Clinical decision support: distance-based, and subgroup-discovery methods in intensive care

Nannings, B.

Citation for published version (APA):
Chapter 2. _Applying PRIM (Patient Rule Induction Method) and Logistic Regression for Selecting High-Risk Subgroups in Very Elderly ICU Patients._


Barry Nannings, Ameen Abu-Hanna, Evert de Jonge
2.1. Abstract

2.1.1 Purpose
To apply the Patient Rule Induction Method (PRIM) to identify very elderly Intensive Care (IC) patients at high risk of mortality, and compare the results with those of a conventional logistic regression model.

2.1.2 Methods
A database containing all 12,993 consecutive admissions of patients aged at least 80 between January 1997 and October 2005 from intensive care units (n=33) of mixed type taking part in the National Intensive Care Evaluation (NICE) registry. Demographic, diagnostic, physiologic, laboratory, discharge and prognostic score data were collected. After application of the SAPS II inclusion criteria, 6,617 patients remained. In this data we searched PRIM subgroups requiring at least 85% mortality and coverage of at least 3% of the patients. Equally-sized subgroups were derived from a recalibrated (second level customization) Simplified Acute Physiology Score II model. Subgroups were compared on an independent validation set using the Positive Predictive Value (PPV), equaling the subgroup mean mortality.

2.1.3 Results
We identified four subgroups with a positive predictive value (PPV) of 92%, 90%, 87% and 87%, covering respectively 3%, 3.5%, 7% and 10% of the patients in the validation set. Urine production, lowest pH, lowest systolic blood pressure, mechanical ventilation, all measured within 24 hours after admission, and admission type and Glasgow Coma Score were used to define these subgroups. SAPS and PRIM subgroups had equal PPVs.

2.1.4 Conclusions
PRIM successfully identified high-risk subgroups. The subgroups compare in performance to SAPS II, but require less data to collect, result in more homogenous groups and are likely to be more useful for decision makers.
2.2. Introduction

Aging of the population has increased the proportion of very elderly (80+) patients being admitted to the ICU. These patients form an important group with high resource usage and a relatively low probability of survival [1]. However, old age alone is not a good predictor for patient survival [2-6]. It is important to discern subgroups within this population with very high chances of not surviving IC treatment.

There are various reasons for seeking such groups. First, subgroups may reveal determinants that provide insight into the patient subpopulations. Some of these determinants may be risk factors that can be acted upon. Second, much research on the efficacy and efficiency of therapeutic interventions relies on the enrolment of high-risk patients to maximize the likelihood of finding a beneficial treatment effect. Third, the groups can be used to improve case-mix adjustments in order to better compare the quality of care of different ICUs. Fourth, information about the patient’s probability of survival can be communicated with the patients and their families. Lastly, such information can support informed decisions about (withholding) treatment e.g. when the expected quality of life is very low and the therapy the patient is receiving is very aggressive. This is especially relevant for the very elderly. It should be noted, however, that the unconditional use of models for this latter reason has raised much resistance in the intensive care community [7].

The most commonly used models in IC for predicting hospital mortality include the Acute Physiology and Chronic Health Evaluation (APACHE) II, III and IV [8], and the Simplified Acute Physiology Score (SAPS) II and III [9-10]. These are parametric models that rely on severity of illness scores: the higher the score, the higher the associated mortality. The scores are based on demographic and diagnostic information, and also on physiological data from the first 24 hours after ICU admittance. Although these models were originally designed for case-mix adjustments, they have also been used for high risk-group detection, e.g. in [11]. A disadvantage of using these models for subgroup identification is that the subgroups are not homogeneous in terms of patient characteristics and hence provide less insight into the makeup of the patient risk groups.

