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 5

Statistical process control for validating a classification tree model for predicting mortality
A novel approach towards temporal validation

Minne L, Eslami S, de Keizer NF, de Jonge E, de Rooij SE, Abu-Hanna A.
Statistical process control for validating a classification tree model for predicting mortality – A novel approach towards temporal validation.
The original publication is available at
Abstract
Prediction models are postulated as useful tools to support tasks such as clinical decision making and benchmarking. In particular, classification tree models have enjoyed much interest in the Biomedical Informatics literature. However, their prospective predictive performance over the course of time has not been investigated. In this paper we suggest and apply statistical process control methods to monitor over more than 5 years the prospective predictive performance of TM80+, one of the few classification tree models published in the clinical literature. TM80+ is a model for predicting mortality among very elderly patients in the intensive care based on a multi-center dataset. We also inspect the predictive performance at the tree’s leaves. This study provides important insights into patterns of (in)stability of the tree’s performance and its “shelf life”. The study underlies the importance of continuous validation of prognostic models over time using statistical tools and the timely recalibration of tree models.

Introduction
Prognostic models are postulated as useful tools in medicine to support tasks such as benchmarking, identification of patients at risk, workload planning and even individual clinical decision making [1–3]. For example in benchmarking, which has a long tradition in the intensive care [8,9], hospitals are compared to each other on the quality of delivered care. In this task prognostic models are first used to provide mortality predictions that are adjusted for the severity of illness of individual patients in each hospital. Then for each hospital the total predicted mortality is compared to the actual observed mortality proportion. The lower the observed mortality (compared to the predicted one), the better the quality of care.

The actual use of prognostic models for any application, however, requires trust in it and this trust is in turn influenced by the clinical credibility and the validity of the models [3,4]. There are different methods and types of prognostic models including logistic regression, neural networks, and classification trees (also called decision trees) [2]. Ideally a model is easily interpretable, clinically credible and statistically valid. In terms of interpretability, the collection of symbolic ‘if-then rules’ that the classification tree represents can provide easier interpretation of the model than purely quantitative models such as neural networks, or even logistic regression. One should however note that the inherent hierarchical structure of classification trees may also limit the interpretability of some rules (for example when predicting adverse events the terms in a rule do not have to necessarily correspond to risk factors and there is a need to simplify the rules first) [5]. On the whole, however, the interpretability of the tree has the potential to support the clinical credibility of the model especially when it provides transparency in the nature and reason of a prediction [6].

In contrast to clinical credibility, which is essentially subjective, there are standard performance measures to assess model validity in terms of discrimination, calibration and accuracy of the predictions [7]. Although performance of developed models is often reported on a held out test set from the available dataset to researchers, there is little work on model validation on prospectively collected data from the same source [3] and we are unaware of work describing prospective validation over the course of time of any published tree-based prognostic model.
Temporal validation is important in medicine due to the effects of population drift [8] implying changes in patient mix, technologies and treatment policies. These changes can affect model future performance [9,10]. This is especially relevant for classification trees as their non-parametric nature renders them more sensitive to overfitting than parametric models, and their performance is affected by variations in the training and test sets [11].

Investigating the temporal performance of prognostic models is often performed once on a prospective data set. An important added value to this can be the continuous monitoring of performance over time. Insights obtained into the continuous temporal validation of classification trees forms an important contribution to prognostic research in Biomedical Informatics.

The aim of this study was to assess and scrutinize the temporal validity of various performance aspects of a classification tree model developed on multi-center data, over the course of a long period of time. The tree-model [6], here referred to as TM80+, was designed to predict mortality in patients aged 80 years and older admitted to the Intensive Care Unit (ICU) and also to identify patient subgroups at very high risk. TM80+ was developed on 6867 consecutive patients of 21 ICUs of university, teaching, and nonteaching hospitals in the Netherlands and is the only mortality prediction model designed specifically for all types of ICU patients aged 80 years and older available in literature [12].

