Information for intensive care evolution: methods to assess and improve data quality and data processing

Arts, D.G.T.

Citation for published version (APA):
Chapter 7

Prognostic models: External validation and the effect of sample size

Danielle G.T. Arts, Niels B. Peek, Rob J. Bosman, Peter H.J. van der Voort, Nicolette F. de Keizer
Chapter 7

Abstract

With this study we aimed to validate four prognostic models (SAPS II, APACHE II, MPM₀ II and MPM₂₄ II) on the Dutch National Intensive Care Evaluation (NICE) registry database. In addition we analysed the effects of sample size on measured performance.

For each model discrimination (AUC) and accuracy (mean squared error (MSE) and Mean Log Likelihood (MLL)) statistics, and three calibration statistics (Hosmer-Lemeshow H and Č, and Copas Z) were measured on data from 41,239 ICU admissions. We simulated the validation process with smaller datasets (n = 100, 250, 500, 750, 1000, 2500, 5000), randomly drawn from the database. The random selection process and the validation were repeated 500 times for each sample size.

Differences in performance between the models, except MPM₀ II, are small. The AUC, MSE and MLL showed large variation with small sample sizes. The averages of these three measures were not influenced by the sample size. With larger sample sizes the calibration statistics increased and lack-of-fit appeared more frequently. With smaller sample sizes there is considerable variation in the model that appears to perform best.

We concluded that the applied calibration statistics are highly influenced by the size of the validation sample. When using a small validation dataset, it is largely a matter of chance which of the four models outperforms the others. Based on these results validation of such models should not be performed on datasets smaller than 2500 records.
Introduction

Prognostic logistic regression models are important tools to provide estimates of patient outcome probabilities. Within the field of intensive care medicine considerable research is aimed at developing and testing these models. Some well known prognostic models that resulted from these initiatives are the Simplified Acute Physiology Score II (SAPS II) [1], the Acute Physiology and Chronic Health Evaluation II (APACHE II) [2] and the Mortality Probability Model II system (MPM II) [3]. By means of a logistic regression equation these models provide for each ICU admission an estimate of the probability of in-hospital death. The MPM II system consists of four submodels, which were designed to provide an estimate of the probability of in-hospital death at four different times during the ICU stay: at admission (MPM₀ II); 24 hrs (MPM₂₄ II); 48 hrs (MPM₄₈ II); 72 hrs (MPM₇₂ II).

Mortality predictions provided by these prognostic models enable, for example, the stratification of patients for enrolment in clinical trials and controlling for severity of illness in auditing [4,5]. As an example of the latter application, prognostic models are being used for the calculation of the ratio of observed to expected mortality for assessment of the clinical performance of ICUs.

Before a prognostic model can be used, its generalizability must be assessed. Therefore, its performance needs to be measured on an external data set, different from the one that was used to develop the model. Such a measurement may cover several aspects of the performance of a prognostic model. They include the ability to distinguish between patients who do or do not experience the event of interest (discrimination) [5] and the extent to which the predicted event rates for groups of patients (calibration) or individuals (accuracy) are in concordance with the observed events. These aspects will be further explained in this paper. According to Justice et al. [6] it is most desirable to have prediction models validated across multiple independent investigators, geographic sites, and follow-up periods. Each of the validation results then provides evidence for the model's ability to generalize to new settings.

Since their development, the APACHE II, SAPS II and MPM II models have been validated in a number of different settings. The results of performance measurements that were published in the past 10 years differ with respect to the best performing model and the decision whether to apply a model or not. Whereas one study [7] concludes that the discrimination of the SAPS II model is superior to that of the APACHE II model, another [8] does not find a difference in discriminative ability between the two models. Similarly, some studies conclude that based on the measured calibration the SAPS II model is insufficient [9,10], whereas another concludes that calibration of the SAPS II is sufficient [11]. The variation in these results might be caused by temporal or geographical differences between the datasets that were used. However, the
variation in results might also be explained by the fact that some studies used a relatively small dataset (e.g. 300 observations) whereas others used a relatively large dataset (e.g. 16,000 observations).

In this study we validate the APACHE II, SAPS II, the MPM₀ II and the MPM₂₄ II model on the Dutch National Intensive Care Evaluation (NICE) registry database⁴. We use a larger array of validation measures than is commonly employed in the literature. Furthermore, we analyse the effects of sample size on measured performance, by simulating the validation process with smaller datasets. These smaller datasets were randomly drawn from the NICE registry database.

