Prognosis in intensive care: inductive methods using sequential patterns of organ dysfunction scores
Toma, T.

Citation for published version (APA):

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 6

Assessing and Combining Repeated Prognosis of Physicians and Temporal Models in the Intensive Care

Lilian Minne, Tudor Toma, Evert de Jonge, Ameen Abu-Hanna

*Artificial Intelligence in Medicine, 2012. In press*
Abstract

Objective: Recently, we devised a method to develop prognostic models incorporating patterns of sequential organ failure to predict the eventual hospital mortality at each day of intensive care unit (ICU) stay. In this study, we investigate using a real world setting how these models perform compared to physicians, who are exposed to additional information than the models.

Methods: We developed prognostic models for day 2 to day 7 of ICU stay by data-driven discovery of patterns of sequential qualitative organ failure (SOFA) scores and embedding the patterns as binary variables in three types of logistic regression models. Type A models include the severity of illness score at admission (SAPS-II) and the SOFA patterns. Type B models add to these covariates the mean, max and delta (increments) of SOFA scores. Type C models include, in addition, the mean, max and delta in expert opinion (i.e. the physicians’ prediction of mortality).

Results: Physicians had a statistically significantly better discriminative ability compared to the models without subjective information (AUC range over days: 0.78-0.79 vs. 0.71-0.74) and comparable accuracy (Brier score range: 0.15-0.18 vs. 0.16-0.18). However when we combined both sources of predictions, in type C models, we arrived at a significantly superior discrimination as well as accuracy than the objective and subjective models alone (AUC range: 0.80-0.83; Brier score range: 0.13-0.16).

Conclusion: The models and the physicians draw on complementary information that can be best harnessed by combining both prediction sources. Extensive external validation and impact studies are imperative to further investigate the ability of the combined model.

Keywords: Prognostic models, calibration, discrimination, human intuition, temporal patterns.
6.1 Introduction

In the Intensive Care Unit (ICU), physicians are daily challenged with the complex task of prognosis. Prognosis can be defined as the prediction of (the probability of) an event, such as death, before its possible occurrence [1, 2]. Patient preferences to undergo intensive life sustaining treatments highly depend on their prognosis. In fact, they often prefer palliative care aiming at comfort and relief of pain if their chances of survival are very low [3, 4] and clinicians’ perceptions of survival chances strongly influence provision of life support [5]. Survival is usually defined as hospital survival, which is the probability that a particular patient will be discharged alive from the ICU or hospital ward after an ICU admission. 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. Predictions of survival can be obtained either subjectively (e.g. expert opinion) or objectively (e.g. mathematical models). The main advantage of an objective approach is that mathematical models are extremely consistent and able to optimally combine information into a global judgment [6]. Commonly used mathematical models for predicting mortality in the ICU are the Acute Physiology and Chronic Health Evaluation (APACHE) I-IV [7] model, the Simplified Acute Physiology Assessment (SAPS) I-III [8] models, and the Mortality Prediction Model (MPM) I-II [9]. These models are intended to provide the expected mortality risk for a given patient by adjusting for the patient’s severity of illness at admission to the ICU. These models are mainly used in a comparative audit among ICUs: the observed mortality in an ICU is compared to the expected mortality in that ICU. In comparison to objective predictions, subjective predictions may be inexact or even inconsistent [10, 11]. There can be substantial disagreement between different clinicians, even if they are both very experienced [12], and prognosis is the part of medical practice they feel most insecure about [13]. On the other hand, clinicians are able to incorporate important implicit knowledge outside the scope of mathematical models and react quickly to changing situations [6].
Several mathematical models have been compared to clinicians (i.e. expert opinion). In general, clinicians seem to have good discriminative ability [11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] that is superior to objective models [24]. Objective models tend however to have better calibration [16, 17]. Capitalizing on the individual strengths of both subjective and objective predictions by using a combined approach seems to yield superior discrimination and calibration over either one alone [11, 16, 18, 22, 25, 26]. However, the problem of comparing repeated predictions for the same patient over time between the two approaches has not been studied before, and in fact there is paucity of work on repeated predictions in either approach (subjective or objective). Repeated predictions are important to better individualize the prognosis for a patient as more information obtained over time becomes available. For example, the probability of survival of patients suffering major complications during their ICU stay will decrease. The current admission-based severity-of-illness ICU models are less suitable for individual prognosis because they are based on only physiologic data from the first 24 hours after ICU admission. We recently suggested a method for developing “temporal” prognostic models for providing hospital survival predictions at each day of ICU stay [27]. Specifically, we showed that the use of specific patterns of sequential data (e.g. the transition from normal renal function to renal failure) lead to models that can better discriminate between survivors and non-survivors and are more accurate than the current “static” models [28, 29]. The aim of this study is to investigate and compare daily prediction in adult ICU patients based on subjective and objective prognostic information. In particular we investigate the following sources of predictions: 1) Expert opinion (daily survival estimates by physicians) and 2) Three types of temporal models including different levels of objective and subjective information.
6.2 Methods