For monitoring predictive performance over time we propose and make use of methods originating from the statistical process control (SPC) methodology. SPC is attractive because it combines intuitive graphical tools for inspecting changes and variation over time with rigorous statistical inference capabilities. We resort to the predictive performance measures of the area under the ROC curve (AUC) and the Brier score. In addition we monitor the mean mortality in patients assigned to the tree’s leaves especially those identified as very high-risk groups by TM80+. This addresses the question of the validity of using the model for the identification of high-risk patients. Finally we also inspect the changes over time in the proportion of patients classified in the tree’s leaves. This proportion is not indicative of the validity of the model, but is solely directed towards understanding the fluctuations in the population over time.

In particular in this paper we address the following research questions: (1) What is the behavior over the course over time of the predictive performance of TM80+? (2) What is the behaviour over time of the proportions of patients classified in the tree leaves and the mortality rate in each leaf? and (3) Can TM80+ reliably identify high risk subgroups? Based on this case-study we reflect on the implications for the use of classification tree models as prognostic instruments in biomedical research.

Background

Classification trees

A classification tree is a non-parametric, data-driven model obtained by recursive partitioning of instances in a training data sample into mutually exclusive subsets. An internal node in the tree represents a predictor variable, and each value of this variable is represented as a branch along which instances are sorted (e.g. for the variable “sex”, instances for men and women are sorted apart, rendering a partition). Tree-models classify new instances by sorting them from the root to some leaf. At the leaf the instance is either assigned the majority class of the training instances (e.g. “survivor” or “non-survivor”) or the instance is assigned a probability estimate of that class. Algorithms to fit tree-models
such as the commonly known ID3 algorithm [13]) grow their branches in an attempt to perfectly classify the training instances. If the size of the tree (e.g. in terms of number of splits) is not controlled, however, the tree model can easily overfit the data. This is especially the case when there is noise in the data (e.g. measurement errors) or when the training set is small. Overfitting occurs when there exists some alternative tree model that may have worse performance (e.g. classification error) than the final tree on the training sample, but better performance in the entire population. Strategies to control the size of the tree vary from deciding on preventing specific nodes to branch out during fitting the tree, to post-pruning an overfitted tree based on a measure of the estimated optimal tree complexity.

Tree-models can easily be represented as sets of if-then rules to improve human readability (a rule represents a path in the tree). Tree-models enjoy a wide popularity in Biomedical Informatics and Machine Learning, partly due to their symbolic representation that facilitates their easy interpretability.

Validation of prognostic models

For assessing a model’s performance one may distinguish between four validation designs: apparent, internal, temporal, and external performance validation [3,9]. In apparent validation the performance of the model is measured on the training set itself. However, this performance tends to be overly optimistic [11] due to overfitting the data. To evaluate the ability of a prognostic model to generalize beyond the training sample, its performance needs to be validated on other data than the set used to develop the model (i.e. the training set). In internal validation, data from the sample available to the researchers is randomly selected and held out for performance evaluation; it is hence not used for developing the model. When data are scarce, resampling techniques like cross-validation or bootstrapping are used instead [3]. Internal validation itself may still be optimistic in the sense that it disregards the population drift that may have occurred over time and hence the model may not generalize beyond the temporal frame it was developed in. Temporal validation refers to evaluating the performance on subsequent data, collected later in time, but from the same data sources that provided the training sample. External validation is defined as validation on a test sample obtained from other sources of data (e.g. other hospitals) than those providing the training sample.

