Terminological systems and prognostic models as instruments for quality assessment in intensive care
Raiez, F.

Citation for published version (APA):
Raiez, F. (2010). Terminological systems and prognostic models as instruments for quality assessment in intensive care

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
The Impact of Different Prognostic Models and their Customization on Institutional Comparison of Intensive Care Units

Ferishta Bakhshi-Raiez, Niels B. Peek, Robert J. Bosman,
Evert de Jonge, Nicolette F. de Keizer

Critical Care Medicine 2007; 35(11):2650-1
ABSTRACT

Objectives: To evaluate the influence of choice of a prognostic model and the effect of customization of these models on league tables (i.e. rank-order listing) in which Intensive Care units (ICUs) are ranked by SMR using APACHE II, SAPS II and MPM24II models.

Design: Retrospective analysis of prospectively collected data on ICU admissions.

Setting: A dataset from a national registry of 86,427 patients from 40 Dutch ICUs over the period of January 2002 to October 2006.

Measurements: The league tables associated with the different models were compared to evaluate their agreement. Bootstrapping was used to quantify the uncertainty in the ranks for ICUs. First, for each ICU the median rank and its 95% confidence interval (CI) were identified for each model. Then, for a given pair of models, for each ICU the median difference in rank and its associated 95%CI was computed. A difference in rank for an ICU for a given pair of models was considered relevant if it was statistically significant and one of the models would categorize this ICU as performance outlier (“excellent performer” or “very poor performer”) while the other did not.

Main results: For 20 ICUs, there was a significant difference in rank (2-19 positions) between one or more pairs of models. Three ICUs were rated as performance outlier by one of the models while the other excluded this possibility with 95% certainty. Furthermore, for ten ICUs, one or more pairs of models classified these ICUs as performance outlier while the other model did not do so with certainty. Regarding the agreement between the original models and their customized versions, in all cases the median change in rank was three positions or less and the models fully agreed with respect to which ICUs should be classified as performance outliers.

Conclusions: Institutional comparison based on case mix adjusted league tables is sensitive to the choice of prognostic model but not to customization of these models. League tables should always display the uncertainty associated with institutional ranks.
6.1 Introduction

Over the last two decades there has been an increasing interest in measuring the quality of health care and, in particular, in quantitative comparisons of performance between health care institutions (1;2). Such performance comparisons are usually based on a performance indicator, e.g. an outcome summary which is believed to reflect the quality of care. A prognostic model is often used to adjust observed raw outcomes for case mix differences between institutions. The aim of comparing case-mix adjusted outcomes across different institutions is to identify high- and low performance institutions and thereby to identify the processes of care that result in better or worse outcomes (3-5). In recent years there is also an increased trend towards publishing performance data and performance-based league tables (i.e. rank-order listings). Such publications can lead to disciplinary measures against hospital organizations and to changes in the behaviour of patients and providers. Therefore, the reliability of procedures for performance assessment and performance comparison is increasingly relevant.

The use of performance indicators as a basis for performance assessment and the use of prognostic models to obtain case-mix adjusted performance estimates has been the subject of discussion in many studies (1;2;4;6-10). Research has shown that different prognostic models may produce different impressions about care performance, especially when prognostic reliability is poor (6;11-14). It has been argued that poor prognostic reliability of the models can be avoided by customizing the models to the concerned population (15-18). Compared to others medical domains, the field of Intensive Care makes most frequently use of prognostic models. However the influence of model choice and its customization on performance assessment of Intensive Care Units (ICUs) is not yet fully explored.

