Mortality prediction in the intensive care: Role of mathematical models in benchmarking and decision-making
Minne, L.

Citation for published version (APA):
Minne, L. (2013). Mortality prediction in the intensive care: Role of mathematical models in benchmarking and decision-making

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.
Chapter 8

Characteristics of daily survival predictions of nurses and physicians in the Intensive Care and their predictive value

Minne L, Dongelmans DA, de Jonge E, Abu-Hanna A.
Characteristics of daily survival predictions of nurses and physicians in the Intensive Care and their predictive value.
Under submission.
Abstract

Purpose: To evaluate 1) characteristics of daily predictions of hospital survival of nurses and physicians in the ICU, and 2) their daily predictive value in terms of discrimination and precision.

Methods: We prospectively collected and validated nurses’ and physicians daily predictions of hospital survival of all consecutive patients admitted between January 23rd 2008 and June 14th 2010 to a 28-bed multidisciplinary mixed adult ICU of a university hospital in Amsterdam.

Results: The data included 7955 predictions for 2221 patients. Mortality was 17.9%. Inter-observer agreement by Cohen’s weighted kappa between nurses and physicians was fair to moderate (ranged 20.4-56.1%). Mean inter-observer variance of different nurses was 2.22 categories. Variance over time was higher for physicians (0.69-0.81 categories) than for nurses (0.61-0.68 categories). On the day of admission, both nurses and physicians performed statistically significantly better (p<0.05) than SAPS-II in terms of discriminative ability (AUC = 0.889 for nurses, 0.882 for physicians and 0.801 for SAPS-II) and precision (Brier = 0.085 for nurses, 0.082 for physicians and 0.112 for SAPS-II). There were no significant differences in performance between nurses and physicians, but their performance gradually decreased over time, although this increased in patients with a minimum stay of 8 days.

Conclusion: Although nurses and physicians have good predictive ability which is superior to the SAPS-II model, they do not always agree and are not always consistent over time. Future work should assess the possible role of nurses and objective prognostic models in the end-of-life decision-making process.

Introduction

In the Intensive Care Unit (ICU), clinicians are continuously faced with the difficult task of prognosis, which is the part of medical practice physicians feel most insecure about [1,2]. Prognosis is the prediction of (the probability of) an event, such as death, before its possible occurrence [3,4]. Prediction of survival status is important because it influences clinical decision-making including provision of life support, as well as patient preferences to undergo or continue intensive life sustaining treatments [5,6]. In fact, patients often prefer palliative care aiming at comfort and relief of pain if their chances of survival are very low [7,8]. The relevant prognostic question to ICU clinicians is to predict hospital survival chances of their patients, which is the probability that a particular patient will eventually be discharged alive from the hospital. This means that a patient that was discharged alive from the ICU to a hospital ward but died in the ward will be considered as a non-survivor. Survival until ICU discharge is hence included within hospital survival.

On the first day of ICU admission, nurses and physicians’ predictions of patient survival seem to have a good discrimination (between survivors and non-survivors) and moderate calibration (between predictions and observed outcomes) [9-22]. There is however substantial disagreement between different physicians [15]. Their predictions are influenced by several factors, such as ego bias (i.e. more trust in self than in others), reverse ego bias (i.e. less trust in self than in others) and level of experience [13,15-18,21-24], and they are hypothesized to be inconsistent [15,19]. The term consistency in this context refers to either 1) consistency between assessments made by different (types of) observers of the same phenomenon (inter-observer reliability) or 2) consistency of assessments...
made by the same (type of) observer at different times of the same phenomenon (test-retest or intra-observer reliability) [11]. Knowledge in the literature about clinicians’ (nurses and physicians) predictions on hospital survival is however mostly confined to the first day of ICU admission. The consistency of their daily predictions of survival and their daily predictive performance in terms of discrimination and precision (i.e. calibration) have not been investigated yet. Such investigation increases our understanding of clinicians’ prognostic behaviour over time and sheds light on opportunities for improving the decision making process. Therefore, the aim of this paper is to evaluate 1) characteristics of daily predictions of hospital survival of nurses and physicians in the ICU, and 2) their daily predictive value in terms of discrimination and precision.