Which validation design is most useful depends on the intended use of the model. In health care, if one intends to use a model for a specific set of participating centers, then temporal validation probably strikes the best balance between the desired generalizability over time and the specificity for the participating centers. In this study we address a new variant of temporal validation: not only that the performance of the model is evaluated on a prospectively collected sample from the participating centers, the performance is also continuously monitored. The advantage of this design is that it allows for inspecting the variability of the performance measures over the course of time, allowing for detecting behavioral patterns over time and indicating a model’s “shelf-life”. We will refer to this design as cTEMP in the sequel. In literature, several studies are available that attempted internal [14–18], temporal [19], or external validation [14,20,21] of classification trees, but to the best of our knowledge there are no studies monitoring model performance continuously over time.
Performance measures
Predictive performance of prognostic models can be measured in terms of discrimination, calibration and/or accuracy. These performance measures will be explained here for the outcome denoting patients’ non-survival. Discrimination, usually measured by the area under the receiver operating characteristic curve (AUC; which is equivalent to the c-index), refers to a model’s ability to distinguish survivors from non-survivors. The AUC is a normalized Mann–Whitney U statistic applied to the model’s predictions, grouped by observed outcomes. It is the probability that an arbitrary patient who died was indeed assigned a higher predicted risk than an arbitrary patient who survived. An AUC of 0.5 indicates that the model does not predict better than chance. An AUC of 1 indicates that the model discriminates perfectly between survivors and non-survivors. The AUC, however, is a non-strictly proper scoring rule [22] meaning that its optimal value can be obtained even if predicted probabilities are not close to the true ones (estimated by the observed proportion of death). Calibration refers to the agreement between predicted and true probabilities. (Mis)calibration is sometimes measured by the Hosmer–Lemeshow statistics, which rely on creating probability groups and measuring the difference between the mean probability in each group with the actual proportion of death in that group of patients. However these statistics are strongly influenced by sample size and by the distribution of the probabilities around the cutoff points used to group the probabilities [23]. Instead, the Brier score is often used, which is the mean squared individual residuals (between each actual outcome and the predicted probability of this outcome):

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$

where $N$ is the number of instances, $p_i$ is the predicted probability for instance $i$ to have the true outcome $o_i$ (0 for survivals and 1 for non-survivals). The Brier score is technically an accuracy measure because it contrasts actual outcomes with predicted probabilities but it includes both discrimination and calibration aspects. In addition it is a strictly proper scoring rule meaning that it gets its minimal (best) value only when the predicted probability equals the actual one.

Another measure of accuracy is the positive predictive value PPV, which is the percentage of patients that died among those predicted to die (because their predicted probability of non-survival exceeds a given threshold probability).

Statistical process control
Statistical process control (SPC) combines rigorous time series analysis and graphical data presentation, which is used to identify structural changes in a process [24]. Its primary tool, the process control chart, is a plot of the data over time with three additional lines; the center line (usually the mean) and upper and lower control limits, typically set at ±3 adjusted standard deviations (called sigma) from the mean. When the process has only inherent variation then the data points exhibit no special patterns and are within the control limits. The process is said to be “in-control” and stable (e.g. the process is predictable within certain limits). Special cause variation, on the other hand, signifies that the process is no longer stable or predictable and has changed either for better or for worse [25]. The most common type of control chart used for individual measurement (i.e. continuous) data is the XMR (Individual Moving Range) chart. XMR is suited for normally distributed
data and detects large sudden changes in processes. The exponentially weighted moving average (EWMA) control chart, in contrast, is more adequate for handling non-normally distributed data [26] and accentuates trends rather than individual points because it smoothes the data points over time, ameliorating the effects of outliers.

**Exponentially weighted moving average (EWMA) control charts**

Exponentially weighted moving average (EWMA) control charts smooth data by combining current and historical observations to rapidly detect possibly small changes in the mean of the monitored statistic [27]. A weighting factor \( w \) between 0 and 1, determines the weight (importance) of each observation \((\infty)\), where more weight is provided to the more recent observations and then decays when moving “backwards” in time. The weighted average for the \( t \)th period \( t = 1, 2, \ldots \) is given by the equation \( y_t = w y_{t-1} + (1 - w) y_{t-1} \) where \( y_0 = m_0 \) which is the mean (this will be the centerline). The weights sum to unity and decrease geometrically.

The upper and lower control limits (UCL and LCL) are defined by

\[
UCL = m_0 + L \cdot s \sqrt{\frac{w}{2 - w} \left[1 - (1 - w)^2t\right]} \quad \text{and} \\
LCL = m_0 - L \cdot s \sqrt{\frac{w}{2 - w} \left[1 - (1 - w)^2t\right]}
\]

where \( s \) denotes sigma, \( L \) is a multiplier that depends on \( w \) and the desirable in-control chart performance. The term \( [1 - (1 - w)^2t] \) converges to 1 with increasing \( t \). Therefore, the control limits approach [27]:

\[
UCL = m_0 + L \cdot s \sqrt{\frac{w}{2 - w}} \quad \text{and} \\
LCL = m_0 - L \cdot s \sqrt{\frac{w}{2 - w}}.
\]

In EWMA charts, a process is defined as unstable if one or more points fall above or below the process control limits (called outliers) [28].