The aim of our study was twofold: firstly, to evaluate the influence of the choice of a prognostic model on league tables in which (Dutch) ICUs are ranked by Standardized Mortality Ratios (SMR), and secondly, to analyze the effect of customization of these models on the resulting league tables. The Acute Physiology and Chronic Health Evaluation II (APACHE II) model, Simplified Acute Physiology Score II (SAPS II) model and Mortality Probability Model II (MPM2II) (19-21), were used for case-mix adjustment. From a statistical prospective, ranks in league tables have a sampling error that is partly related to the number of observations on which they are based. However, in daily practice of institutional comparisons, this sampling error is not accounted for. Therefore, the resulting ranks may then be governed by chance and not by actual performance. Recently developed statistical techniques make it possible to compute confidence intervals for the ranks given to institutions (8;22). By doing so, it becomes possible to reliably judge the influence of the used prognostic model (and its customization) on league tables, and therefore these techniques were applied here.
6.2 Materials and Methods

6.2.1 Patients

In 1996 the Dutch National Intensive Care Evaluation (NICE) foundation started collecting data on patients admitted to Dutch ICUs (23). The data has been encrypted in a way that all patient identifying information, such as name and patient identification number, has been removed. In the Netherlands, there is no need to obtain consent to make use of registries without patient identifying information. The NICE initiative is officially registered according to the Dutch Personal Data Protection Act. At the moment of the study 47 ICUs were voluntarily participating in NICE which amounts to about 50% of all ICUs in the Netherlands. The participating ICUs are a mixture of tertiary, teaching and non-teaching settings in urban and non-urban hospitals. The registry contains all variables necessary to calculate among others the APACHE II, SAPS II and MPM24II prognostic models (19-21). The choice of these models for this study was based on the fact that they correspond closely to each other since all three models concern the same time span and because the models show a comparable fit within the Dutch ICU patients. Audit mechanisms were in place to ensure the validity of the data (24).

This study used a dataset from the NICE registry with data on patients admitted to the Dutch ICUs from January 1st 2002 to October 1st 2006. ICUs were eligible in case they had a minimum participation period of 12 months and at least 100 admissions. The general exclusion criteria defined by the original models were applied to the data (19-21): all readmissions within the same hospital stay, admissions after cardiac surgery, burns and admissions with missing admission types were excluded from the analysis. Furthermore, additional exclusion criteria were applied according to the specific models (19-21).

6.2.2 Customization of the models

The models were customized at two levels as described elsewhere (17;25;26). The first level customization involved fitting a new logistic regression equation with observed in-hospital death as dependent variable and the logit-transformed original prediction as the independent variable. Thus, first level customization did not change the influence of individual covariates included in the model, but calibrated their joint influence to the observed mortality in the external dataset. In contrast, second level customization did involve fitting a new logistic regression equation with in-hospital mortality as dependent variable and with all original covariates as independent variables. An extensive description of these customization procedures is provided in appendix 6.1.
6.2.3 Performance assessment

Measures of discrimination and calibration were used to assess the performance of the original and the customized models. Discrimination of the models was assessed by calculating the Area Under the Receiver Operating Characteristic curve (AUC) (27). Calibration was assessed by calculating the Hosmer-Lemeshow goodness-of-fit statistic $\hat{C}$ (28). The value of the $\hat{C}$ statistic was $\chi^2$ (chi-square) tested with a significance level of 0.01. Ten-fold cross validation was used to obtain performance estimates of the customized models (29;30).

6.2.4 ICU ranking

To evaluate the influence of the choice of the prognostic model on institutional comparison, SMR values were calculated for each ICU based on each of the three original models. Similarly, to evaluate the effect of model customization on this procedure, new SMR values were calculated for each ICU using the customized models. Subsequently, for each prognostic model, the ICUs were ranked by increasing SMR value. The uncertainty associated with the ranks was calculated using the simulation procedure developed by Marshall and Spiegelhalter (8). In this procedure, the steps of model customization, calculating SMR values and ranking ICUs are repeated for 10,000 times on bootstrap samples of the original dataset. Based on the simulated ranks, the median rank and its 95% empirical confidence interval (CI) are subsequently identified for each ICU. The procedure was carried out for the three original models and for their (six) customized versions.