Materials and Methods

Data collection

In this prospective study, our cohort consisted of all consecutive patients admitted between January 23rd 2008 and June 14th 2010 to a 28-bed multidisciplinary mixed adult ICU of a university hospital in Amsterdam. We included only patients for whom both nurse and physician predictions were available from their day of admission. We extracted demographics, patient outcomes and data necessary to calculate severity of illness scores (e.g. SAPS-II; Simplified Acute Physiology Score-II) from the Dutch National Intensive Care Evaluation (NICE) registry [25]. SAPS-II is an integer quantifying the severity of illness of the patient based on data collected in the first 24 hours of admission to the ICU.

Intervention

We developed a software module that elicits clinicians’ estimates of the likelihood of patient survival until hospital discharge. This module was integrated into the Patient Data Management System (PDMS) that is used in the ICU. Triggering of this module was seamlessly intertwined in the clinician’s workflow: after entry of the daily routine clinical patient data, a self-explaining question popped up automatically in a window on the computer screen for which nurses and physicians could choose between 10 probability categories: 1) 0-10%; 2) 10-20%; 3) 20-30%; 4) 30-40%; 5) 40-50%; 6) 50-60%; 7) 60-70%; 8) 70-80%; 9) 80-90%; and 10) 90-100%. Or they could state that they had no idea regarding the survival chance of a particular patient. Categories were indicated with red (low survival probabilities) and green (high survival probabilities) color-scales. Eliciting these survival probability categories took typically place between 9 and 12 am, but estimates could be changed until midnight (12 pm). We estimated that this set up would result in one physician and three nurses’ predictions of a patient’s hospital survival every 24 hours. Our institutional medical ethical committee waived the need for informed consent from individual patients, as no patients are subjected to additional actions.

Clinicians

Our ICU employs 120 nurses, 10 intensive care attendings, 8 fellows and approximately 14 residents whom all participated in this study. Critical care attendings are medical specialists (e.g. anaesthesiologists, internists or cardiologists) who have completed an additional intensive care specialization of 2 years. Fellows and residents have completed at least 2 years of post-MD training. Fellows are in training for an intensive care specialization and residents are in training for anaesthesiologist or another specialization. Clinicians
were unaware of their colleagues' assessments. They were not trained to estimate survival probabilities and did not receive feedback on their estimates. They were blinded for severity of illness scores, but not for scores on the degree of organ failure. Clinicians were notified of this study by email and by a single announcement (by EdJ) during their staff meeting.

**Statistical analysis**

The statistical analysis was conducted, for all patients and in a subset of patients with a minimum stay of eight days, in two steps as described below. Missing predictions of physicians and nurses were imputed by taking the mean value of the adjacent predictions of the previous and next day. When these were not available, the complete patient record was excluded. A second reason for patient exclusion was not having at least one prediction from a physician and one from a nurse for the same patient. We analysed the data using the statistical environment R version 2.10.1 [34]. For part of the analyses, we conducted subgroup analyses on admission type (planned vs. unplanned), survival status and minimum ICU stay of 8 days. The latter is important to get insight into changes over time for the same cohort of patients over time (because in general the length of stay varies among all patients).