6.2.1 Data collection

We designed a prospective comparative study in which we included all consecutive patients admitted between October 2007 and October 2008 to a 28-bed multidisciplinary mixed adult ICU of a 1002-bed university hospital in Amsterdam. We extracted demographics, patient outcomes and all data necessary to calculate the severity of illness score in the first 24 hours (SAPS-II, which ranges between 0 and 163) and the daily organ failure scores (SOFA scores, ranging between 0 and 24) from the Dutch National Intensive Care Evaluation (NICE) registry [30]. The value of a SOFA score is the summation of 6 individual organ system failure scores each ranging from 0 to 4 indicating the degree of organ failure of this particular system where 0 indicates normal organ functioning and 4 indicates complete failure). Every 24 hours, physicians' estimates of the probability of survival up to hospital discharge were elicited and recorded in a software module that we developed and integrated into the Patient Data Management System (PDMS). A PDMS is a computerized system that automatically collects and stores vital parameters from the patient monitor and provides digital patient charts [31]. After completion of routine data collection, a question regarding the patient's probability of survival popped up automatically in which physicians could choose between 10 probability categories (0–10%, 10–20%, 20–30%, 30–40%, 40–50%, 50–60%, 60–70%, 70–80%, 80–90% and 90–100%) or state that they did not have any clue. Categories were indicated by red (low survival probabilities) and green (high survival probabilities) color-scales. The scoring moment was typically between 9 and 12 am, but estimates could be changed until 12 pm.

6.2.2 Physicians

At the time we started this study, our ICU employed 10 critical care attending physicians, 8 fellows and approximately 14 residents, all of which participated in the study. Critical care attending physicians are specialists (e.g. neurologists or cardiologists) who have
completed an additional intensive care specialization of two years. Fellows and residents have completed at least two years of post-MD training. Fellows are in training for the intensive care specialization and residents are in training to become anaesthesiologists. Physicians were unaware of both their colleagues’ assessments and the predictions given by the objective models. They were not trained to estimate survival probabilities and did not receive feedback. Physicians were notified of this study by email and by an announcement during their staff meeting.