6.2.5 Rank comparisons

The different league tables of median rank based on the three original models and their customized versions thus obtained, were subsequently compared to determine the influence of model choice and customization on the rank assigned to individual ICUs. Given a pair of league tables, two methods were used to quantify their differences. First, for each ICU the median change in rank and its associated 95% CI was computed based on the 10,000 bootstrap samples from the simulation procedure. By doing so, it was possible to determine whether there were significant changes in ICU rank when one model would be used instead of the other.

Although changes in rank between two league tables indicate that the associated models rate performance differently, this is not dramatic when the differences are small. For the purpose of performance assessment, it is not the precise estimation of the rank that is of practical relevance, rather than the fact whether an institution is categorized as a performance outlier, i.e. as either an excellent or a very poor performer. The second method therefore determined for each ICU, whether one or more models would unambiguously
categorize it as “likely excellent performer”, meaning that the model would place that ICU among the top 25% performers with 95% certainty or as “likely very poor performer”, meaning that the model would place the ICU among the worst 25% performers with 95% certainty.

In case there was a significant difference in rank for a certain ICU between two models and one model would rate this as likely excellent (or very poor) performer while the other model excluded the possibility with 95% certainty, the models were said to “strongly disagree”. In case there was a significant difference in median rank and one model rate this ICU as likely excellent (or very poor) performer while the other did not do so without excluding the possibility (only a part of the 95% CI lays in the top/bottom 25%), the models were said to “weakly disagree”.

S-Plus 6.2 for Windows (Insightful Corporation, Seattle) was used to perform the statistical analyses.

### 6.3 Results

From January 1st 2002 to October 1st 2006, 86,427 patients were admitted to 40 Dutch ICUs participating in the NICE registry who met the general inclusion criteria. The median number of participation months for these ICUs was 52 (Inter Quartile Range: 41-57). All these ICUs are mixed medical surgical units.

As shown in Table 6.1, for the APACHE II model, the SAPS II model and MPM24II model respectively 77149, 80721 and 45941 patients were included for analyses.

The demographics of the populations included for analysis are shown in Table 6.2.

#### 6.3.1 Customization of the models

All models showed overprediction, meaning that the predicted mortality risks were structurally too high, and underdispersion, meaning that the predictions were not extreme enough (i.e. too close to the mean).

Table 6.3 describes the results of the evaluation of the performance of the three prognostic models before and after customization. The Hosmer-Lemeshow Č-statistics revealed imperfect calibration of the original models. However, all original models showed a good discrimination (AUC ≥ 0.78). Customization had a small effect on discriminative power of the models, but it did improve the calibration of all models.
6.3.2 ICU ranking

Comparison of the original models

The SMR values for individual ICUs ranged from 0.39 (CI: 0.32-0.48) to 1.07 (CI: 0.91-1.26) using APACHE II model, from 0.71 (CI: 0.50-0.99) to 1.72 (CI: 1.61-1.84) using SAPS II model and 0.60 (CI: 0.38-0.91) to 1.56 (CI: 1.17-2.05) using MPM24II model. Figures 6.1a to 6.1c show for each pair of prognostic models (three pairs in total) the changes in median rank based on these SMR values and their associated confidence intervals. When SAPS II was used instead of APACHE II for case-mix adjustment, a significant change in rank occurred for twenty-one ICUs (52.5%). The median change in rank among these ICUs ranged from four to eighteen positions. As illustrated in Figure 6.1a, ICU number 13 for example had a median change in rank of 18 positions (95% CI 13 - 22).

Table 6.1 Number of admissions satisfying each of the exclusion criteria for the published models (multiple exclusion criteria possible). Total number of included admissions is “86427”.