Commonly, the values of \( m_0 \) and \( s \) are calculated based on the mean and sequential changes, respectively, of the time-series that is monitored. However, one may also determine \( m_0 \) and \( s \) based on “external” factors (for example one may use some industrial standards). In the context of our work, the prognostic model TM80+ has been proposed including an estimation of its performance and its variability. An important question that potential prospective users of TM80+ should pose is how TM80+ would fare in comparison to the reported mean and variability of TM80+’s performance. To address this question, \( m_0 \) and \( s \) should be calculated from the retrospective period when the model was first internally validated.

**Material and methods**

**Statistical analysis**

In this study we assessed the temporal validity of a classification tree over the course of time in terms of its AUC and Brier score using SPC control charts. As the AUC and Brier
score are both non-normally distributed and because we were interested in structural persistent changes, rather than in sporadic maxima or minima of performance, we used EWMA control charts. We split our dataset in 20 consecutive subsets of equal size $N_g$, which is adequate for valid SPC analysis (the number of groups should be between 12 and 36 [29]). The number of groups was chosen to be 20 because it still results in sizable groups ($\geq$600 admissions per group) and in long enough series of performance measures to be scrutinized over time. Group sizes were fixed such that their means were based on the same numbers of patients (the resulting time periods were still similar though, between three to four months in each group). Centerlines and control limits in the SPC chart were estimated from the original internal validation set of the tree-model based on $m_0$ and $s$ (sigma), where $m_0$ is the mean performance and $s$ was taken as the standard deviation. These values should however be calculated for sample size of $N_g$. Hence we took 1000 random samples of size $N_g$ obtained from this original internal validation set and then calculated for each sample the performance measure of interest (this also allows for inspecting the distribution of these statistics). Next, the mean ($m_0$) and sigma ($s$) of these 1000 values were calculated, averaged, and then used in the formulas for UCL and LCL described above to derive the control limits. We selected the commonly used default values for $w$ (which is 0.2) and the desired performance level $L$ (which is 3) [29]. Previous research showed that setting the performance level at 3 sigma ($L = 3$) works best for EWMA [29] and that, in order that small shifts be handily detected, $w$ should lie somewhere between 0.2 and 0.5 [30].

The most frequent choices are 0.25 and 0.33 [31–33]. We assess the robustness of our approach by using different values for $L$ (2, 2.5 and 3) and $w$ (0.2, 0.25, 0.3, 0.33, 0.4 and 0.5). Aside from the AUC and Brier score of the whole tree, we also assessed, in the same manner, the behavior over time of each of the tree-model’s leaves. This is done by calculating the proportion of patients assigned to each leaf and the mean mortality rate assigned to it.

To assess the tree-model’s ability to identify high-risk subgroups, we were specifically interested in the mean observed mortality in the three leaves corresponding to the high-risk subgroups identified in de Rooij et al. [6]: G1) GCS > 6, planned surgery, 24 h urine production < 1.25 liter and bicarbonate < 16.7 mmol per liter (mean mortality in this group was 78%); G2) GCS > 6, unplanned surgery or medical reason of admission, 24 h urine production < 0.75 liter and systolic arterial blood pressure < 72 mm Hg (mean mortality in this group was 88%); G3) Glasgow Coma Score (GCS) < 7 (mean mortality in this group was 89%). When the threshold on non-survival is taken to be lower than the observed mean mortality in a leaf then all patients in the leaf are predicted to die. Hence the mean mortality in a leaf is equivalent to the PPV. We both monitored mortality in each of these subgroups individually and in the three of them combined. All analyses were conducted in the statistical environment R version 2.10.1 [34].