1) **Characteristics of daily survival predictions**

The first step in our analysis was to assess general characteristics of daily survival predictions by nurses and physicians, irrespective of the true patient outcome. We started by visualizing frequency graphs of the two prediction distributions, of nurses and physicians, among survivors and non-survivors. Next, we assessed consistency of physicians’ and nurses’ predictions by using the following measures (explained in Table 8.1 and Appendix 8.1). First, we measured inter-observer reliability between nurses and physicians, and among nurses. Inter-observer agreement between nurses and physicians was obtained by Cohen’s weighted kappa coefficient [26] for each of the first eight days of prediction. In addition, we measured inter-observer variance among nurses. Inter-observer variance among physicians cannot be obtained, as only one physician prediction was available on each day. Note that higher inter-observer variance means higher agreement and vice versa. Second, we assessed intra-observer reliability by measuring 1) intra-observer variance over time for both nurses and physicians separately, and 2) mean pairwise correlation and mean pairwise absolute difference between the predictions of the first eight prediction days. Note that variance over time is not necessarily bad or illogical as the patient’s condition may change. Finally, the overall standard deviation of the predictions ($\sigma_{\pi}$) was used as a measure of model sharpness, where larger values indicate a greater level of sharpness than small values [27,28]. For example, the standard deviation would be large if predictions fell in either category 1 or 10, while it would be minimal if all predictions fell in the same category.

2) **Predictive value of daily survival predictions**

The above-mentioned measures are properties of the predictions only. The second step of our analysis was to assess the prediction-observation pairs, i.e. assessing the predictive value of the daily predictions by comparing them to the actual patient outcome (e.g. death or alive at hospital discharge). We measured performance by the Area Under the receiver
### Table 8.1. Measurements of consistency

<table>
<thead>
<tr>
<th>Measure</th>
<th>Groups and unit of time</th>
<th>Calculation</th>
<th>Interpretation</th>
</tr>
</thead>
</table>
| **Inter-observer agreement** | Nurse versus physician on a given day \(d\) | Cohen’s Weighted Kappa for day \(d\):  \(1 - \frac{\sum(O_i - E_i)}{\sum(E_i)}\)  
\(O_i\): observed agreement  
\(E_i\): expected agreement  
\(w_{ij}\): weight for each combination \((i,j)\)  
\(P_n\): prediction by nurse \(i\)  
\(P_p\): prediction by physician \(j\)  |  
0% : no more agreement than chance,  
<20%: poor agreement,  
20 - 40%: fair agreement,  
40 - 60%: moderate agreement,  
60 - 80%: good agreement,  
>80%: very good agreement [35],  
100%: complete agreement  |
| **Inter-observer variance**  | Nurse versus other nurses over all days | \(V_i = \frac{1}{N_p} \sum(V_{id})\)  
\(V_{id}\): Variance of nurse predictions for each patient \(i\) on day \(d\)  
\(V_i\): Mean of \(V_{id}\) for each patient \(i\)  
\(IOV = \frac{1}{N_p} \sum(V_i)\)  | IOV = 0 : predictions are identical  
High value indicates large spread around the mean  |
| **Mean pairwise correlation** | Nurse or physician over all days | \(M_{id} = \frac{1}{N_p} \sum(M_{i})\)  
\(M_{id}\): Mean nurse or physician prediction for each patient \(i\) on day \(d\)  
\(M_{i}\): Mean of \(M_{id}\) for all patients  
\(C_{dd}\): Pairwise correlations of \(M_{id}\) between any two days  
\(MPC = \frac{1}{N_p} \sum(C_{dd})\)  | MPC = 0 : no correlation between different prediction days  
1 : perfect positive correlation  
-1 : perfect negative correlation  |
| **Mean absolute difference** | Nurse or physician over all days | \(M_{id} = \frac{1}{N_p} \sum(D_{id})\)  
\(M_{id}\): Mean nurse or physician prediction for each patient \(i\) on day \(d\)  
\(D_{id}\): Pairwise absolute differences in \(M_{id}\) between any two days per patient  
\(MAD = \frac{1}{N_p} \sum(D_{id})\)  | MAD = 0 : no differences between different prediction days  
High values indicate large difference (max. difference is 9 categories)  |

\(N_c\) refers to the number of probability categories. For example, if on a given day the nurses selected category 4, while the physicians chose category 6 for the same patient, then the weight for this observation is \((6-4)/(10-1) = 2/9\).  