<table>
<thead>
<tr>
<th></th>
<th>APACHE II</th>
<th>SAPS II</th>
<th>MPM24II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &lt; 16 yrs</td>
<td>836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of stay &lt; 8 hrs</td>
<td>7953</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing diagnostic category</td>
<td>3796</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing APACHE II score</td>
<td>8122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age &lt;18 yrs</td>
<td>1736</td>
<td>1536</td>
<td></td>
</tr>
<tr>
<td>Transfer to an ICU in another hospital</td>
<td>3408</td>
<td>1871</td>
<td></td>
</tr>
<tr>
<td>Missing SAPS II score</td>
<td>1596</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing MPM24II probability</td>
<td></td>
<td>175</td>
<td></td>
</tr>
<tr>
<td>Length of stay &lt; 24 hours</td>
<td></td>
<td>38803</td>
<td></td>
</tr>
<tr>
<td>Total included</td>
<td>78269</td>
<td>80146</td>
<td>45941</td>
</tr>
</tbody>
</table>

When APACHE II was used instead of MPM24II, a significant change in rank occurred for nineteen ICUs (47.5%). The median change in rank among these ICUs ranged from two to nineteen positions. When SAPS II was used instead of MPM24II, for seventeen ICUs (42.5%) the rank number significantly changed with a median change in rank of six to sixteen positions. However, as mentioned before, it is not the precise estimation of the rank that is relevant in institutional performance comparison, but more the question whether an institution is unambiguously categorized as a performance outlier.
Table 6.2 Basic Demographics in the used samples.

<table>
<thead>
<tr>
<th></th>
<th>APACHE II</th>
<th>SAPS II</th>
<th>MPM24II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital mortality (%)</td>
<td>19.4</td>
<td>21.4</td>
<td>25.4</td>
</tr>
<tr>
<td>ICU Length of stay (Mean days ±SD)</td>
<td>4.6 (9.0)</td>
<td>4.3 (8.7)</td>
<td>6.8 (10.7)</td>
</tr>
<tr>
<td>Male (%)</td>
<td>58.4</td>
<td>58.1</td>
<td>58.8</td>
</tr>
<tr>
<td>Admission type (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>46.3</td>
<td>45.6</td>
<td>53.2</td>
</tr>
<tr>
<td>Urgent surgery</td>
<td>17.5</td>
<td>17.4</td>
<td>20.1</td>
</tr>
<tr>
<td>Elective surgery</td>
<td>36.2</td>
<td>36.0</td>
<td>26.7</td>
</tr>
<tr>
<td>Age ( mean ±SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survivors</td>
<td>61.3 (16.9)</td>
<td>61.7 (16.7)</td>
<td>62.4 (16.1)</td>
</tr>
<tr>
<td>Non-survivors</td>
<td>59.8 (17.1)</td>
<td>59.59 (16.8)</td>
<td>60.5 (16.4)</td>
</tr>
<tr>
<td>APACHE II Score (mean ±SD)</td>
<td>21.61 (12.41)</td>
<td>21.70 (12.42)</td>
<td>21.50 (11.6)</td>
</tr>
<tr>
<td>Probability (mean ±SD)</td>
<td>0.38 (0.32)</td>
<td>0.39 (0.32)</td>
<td>0.38 (0.30)</td>
</tr>
<tr>
<td>SAPS II Score (mean ±SD)</td>
<td>33.44 (18.9)</td>
<td>33.25 (19.7)</td>
<td>38.67 (17.9)</td>
</tr>
<tr>
<td>Probability (mean ±SD)</td>
<td>0.23 (0.26)</td>
<td>0.23 (0.27)</td>
<td>0.29 (0.27)</td>
</tr>
<tr>
<td>MPM24II Probability (mean ±SD)</td>
<td>0.21 (0.22)</td>
<td>0.22 (0.22)</td>
<td>0.28 (0.24)</td>
</tr>
</tbody>
</table>

Table 6.3 Discrimination and calibration of the models before and after customization.