**Case study**

The TM80+ tree-model [6] (Appendix 5.1) was developed on the data of the 21 university, teaching and non-teaching hospitals in the Netherlands participating in the National Intensive Care Evaluation (NICE) registry [35] from January 1997 to December 2003. The data consisted of 6867 patient records aged 80 years and older admitted between January 1997 and December 2003. TM80+ was internally validated by randomly selecting two thirds of the dataset for training and one third for testing [6]. Its performance in the
internal validation set in terms of AUC and Brier score (± standard deviation) was 0.77 ± 0.01 and 0.16 ± 0.005, respectively. Our temporal validation set, used in this paper, consisted of all consecutive patients aged 80 years and older admitted from January 2004 to July 2009 to the same 21 ICUs which data were used for model development.

Results of the case study

During our study period 12,143 patients were admitted to the participating centers and included in this study. Patient characteristics are summarized in Table 5.1. The overall AUC and Brier score ± standard deviation of the classification-tree model in this set of patients were 0.76 ± 0.005 and 0.17 ± 0.002, respectively.

| Table 5.1. Patient characteristics in the developmental, temporal and external sets. |
|---------------------------------|----------------|
| **Internal validation set**     | **Temporal validation set** |
| N                               | 2,289          | 12,143         |
| Age (range)                     | 80-103         | 80-108         |
| Age (mean +/− sd)               | 83.5+3.6       | 82.5+13.7      |
| Male(%)                         | 48.0           | 50.2           |
| Died (%)                        | 30.5           | 32.0           |
| APACHE II score (mean +/− sd)   | 18.4+7.2       | 19.1+7.4       |
| SAPS II score (mean +/− sd)     | 41.6+17.2      | 43.9+17.4      |
| LOS ICU, days (median (IQR))    | 1.1 (0.8-3.2)  | 1.4 (0.8-3.7)  |
| Admission type (%)              |                |                |
| Medical                         | 33.8           | 38.7           |
| Unplanned surgery               | 19.3           | 20.0           |
| Planned surgery                 | 46.8           | 41.3           |

![AUC of Tree-Model](image1.png) ![Brier Score of Tree-Model](image2.png)

**Figure 5.1.** Temporal validation of tree-model: A: AUC of tree-model (all points fall within the upper and lower limits, hence the process is stable); B: Brier score of tree-model (process is unstable as some points fall above the upper limit, indicating increased inaccuracy). Each of the 20 (time) groups consists of about 607 consecutively admitted patients (points 1-3 [2004]; points 3-7 [2005]; points 7-11 [2006]; points 11-14 [2007]; points 14-18 [2008]; points 18-20 [2009]. The plus-signs represent the original (non-smoothed) data points.
Figure 5.1 demonstrates that although the AUC of the tree-model was stable, its Brier score was not (the Brier chart showing 9 outliers). Table 5.2 shows the results of sensitivity analyses with different values for $w$ and $L = 3$. The AUC was stable for each value of $w$, while the Brier score was unstable for each $w$ except for $w = 0.5$. We repeated the analyses for $L = 2$ and $L = 2.5$ (data not shown). For $L = 2.5$ the results were essentially the same as in the table. For $L = 2$, as may be expected, more instability was encountered. This was however only the case for the AUC, but not for the Brier score, and only for the lower $w$ values. Smaller values of $L$ (which narrow the control limits) and smaller values of $w$ (which give more weight to past data and thus more smoothed data points) result in more sensitive charts but increase the risk of false-positives.

**Table 5.2. Sensitivity Analysis, based on the recommended value of $L=3$ [29].**

<table>
<thead>
<tr>
<th>$w$</th>
<th>Stable</th>
<th>Unstable</th>
<th>Stable</th>
<th>Unstable</th>
<th>Stable</th>
<th>Unstable</th>
<th>Stable</th>
<th>Stable</th>
<th>Stable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brier score</td>
<td>0.2</td>
<td>0.25</td>
<td>0.3</td>
<td>0.33</td>
<td>0.4</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality in combined high-risk subgroup</td>
<td>Stable</td>
<td>Stable</td>
<td>Stable</td>
<td>Stable</td>
<td>Stable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.3. Stability in tree-model’s leaves.**