**Methods**

**Data**

In 1996 the Dutch National Intensive Care Evaluation (NICE) foundation has started the (voluntary) registration of data of admissions to Dutch ICUs. The NICE registry database contains for each ICU admission 108 demographic, diagnostic and physiologic variables collected within the first 24 hours of ICU admission and outcome data, such as length of stay on ICU and in-hospital mortality.⁴ Data collected include all raw data values necessary to calculate the original SAPS II [1], APACHE II [2], MPM₀ II and MPM₂₄ II [3] mortality probabilities. APACHE II, SAPS II, MPM₀ II and MPM₂₄ II mortality probabilities are calculated in the national database, at the NICE data coordinating centre. MPM₄₈ II and MPM₇₂ II were not calculated as physiologic data gathering was limited to the first 24 hours of IC treatment. Stringent measures are taken to control the data quality and uniformity of data collection procedures in the participating ICUs [12,13].

The validation dataset used in this study consisted of data from 83,824 admissions to 29 Dutch ICUs between January 1, 1999 and December 31, 2003 registered in the NICE database. The developers of the APACHE II, SAPS II, MPM₀ II and MPM₂₄ II models have defined criteria for populations on which the models can be applied. We combined the criteria of all four models to obtain one dataset that satisfied all criteria. According to the combined criteria we excluded patients with age < 18, patients with an ICU length of stay < 8 hours, acute coronary care and cardiac surgery patients, burn patients, re-admitted patients, patients with missing severity of illness scores or probabilities, and patients with missing (hospital) survival status. The characteristics of the

---

³ [www.stichting-nice.org](http://www.stichting-nice.org)
remaining dataset were compared to those used to develop the original APACHE II, SAPS II, MPM₀ II and MPM₂₄ II models.

**Validation measures**

**Discrimination**

The term discrimination refers to a model’s ability to distinguish survivors from non-survivors. As a measure of discrimination we calculated the area under the Receiver Operating Characteristic (ROC) Curve [14]. This Area Under the Curve (AUC) represents the probability that a patient who died has a higher predicted probability of dying than a patient who survived. An AUC of 0.5 indicates that the model does not predict better than chance. An AUC of 1 indicates that the model discriminates perfectly. The AUC of a model gives no indication of how close the individual predicted probabilities are to the observed outcome. To take this aspect of a models performance into account, we have to look at the accuracy of the predictions.

**Accuracy**

Accuracy refers to the difference between predictions and observed outcomes at the level of individuals. In this study we applied two different accuracy measures, the Mean Squared Error (MSE) and the Mean Log-Likelihood (MLL). The MSE is calculated as

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2,
\]

where \( y_i \) is the observed outcome and \( \hat{\pi}_i \) is the predicted probability of dying given by the model for patient \( i \), for each of \( n \) independent observations. Note that \( \hat{\pi}_i \) is not estimated from the current validation dataset, but from the dataset that was originally used to develop the prognostic model. The MSE is closely related to the Brier score [15].

The MLL is calculated as

\[
MLL = -\frac{1}{n} \sum_{i=1}^{n} y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)
\]

The MLL is sometimes referred to as cross-entropy, and is closely related to the model's deviance.

Both these accuracy measures assign a penalty to each individual prediction, based on the disagreement with the observed outcome, and compute the mean of the assigned penalties. If there is good agreement, the penalty is close to 0. If the agreement is very poor, the MSE-penalty is close to 1, while the penalty according to the MLL can become much greater. Thus the MLL gives higher
penalties in case of extremely poor agreement between predictions and observed outcome. Low MSEs and low MLLs both indicate high accuracy of the predictions made by the model.