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Ĉ-Statistic</th>
<th>SMR (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>APACHE II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Model</td>
<td>0.821</td>
<td>2627.81</td>
<td>0.76 (0.71-0.77)</td>
</tr>
<tr>
<td>1st Level Customization</td>
<td>0.821</td>
<td>77.52</td>
<td>1.00 (0.98-1.03)</td>
</tr>
<tr>
<td>2nd Level Customization</td>
<td>0.824</td>
<td>148.84</td>
<td>1.00 (0.98-1.03)</td>
</tr>
<tr>
<td>SAPS II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Model</td>
<td>0.85</td>
<td>920.71</td>
<td>0.94 (0.90-0.95)</td>
</tr>
<tr>
<td>1st Level Customization</td>
<td>0.85</td>
<td>60.22</td>
<td>0.99 (0.97-1.01)</td>
</tr>
<tr>
<td>2nd Level Customization</td>
<td>0.85</td>
<td>55.33</td>
<td>0.99 (0.97-1.01)</td>
</tr>
<tr>
<td>MPM24II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Model</td>
<td>0.78</td>
<td>502.58</td>
<td>0.97 (0.94-1.00)</td>
</tr>
<tr>
<td>1st Level Customization</td>
<td>0.78</td>
<td>98.33</td>
<td>0.99 (0.98-1.03)</td>
</tr>
<tr>
<td>2nd Level Customization</td>
<td>0.78</td>
<td>139.61</td>
<td>1.00 (0.99-1.04)</td>
</tr>
</tbody>
</table>

AUC: Area under receiver operating characteristics curve; Ĉ-statistics: Hosmer and Lemeshow goodness-of-fit statistics Ĉ and the related p-value; SMR: Standardized Mortality Ratio and CI: Confidence Interval.

Figures 2a to 2c display, for each of the three prognostic models, the median SMR-based rank and its 95% CI per ICU (the higher the rank, the better is the ICU performance). Figures 2a to 2c also visualize the categories of performance. An ICU is classified as a
performance outlier (i.e. as an excellent performer or as a very poor performer) when its entire CI is contained in the corresponding region (respectively the upper quarter and the lower quarter of the league table).

In total, twenty-one ICUs are marked as performance outlier by one or more models: APACHE II marked eleven ICUs (numbers: 1, 5, 18, 21, 22, 24, 25, 26, 30, 34, and 36) as performance outlier, SAPS II marked seven ICUs (numbers: 10, 13, 14, 18, 23, 34, and 35) as performance outlier and MPM24II marked twelve ICUs (numbers: 1, 10, 13, 16, 18, 20, 24, 26, 31, 33, 35, and 40) as performance outlier. The three models agreed on only one of these ICUs (number 18) as being a performance outlier with 95% certainty.

For the remaining ICUs which were marked as performance outlier by one or more models and had a significant difference in rank between some pair of models, there were considerable disagreements.

APACHE II and SAPS II models strongly disagreed for two ICUs (numbers 13 and 30). For instance ICU number 13 was rated as likely very poor performer by the SAPS II model while the APACHE II excluded a bottom 25% ranking with 95% certainty. Furthermore, the two models showed weak disagreement for eight of these ICUs (numbers 1, 10, 14, 22, 23, 24, 26 and 35). APACHE II and MPM24II models strongly disagreed for ICU number 13 which was rated as likely very poor performer by the MPM24II model while the APACHE II excluded a bottom 25% ranking with 95% certainty. Furthermore, weak disagreement was observed for eight ICUs (numbers 10, 21, 22, 30, 31, 34, 35, and 40).

SAPS II and MPM24II models strongly disagreed for ICU numbers 40 which was rated as likely excellent performer by the MPM24II model while the APACHE II excluded a top 25% ranking with 95% certainty. Additionally, weak disagreement was observed for two ICUs (numbers 24 and 34).