<table>
<thead>
<tr>
<th>Leaf Description</th>
<th>Orig. N</th>
<th>Obs. N</th>
<th>GS</th>
<th>Stable</th>
<th>Orig. P</th>
<th>Obs. O</th>
<th>GS</th>
<th>Stable</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. GCS &gt; 6 AND urgent surgery AND 24h urine &lt; 750 ml AND Systolic ABP &gt;= 72 mmHg AND pH &gt;= 7.35</td>
<td>1.6</td>
<td>2 [0-4]</td>
<td>607</td>
<td>No</td>
<td>29 [20-40]</td>
<td>44 [14-62]</td>
<td>14</td>
<td>Yes</td>
</tr>
<tr>
<td>5. GCS &gt; 6 AND urgent surgery AND 24h urine &gt;= 750 ml AND urea &gt;= 10.8 mmol/L AND no mechanical ventilation 24h after adm</td>
<td>8.7</td>
<td>10 [7-13]</td>
<td>607</td>
<td>No</td>
<td>33 [28-38]</td>
<td>31 [22-41]</td>
<td>63</td>
<td>Yes</td>
</tr>
<tr>
<td>6. GCS &gt; 6 AND urgent surgery AND 24h urine &gt;= 750 ml AND urea &lt; 10.8 mmol/L AND mechanical ventilation 24h after adm AND bicarbonate &gt;= 22.6 mmol/L</td>
<td>4.1</td>
<td>4 [2-6]</td>
<td>607</td>
<td>No</td>
<td>34 [28-42]</td>
<td>49 [35-74]</td>
<td>23</td>
<td>Yes</td>
</tr>
<tr>
<td>7. GCS &gt; 6 AND urgent surgery AND 24h urine &gt;= 750 ml AND Systolic ABP &gt;= 72 mmHg AND mechanical ventilation 24h after adm AND bicarbonate &lt; 22.6 mmol/L</td>
<td>8.5</td>
<td>9 [6-11]</td>
<td>607</td>
<td>No</td>
<td>57 [52-62]</td>
<td>57 [45-69]</td>
<td>52</td>
<td>Yes</td>
</tr>
<tr>
<td>8. GCS &gt; 6 AND urgent surgery AND 24h urine &lt; 750 ml AND Systolic ABP &gt;= 72 mmHg AND pH &lt; 7.35</td>
<td>7.2</td>
<td>4 [3-6]</td>
<td>607</td>
<td>No</td>
<td>67 [58-75]</td>
<td>61 [40-77]</td>
<td>26</td>
<td>Yes</td>
</tr>
<tr>
<td>9. GCS &gt; 6 AND planned surgery AND 24h urine &lt; 1.25 L AND bicarbonate &lt; 16.7 mmol/L</td>
<td>1.1</td>
<td>1 [0-2]</td>
<td>607</td>
<td>No</td>
<td>78 [64-87]</td>
<td>59 [17-100]</td>
<td>6</td>
<td>Yes</td>
</tr>
<tr>
<td>10. GCS &gt; 6 AND urgent surgery AND 24h urine &lt; 750 ml AND Systolic ABP &lt; 72 mmHg</td>
<td>3.0</td>
<td>4 [2-6]</td>
<td>607</td>
<td>No</td>
<td>88 [81-92]</td>
<td>87 [77-96]</td>
<td>26</td>
<td>Yes</td>
</tr>
</tbody>
</table>

ABP = Arterial Blood Pressure; adm = admission; GCS = Glasgow Coma Scale. Orig. N = proportion % of patients classified to leaf in original study. Obs. N = proportion % of patients classified to leaf in this study; mean [range]. GS = Group size of SPC data points. Orig. P = predicted mortality % from original study; mean [inter quartile range]. Obs. P = observed mortality % in this study; mean [range].

* The group size was too low to reliably assess the stability of this leaf.
All leaves of the tree-model had an unstable proportion of patients over the course of time, but stable mortality means (Table 5.3; Appendix 5.2). Mortality means were stable in all three leaves corresponding to the three high-risk subgroups, and remained high in leaves 10 and 11 (G2 and G3 in Figure 5.2). Mean mortality in the combined high-risk subgroup in the temporal validation set was 83% (significantly higher than in the total sample of 32.0% (p < 0.01), but significantly lower than the mean estimated risk by the tree-model of 87% (p < 0.01) and stable (Figure 5.2). Sensitivity analysis (Table 5.2) shows that mean mortality in the combined high-risk subgroup was stable for each value of \( w \) between 0.2 and 0.5. For \( L = 2 \) and \( L = 2.5 \), however, instability was found in the lower values of \( w \) (data not shown).