In this paper we apply a relatively new non-parametric method for subgroup discovery called the Patient Rule Induction Method (PRIM) [12] for the identification of subgroups at very high risk of dying. We compare it to the SAPS II conventional parametric logistic regression model. PRIM was chosen because it was designed to work with high dimensional data, is parsimonious with data, handles missing values in a non adhoc manner, is based on solid statistical ideas, and has a computer implementation available to the public. We compared the PRIM subgroups with subgroups derived using SAPS II, as SAPS II is the prognostic model of preference of NICE. We also made use of APACHE II, but only to categorize our continuous variables into scores, that were also used as input for PRIM sub-group discovery. APACHE II was chosen for this because it covered most of the variables we needed to categorize.
It should be noted that subgroups that are discovered using PRIM are always specific for the data used to find them. The factors used to define a subgroup are thus specific for the specific ICU’s tools, staff etc. from which data have been obtained. As an example, consider a staff pre-conception that a certain patient will not be saved and a decision might be made to stop treatment. This group of patients might be recognized as a high-risk group by PRIM and as such, will be a self-fulfilling prophecy. Of course this is true for most research concerning prognostic models. The problem can be partly alleviated when the data is a good reflection of the total population.

To our knowledge this is the first time that PRIM is applied within our domain and the results are therefore also of theoretical interest. In this paper we compare PRIM to a logistic regression model. Below we provide preliminaries required for understanding these approaches.

2.2.1 Subgroup discovery

Subgroup discovery [13-14] aims at finding patterns, corresponding to subgroups with interesting properties, in the data. This is in contrast to developing a global model, such as a classification tree or logistic regression model, aiming at a global good performance. Subgroup discovery approaches can be characterized by the type of the target variable and covariates, subgroup description language, subgroup quality function, and search strategy. Algorithms originating from the Machine Learning and Data Mining literature tend to focus on discrete variables. These algorithms use search heuristics and they often employ a beam search to mitigate the consequences of greedy choices. The PRIM algorithm is an example of a subgroup discovery algorithm.

2.2.2 Patient Rule Induction Method (PRIM)

The Patient Rule Induction Method suggested by Friedman and Fisher [12] is referred to as a “bump-hunting” algorithm. Bump-hunting algorithms are used to find regions in the input variable space (or covariate space) that are associated with a relatively high or low mean value for the outcome. This is unlike regression models, which seek to model the whole population by optimizing a likelihood function or a human function such as patient utility. A region is described by conjunctive conditions using the input variables and is associated with the mean value of the output in that region. These rules have the following form:

\[
\text{If condition}_1 \text{ and } \ldots \text{ and condition}_n, \text{ then predicted mean outcome value.}
\]

These conditions can use numeric (e.g. age) or categorical (admission type) attributes. For continuous attributes a condition will have the following form:

\[
\begin{align*}
&\text{variable < value, or} \\
&\text{variable > value, or} \\
&\text{value}_1 < \text{variable} < \text{value}_2
\end{align*}
\]

For categorical attributes, conditions have the following form:
variable = value
variable = value, or … or value_m

A rule defined using such conditions corresponds to a hypercube in the input variable space and is often called a box. It will be a simple rectangle in two-dimensional space.

Rules discovered using PRIM can be applied to a new dataset. To validate a rule one could compare the expected mean associated with the rule to the observed mean on a validation set.

**PRIM Rule induction**

When many input variables are considered, it is not feasible to consider all possible rules in order to choose the best one. Hence, PRIM uses heuristics to constrain the search for the rules. PRIM starts with a box containing all given observations. For each continuous variable it considers removing ("peeling") a small portion of observations with the highest and, separately, lowest values of the variable. For example if the dataset consists of the attributes age and height then PRIM will consider 4 operations corresponding to removing the observations with the highest and lowest values of each variable. It chooses the peel that results in the remaining box with the highest outcome mean. In this research we are not interested in finding regions with a low outcome value, but to achieve this one would simply have to inverse the outcome and perform the same analysis. The other candidate peels are discarded. The process is reiterated on the obtained sub-box until no additional peels seem to improve the outcome mean or until a resulting sub-box would include too few observations, where this minimum threshold is specified by the analyst.