6.2.3 Development of mathematical models

A model was developed for each day of ICU stay from day 2 to day 7 by the procedure described in [28]. This procedure involves two steps: 1) data-driven discovery of temporal patterns of sequential organ failure scores, and 2) embedding them in the familiar logistic regression model. As in the original work in [28], SOFA scores were categorized as low (L) if SOFA 0, ..., 6, medium (M) if SOFA 7, 8 or high (H) if SOFA 9, ..., 24 and patterns were aligned to the day of prediction. For example, the pattern M,H on day 4 (the day of prediction) means a high SOFA score on day 4 preceded by a medium score on day 3. In our systematic review on SOFA-based models [32], we found two promising strategies leading to good performing models (not necessarily for repeated predictions): the first was our approach based on temporal patterns (described above) and the second was the combination of SOFA abstractions with admission scores (e.g. SAPS-II). Commonly used abstractions of SOFA are its mean, maximum (both from the day of admission \(d_0\) to day of prediction \(d_i\)), and delta (difference in scores between day \(d_i\) and \(d_0\)). In this study, we developed logistic regression models containing three levels of information: Type A models which use SAPS-II and SOFA patterns; Type B models which in addition use mean, max and delta SOFA scores; and Type C models which in addition to the variables in Type B models also use mean, max and delta of expert predictions (by translating the survival predictions, e.g. “60-70%” into a mortality probability, e.g. 0.35). These variables were then considered candidates for inclusion as binary covariates (0: the patient does
not exhibit the pattern, and 1: the patient exhibits the pattern) in the temporal logistic regression models. The Akaike Information Criterion (AIC) \cite{33} was used to select the optimal subset of covariates yielding the best model (i.e. the one with the lowest possible AIC). The AIC is defined as $-2\ln(L) + 2k$ where $L$ is the maximized likelihood of the model and $k$ is the number of free parameters in the model. The AIC trades off predictive performance for parsimony by penalizing for the number of variables included in the model in order to avoid overfitting \cite{28}. The variable selection procedure iteratively eliminates the least predictive covariate (i.e. the one associated with the highest AIC) of a set of considered covariates until the AIC cannot be decreased any further. To develop a Type A model, we first only included patterns of length 1. Only those that survived the AIC selection procedure were included together with patterns of length 2 etc. To prevent collinearity, we did not allow for logically entailed patterns to be selected in the model: we sought the model with the lowest AIC that had a subset of variables without logical entailment. For example, consider the three patterns L,H; M,H; and H. The presence of either L,H or of M,H entails H. In order to avoid entailment among variables we compare (using the AIC) a model including L,H and M,H without H, with a model including H alone. For the Type B models, we selected the model with the lowest AIC from the following four models: Type A, Type A + mean SOFA, Type A + max SOFA, Type A + delta SOFA. Then we did the same for the Type C models with the Type B models as a starting point. The data extraction and model development process is shown in Figure 6.1. Finally, for comparison reasons, we developed classification trees based on type C information for each day of prediction. We first completely overgrew the tree to overfit the data and then pruned back the overgrown tree to its optimal size based on its minimum 10-fold cross-validation error.

6.2.4 Statistical analysis

Discrimination (i.e. the ability of the model to distinguish between survivors from non-survivors) was measured by the Area Under the Receiver Operating Characteristic Curve
Figure 6.1: Data extraction and model development including an example with hypothetical patient data for a model at day 3.

(AUC). An AUC ranges between 0 to 1 with higher values indicating better discriminative ability. Although the AUC is dependent on the prevalence of the event in the sample, it is common to consider an AUC between 0.6 and 0.7 as poor, between 0.7 and 0.8 as fair, between 0.8 and 0.9 as good and above 0.9 as excellent. The Brier score is a measure of accuracy, which has both aspects of discrimination and calibration (i.e. the degree of similarity between predicted risks and patients’ actual risks of death). Note that, according to the terminology of [34], the term accuracy in this paper refers to the Brier score and not the percentage of correctly classified cases as it is sometimes used in machine learning. The formula of the Brier score is:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$
where \( N \) is the number of observations, \( p \) the predicted probability and \( o \) the observed outcome (0: survival, 1: non-survival). The larger the Brier score the worse the accuracy. Parameter estimates and confidence intervals were obtained by calculating the bootstrap sampling distributions of the respective statistics based on 1,000 bootstrap samples. Records with missing values of SAPS-II were excluded from the analysis, while missing physicians' predictions were imputed by taking the value of the previous day or, when not available, the mean of all predictions available for the patient involved. Analyses were conducted in the R statistical environment.