**Figure 5.2.** Mortality in high-risk subgroup of the tree-model. There are no indications of instability in the process but note the wide ranges between the upper and lower limits. Each of the 20 (time) groups consists of about 6 patients in G1; 26 patients in G2; and 21 patients in leaf G3. In the combined graph, each group consists of about 53 patients. The plus-signs represent the original (non-smoothed) data points.

**Discussion**

Our work demonstrates the intricacy involved in validating prognostic models, in this case classification tree models, once the dimension of time is taken into account. In particular our findings show that, in a large temporal validation set, the performance of the tree-model in predicting hospital mortality in ICU patients aged 80 years and older is stable over time in terms of discriminative ability, but at the same time unstable in terms of
accuracy. Although the actual mean mortality was lower than predicted, mortality was stable and the tree-model was still able to identify subgroups of patients with high risk of hospital mortality. Mortality was stable in all three individual high-risk subgroups and also in the combined one, although the ranges between their upper and lower limits were very wide (0.18–1.00 for G1, 0.65–1.00 for G2 and 0.80–1.00 for G3). In terms of case-mix, we saw that the proportion of patients assigned to the tree’s leaves was not stable. This in itself is not indicative of model performance but of case-mix changes over time.

This study has the following strengths. It is a large study with real-world data obtained from many participating centers. The model refers to an already published model (and the only model in this population) with a natural starting moment of the temporal validation. We use both discrimination as well as accuracy performance measures (including a calibration element). While various studies internally validate tree-models [14–18], temporal [19] and external validations [14,20,21] are less often conducted [3]. One study assessed the stability of an existing tree-model in pneumonia patients (by comparing its structure to a newly developed model with the same predictor variables but on a different training set) [14]. In general, these studies used smaller datasets and more restrictive case-mix, and none of them monitored model behaviour over the course of time.

As far as we know, there is no published work of other groups using statistical process control (SPC) methodology for tracking predictive ability of prognostic models over the course of time. In a master’s thesis [26], SPC was used in combination with prognostic model performance in a neonatal ICU environment. Instead of validating model performance over time, however, the authors aimed to use SPC for risk adjustment to assess a model’s ability (in terms of the AUC) to recover from conditions in a manipulated sample (e.g. changing predictor or outcome values, deleting predictors from the model or introducing new ones). Moreover, we conducted sensitivity analysis with different values for $w$ and $L$ to assess the robustness of our choices, which showed the stability of the patterns found. Using different numbers of groups (in the range of 12–36) and fixed time intervals instead of fixed number of patients per group also yielded the same patterns (data not shown). Note that in cases unlike ours, where the number of patients may sharply fluctuate among time intervals and one uses the approach prospectively then the analysis should be adapted because the EWMA limits are based on a specific sample size while the obtained statistic (AUC or Brier) will be based on a different number of patients. One should use fixed time intervals because when there is a very large number of patients in one period it is not realistic to view them as associated with a sequence of points evolving over time: it just happened that we had many samples in one period reflecting the state of the process at that time. This time interval should be chosen to allow for some minimal number of patients, $M$, which is used to calculate the upper and lower control limits of the EWMA chart. For any fixed-time interval $i$ containing $S_i$ patients one can take $R_i$ (re-)samples of $M$ out of $S_i$ patients to calculate the statistic of interest ($AUC$ or Brier). A practical resampling strategy is to decide on the maximum tolerated overlap between resamples for calculating $R_i$. If this maximum is set to zero then all resamples are disjoint, for as much as $S_i$ allows for. In this strategy one uses the EWMA formulas with the variable number of observations, specifically $R_i$ observations (e.g. number of AUC values) in time interval $i$, as described in [36].

Our study has also limitations. We used one case study validating a classification tree. Therefore we were not able to compare our findings to other types of prognostic models.
(e.g. logistic regression) or even other tree-models. Recent work showing the abilities of probabilistic models [37] merits future work on the monitoring of these and other types of prediction models. We considered one strategy to use SPC, while it is not known which strategy (e.g. choice of number and size of groups, applied rules, type of chart and calculation of upper and lower limits) is the most reliable in our situation. However, our sensitivity analysis demonstrated the robustness of our results. Other SPC procedures, such as Cumulative Sum Chart (CUSUM), may be considered, but we chose EWMA because it intuitively appeals to the idea that the immediate past merits more weight than the distant one when one monitors change. Although we used a large number of participating centers, the generalizability of our findings to other (non-Dutch) ICUs may be limited. The lesson learned in this paper however pertains to behavior of prognostic models in general. Finally, in some leaves, the number of patients was insufficient to reliably assess the stability of the particular leaf.

A recent series of publications on prognostic models [3] stated that temporal and external validations of prognostic models are lacking in general and that this forms one of the main barriers for using prognostic models in practice. Our findings show on a large real-world dataset obtained from clinical practice the instability and variable predictive performance of the tree-model in terms of accuracy. This implies that the tree cannot be reliably used over time for benchmarking purposes and underlines the importance of continuous temporal validation.

However for the purpose of the identification of high-risk subgroups at least two subgroups of the three pre-identified subgroups could be used. Identification of high-risk subgroups might be useful for selecting patients in clinical trials’ arms or for comparing the size of these groups in case mix among ICUs. Tree-models have the advantage that these subgroups are explicitly defined by patient characteristics rather than solely by severity of illness scores, thus making the communication about prognosis more transparent.

This study also shows the added value of time series analyses in providing insights in “concept drifts”, i.e. the changes in the underlying data distribution [8,10]. We found instabilities in the proportion of patients falling in almost half of the leaves and in their outcomes. Many of these changes can probably be explained by changes over time in either healthcare policies or medical technologies (e.g. admittance policy to the ICU, medication prescriptions, and surgical procedures in older age). At least the SPC approach can allow for informed discussions about changes over time. Instead of lumping up a long period of time in one category, by using SPC one is able to scrutinize patterns in performance behavior.

This can also be used to help inspect the “shelf-life” of a model due to concept drift and identify the need for model recalibration or redevelopment [38,39]. Two general approaches may be followed to identify the need for recalibration. In the “reactive approach” one may check whether model performance is stable for the new data before applying the model (e.g. for quality assessment). If model performance is stable, one can continue using the original model, but if it is unstable the model should be recalibrated on the old data before its application for the new data. In contrast, the “pro-active approach” implies monitoring warning signs of performance over time. For example one may rely on variants of the common rules of instability of the general run-chart [27] by e.g. inspecting the number of points trending up or down, or the number of points at the same side of
the center line. In general, p-values associated with the EWMA points will not be exact (as their values are correlated), but the general strategy to re-calibrate in the face of such warnings is valid.

Future research should address questions regarding (a) how different types of prediction models (e.g. parametric probabilistic models) will fare when monitored over the course of time; (b) what would be the optimal strategy of using SPC for tracking a prognostic model’s performance over time (i.e. number of points, type of chart and calculation of upper and lower limits); (c) how to decide when a prognostic model has reached its expiry date and needs to be recalibrated; and (d) what is the best approach for recalibrating models (including the amount of data required for recalibration, weighting of recent and older data, and level of recalibration).

**Conclusions**

Our findings show that tree-models can show different contemporaneous patterns of behavior: they can have stable discrimination, instable accuracy and stable high risk means in high risk subgroups. This means that the adequacy of the models depends on their intended use. The instable accuracy indicates insufficient ability to react to a changed environment (e.g. changes in patient case mix). We stress the importance of validating a (tree-) model in new, independent data sets, preferably over the course of time. Statistical process control appeared useful for this purpose, as one is able to scrutinize patterns in performance behavior and concept drift, and thereby identify the need for model recalibration or redevelopment. Tree leaves provide a natural unit for close inspection, and we hence advocate assessing performance behavior, and incidence of subgroup membership, at the leaves over time, which may provide useful insights in concept drifts.

**Appendices**

Appendix 5.1 – Classification-tree model of the Rooij et. al (2007) (available online)
Appendix 5.2 – Stability of the tree-model’s leaves (available online)

**References**