\(N_{pr}\) refers to the total number of predictions.  

\(N_p\) refers to the total number of patients. \(N_d\) refers to the number of prediction days.  

These measures are calculated separately for both groups.
operating characteristic Curve (AUC), the Brier score, and positive predictive values (PPV) at cut-off points 1, 2 and 3, i.e. the three lowest probability categories (the PPV is the percentage of non-survivors among those who were predicted to die). The AUC is a measure of discrimination (i.e. the ability to distinguish survivors from non-survivors) and is equivalent to the probability that a randomly selected patient who died had a higher predicted risk assigned to him/her than another randomly selected patient who survived. An AUC of 0.5 indicates that the model does not predict better than flipping a fair coin. An AUC of 1 indicates that the model discriminates perfectly between survivors and non-survivors. The Brier score contains information on both discrimination as well as precision [29]. The lower the Brier score the better. We subtracted the Brier score of the physicians’ predictions from the Brier score of the nurses’ predictions for the same patients to measure the pure difference in imprecision between the two [29]. We calculated 1) the Brier score of the nurses, $BS_N$, and the Brier score of the physicians $BS_P$, per patient per day, 2) the mean of the difference $BS_N - BS_P$ for each patient, and 3) the mean of the means of the total population. Finally, on the day of admission, nurses’ and physicians’ predictions were compared to those obtained by an objective mathematical model based on a recalibrated SAPS-II model obtained by first-level recalibration. First-level recalibration refers to fitting a new logistic regression model using the outcome from the new dataset and the (log odds of the) original probabilities as the sole input variable. We tested if outcome measures were statistically significantly different between nurses and physicians by calculating the differences in 1000 bootstrap samples and examined whether their 95% confidence interval contains zero.

**Results**

Of 4953 patients admitted during our study period, 3837 patients (77.5%) had at least one nurse and one physician prediction. After imputation and removal of cases with at least two consecutive missing values 2221 patients (57.9%) were included (Figure 8.1). For these patients, 7955 prediction days were available (including 4% imputed nurse predictions and 1% imputed physician predictions). Mortality was 17.9%. Patient characteristics of included and excluded patients, for either of the two reasons above, are summarized in Table 8.2. No significant differences between included and excluded patients were found.

<table>
<thead>
<tr>
<th>Table 8.2. Patient characteristics.</th>
<th>Validation set (study cohort)</th>
<th>Excluded patients due to missing values</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>2,221</td>
<td>2,732</td>
</tr>
<tr>
<td>Age (median [range])</td>
<td>64 [17-97]</td>
<td>64 [0-111]</td>
</tr>
<tr>
<td>Age (mean +/- sd)</td>
<td>61.2 +/- 16.2</td>
<td>61.2 +/- 16.7</td>
</tr>
<tr>
<td>Male(%)</td>
<td>63.4</td>
<td>63.2</td>
</tr>
<tr>
<td>Died (%)</td>
<td>17.1</td>
<td>16.8</td>
</tr>
<tr>
<td>APACHE II score (mean +/- sd)</td>
<td>18 +/- 7.1</td>
<td>17.7 +/- 7.1</td>
</tr>
<tr>
<td>SAPS II score (mean +/- sd)</td>
<td>47.2 +/- 16.8</td>
<td>47.0 +/- 17.3</td>
</tr>
<tr>
<td>Admission type (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>41.8</td>
<td>41.8</td>
</tr>
<tr>
<td>Unplanned surgery</td>
<td>16.3</td>
<td>15.8</td>
</tr>
<tr>
<td>Planned surgery</td>
<td>41.9</td>
<td>42.4</td>
</tr>
</tbody>
</table>
Characteristics of daily survival predictions of nurses and physicians in the Intensive Care and their predictive value