### 6.3 Results

#### 6.3.1 Patient characteristics

The baseline characteristics of the 912 patients admitted during the study period are shown in Table 6.1. 3955 physicians-predictions of survival were registered, and 623 values (13.6%) were missing. Physicians indicated in 77 cases (1.5%) that they had no idea regarding a patient’s chance of survival. There were no missing values in SOFA scores and only two missing values in SAPS-II (0.2%). Hospital mortality was 19.2%.

#### 6.3.2 Distributions of SOFA scores and expert opinion

Figure 6.2 shows histograms of expert opinion (i.e. the physician’s predictions of mortality) and SOFA scores. Mean +/- standard deviation of the SOFA score was 7.87 +/- 1.85. Mean predicted mortality by physicians was 37.6% (1.96 times the actual mortality). Mean standard deviation of expert opinion per patient (over the days) was 13.3%. To assess whether expert opinion could reliably identify high-risk patients, all patients with predicted mortality above 90% were obtained yielding 135 patients. Of these 74.8% actually died (this is the positive predictive value at 0.9).

#### 6.3.3 Mathematical models

The logits (log odds) of the best temporal models for days 2-6 are summarized in Table 6.2.
Survivors  Non-survivors  Total population
\( n=737 \)  \( n=175 \)  

<table>
<thead>
<tr>
<th>Admission type (%)</th>
<th>Survivors</th>
<th>Non-survivors</th>
<th>Total population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical</td>
<td>34.9</td>
<td>62.5</td>
<td>40.7</td>
</tr>
<tr>
<td>Urgent</td>
<td>16.7</td>
<td>25.1</td>
<td>18.3</td>
</tr>
<tr>
<td>Planned</td>
<td>48.4</td>
<td>9.7</td>
<td>41.0</td>
</tr>
</tbody>
</table>

| Mean age (SD) (years) | 60.3 (16.5) | 64.0 (15.4) | 61.0 (16.4) |
| Male/female (%)       | 64.3/35.7   | 56.6/43.4    | 62.8/37.2    |
| Median LOS [IQR] (days) | 2 [0-5]    | 3 [1-9]      | 2 [1-6]      |
| Mean SAPS-II (SD)     | 44.0 (15.3) | 57.2 (18.8)  | 46.5 (16.8)  |
| Mean APACHE-II ± SD   | 18.2 (6.7)  | 24.7 (7.4)   | 19.5 (7.3)   |
| Mean SOFA (SD)        | 7.0 (3.7)   | 10.1 (4.1)   | 7.9 (4.0)    |
| Mean expert opinion (SD) (%) | 29.5 (24.2) | 57.7 (27.7) | 37.6 (28.2) |

Table 6.1: Patient characteristics. IQR = Inter-quartile range, LOS = Length of Stay, SAPS = Simplified Acute Physiology Score, SD = Standard deviation, SOFA = Sequential Organ Failure Assessment.

For example the best temporal model of type A on day 4 is described by the following logistic regression model:

\[
e^{-2.06 + 0.022 \cdot \text{SAPSII} + 0.47 \cdot H - 0.95 \cdot L,L} \cdot \frac{1}{1 + e^{-2.06 + 0.022 \cdot \text{SAPSII} + 0.47 \cdot H - 0.95 \cdot L,L}}
\]

where \( H \) means high SOFA score on day 4 (the day of prediction), and \( L,L \) means low SOFA scores on days 3 and 4. Consider for example the hypothetical patient from Figure 6.1 which has a SAPS-II score of 25. If this patient would have a medium SOFA score on day 4, the probability of dying according to this model would be:

\[
e^{-2.06 + 0.022 \cdot 25 + 0.47 \cdot 0 - 0.95 \cdot 0} \cdot \frac{1}{1 + e^{-2.06 + 0.022 \cdot 25 + 0.47 \cdot 0 - 0.95 \cdot 0}} = 18.1\%
\]

as both patterns are absent. If the SOFA pattern was \( L,L \), however, the probability of dying would be reduced to:
Note that Table 6.2 shows no type B models after day 2 as type A models had a lower AIC than all of the following models: type A + mean SOFA, type A + max SOFA and type A + delta SOFA. This means that, after day 2, mean, max and delta SOFA had no significant added prognostic value over the SOFA patterns.