Calibration
Calibration refers to the agreement between predicted probabilities and the 'true probabilities’. Of course the true probability of a patient’s outcome is not known, otherwise we would not need to develop prognostic models. However, we can approximate the true probabilities by taking the mean of the observed outcomes within predefined groups of patients. In this study, calibration was measured through Goodness-of-fit (GOF) statistics. GOF statistics provide numerical evidence for a model’s calibration and enables us to test the hypothesis whether the model fits on the data that was used. The null hypothesis says that the model is correct, i.e. the predicted probabilities are equal to the true probabilities. We applied the two goodness-of-fit statistics, H and Ĉ, proposed by D. Hosmer and S. Lemeshow [16]. The H and Ĉ statistics differ in the way the groups of patients are composed. Grouping for the H statistic is based on partitioning of the probability interval (0-1) into ten equally sized ranges. The Ĉ goodness-of-fit statistic sorts observations according to their expected probability and partitions the observations into ten groups of equal size. Given the grouping strategy, the calculations of the statistics C and H are the same:

\[
\hat{C} = \sum_{g=1}^{10} \frac{(O_g - n_g \bar{p}_g)^2}{n_g \bar{p}_g (1 - \bar{p}_g)},
\]

where \(g\) is the number of the groups, \(O_g\) is the sum of the observed outcomes in group \(g\), \(n_g\) is the number of ICU admissions in group \(g\), and \(\bar{p}_g\) is the mean of the predicted probabilities in group \(g\).

It has been shown that H and Ĉ follow a \(\chi^2\) distribution. Small values for H and Ĉ relate to high p-values, implying acceptance of the Null hypothesis which says that the model fits to the data. High values of H and Ĉ indicate poor fit of the model; It is known that the values of H and Ĉ increase with larger sample sizes when there is a lack of fit. In addition to testing goodness-of-fit, in this study we also use the statistics Ĉ and H for comparing the models and selecting the best calibrated model.

A disadvantage of the Hosmer-Lemeshow tests is that the value of the statistic is sensitive to the choice of the cut-off points that define the groups. Therefore various alternative goodness-of-fit tests for the logistic regression model have been proposed in the literature. In this study we have used the test proposed by Copas [17], which is based on the sum of squared residuals:
\[ R_c = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2 \]

The distribution of \( R_c \) is approximately normal with mean value
\[ \sum_{i=1}^{n} \hat{\pi}_i (1 - \hat{\pi}_i) \]

The observed value of \( R_c \) is transformed into a test statistic \( z \) that follows the standard normal distribution. A large deviation of \( z \) from 0 is related to a small p-value, indicating lack of fit of the model. As the Hosmer-Lemeshow goodness-of-fit statistics, the value of \( z \) increases with larger sample sizes when the model does not fit to the data.

Based on the measured discrimination and accuracy one can choose the best discriminating or the most accurate model out of two or more models. Measures for discrimination and for accuracy do not provide statistical evidence whether to accept or to reject a model for use in practice, whereas measures for calibration do.

**Sample sizes**

The entire validation dataset was used to draw validation datasets of smaller sizes. Reflecting the different sizes of validation samples used in other studies, we composed datasets containing data of 250, 500, 750, 1000, 2500 and 5000 ICU admissions. ICU admissions for these datasets were randomly selected from the entire dataset. To reduce the effects of chance we repeated this random selection process 500 times for each sample size. For each sample size we validated the four models on all these 500 datasets.

**Analyses**

We calculated all validation statistics mentioned above for each of the models, using the entire dataset with all eligible ICU admissions. We will refer to the results from the entire dataset as the ‘gold standard results’. Based on the gold standard results we determined for each validation measure which of the four models performed best. These models were considered the ‘gold standard best model’. To assess the effect of sample size we have applied each of the validation measures on the subsamples of different sizes (\( n=(100, 250, 500, 750, 1000, 2500, 5000) \)). For each sample size, for each validation measure the mean (or the median) of the results of the 500 runs and the 95% confidence interval of the mean were determined. For the three calibration tests we analyzed for each sample size how often the test statistic was significant. Because of the large number of runs (500) for each sample size we applied a significance level of \( \alpha=0.01 \). For each validation measure, we analyzed how often the ‘gold standard best model’ was indeed found as the best performing model in samples of varying size.
Results

Results based on the entire dataset

The validation dataset used in this study consisted of data from 83,824 admissions. According to the specific exclusion criteria of the SAPS II, the APACHE II and the MPM II model respectively 38,345, 38,360 and 38,849 patients needed to be excluded. No significant differences in outcome or severity of illness were observed between the different sets of eligible patients. We combined the criteria of the models to obtain one dataset that satisfied all criteria. The remaining dataset consisted of 42,139 ICU admissions. The observed ICU mortality within this dataset was 13.0%. The observed hospital mortality was 19.8%. Table 7.1 compares the dataset used in this study to those used to develop the original models [1-3].