Figure 8.1. Data collection. \( N_{\text{Pat}} \) = number of patients, \( N_{\text{NP}} \) = Number of nurse predictions, \( N_{\text{PP}} \) = Number of physician predictions, \( NA_1 \) = missing values in dataset, \( NA_2 \) = missing records of days (including sofa)

Figure 8.2. Frequency of use (\( y \)-axis) of probability categories (\( x \)-axis) by nurses and physicians.
1) Characteristics of daily survival predictions

Figure 8.2 shows the distribution of probabilities given by nurses and physicians in survivors and non-survivors. For survivors, both nurses and physicians often chose the three highest survival probability categories (category 8, 9 or 10 was chosen in approximately 76% of the survivors). For non-survivors, physicians more often selected the lowest probability category (i.e. category 1) than nurses. The probability distribution followed a similar pattern in planned admissions (Figure 8.3a). In unplanned admissions (Figure 8.3b), nurses and physicians selected the lowest probability category for almost half of the non-survivors.

![Frequency of use (y-axis) of probability categories (x-axis) by nurses and physicians. A: Planned admissions. B: Unplanned admissions.](image)

Inter-observer agreement between nurses and physicians ranged from 20.4% in long term non-survivors on day 1 to 56.1% in non-survivors on day 2 (Table 8.3). Although inter-observer agreement decreased over time in non-survivors (49.7% on day 1 to 42.3% on day 8), it increased in survivors (33.7% on day 1 to 37.5% on day 8) and in long term patients (29.7% on day 1 to 49.7% on day 8). Mean inter-observer variance of different nurses was 2.22 categories. If predictions of different observers would be perfectly consistent, the inter-observer agreement would be 100% and the variance would be 0 (Table 8.1; Appendix 8.1).

| Table 8.3. Kappa agreement between nurses and physicians. |
|-----------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                 | Day 1  | Day 2  | Day 3  | Day 4  | Day 5  | Day 6  | Day 7  | Day 8  |
| All             | 52.6   | 55.1   | 54.0   | 52.7   | 50.1   | 49.6   | 47.3   | 49.7   |
| Survivors       | 33.7   | 33.4   | 35.6   | 35.0   | 37.2   | 37.6   | 32.6   | 37.5   |
| Non-survivors   | 49.7   | 56.1   | 54.4   | 51.4   | 45.5   | 45.0   | 45.0   | 42.3   |
| Long term       | 29.7   | 37.1   | 38.1   | 43.2   | 38.3   | 40.7   | 45.9   | 49.7   |
| Long term survivors | 29.5   | 31.9   | 30.6   | 37.1   | 31.0   | 33.1   | 35.0   | 37.5   |
| Long term non-survivors | 20.4   | 35.5   | 35.6   | 36.9   | 36.0   | 36.5   | 41.1   | 42.3   |

Variance over time was higher for physicians (0.69-0.81 categories) than for nurses (0.61-0.68 categories), higher in non-survivors compared to survivors, and higher in patients with a minimum stay of 8 days compared to the total population (Table 8.4).
Characteristics of daily survival predictions of nurses and physicians in the Intensive Care and their predictive value

Pairwise correlation of nurse predictions between different days was 0.65; mean pairwise absolute difference was 1.5 categories. For physicians the mean pairwise correlation between different days was 0.62 and the mean absolute difference 1.61 categories (Table 8.5). The standard deviation, i.e. sharpness of the predictions was 2.6 categories for nurses and 2.8 categories for physicians. Although nurses and physicians agreed in most patients with very high or very low chances of survival, there were 62 cases (0.01%) in which one of them thought the chance of survival was lower than 30%, while the other thought it was higher than 70% (Figure 8.4). Overall patterns were the same in planned admissions and unplanned admissions, but inter-observer variance and variance over time were generally lower in planned admissions, while inter-observer agreement was generally higher (data not shown).

Table 8.4. Variance of individual observers over time.