For continuous variables the amount of data to be removed in each peel can be controlled by the data analyst and is specified as a percentage (alpha), usually 5%, of the observations in the current box. Choosing a high alpha risks missing an optimal box: in each iteration, PRIM makes a choice to remove a big chunk of data based only on one variable in that iteration. If this choice is not the optimal one, then PRIM may not be able to recover from this mistake. Choosing a small alpha makes PRIM more "patient": it will need more steps to arrive at an answer but it is much less at risk to get a suboptimal result. For categorical attributes, PRIM considers removing observations corresponding to one value of the variable at a time.

The final box after peeling may not be optimal because of past greedy suboptimal choices. PRIM aims to recover from these mistakes by trying to expand the box in a process inverse to ‘peeling’ called ‘pasting’, in which the box is iteratively enlarged as long as the outcome’s mean increases. The result of peeling and pasting is a sequence of boxes, consisting of all the boxes obtained in the process: from the initial box containing all the data to the box that is obtained after pasting.
As any non-parametric algorithm that learns from data, the boxes derived with PRIM may overfit the data. To avoid overfitting, PRIM uses cross-validation: it reports the mean for each obtained box not only on the data that was used to derive the box but also on a held-out set obtained from the developmental set itself, and is thus not part of the independent testset. A significant difference in the outcome mean on the held-out set usually indicates overfitting and the analyst is advised not to trust such boxes.

When a box is finally chosen and noted, its associated observations are removed and the search for a new box can be started by repeating the whole peeling and pasting process in the remaining data. Sub-boxes are always conditioned on those obtained earlier: to estimate a mean outcome of a box, one should first remove the data corresponding to the earlier boxes.

PRIM provides a number of tools to post-process the rules that were discovered, such as the removal of redundant variables, assessment of inter-box dissimilarity and plotting relative frequency ratio plots, but these are outside the scope of this paper. The interested reader is referred to [12].

2.2.3 Related work

Besides PRIM, other subgroup discovery algorithms exist. The Data Surveyor algorithm for subgroup discovery by Holsheimer et al. [15] considers one variable at a time and seeks the value interval having the highest target mean. Directly targeting the (at that iteration) optimal interval can potentially make it much more greedy than PRIM, as a final box can be reached after only very few iterations. A subgroup is expressed as a conjunction of interval-based constraints. The CN2-SD [16] algorithm resembles Data Surveyor in the subgroup description language and the search strategy. It is an adaptation of the CN2 classification rule learner to subgroup discovery. The algorithm develops constraints on the value ranges of variables and uses a quality function which is a tradeoff between the generality of the rules and the relative accuracy of the rules. CN2 requires both the outcome and the covariate variables to be discrete. The SD-Map algorithm [17] is an extension of the FP-tree algorithm (frequent pattern discovery) for subgroup discovery. It is efficient because it bypasses the generate-and-test hypotheses cycle. It is one of the few subgroup discovery algorithms explicitly dealing with missing data. However, it only works with discrete attributes (covariates and target variable).

We chose to use PRIM, as opposed to the other algorithms, because we valued its patience and its ability to deal with continuous attributes, as well as it being publicly available. The following section introduces logistic regression models.

2.2.4 Logistic regression models in Intensive Care

A logistic regression model (LRM) is a probabilistic parametric model. For a given set of covariate values, the model predicts the probability of a binary outcome variable $Y$. $Y=1$ indicates the occurrence of the event, such as death. The model has the following form:
where

\[ p(Y = 1|x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \]

and is linear in the \( \beta_i \) coefficients. LRMs are used in most IC predictive models where \( x \) commonly includes one or more severity of illness scores and sometimes also diagnostic categories. For example the logit of the SAPS II model is:

\[ -7.7631 + 0.0737 \cdot \text{SAPS} + 0.9971 \cdot \ln(\text{SAPS} + 1) \]

where SAPS quantifies the severity of illness score (the higher the score, the worse the patient’s condition).