Figure 6.3 shows the classification tree that was fit for day 4. Expert opinion abstractions were always selected as the root of the classification tree (mean on days 2 and days 5 to 7, and maximum on days 3 and 4). Mean SOFA was included as one of the variables in three of the six tree-models, while SAPS-II (on day 2), the pattern L,L (on day 4) and delta expert opinion (on day 6) were included in one of the six tree-models. In our clas-
Table 6.2: Description of the developed logistic regression models. SAPSII = Simplified Acute Physiology Score-II, L = Low (score of 0-7), M = Medium (score of 7-8), H = High (score of 9-24), expert = expert opinion, i.e. the physician’s predictions of mortality. Note that no type B models are shown after day 2 as for these days type A models had a lower AIC than type A + mean SOFA, type A + max SOFA and type A + delta SOFA.

6.3.4 Validation

Predictive performance of the temporal models and expert opinion is summarized in Figure 6.4. The AUC of expert opinion ranged from 0.78 (on day 7) to 0.79 (on days 2-6), while the Brier score ranged from 0.15 (day 2) to 0.18 (day 5-7). Type A/B models were comparable to physicians in terms of the Brier score (range: 0.16-0.18), but significantly worse in terms of the AUC (range: 0.71-0.74). Type C models had higher (better) AUCs
Figure 6.3: Classification tree for day 4. The candidate variables were the same as for the Type C temporal logistic regression models. Mxexp = max expert opinion (i.e. predicted probability of mortality by the physician, mnexp = mean expert opinion, mnsoga = mean SOFA score (all three from the day of admission until day 4), X2 = pattern L,L (i.e. low SOFA score on days 3 and 4).

(range: 0.80-0.83) than physicians for all days of prediction as well as lower (better) Brier scores (range: 0.13-0.16). The differences were statistically significant on day 2 for the AUC and on days 2-4 for the Brier score. Type C models were also statistically significantly better compared to Type A/B models for all days of prediction. The classification tree had worse AUCs (range: 0.67-0.79) and Brier scores (range: 0.13-0.17) compared to
Figure 6.4: Predictive performance of physicians and temporal models. AUC = Area Under the Receiver Operating Curve. Note that no predictive performance of type B models is shown after day 2 as for these days type A models had a lower AIC than type A + mean SOFA, type A + max SOFA and type A + delta SOFA.

The differences were significant for all days except for day 6 for the AUC and for days 3 and 5 for the Brier score.
6.4 Discussion

