical analysis to evaluate the accuracy of two of these models, SAPS II and APACHE II when applied in an independent database. The studied population comprises data collected during a period of three months in 1094 patients consecutively admitted to 19 mixed medical-surgical ICUs in Portugal. After application of the exclusion criteria described by the developers of the models, the analysed population comprised 982 patients.

In this cohort, discrimination was better for SAPS II than for APACHE II (SAPS II: area under the receiver operating characteristic curve 0.817, standard error 0.015; APACHE II: respectively 0.787, 0.015; \( p < 0.001 \)); however, both models presented a poor calibration, with significant differences between observed and predicted mortality (Hosmer-Lemeshow goodness-of-fit tests H and C, \( p < 0.001 \)). In a stratified analysis, this study was unable to demonstrate any definite pattern of association between the poor performance of the models and specific subgroups of patients except for the most severely ill patients, in which both models overestimated mortality.

We conclude that SAPS II performed better than APACHE II in this independent database. However, the results shows that they can not be used to analyse quality of care or performance among ICUs in the target population, at least without appropriate customisation.

In chapter 3, a similar methodology is applied to a large international database (EURICUS-I). The performance of SAPS II and MPM II \( _6 \) was assessed on data collected during a period
of four months from 16060 patients consecutively admitted to 89 ICUs from 13 European areas. After application of the original exclusion criteria, the analysed group comprised 10027 patients.

We observed that, despite having a good discriminative capability, as measured by the area under the receiver operating characteristic curves (SAPS II: 0.822, standard error 0.005; MPM II: 0.785, standard error 0.006) both models presented a poor calibration, with significant differences between observed and predicted mortality (Hosmer-Lemeshow goodness-of-fit tests H and C, \(p < 0.0001\)). Both SAPS II (predicted risk > 40 \%) and MPM II (predicted risk > 30 \%) overestimated the risk of death. The evaluation of the uniformity of fit of SAPS II and MPM II demonstrated large variations across the various sub-groups of patients.

We conclude that the original SAPS II and MPM II models did not accurately predict mortality in the studied population. This study calls for the need of previous validation prior to the application of general outcome prediction models in distinct populations. Furthermore, in stratified analysis, there was evidence that some of the inaccuracies observed could be linked to specific patient baseline characteristics.

To formally test this hypotheses, two complementary studies were done.

In chapter 4 we studied the relations between patient and hospital baseline characteristics and the length of stay (LOS) in the ICU; furthermore, the effects of the LOS on the performance of SAPS II were evaluated.

The study used EURICUS-I database, as described before. After applying SAPS II exclusion criteria, the analysed population comprised data of 11574 patients.

Overall, LOS was greater in medical patients, patients admitted from other ICUs/other hospitals or from the ward, patients with non-operative respiratory and trauma (both non-operative and post-operative) diagnoses, and in non-survivors. Greater LOS was observed in university hospitals. The presence of a step-down unit in the hospital did not reduce the LOS in the respective ICU.

SAPS II predicted risk of death was significantly higher in patients with longer stay in the ICU, with values ranging from 16.6 \% for patients with a LOS less than 1 day to 45.8 \% for patients with a LOS greater than one month. The relation between SAPS II score and LOS was not linear, with lower LOS on both extremes of severity (low/high SAPS II). This non-linearity could be explained by different relationship between SAPS II score and LOS in survivors and non-survivors. The performance of SAPS II in this population was substantially affected by the length of stay in the ICU of the analysed patients, both in discriminative power (area under ROC curve ranging from 0.888 in patients with a LOS less than one day to 0.594 for patients with a LOS greater than two weeks) and calibration (mainly in the patients with longer stays).
In conclusion, the results shows that the length of stay of ICU patients is influenced by patient and hospital characteristics other than solely by the severity of illness of the patients. The mortality predictive performance of the SAPS II loses significantly its accuracy when concerning groups of patients with LOS greater than one week. The results of this study could not document that the existence of step-down units do have a beneficial influence on the use of the ICU facilities.

Chapter 5. Using the EURICUS-I database, a formal statistical analysis of the performance of SAPS II and MPM II₀ models was performed across relevant subgroups of patients (uniformity of fit). The analysed group comprised data of 10027 patients. Stratification was made using two different strategies: a) patient-related criteria included in the models and for which we should expect that the models take them into account (age group and type of patient) and b) other case mix criteria not included in the models (location in the hospital before ICU admission and diagnostic category). Together with severity of illness, this set of variables comprehend the basic baseline categories of patient characteristics. Afterwards, the performance of the models was evaluated in each of these mutually exclusive subgroups using formal tests of discrimination, calibration and observed/expected mortality ratios.

Better predictive accuracy was achieved in elective surgery patients admitted from the operative room/post-anaesthesia room with gastrointestinal, neurologic or trauma diagnoses, and younger patients with non-operative neurological, septic or trauma diagnoses. All these characteristics appear to be linked to a lower severity of illness, with both models overestimating mortality in the more severely ill patients.

In conclusion, very large differences were apparent in relevant subgroups, varying from excellent to almost random predictive accuracy. These differences can explain some of the difficulties of these models to accurately predict mortality when applied to different populations with distinct patient baseline characteristics. This study stresses the importance of evaluating multiple diverse populations to generate the design set and of methods to improve the validation set before extrapolations can be made to new populations. It also underlines the necessity of a better definition of the patients baseline characteristics in the samples under analysis and the formal statistical evaluation of the application of the models to specific subgroups.

Is customisation able to solve the problems detected? This technique has been suggested as an important tool to improve the accuracy of general outcome prediction models in case an unbiased estimation of the probabilities of mortality is required.

In chapter 6 two different strategies to improve the fit of MPM II₀ were explored, testing their effects on the overall goodness of fit and in the uniformity of fit of the model. The application of the exclusion criteria for MPM II₀ to the original 16060 patients data resulted in a study group comprising data of 10397 patients.

After splitting the database randomly in two sub-sets (development and validation) two new
logistic regression equations relating MPM II_0 to mortality were fitted; the first with the original logit of MPM II_0 as independent variable (first level customisation) and the second with all the 15 original variables (second level customisation). Discrimination (area under ROC curve), Hosmer-Lemeshow goodness of fit tests H and C and observed/expected mortality ratios were evaluated in both samples and within relevant subgroups.

The discriminative capability of the models was only slightly affected by customisation (0.810 vs 0.803), remaining lower than in the original description of the MPM II_0 (0.824). Calibration improved, with Hosmer-Lemeshow goodness of fit tests H and C showing a good calibration of the models. However, the formal evaluation of discrimination, calibration and observed/expected mortality ratios across relevant subgroups showed to be poor in some groups.

In this ICU patients database, second level customisation was more effective than first level customisation in improving the overall goodness of fit of MPM II_0, and it should probably be chosen as a preferential strategy to improve the fit of a model when the sample size is large enough. However, it had only a slight impact on discrimination. Moreover, its effect on the uniformity of fit is insufficient to overcome the problems that can arise when the model is applied in populations in which the case mix is distinct from that of the population where it was originally developed.

Finally, in chapter 7 the implications of the findings for the medical profession are discussed. This work demonstrates clearly the inappropriateness of the standardised mortality ratio in the evaluation of performance. Many of the differences between predicted and observed mortality can be assigned to differential performance of the instruments across relevant subgroups. We tested formally this hypothesis and conclude that, even with better calibrated instruments, different case-mix characteristics in different ICUs will have an important impact on the relation between predicted and observed mortality. These conclusions preclude actually their use for performance evaluation outside an experimental setting.