One reason for the popularity of the LRM is the interpretation that is given to a covariate coefficient \( \beta_i \) in terms of odds ratio. For an event with probability \( p \) its odds are \( p/(1 - p) \). The odds ratio is defined as the ratio of the odds of an event occurring in one group (e.g. smokers) to the odds of it occurring in another group (e.g. non-smokers). For a binary covariate with coefficient \( \beta_i \), \( e^{\beta_i} \) turns out to be equal to the odds ratio of the groups that the covariate defines. For a continuous variable such as SAPS the quantity \( e^{\beta_i} \) is equal to the odds ratio of a group of individuals having a SAPS of one unit more than the other group.

### 2.3. Materials and methods

The Dutch National Intensive Care Evaluation (NICE) comprises a continuous and complete registry of all patients admitted to the intensive care units (ICUs) of the participating hospitals in the Netherlands. This NICE is not to be confused with the (British) National Institute for Health and Clinical Excellence. The data used in this study consisted of all 12,993 consecutive admissions of patients 80 years and older between January 1997 and October 2005. The data originated from all 33 adult ICUs, of mixed type, that were participating in NICE when the research project was initiated (January 2004). To facilitate comparison with the SAPS II model we applied the SAPS II exclusion criteria: no readmissions, no cardio-surgical patients, and no patients with burns, resulting in 6,617 patients. The dataset was split randomly in a developmental set containing 66% of the patients and a validation set containing the rest. Fig. 1 shows the number of patients in the exclusion, inclusion, developmental, and validation sets. Details concerning the quality of the data used in this study were published elsewhere [18].
Figure 1. Flowchart showing the number of patients in the exclusion, inclusion, developmental, and validation sets.

The database included the following variables, all related to the first 24 hours of stay: age, gender, length, weight, Body Mass Index (BMI), admission type (medical, scheduled, unscheduled), cardiopulmonary resuscitation, gastrointestinal bleeding, intracranial mass effect, dysrhythmia, cerebrovascular accident, acute renal failure at admission to the ICU, chronic renal insufficiency, chronic dialysis, metastasized cancer, aids, hematological malignancy, cirrhosis of the liver, cardiovascular insufficiency, respiratory insufficiency, immunological insufficiency, confirmed infection, burns, sepsis, mechanical ventilation at 0/24 hours, Glasgow Coma Score (GCS) and sub-scores, urine, vasoactive drugs, arterial partial oxygen pressure (PaO2), fraction inhaled oxygen (FiO2), arterial CO2 pressure (PaCO2), PaO2/FiO2 ratio, alveolar-arterial oxygen difference (AaDO2), prothrombin time, urea, bilirubin, severity of illness score (SAPS II), predicted mortality probability (SAPS II); lowest and highest value of respiratory rate, blood pressure, temperature, white blood cell count, creatinin, potassium, sodium, bicarbonate, hematocrit, albumin, glucose; the admission, lowest, highest value of heart rate and systolic blood pressure, the lowest pH value, and ICU- and hospital mortality. Description of these variables can be found on the NICE website (unpublished data, http://www.stichting-nice.nl).
### Table 1. Description of the patient population. Data are reported as interquartile range (median).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Developmental set (n = 4413)</th>
<th>Validation set (n = 2204)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yrs</td>
<td>81-85 (83)</td>
<td>81-86 (83)</td>
</tr>
<tr>
<td>Admission type, %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>46.0</td>
<td>44.2</td>
</tr>
<tr>
<td>Surgical unscheduled</td>
<td>23.2</td>
<td>21.8</td>
</tr>
<tr>
<td>Surgical scheduled</td>
<td>30.7</td>
<td>34.0</td>
</tr>
<tr>
<td>Male, %</td>
<td>46.5</td>
<td>45.3</td>
</tr>
<tr>
<td>SAPS II Score</td>
<td>30-53 (40)</td>
<td>30-53 (40)</td>
</tr>
<tr>
<td>APACHE II Score</td>
<td>14-23 (18)</td>
<td>14-23 (17)</td>
</tr>
<tr>
<td>GCS 24 hrs after admission</td>
<td>15-15 (15)</td>
<td>15-15 (15)</td>
</tr>
<tr>
<td>CPR, %</td>
<td>8.3</td>
<td>7.7</td>
</tr>
<tr>
<td>ICU LOS</td>
<td>0.8-3.7 (1.3)</td>
<td>0.7-3.5 (1.2)</td>
</tr>
<tr>
<td>ICU mortality, %</td>
<td>20.4</td>
<td>21.1</td>
</tr>
<tr>
<td>Hospital LOS</td>
<td>6.2-28 (14)</td>
<td>7.0-27.0 (14)</td>
</tr>
<tr>
<td>Hospital mortality, %</td>
<td>34.5</td>
<td>34.5</td>
</tr>
</tbody>
</table>