Comparison of the original models and their customized versions
Regarding the agreement between the original models and their customized versions, the detailed results are not shown because of space restriction. No significant changes in rank occurred when a customized version of APACHE II or SAPS II was used instead of the original model. Consequently, the original models and their customized versions fully agreed with respect to which ICUs should be classified as performance outliers. For the MPM24II model, there were some small differences in the league tables obtained with the original and the customized versions. In all cases the median change in rank was three positions or less. The original model and its customized versions again fully agreed with respect to which ICUs should be classified as performance outliers.
Figure 6.1a Estimated change in rank (and 95% Confidence Intervals) for each ICUs between APACHE II model and SAPS II model.

Figure 6.1b Estimated change in rank (and 95% Confidence Intervals) for each ICUs between APACHE II model and MPM24II model.

Figure 6.1c Estimated change in rank (and 95% Confidence Intervals) for each ICUs between SAPS II model and MPM24II model.
The Impact of Different Prognostic Models on Institutional Comparison

Figure 6.2a Median rank (and 95% Confidence Intervals) for each ICU based on APACHE II original model. Vertical lines divide ICUs into quarters according to rank. Numbers of admissions are given on the right side of the figure.

Figure 6.2b Median rank (and 95% Confidence Intervals) for each ICU based on SAPS II original model. Vertical lines divide ICUs into quarters according to rank. Numbers of admissions are given on the right side of the figure.

Figure 6.2c Median rank (and 95% Confidence Intervals) for each ICU based on M24II original model. Vertical lines divide ICUs into quarters according to rank. Numbers of admissions are given on the right side of the figure.
6.4 Discussion

There is a broad consensus on the need for risk adjustment when comparing outcomes across different institutions by league tables. However, as Iezonni stated, “there is a “risk” to risk adjustment” (6) because different risk adjustment models will not always agree on the identity of outlier performing institutions and on ranking the institutions. The absence of a gold standard makes it impossible to measure the validity of the different prognostic models used within the same field of care. Nevertheless, it is possible to determine how strongly different prognostic models agree on performance assessment.

In agreement with earlier observations in other domains of medicine (8;22), we found that ranking ICUs based on SMRs may be largely influenced by sampling errors. Reporting ranks thus may provide misleading information. This has important implications as payers and consumers of health care are increasingly seeking information on quality of care of institutions from rankings based on outcomes. Consequently, the level of uncertainty of a ranking should be quantified by providing the 95% confidence interval.

One disadvantage of rank ordering institutions by performance is that the resulting league table conceals absolute differences in performance. When these differences are small, ranks will be merely the product of chance, irrespective of the underlying sample size. In this study, some of the rank confidence intervals were very wide, indicating that substantial uncertainty was associated with the ranks. As was also shown in (8), this occurs more frequently with ICUs that are placed in the middle of the league tables, because a large group of such ICUs perform similarly without significant differences in SMR. Consequently the ranks of these ICUs do not differ significantly and their confidence intervals are wide, even though they are calculated from large numbers of ICU admissions.

We also show in agreement with previous studies (11-14), that judgement about relative performance of institutions based on case mix adjusted league tables is sensitive to the choice of prognostic model. There were substantial differences in rank given to ICUs by the three models resulting in disagreements in marking these ICUs as performance outliers.

An SMR can be applied in two manners when assessing performance. One method is to assess the absolute performance of an ICU by comparing its SMR value to a reference value. Glance et al. used this method in order to evaluate the effect of model choice (APACHE II, SAPS II and MPM0II) on detecting performance outliers. Performance outliers were defined as ICUs that did not incorporate the value 1 in the 95% CI of their estimated SMR (31). They concluded that the evaluation of ICU performance with this method is not influenced by the choice of prognostic model. However, since the original models were developed up to twenty years ago and based on data from patients in other parts of the world, comparing SMRs to the reference value of 1 may not be accurate. In our case, for example, all three original models overpredict mortality in Dutch ICU patients.
resulting in SMRs of less than one. Within this study, we therefore used the other performance assessment method, which focuses on relative performance of ICUs and does not depend on a reference value based on outdated models.