Table 7.2 provides the results of all six validation measures, as applied on the original models. According to the goodness-of-fit tests the calibration of all four models was insufficient (p<0.01). In table 7.2 we highlighted for each validation measure the model that provided the best result. The models with the best results are here called the ‘gold standard best models’. The MSEs and MLLs of the original models showed no difference in performance between the SAPS II and the MPM_{24} II model. Therefore in these cases both models were considered to be ‘gold standard best models’.

Table 7.1 – Basic demographic data; NICE dataset compared to datasets used for development and evaluation of original APACHE II, SAPS II and MPM II models.

<table>
<thead>
<tr>
<th></th>
<th>NICE</th>
<th>SAPS II</th>
<th>APACHE II</th>
<th>MPM II</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of admissions</td>
<td>42139</td>
<td>12997</td>
<td>5030</td>
<td>19124</td>
</tr>
<tr>
<td>Hospital mortality (%)</td>
<td>19.8</td>
<td>21.8</td>
<td>19.7</td>
<td>20.8</td>
</tr>
<tr>
<td>Reason for admission to ICU</td>
<td>Medical (%)</td>
<td>43.2</td>
<td>48.4</td>
<td>46.0</td>
</tr>
<tr>
<td></td>
<td>Unscheduled surgery (%)</td>
<td>17.4</td>
<td>19.6</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td>Scheduled surgery (%)</td>
<td>39.4</td>
<td>31.2</td>
<td>27.0</td>
</tr>
<tr>
<td>Mean age survivors (years ± SD)</td>
<td>59.5 ± 16.9</td>
<td>57.2 ± 18.5</td>
<td>55.4</td>
<td></td>
</tr>
<tr>
<td>Mean age non-survivors (years ± SD)</td>
<td>66.86 ± 14.9</td>
<td>62.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male (%)</td>
<td>59.5</td>
<td>59.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean LOS^1 ICU (days ± SD)</td>
<td>4.7 ± 9.6</td>
<td>6.6 ± 9.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean LOS^1 hospital (excl pre-IC^2)</td>
<td>18.4 ± 26.1</td>
<td>19.1 ± 18.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAPS II (mean score ± SD)</td>
<td>33.3 ± 18.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(mean probability ± SD)</td>
<td>0.22 ± 0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APACHE II (mean score ± SD)</td>
<td>16.0 ± 8.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(mean probability ± SD)</td>
<td>0.24 ± 0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPM II_0 (mean probability ± SD)</td>
<td>0.20 ± 0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPM II_{24} (mean probability ± SD)</td>
<td>0.22 ± 0.22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^1LOS = Length of stay, ^2 excl pre-IC = hospital stay before ICU admission not included.

104
Table 7.2 – Results of validation of models based on the entire dataset (n=42,139)

<table>
<thead>
<tr>
<th>Validation measure</th>
<th>Model</th>
<th>Mean AUC (SD)</th>
<th>Mean MSE (SD)</th>
<th>Mean Log-likelihood (SD)</th>
<th>Hosmer-Lemeshow H</th>
<th>Hosmer-Lemeshow C</th>
<th>Copas z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination AUC (SD)</td>
<td>APACHE II</td>
<td>0.818 (0.0051)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAPS II</td>
<td>0.831 (0.0049)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM0 II</td>
<td>0.796 (0.0053)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM24 II</td>
<td>0.822 (0.005)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy Mean Squared Error (SD)</td>
<td>APACHE II</td>
<td>0.125 (0.00098)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAPS II</td>
<td>0.120 (0.00107)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM0 II</td>
<td>0.129 (0.00112)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM24 II</td>
<td>0.120 (0.00105)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Mean Log-Likelihood (SD)</td>
<td>APACHE II</td>
<td>0.396 (0.00268)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAPS II</td>
<td>0.384 (0.00308)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM0 II</td>
<td>0.406 (0.00316)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM24 II</td>
<td>0.384 (0.00312)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Hosmer-Lemeshow H</td>
<td>APACHE II</td>
<td>969.74 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAPS II</td>
<td>1079.02 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM0 II</td>
<td>440.15 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM24 II</td>
<td>214.53 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Hosmer-Lemeshow C</td>
<td>APACHE II</td>
<td>881.31 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAPS II</td>
<td>879.44 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM0 II</td>
<td>371.13 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM24 II</td>
<td>206.13 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration Copas z</td>
<td>APACHE II</td>
<td>-19.62 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAPS II</td>
<td>69.60 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM0 II</td>
<td>70.04 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPM24 II</td>
<td>4.17 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p < 0.01, □ = best performance