Interquartile range is the range between the 25\(^{th}\) to 75\(^{th}\) percentile. SAPS = Simplified Acute Physiology Score, APACHE = Acute Physiology And Chronic Health Evaluation, GCS = Glasgow Coma Score, CPR = Cardiopulmonary resuscitation, ICU = Intensive Care Unit, LOS = Length Of Stay.

PRIM considers only conjunctive rules on continuous variables and hence cannot generate a condition using disjunctions, such as "blood pressure > 90 or heart rate > 110" nor on the same continuous variable "blood pressure < 70 or > 90". However, the latter type of conditions represents a relevant variable-outcome relationship in which a low and a high value of a variable, such as body temperature or blood pressure, are associated with a high risk. PRIM can in principle discover two high risk subgroups in different runs, one for the low and one for the high values of the variable, but this would be unintuitive. To capture such a covariate-outcome relationship in a single rule we also include severity of illness scores associated with each continuous variable. Such a score will receive a high value for low as well as for high values of the variable under consideration. The scores were obtained by applying the APACHE III scoring scheme [8] because it covers most used variables and discerns relatively many score values. Variables in our data that were not included in the APACHE III scoring scheme were scored according to the APACHE II or SAPS II schemes, in this order. Following common practice, we scored missing values as 0 (i.e. the value is assumed to be normal in the normal range). An example of a rule that PRIM can discover using a continuous variable that is scored using the APACHE III scoring scheme is: if APACHE3 heart rate score > 15, then predicted mortality is 0.80. It should be noted that this means the actual heart rate would be equal to or higher than 155 beats per minute. Scores are used in addition to the original (continuous) variables. This means that both continuous variables as well as their scored counterparts can be part of the same subgroup definition.
To make the comparison between PRIM and the SAPS II logistic regression model we recalibrated the SAPS II model on our developmental set using second-level customization (the coefficients of the model are fitted anew) [19-20].

In this study we used the SuperGEM™ 1.0 software that implements PRIM (unpublished data: http://www-stat.stanford.edu/~jhf/SuperGEM.html). Using the developmental set, we searched for the largest subgroups having (a mean of) at least 85% hospital mortality on the developmental set and the held-out set. Furthermore, we required that each subgroup should include at least 3% of the patients in the training set. To allow for alternative overlapping subgroups, we applied PRIM with different parameter settings, each time starting with the whole developmental set. Searching for new subgroups was stopped when the total number of unique patients covered by the subgroups approached our pre-determined threshold of 10% of the population in the developmental set. For comparison, each PRIM group was compared to an equally sized group containing patients with the highest SAPS II scores and consequently, the highest SAPS II predicted mortality.

2.3.1 Statistics

We calculated the positive predictive value (PPV) of the PRIM and corresponding (equally-sized) SAPS II subgroups on an independent validation set. For each subgroup we constructed 1000 bootstrap samples to calculate the 95% Confidence Interval (CI) for the difference between the PPVs obtained by PRIM and SAPS II. Statistical analysis was performed with S-PLUS® 6.2 (Insightful, Seattle, WA). Data are reported as interquartile range and median. The level of significance was set at p < 0.05.