In this study, we compared the performance of physicians and mathematical models in predicting mortality in the ICU. Physicians had fair to good discriminative ability (AUC range: 0.78-0.79), which was statistically significantly better compared to the mathematical models (AUC range: 0.71-0.74). The accuracy of physicians (Brier score range: 0.15-0.18) was comparable to the accuracy of the mathematical models without subjective information (Brier score range: 0.16-0.18). Nevertheless, physicians seem to markedly overestimate mortality. The mean predicted mortality by physicians (37.6%) was much higher than the observed mortality (19.2%) and patients estimated as high risk patients (> 0.9 probability of death) still had a realistic chance to survive (25.2%). Mathematical models were significantly improved by adding subjective information (AUC range 0.80-0.83 and Brier score range: 0.13-0.16), which made them superior to physicians. Remarkably, for most prediction days no models were found that included abstractions of the SOFA score. This shows that the SOFA patterns have superior predictive value than the SOFA abstractions. The combined model only included the SAPS-II score on day 2, while it was included in the other model types until day 6. Similarly, the tree-model included the SAPS-II score only on day 2. Apparently, the SAPS-II score looses its predictive value with increasing ICU stay days as it only contains information pertaining to the first day of admission. This explanation is confirmed by the increasingly reduced value of the SAPS-II coefficient over days. As expected, the expert opinion coefficient in the combined models is positive, whereas SOFA patterns that only include the “low” category have a negative coefficient (indicating that these patterns have a protecting effect in respect to mortality). No other patterns have been selected. These other patterns may either have low frequency in the dataset or low additive predictive value. In recursive partitioning, expert opinion was always chosen as the root of the classification tree. The high coefficients of expert opinion in the logistic regression models also confirm that this may be the most important predictor of mortality. This can be partly explained as physicians are the ones
who are responsible for making end-of-life decisions. Of note, only one tree-model used a SOFA pattern. Compared to the tree-models, the logistic regression models had however superior predictive performance in terms of both accuracy and discrimination. It is likely that once the tree had made use of the most predictive variables it was left with relatively too little data to render the SOFA patterns as significant. To our knowledge the present pilot study is unique, and the first in comparing models with daily predictions of physicians. The idea of models with repeated predictions is relatively new, and these models have not been combined with subjective information before, at least not in the ICU. Other strengths of this study are the use of bootstrapping techniques to correct for optimistic estimates of predictive performance because of similarities in the training and testing datasets, the AIC based incremental variable selection and the comparison of the techniques of logistic regression and recursive partitioning. Weaknesses of this study include the use of a relatively small dataset and the fact that different physicians provided estimates for the same patients. Other studies scrutinizing expert opinion on a daily basis either validated the most recent prediction only [26] or did not provide direct measures of predictive performance, such as the AUC or error rates [5, 35]. In general, other studies validating predictive ability of physicians also found fair to good discrimination [11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] and superior predictive performance of a combined approach of subjective and objective information [11, 16, 18, 22, 25, 26]. We found that the mathematical models can be significantly improved by the inclusion of subjective information. These models can provide some patient groups with information about (non-)survival probabilities, and can be potentially useful to support individual decision-making. Although some of these patients will have very high probabilities to die and may prefer palliative care, the models might also help physicians not to withdraw therapy. As physicians tend to overestimate mortality, the models might be able to prevent unjustified decisions leading to withdrawal of treatment in patients who would otherwise have a good chance to survive. Although one may argue that a patient would benefit more from an accurate prediction on the day of admission, than on the day before his
or her non-survival, it is often hard to say at an early stage which patients will survive ICU admission. Only patients that may benefit from ICU admission are admitted, but it is likely that their chances for survival during admission will change as their condition will change for either better or worse. Indeed we found in this study that SAPS-II may lose its predictive value with increasing ICU days of stay. Our models provide additional information on the patient’s probability of survival during ICU admission that may be used in addition to the admission model. Moreover, an advantage of including expert opinion in the final prognosis is that it may lead to higher acceptance of the models by physicians and patients or their families. Before they can make their way into clinical practice, however, extensive external validation and studies on their impact on clinical decisions and patient outcomes are imperative [36]. Future research needs to focus on questions about how to optimize these temporal models, the potential value of including nurses’ predictions of mortality, their external validity, their acceptability by clinicians and their potential impact on clinical decisions and patient outcomes. Note also that in our study the AUC and Brier scores were generally better on the first days of admission, but that the patients in later days form subsets of those in earlier days because on each day some patients leave the ICU (they are either discharged alive or die). Investigating the behavior of the AUC and Brier scores over time for the same patient group merits future research.

6.5 Conclusion

Mathematical models have fair but worse discrimination than physicians, but better calibration. They can be significantly improved by the inclusion of subjective information. This may lead to higher acceptance of the models by physicians and patients or their families to support individual decision-making in patient groups with very high risk of dying. Before they can make their way into clinical practice, however, extensive external validation and studies on their impact on clinical decisions and patient outcomes are imperative [36].
Acknowledgements

This work was partially supported by the Netherlands Organization for Scientific Research (NWO) under the I-Catcher project number 634.000.020.

Bibliography