A recent study showed that the accuracy of prognostic models changes over time and should be periodically retested, customized and updated (18). Customization as used in our study will repair the model’s general tendency to overpredict (or underpredict) death leading to systematic changes in SMRs. Subsequently, it will have an important influence on the performance assessment method used by Glance et al. in which the SMR is compared to a reference value (10;31). We showed that customization had a large impact on the performance of the models, improving in particular its calibration. However, it did not influence the assessment of relative quality performance of the ICUs based on SMR.

The agreement between league tables based on the SAPS II and MPM24II was larger than the agreement between those based on APACHE II and SAPS II and between APACHE II and MPM24II. This might be attributed to the fact that the SAPS II model and MPM24II model were developed on the same (more recent) database and use more similar covariates. Moreover, APACHE II uses additional diagnostic information and information on co-morbidities which leave room for inter observer differences in interpretation (10;24;31).

There are some other potential explanations for the differences in rank order using different prognostic models. The three prognostic models used within this study emphasize different aspects of the case mix resulting in divergent predictions for individual patients. There are also case mix differences which are not corrected for by any of the three prognostic models used. Lead time bias for example, referring to the treatments given just before ICU admission, may influence predicted mortality and SMRs (32). Systematic differences in data definitions and data collection by individual ICUs, e.g. the presence of a Patient Data Management System, may lead to overprediction or underprediction of mortality and thereby to lower or higher SMRs (33). Also, if an ICU treats many patients with a specific diagnosis, for example metastasized carcinoma, its SMR will probably be higher for all prognostic models used in this study (34;35). Thus these differences will influence the absolute performance of ICUs by each model. Although all models used within this study suffer from these limitations, the influence may differ between models, contributing to the differences in rank order using individual models.

Furthermore, we can not completely exclude that because of voluntarily participation, the ICUs included in our study form a subset of the “best” performing ICUs in the Netherlands and are not representative for the general Dutch ICU population, resulting in volunteers bias. Nevertheless, we believe that volunteers bias might be important in case we are interested in the assessment of absolute performance of individual ICUs by the three models. It is unlikely that volunteers bias will affect the evaluation of the choice of the
prognostic model on the assessment of relative performance of ICUs. Therefore, we believe
that our results can be generalized to other ICU sub-populations.
We believe that the importance of disagreements between the models depends on the
purpose for which the performance data are used. If rank data are used by ICUs internally
as an indication of their performance compared to other ICUs, the (weak) disagreements
between these models may be of limited importance. However, external use of performance
data to judge quality of care or to steer national health care decisions make these
disagreements between the models very important. Therefore, any application of league
tables must be of great caution and performance information in these league tables should
not be used to draw definite conclusions.

Conclusions
We conclude that assessments of institutional performance in case-mix adjusted league
tables are sensitive to the model that is used for case-mix adjustment. Furthermore, the
level of uncertainty of ranking should be indicated by providing the 95% confidence
intervals. Therefore, performance data in league tables should only be used as a rough
indication of the quality of care in an individual ICU. Additional information on the
practices of care are needed before firm conclusions can be drawn. Customization of
prognostic models to the population concerned does not affect league tables.

Bibliography
1. DesHarnais SI, Forthman MT, Homa-Lowry JM, Wooster LD: Risk-adjusted quality outcome
measures:  indexes for benchmarking rates of mortality, complications, and readmissions. Qual
2. Goldstein H, Spiegelhalter DJ: League tables and Their Limitations: Statistical Issue in
385-443.
8.
critical care units: a preliminary assessment of their policies and practices. J Crit Care
2003;18(2):76-86.
5. Lilford R, Mohammed MA, Spiegelhalter D, Thomson R: Use and misuse of process and
8.
8. Marshall EC, Spiegelhalter DJ: Reliability of league tables of in vitro fertilisation clinics: