Data-driven methods to improve quality assessment of intensive care units
Koetsier, A.

Citation for published version (APA):
Koetsier, A. (2014). Data-driven methods to improve quality assessment of intensive care units

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.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Chapter 6: 
A Modified Real AdaBoost algorithm to discover Intensive Care Unit subgroups with a poor outcome

Antonie Koetsier, Nicolette F. de Keizer, Ameen Abu-Hanna, Niels Peek

6.1 Abstract

The Intensive Care Unit (ICU) population is heterogeneous. At individual ICUs, the quality of care may vary within subgroups. We investigate whether poor outcomes of an ICU can be traced back to excess deaths in specific patient subgroups, by discovering candidate subgroups, with a modified adaptive decision tree boosting algorithm applied to 80 Dutch ICUs. Genuine subgroups were selected from candidate subgroups when the case-mix adjusted outcomes were poorer than those of the five top performing ICUs. For 59 ICUs we discovered 122 genuine subgroups and most were defined by one to four variables, with a median of three [2-4]. Variables Glasgow Coma Scale and age were used most. There were 29 ICUs with overall poor outcomes, and for 22 our algorithm found all excess deaths. A new method based on adaptive decision tree boosting discovered many subgroups of ICU patients for which there is potentially room for outcomes improvement.
6.2 Introduction

One of the goals of many registries in health care is to assist in improving the quality of care [1]. These registries collect data from patients with a specific disease or treatment, and from multiple centers. A principal measure of quality, particularly in surgery and critical care, is patient mortality. However, the Intensive Care Unit (ICU) population is heterogeneous reflected by many different subgroups of patients such as surgical and non-surgical, acute and planned, old and young patients. Therefore it is necessary to adjust observed death rates by predicted mortality risk in order to make valid judgments about quality of care. Predicted mortality risks are typically obtained from a predictive model, such as the Acute Physiology and Chronic Health Evaluation IV (APACHE IV) [2] model for ICU patients. Risk-adjusted mortality rates are typically calculated for an entire hospital or ICU over a given time period thus reflecting the overall quality of care in that center or unit during that period. This approach does however not identify if there are specific subgroups of patients which are responsible for unexpected high mortality rates. Furthermore, low or intermediate overall mortality rates can falsely suggest that there are no subgroups of patients where quality is poor.

The field of machine learning and data mining has provided methods for 'boosting' the performance of predictive models, by searching for subgroups where model predictions systematically deviate from observed outcomes. One of the most popular methods is AdaBoost, short for Adaptive Boosting, developed by Freund and Shapire [3]. In this paper, we develop a modified Real AdaBoost algorithm for identifying subgroups of ICU patients whose outcomes are systematically worse than predicted by the APACHE IV model. While the original AdaBoost algorithm uses the identified subgroups to iteratively improve the predictive model, our method does not alter the APACHE IV model but reports the subgroups to clinical users as an impetus for quality improvement. In our algorithm, decision trees are used as the underlying machine learning method for subgroup discovery. To prevent false-positive warnings all subgroups are further selected using the Achievable Benchmarks of Care (ABC) method [4].

The objective of this paper is to investigate whether unexpected high mortality rates in individual ICUs can be traced back to excess deaths in specific patient subgroups. To this end, our algorithm was applied to a large dataset of ICU patients from the Dutch National Intensive Care Evaluation (NICE) registry, and the quality of the resulting subgroups were systematically analyzed.

6.3 Methods

In this section, we first discuss the dataset that was used in the analysis. Then we introduce the first step of our algorithm, the modified Real AdaBoost algorithm that was used to discover, for each ICU that contributed to the dataset, candidate subgroups with unexpected high mortality. Subsequently we describe the second step of our algorithm, where the ABC method was used to select from these candidates those subgroups with genuine higher mortality rates than are seen in other ICUs, referred to as genuine subgroups. We finally
describe the methods that were used to analyze the quality of the resulting genuine subgroups for this study.

6.3.1 Data

Data of the NICE registry from 2011 was used, consisting of all patients admitted to one of the 80 participating ICUs at that time. The NICE registry collects severity of illness data on the first 24 hours that patients are admitted to the ICU, as well as ICU and in-hospital mortality and length of stay. We used the APACHE IV model to calculate the predicted mortality risk for each patient, and therefore excluded patients according to APACHE IV’s exclusion criteria: length of stay less than four hours, readmitted patients, patients with burns, patients aged younger than 16, patients admitted after transplant operations, and patients coming from another ICU. Cardiac surgery patients were excluded because they have other diagnostic characteristics, are less heterogeneous, and only exist in a minority of the Dutch ICUs. All patient identification data, such as name and patient number, have been removed. In the Netherlands, there is no need to obtain consent to make use of registries without patient identifying information. The NICE registry is officially registered according to the Dutch Personal Data Protection Act.

Because the APACHE IV model was developed on different data, we recalibrated the model to our dataset by fitting a logistic regression model on observed mortality outcomes \( o_i \) \((o_i=0,1)\) using the logit-transformed APACHE IV mortality risk \( p_i \) as only covariate [5]. Due to the recalibration, overall observed and predicted mortality in the dataset are equal and therefore the ratio of observed and predicted mortality, the standardized mortality ratio (SMR), for the entire dataset is exactly equal to one. In general, an SMR above one reflects outcomes poorer than predicted while an SMR below that value is indicative for outcomes better than predicted.

6.3.2 Modified Real AdaBoost algorithm

Adaptive boosting [3] is an ensemble-based machine learning method in which different base models are constructed on the same dataset in a series of \( R \) 'rounds'. Each subsequent base model focuses on instances in the dataset that were not correctly predicted by the base models from previous rounds. Eventually, the resulting ensemble of \( R \) base models jointly makes predictions for new instances using weighted majority voting. In principle, any machine learning method can be used to construct the base models but 'weak' (i.e. unstable) learners are known to yield better results. A popular choice, which we will also use here, is decision tree learning [6]. With this choice, each base model effectively identifies subgroups of instances that were poorly predicted so far. That is, the weighted majority vote of all previous base models systematically deviates from observed outcomes within these subgroups.

In the first step of our algorithm, we applied adaptive decision tree boosting to 80 subsets of the data contributed by the participating ICUs. The resulting ensembles of base
models were discarded after round $R$ because we were not interested in building new models. Instead our algorithm returned all subgroups, identified in any round, for which observed mortality rates exceeded those that were predicted by the recalibrated APACHE IV model. Because this model predicts probabilities rather than classes (i.e., survival outcomes), we used the Real AdaBoost algorithm [7], which is a generalization of the original AdaBoost algorithm for real-valued predictions. Starting off initially with weights defined by the APACHE IV predictions, the Real AdaBoost algorithm iteratively reweighs all instances in the dataset in order to focus the decision tree learner in subsequent rounds on regions where predictions deviate from observed outcomes. Eventually, we obtained for each ICU a set of candidate subgroups with unexpected poor outcomes.

In the first round the weights were calculated as follows for instances $i = 1, \ldots, N$:

$$w_i^1 = \begin{cases} \frac{pl_i}{1-p_i} & \text{for } y_i = 0 \\ \frac{1-p_i}{p_i} & \text{for } y_i = 1 \end{cases}$$

(1)

and normalize such that $\sum_{i=1}^{N} w_i^1 = 1$

That is, weights were high for instances $i$ with large residuals $|y_i - p_i|$ and low for instances with small residuals. Let $S_1, \ldots, S_k$ be the set of candidate subgroups defined by the leaf nodes of the tree that resulted from the weighted decision tree analysis, and $q_1, \ldots, q_k$ be the corresponding class probabilities predicted by the tree. Each candidate subgroup $S_j$ is defined by the conjunction of conditions which are associated with the path from the root of the tree to the leaf node. For each candidate subgroup $S_j$, $\text{SMR}(S_j)$ was calculated to determine whether the outcomes of the ICU were poor for that subgroup, reflected by a ratio above one.

The SMR for a given candidate subgroup $S_j$ of patients is defined as:

$$\text{SMR}(S_j) = \frac{\sum_{s \in S_j} o_s}{\sum_{s \in S_j} p_s}$$

(2)

where $o_s$ and $p_s$ are the observed and predicted mortality outcomes in candidate subgroup $S_j$.

In subsequent rounds, weights were adjusted only for candidate subgroups of patients where the outcomes of the ICU was poor (e.g. $\text{SMR}(S_j)$ above one), to prevent that the same subgroups were found over and again, as follows:

$$w_i^m = w_i^{m-1} \times \begin{cases} \frac{q_i}{1-q_i} & \text{for } y_i = 0 \\ \frac{1-q_i}{q_i} & \text{for } y_i = 1 \end{cases}, \quad m = 2, \ldots, R$$

(3)
For the other candidate subgroups, the weight remained equal to that of the previous round. As before, the weights were normalized prior to constructing the decision tree.

We performed ten iterations ($R=10$) of adaptive boosting for each of the 80 ICUs. For each decision tree the minimum number of instances per node was set to 20 and the maximum depth of the tree was 20.

### 6.3.3 ABC Method

Due to chance or inaccurate APACHE IV model predictions, there is a possibility that the candidate subgroups found by the first step do not describe true poor outcomes of the ICUs. Therefore, in the second step of our algorithm, we used the ABC method [4] for filtering candidate subgroups of patients. Thus, candidate subgroups of patients were removed when there was insufficient evidence that the outcomes of the ICUs for the subgroups were poor compared to the best performing ICUs for that same subgroup, resulting in a new set of subgroups referred to as genuine subgroups.

For each candidate subgroup, we have the SMR($S_j$), 95% confidence interval (CI), and the corresponding ICU for which the subgroup was found. To be able to remove candidate subgroups, we first selected all the patients in the dataset, containing data from all ICUs, which adhered to the subgroup description $S_j$ resulting in the subgroup dataset. Secondly, for all ICUs in the subgroup dataset, the SMR was calculated and sorted from lowest to highest based on the value of the SMR. Top performing ICUs were selected based on the following conditions: (1) consist of at least five unique ICUs, (2) and cover at least ten percent of the subgroup dataset. The data of the top performing ICUs was aggregated into a new dataset, and accordingly SMR($S_j,abc$) and 95% CI was calculated. In the last step we compared the SMR($S_j,abc$) and 95% CI with the SMR($S_j$) and 95% CI of the ICU for which the candidate subgroup $S_j$ was initially found by the modified Real Adaboost algorithm. If the ICU’s CI of SMR($S_j$) crossed or was below the CI of SMR($S_j,abc$) we removed the candidate subgroup of the ICU. In all other situations the candidate subgroup of the ICU was not removed, as the conclusion was that the outcomes were poor for the candidate subgroup.

### 6.3.4 Quality of the genuine subgroups

To further investigate the subgroups found, we recorded per ICU the number of candidate subgroups, number of genuine subgroups and per genuine subgroup the number and type of variables describing the demographic and diagnostic characteristics of each patient. Each ICU’s overall outcome was categorized as good (95% CI of overall SMR below one), intermediate (95% CI of overall SMR crossing one), or poor (95% CI of overall SMR above one). The complexity of the genuine subgroups were assessed by determining the number of variables used for each subgroup, which variables occur the most, and describing the five genuine subgroups with the poorest outcomes. Furthermore, the percentage overlap
of patients between different genuine subgroups within ICUs was calculated with the ratio of the number of patients \( n_p \) existing in all the different genuine subgroups and total number of patients of the ICU:

\[
\left( \frac{\sum_{S_j=k}^{S_j} n_p}{\sum n_p} - 1 \right) \times 100
\]  

(4)

The overlap is a measure indicating whether each genuine subgroup consists of unique patients for the ICU, or whether some genuine subgroups are subsets of other genuine subgroups. We also calculated how many excess deaths there were for all the patients \( i = 1, \ldots, L \) of each ICU in the dataset and the number of excess deaths found for all genuine subgroups combined for each ICU. Excess deaths per ICU in the dataset were calculated by:

\[
\sum_{i=1}^{L} o_i - \sum_{i=1}^{L} p_i
\]  

(5)

And excess deaths for each genuine subgroup \( S_j \) per ICU were calculated by:

\[
\sum_{S_j=k}^{S_j} o_i - \sum_{S_j=1}^{S_j=k} p_i
\]  

(6)

In the ideal situation, we would find at least the same number of excess deaths in the genuine subgroups for all ICUs combined from our algorithm as in the excess deaths for all ICUs combined from the original dataset.

6.4 Results

Our recalibrated APACHE IV model had an Area Under the Curve (AUC) of 0.88 and the \( R^2 \)-value was 0.43. With the use of our modified Real AdaBoost algorithm, 65 of the 80 ICUs had at least one candidate subgroup of patients with an unexpected poor outcome. A total of 247 candidate subgroups were found. After applying the ABC method 122 genuine subgroups remained where the ICU’s outcome for that subgroup was poor, comprising 59 ICUs. Of the 59 ICUs, there was a median of two genuine subgroups with a range of 1-6. SMR(\( S_j \)) of the genuine subgroups ranged from 1.0-4.3 whereas for the top five performing ICUs the range was 0.0-0.9. All genuine subgroups combined consisted of 21464 unique patients (44.3% of all national ICU patients) and 3915 unique deaths (43.7% of all national ICU deaths). At least one genuine subgroup was found for one ICU with overall good, 48 with intermediate, and ten with poor outcomes. Table 1 shows an overview of the overall outcomes of the ICUs and whether genuine subgroups were found. The 21 ICUs where no genuine subgroups were found consisted of eight ICUs with good, twelve with intermediate, and one with poor outcomes.
Table 1. Overall outcomes of ICUs and whether genuine subgroups were found.

<table>
<thead>
<tr>
<th>ICU outcomes</th>
<th>Number of ICUs</th>
<th>Number of ICUs with genuine subgroup(s) found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Intermediate</td>
<td>60</td>
<td>48</td>
</tr>
<tr>
<td>Poor</td>
<td>11</td>
<td>10</td>
</tr>
</tbody>
</table>

Most genuine subgroups used one (13.9%), two (29.5%), three (24.6%), or four (20.5%) variables to describe the demographic and diagnostic characteristics of the patients, with a median of three [2-4]. The variable age, lowest Glasgow Coma Scale in first 24 hours of ICU admission (GCS, a scale for neurological functioning where a value of 15 is the best outcome), and admission type (medical, emergency surgery, or planned surgery) occurred the most in the genuine subgroups (30.4%, 29.0%, and 7.3% respectively). For 42 of the genuine subgroups, the GCS was above 10 points. The most common range in age was between 47 and 76. Admission type varied evenly among the genuine subgroups. Other variables, such as chronic or acute diagnosis, occurred in five percent or less of the genuine subgroups. Table 2 shows the five genuine subgroups which had the highest SMR, and with the 95% CI above one. Four of the five genuine subgroups belonged to ICUs with overall intermediate outcomes (ICU A, B, and D), and one subgroup was from an ICU with overall poor outcomes (ICU C). All five genuine subgroups used the GCS and age.
<table>
<thead>
<tr>
<th>ICU</th>
<th>Number of patients in 2011</th>
<th>Genuine subgroup description</th>
<th>Number of patients subgroup</th>
<th>Number of deaths in subgroup</th>
<th>Genuine SMR [95%CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1094</td>
<td>Lowest GCS in first 24h ≥ 7 AND Age ≥ 45 AND Chronic Renal Insufficiency = False AND Neoplasm = False AND Confirmed Infection = False AND Chronic Respiratory Insufficiency = True</td>
<td>31</td>
<td>7</td>
<td>4.4 [1.7-8.6]</td>
</tr>
<tr>
<td>B</td>
<td>497</td>
<td>Lowest GCS in first 24h ≥ 13 AND Age ≥ 67 AND Planned ICU Admission = False AND COPD = True</td>
<td>27</td>
<td>8</td>
<td>3.3 [1.4-6.2]</td>
</tr>
<tr>
<td>C</td>
<td>1121</td>
<td>Lowest GCS in first 24h ≥ 14 AND Intracranial Mass Effect = True</td>
<td>55</td>
<td>12</td>
<td>3.3 [1.7-5.9]</td>
</tr>
<tr>
<td>B</td>
<td>497</td>
<td>Lowest GCS in first 24h ≥ 13 AND Age ≥ 44 AND Age &lt; 66 AND Vasoactive Drugs = False AND Gender = Male</td>
<td>75</td>
<td>6</td>
<td>3.1 [1.1-6.3]</td>
</tr>
<tr>
<td>D</td>
<td>281</td>
<td>Lowest GCS in first 24 hours = 15 AND Age ≥ 57 AND Admission Type = Medical AND Gender = Female</td>
<td>41</td>
<td>10</td>
<td>3.0 [1.4-5.3]</td>
</tr>
</tbody>
</table>
For 18 ICUs we found patients overlapping between the different genuine subgroups. The overlap ranged from 5.8% to 219.4% with a mean of 54.1% and a median of 32.5% implying that there were 32.5% more patients in the genuine subgroups than in the original ICU data. Thus, there are genuine subgroups which are subsets of other more general genuine subgroups. In 22 of the 29 ICUs with excess deaths, at least the same number of excess deaths were reflected in the genuine subgroups. For the seven other ICUs, the percentage of excess deaths found by the genuine subgroups was between 80.1% - 96.3%.

6.5 Discussion

Our paper describes a modified Real AdaBoost algorithm to detect genuine subgroups of patients where ICU performance is poor. Additionally we investigated the suitability of our algorithm for use in ICU practice by describing the genuine subgroups found and how many excess deaths in the ICU population were reflected by the genuine subgroups. The modified Real AdaBoost algorithm detected genuine subgroups in the majority of the ICUs. For ten ICUs of the eleven with poor outcomes genuine subgroups were found. Furthermore, in most cases the total number of excess deaths in all ICUs were also found by the genuine subgroups. This suggest that we found genuine subgroups of patients which were responsible for the majority of the total excess deaths in the ICU population.

In our algorithm the possibility exists that the APACHE IV model systematically over- or underestimates the mortality for some candidate subgroups. However, candidate subgroups with underestimated predictions found in our algorithm will be removed by subsequently performing the ABC method. Due to the good AUC and R^2-value value of the APACHE IV model we expect that the number of times this occurred was kept to a minimum. However, our modified Real AdaBoost algorithm will perform worse when using predictive models which frequently over- or underestimate the mortality predictions, e.g. have a lower AUC and R^2-value. This leads to many false positive candidate subgroups found, which are subsequently removed by the ABC method ultimately leading to no genuine subgroups found. To overcome this, the number of R rounds can be increased. Another possibility is to first apply the adaptive boosting algorithm to the whole study population, leading to new more accurate predictions, and then performing subgroup discovery for individual ICUs.

A limitation is the possible unreliability of the data. A variable such as the GCS is a subjective variable, meaning that ICUs may differ in how they score this. Therefore, we may have found GCS in many of the genuine subgroups because of these scoring differences. This may also account for other variables thereby influencing the subgroups found. However, data quality of the NICE registry is frequently assessed using audits and errors are comparable with other registries [8].

Another limitation is that there is some overlap between genuine subgroups meaning that specific genuine subgroups, with large number of descriptive variables, are subsets of other more general genuine subgroups with few descriptive variables. Future research should aim at improving the algorithm by limiting this overlap.
A last limitation are the limits or boundaries we set in our modified Real AdaBoost algorithm. We only adjusted weights for candidate subgroups where the SMR was above one. Lowering this limit may result in more candidate subgroups detected but an increase in false positive candidate subgroups is also possible. Candidate subgroups were only selected as genuine when the subgroup’s lower CI was above the upper CI of the top five ICUs. Candidate subgroups not selected do not necessarily imply that the performance of the ICUs in reality was not poor for these groups. Furthermore, decision trees were chosen as machine learning method, whereas other methods such as the Patient Rule Induction Method [9] may yield other results.

Our algorithm based on decision tree boosting discovered many genuine subgroups of ICU patients, for which there is potentially room for outcomes improvement. The resulting genuine subgroups can aid ICUs and management in setting up tailored quality improvement initiatives.

6.6 References