<table>
<thead>
<tr>
<th>Nurse</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
<th>Day 7</th>
<th>Day 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.64</td>
<td>0.61</td>
<td>0.62</td>
<td>0.64</td>
<td>0.63</td>
<td>0.61</td>
</tr>
<tr>
<td>Survivors</td>
<td>0.60</td>
<td>0.58</td>
<td>0.60</td>
<td>0.61</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>Non-survivors</td>
<td>0.75</td>
<td>0.70</td>
<td>0.68</td>
<td>0.69</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>Long term</td>
<td>0.68</td>
<td>0.64</td>
<td>0.65</td>
<td>0.64</td>
<td>0.63</td>
<td>0.61</td>
</tr>
<tr>
<td>Long term survivors</td>
<td>0.64</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>Long term non-survivors</td>
<td>0.76</td>
<td>0.71</td>
<td>0.71</td>
<td>0.69</td>
<td>0.64</td>
<td>0.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physician</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
<th>Day 7</th>
<th>Day 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.71</td>
<td>0.69</td>
<td>0.70</td>
<td>0.72</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>Survivors</td>
<td>0.77</td>
<td>0.77</td>
<td>0.78</td>
<td>0.70</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>Non-survivors</td>
<td>0.89</td>
<td>0.76</td>
<td>0.75</td>
<td>0.78</td>
<td>0.76</td>
<td>0.84</td>
</tr>
<tr>
<td>Long term</td>
<td>0.81</td>
<td>0.77</td>
<td>0.77</td>
<td>0.72</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>Long term survivors</td>
<td>0.65</td>
<td>0.68</td>
<td>0.70</td>
<td>0.70</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>Long term non-survivors</td>
<td>0.89</td>
<td>0.73</td>
<td>0.70</td>
<td>0.78</td>
<td>0.76</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Figure 8.4. Estimated survival probabilities by nurses and physicians for survivors (blue) and non-survivors (red). The sizes of the circles indicate the number of observations.
Table 8.5. Mean pairwise correlation and absolute difference between first 8 days.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>All</th>
<th>Survivors</th>
<th>Non-survivors</th>
<th>Long term</th>
<th>Long term survivors</th>
<th>Long term non-survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurse</td>
<td>0.65</td>
<td>0.54</td>
<td>0.57</td>
<td>0.61</td>
<td>0.57</td>
<td>0.49</td>
</tr>
<tr>
<td>Physician</td>
<td>0.62</td>
<td>0.53</td>
<td>0.51</td>
<td>0.54</td>
<td>0.46</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abs. diff.</th>
<th>All</th>
<th>Survivors</th>
<th>Non-survivors</th>
<th>Long term</th>
<th>Long term survivors</th>
<th>Long term non-survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurse</td>
<td>1.5</td>
<td>1.79</td>
<td>1.38</td>
<td>1.52</td>
<td>1.40</td>
<td>1.77</td>
</tr>
<tr>
<td>Physician</td>
<td>1.61</td>
<td>1.92</td>
<td>1.5</td>
<td>1.81</td>
<td>1.72</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Figure 8.5. Predictive performance (y-axis) at first 8 days of admission (x-axis). A: AUC. B: Brier Score.