Results based on datasets of different sizes

For reasons of space restriction Figures 7.1 and 7.2 display only the results for the SAPS II model. The other three models showed similar results. Figure 7.1 displays the means and the 95% confidence intervals of the measured AUC, MSE and MLL. In all cases the AUC, the MSE and the MLL was unbiased (the average value agreed with the gold standard value) but displayed considerable variation, especially for small sample sizes. For example, in case of a sample size of 1000 observations 80% of the estimated MLLs were outside the 95% confidence interval of the gold standard value for the SAPS II model.

Figure 7.2 displays the medians and the 95% confidence intervals of the three calibration statistics. In contrast to the accuracy and discrimination statistics, calibration statistics were severely biased by sample size in all cases: the calibration statistics increased with larger sample sizes. With samples sizes of 2500 or larger all measured test statistics showed poor fit of the model (p<0.01).
Figure 7.1 – Mean ($\mu$) AUC, MSE and MLL with 95% confidence intervals, for the SAPS II model, using different sample sizes. The dotted lines display the gold standard AUC, MSE or MLL.

Figure 7.2 – Median ($\tilde{\mu}$) calibration statistics and 95% confidence intervals from the Hosmer-Lemeshow $H$ and $\hat{C}$ and the Copas tests, for the SAPS II model, using different sample sizes. The dotted line displays the value of the calibration statistic above which the model is rejected ($\alpha=0.01$). These values differ between the Copas test (2.57) and the Hosmer-Lemeshow tests (21.66).
Figure 7.3 shows per sample size the frequencies by which each of the four models showed superior discrimination over the others, as measured by the AUC. In this figure the model with the best discrimination according to the gold standard, SAPS II, is indicated by an asterisk (*). With a sample size of 100 or 250 observations, the superior discrimination by the SAPS II model is found in less than 50% of the cases. As the sample size increases, the SAPS II model is more often correctly identified as the best model. With a sample of size 5000 the SAPS II model was identified as the best model in nearly all of the 500 runs (97%). The frequencies by which the other models appear as the best performing model decrease with larger sample sizes. This pattern appeared to hold not only for the AUC, but also for the accuracy and the calibration measures.

![Figure 7.3 - Variation in apparent superior discrimination, as measured by the AUC. For each sample size, the frequency (y-axis) by which each of the models (APACHE II, SAPS II, MPM_0 II and MPM_24 II) appeared to be the best discriminating model is shown.](image)

**Discussion and conclusion**

The importance of externally validating prognostic models is commonly recognized in the statistical literature. In the field of intensive care several studies have validated one or more of the existing prognostic models. The sizes of the validation samples that were used in these studies vary widely. The
The general aim of this study is to gain insight into the effect of sample size on the measured performance of a prognostic model. Within this study we compare the results based on samples with different numbers of observations, ranging from one hundred to a few thousands, to a gold standard. The gold standard was based on a very large dataset from the NICE registry, containing tens of thousands of observations. Considering the large number of observations in this ‘gold standard dataset’ and the small confidence intervals resulting from it, we believe that using the results of the large dataset as the gold standard is justified.

According to our gold standard, the differences in performance between the APACHE II, SAPS II and MPM24 II models are small. Performance of the MPM0 II model was less than the other three models. This might be due to the fact that this model is based only on data available at ICU admission, whereas the other models are all based on data from the first 24 hours of ICU admission. In addition, MPM0 II data available in the NICE database only covered the first hour after ICU admission, whereas the data from which the MPM0 II model was derived also covered the last hour before ICU admission.

The APACHE II, SAPS II, MPM24 II and MPM0 II model were validated using samples of varying sizes. We found that the average AUC, MSE and MLL agreed with the gold standard values. However, with small sample sizes (n=100, 250, 500, 750, 1000) the measured discrimination and accuracy showed a large variation. Therefore, with sample sizes of 1000 or less observations one should question the validity of the measured discrimination and accuracy. In contrast to the discrimination and accuracy measurements the calibration test statistics (H, Č, Copas) showed large deviations from the gold standard calibration statistics with all sample sizes. This, combined with the observed large differences in calibration statistics between the different sample sizes, confirmed that the sample size has a major influence on the calibration as measured by the GOF statistics applied in this study. This finding is similar to that of Zhu et al. [18] who examined the impact of hospital mortality on the measured performance of the MPM0 II and the MPM24 II models, for various sample sizes. In contrast to Zhu et al. we here applied a broader range of performance measures on a larger set of generally accepted prognostic models. Furthermore, we analysed the consequences of the effect of sample size for conclusions that are based on the measured performance.