2) Predictive value of daily survival predictions

Figure 8.5 shows the Brier scores and AUCs on the first 8 days of admission. Predictive performance gradually decreased during admission showing the same pattern for nurses
and physicians for both the Brier score and AUC. There were no statistically significant differences in Brier scores and AUCs between nurses and physicians. The combined predictions of nurses and physicians (by taking their mean) achieved better Brier scores (range 0.08-0.15) and AUCs (range 0.84-0.91). These differences were significant (p<0.05) for day 1-5 (AUC) and day 1-6 (Brier score) compared to nurses alone, and for day 1-8 (AUC) and days 1-3, 6 and 7 (Brier score) compared to physicians alone. In patients with a minimum stay of 8 days these patterns were however exactly the other way around, showing increasing predictive performance over the course of admission. These patterns were the same for planned and unplanned admissions (data not shown). Discriminative ability decreased most and increased least for unplanned admissions, especially for nurses. For the cut-off point of category 1, the positive predictive values of nurses ranged from 0.75 (day 8) to 1.0 (days 3 and 7), which means that on day 8, 75% of the patients predicted to have a survival probability of at most 10% by nurses actually died. For physicians, this range was from 0.78 (days 7 and 8) to 0.95 (day 1). Median positive predictive values for nurses at cut-off points 1, 2, and 3, were 0.95, 0.87 and 0.8, respectively. For physicians, these were 0.88, 0.82, and 0.76.

On the day of admission, both nurses and physicians performed statistically significantly better (p<0.05) than the recalibrated SAPS-II model in terms of discriminative ability (AUC = 0.889 for nurses, 0.882 for physicians and 0.801 for the recalibrated SAPS-II model) and precision (Brier = 0.085 for nurses, 0.082 for physicians and 0.112 for the recalibrated SAPS-II model). The mean difference between the physicians’ Brier score and the nurses’ Brier score was -0.004 for the total population (-0.006 for survivors and 0.008 for non-survivors). This means that nurse predictions are slightly closer to the observed mortality risk in non-survivors, while physician predictions are slightly closer to the observed risk in survivors and in the total population.

**Discussion**
In the Intensive Care, daily survival predictions by nurses and physicians showed the same general characteristics, but physicians were more inclined to give very low probabilities of survival in non-survivors than nurses. There is fair to moderate agreement between nurses and physicians. Physicians’ predictions were generally less consistent over time (i.e. there was more variance) than those of nurses, but they both had more difficulties with estimating survival probabilities in non-survivors and in the first couple of days of patients with a longer ICU stay (≥ 8 days).

Physicians and nurses generally had good, comparable predictive ability in terms of discrimination and precision, which was statistically significantly better than SAPS-II on the day of admission. We obtained similar results when correcting for inter-observer variation by calculating the Brier score per observer per patient, taking the mean per patient and then the mean of the total sample (data not shown). Although predictive ability of physicians and nurses in the overall population decreased over time, it in fact improved over time for patients with an ICU stay of at least 8 days. The decrease in their predictive ability in the overall population was conforming expected as patients with shorter ICU stay (which are presumably easier to predict) are already discharged or death. In general, the predictive ability of both nurses and physicians on the last eight days before discharge improved over days (the closer to the day of discharge the better; data not shown). For example, for a patient who stayed at the ICU for 22 days this would be from day 16 to
day 22. A possible explanation for the better predictions the closer to patient discharge is that predictions are influenced by knowledge about planned intervention decisions. For example, there is little challenge in predicting the outcome on the day that a decision has been made to stop treatment or discharge a patient to the ward. In addition, actions and discussions undertaken in preceding days to the decision facilitate making the predictions. Additional sensitivity analysis in which we excluded predictions at day of discharge showed indeed that clinicians performed slightly worse but confirmed the same patterns described above (data not shown).

To our knowledge this is the first study assessing and comparing characteristics of daily predictions of survival in the ICU by nurses and physicians, and their predictive value in terms of discrimination, precision and accuracy. We included a large number of patients and used various complementary measures to assess different aspects of inter-observer reliability, variance over time and predictive performance. We conducted additional sensitivity analyses showing the robustness of our results. Our study has some limitations. Our exclusion criterion of observations with at least two consecutive missing values lead to discarding a considerable number of patients. We compared the characteristics of excluded and included patients and found no statistically significant differences. Nurses had however slightly worse predictive ability in the group for which physician predictions were missing (data not shown). We had insufficient data to compare inexperienced with experienced physicians. Our dataset mainly consisted of predictions of less experienced physicians (<2 years of experience) and more experienced nurses (+/- 20 years of experience). Since in general experienced physicians are the ones taking decisions on withholding or withdrawal of treatment they may perform better than inexperienced ones. We presume that decisions on withholding or withdrawal of treatment are in general taken as a team. Team decisions were not taken into account in our analysis since it is unclear which specific team discussed which specific patient. In any case the teams making predictions and those making decisions are not necessarily the same. Although we may have used some sophisticated statistical techniques that are uneasy to understand, these techniques are necessary for a thorough understanding of daily subjective predictions. For example, rather than selecting arbitrary cut-off points to calculate sensitivity and specificity, the AUC includes all possible cut-off points. In addition, we used techniques that provide insight in levels of agreement and consistency (or variance) over time. Finally, we used 10 probability categories, while more specific predictions might be needed in clinical practice.

Very few studies assessed daily predictions of survival in the ICU [5,6,30], and these did not measure characteristics of daily predictions or predictive performance in terms of discrimination, precision and accuracy, and were based on small patient cohorts. Most of the studies that validated the inter-observer reliability of survival predictions by nurses and/or physicians with different levels of experience were based solely on the day of admission [6,15-18,23]. These studies generally found the same levels of intra- and inter-observer variance between groups of different types of clinicians as in our study, but none of them assessed inter-observer variance within a group of only nurses. In addition, it was shown that clinicians’ prognoses are influenced by several external factors, such as ego bias, reverse ego bias and level of experience [13,15-18,21-24]. No studies measured variance of individual observers over time.

In general, previous studies validating nurses’ and physicians’ predictions on the day of admission also found good, comparable predictive ability [10-21,30], although physicians
were more precise [14,21,31]. In our study, nurses did slightly better than physicians, but the differences were not statistically significant. Similar to the findings in this study, earlier studies found superior predictive ability of physicians over objective mathematical models on the day of admission [32], although one study found better precision in the objective model [12]. Combining the predictions of different sources (e.g. by taking the mean of different clinicians or combining subjective with objective information) however yields superior discriminative ability and precision over either source alone [5,12,14,17,19,20,23,30,33,34].

Serious clinical decisions regarding the withdrawal or continuation of intensive medical treatment are based on subjective predictions of patient survival. Although physicians and nurses have good predictive ability, this study shows that their perceptions of survival status are not always consistent over time and there is only fair to moderate agreement between them (range inter-observer agreement = 20.4-56.1%). These inconsistencies are not necessarily bad, but illustrate the difficulties of prognostication and stress the importance of inter-collegial communication, as “two know more than one” (significant better performance was found after combining the predictions from two sources). One should carefully think about how to improve the end-of-life decision-making process and communication of survival probabilities in the ICU. As nurses have similar predictive ability compared to physicians, but may view the patient from a different perspective, there may be room for improvement in the communication between nurses and physicians in end-of-life decision-making. In addition, although still very premature, there may be a future role for objective prognostic models in supporting inter-collegial communication and communication with patients and their families. We presented the results of this study in our clinical practice to stimulate awareness and self-reflection and used it as a starting point for structuring the end-of-life decision-making process.

Future research should focus on 1) understanding the process of end-of-life decision-making, inter-collegial communication, and communication with patients and their families, 2) identification of areas of improvement, and 3) the possible role of nurses and 4) role of objective prognostic information in this decision-making process.

Conclusions
Although nurses and physicians have good predictive ability which is superior to the SAPS-II model, they do not always agree and are not always consistent over time. This is not necessarily bad, but illustrates the difficulties of prognostication and stresses the importance of inter-collegial communication. (Experienced) nurses have similar predictive ability compared to (less experienced) physicians, but may view the patient from a different perspective. Although still very premature, future work should assess the possible role of nurses in the end-of-life decision-making process and for objective prognostic models in supporting inter-collegial communication and communication with patients and their families.

Appendices
Appendix 8.1 – Explanation of measures and their interpretation (available by request)
References