GOF statistics enable to test the hypothesis that a model’s predictions fit to the observations in a validation dataset. Based on this test one can decide to use a model or not, for example for calculation of observed mortality vs. predicted mortality ratios (Standardized Mortality Ratios). In the literature we found differences between studies with respect to the conclusions regarding the acceptation of models for use in clinical practice, and regarding the model that exhibited the best fit to validation data [7-11]. As mentioned before, these differences might be explained by geographical or temporal differences in the
Validation of prognostic models

validation samples or by differences in the sizes of the validation samples. The current study eliminated geographical and temporal confounders by drawing the samples from one large database. The assumption that the effects measured in this study are merely caused by the varying sizes of the validation samples is justifiable. A limitation of our approach is that when interpreting the results based on the larger sample sizes (e.g. n = 5000) one should keep in mind that the small variation in measured performance might partly be due to correlation between the 500 draws. This can only be solved by a further increase in the reference sample.

Many external validation studies in intensive care medicine restrict to measuring discrimination by means of the AUC and calibration by means of Hosmer-Lemeshow goodness-of-fit statistics. Our results confirm that goodness-of-fit statistics are highly sensitive to the size of the validation sample. With a large number of observations in the validation dataset, it is improbable that the associated calibration tests will accept the model. This structural defect is inherent to calibration tests: Each observed value that deviates from the expected value will increase the calibration test statistic. Thus, a large test statistic is not (just) a consequence of poor calibration of the models on the validation data. When interpreting the findings of the calibration tests one should always consider the fact that they reflect the size of the validation sample that was used.

Even though the differences in performance between the four models on the entire dataset were small, we selected one (or in some cases two) ‘gold standard best model’ for each validation measure. The results of this study show that with smaller validation datasets the variation in the selected best performing models increases. For example, in case of a sample size of 250 the gold standard best discriminating model was selected in less than 50% of the cases. This indicates that in case of small sample sizes the selection of the best performing model might actually be a matter of chance. In all cases the chance of selecting the correct model (gold standard best model) increased with larger sample sizes.

It is helpful to know what minimum sample size would be advisable for future external validation studies. Both Vergouwe et al. [19] and Steyerberg et al. [20] have provided advise considering the sample size required for external validation of prognostic models. Vergouwe et al. [19] studied at which effective sample sizes relevant differences in model performance can be detected with adequate power. They concluded that a validation sample should at least contain 100 events (e.g. deaths) and 100 non-events to obtain adequate power in validation studies. This would mean that if the frequency of events (e.g. hospital mortality) would be 20% one would need a validation dataset containing at least 500 observations. Since in their study Vergouwe et al. applied datasets with a maximum of 500 observations (225 events) it was not possible to analyze the effect of larger databases. Steyerberg et al. concluded that a reliable assessment
of external validity requires a sample which contains at least 20 events per variable. For example in case of the MPM$_0$ II model, which includes 15 variables, a validation set containing 300 events would be required. Thus in case of a hospital mortality rate of 20% the validation dataset should contain 1500 observations. Based on the results shown in figure 7.3 we conclude that in the domain of IC medicine at least 2500 observations are needed to have 85% certainty that the best performing model is used. In the literature concerning validation of IC prognostic models often this number is not reached. A common problem of prediction models is that predicted probabilities in external data sets show structural deviations from the true probabilities. These structural errors negatively influence the performance of prediction models. However, structural errors can be repaired by customizing the model to the new data. Several strategies exist to do so [18, 21]. We have planned to customize the APACHE II, SAPS II, MPM$_{24}$ II and MPM$_0$ II models to the NIC E population and to re-validate these customized versions of the models. Additional studies will evaluate the effect of the size of validation samples on the measured performance after customization.

From the results of this study we conclude that the calibration statistics applied here are highly influenced by the size of the validation sample: using small samples will result in apparent good fit, using large samples will result in apparent poor fit. In addition, when using a validation dataset with a small number of observations the measured performance and the selection of the best performing model will be largely a matter of chance.
References