2.4. Results

Using PRIM we found three subgroups in the developmental set, that we refer to as A, B and C. Subgroup A is defined as patients having:

- 24 hour urine production < 0.83 l
- mechanical ventilation at 24 hours after admission
- lowest systolic blood pressure during the first 24 hours < 75 mmHg
- lowest pH during the first 24 hours < 7.3 and
- medical or unscheduled surgical reason for admission.

The mean outcome for group A on the developmental set was: 94.8%. Subgroup B is defined as patients having:

- lowest systolic blood pressure during the first 24 hours < 70 mmHg
- 24 hour urine production < 0.9 l and
- lowest pH value during the first 24 hours < 7.3 or > 7.6.

The mean outcome for group B on the developmental set was: 91.2%.
Subgroup C is defined as patients having a Glasgow Coma Score < 5. It is associated with mean outcome on the developmental set of 86.6%.

Observe that in Subgroup A, only the original continuous variables turned out to be selected in the definition although both the scores and original variables were available for use. In the definition of Subgroup B, scores of continuous variables were selected, which have been converted back to their approximate original values, as reported above, for readability.

Table 2 provides a description of Subgroups A, B, and C on the validation set in terms of coverage, group makeup and performance. The table also describes a new subgroup obtained by the union of patients covered by Subgroups A, B and C.

The subgroups of PRIM and SAPS II all have a high PPV in the validation set (Table 2). Note that PPV (the proportion of the event within a subgroup) is equivalent to the hospital mortality mean. It is quite coincidental that the mortality means of the PRIM subgroups turned out to be equal to the means in their corresponding SAPS II groups. However, as can be seen in the table, the PRIM and corresponding SAPS II subgroups only partially overlap and hence consist of different patients. Slightly changing the definition of a subgroup would lead to non-identical results. For example, if we would have used the value 0.7 l instead of 0.83 l in the “24 hour urine production” condition in PRIM subgroup A, then it would have lead to a mean mortality of 0.91 and 0.93 for the PRIM and SAPS II subgroups respectively.

Combining the patients contained in any of the three subgroups in one composite group also results in a high PPV while at the same time including considerably more patients than the individual subgroups. This means that although the subgroups may overlap (as a single patient can belong to multiple subgroups), they still differ sufficiently to provide added value when combined. PRIM and SAPS II consider different patients as the highest risk patients, as seen by the low overlap between Subgroups A and B and the corresponding SAPS II subgroups.

The difference in PPV between the PRIM and corresponding SAPS II subgroups was not statistically significant (95% confidence interval -0.015 – 0.076). Although the patients with the highest SAPS II model probability are indeed at high risk, the original (un-recalibrated) SAPS II model greatly overestimates the actual risk, as seen by the SAPS II predicted mortality in Table 2, and is thus not suited for identifying patients with a risk of death exceeding a pre-specified threshold.
### Table 2. Description of the subgroup population and estimates on the validation set.

Data are reported as interquartile range (median), and if after this another number is present, it is the mean value. Interquartile range is the range between the 25th to 75th percentile. PPV = Positive Predictive Value, SAPS = Simplified Acute Physiology Score. In the table SAPS II always refers to the original/un-recalibrated SAPS II model unless noted otherwise. GCS = Glasgow Coma Score, CPR = Cardiopulmonary Resuscitation, ICU = Intensive Care Unit, LOS = Length Of Stay. *That the PPV of the PRIM subgroups and the corresponding SAPS group is equal is coincidental.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>A or B or C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PRIM</td>
<td>SAPS II</td>
<td>PRIM</td>
<td>SAPS II</td>
</tr>
<tr>
<td>PPV (Hospital mortality), %</td>
<td>91.8*</td>
<td>91.8*</td>
<td>89.5*</td>
<td>89.5*</td>
</tr>
<tr>
<td>Patients covered by subgroup, %</td>
<td>2.8</td>
<td>3.5</td>
<td>8.8</td>
<td>9.6</td>
</tr>
<tr>
<td>SAPS II Score</td>
<td>70-95 (80)</td>
<td>91-100 (95)</td>
<td>70-94 (80.5)</td>
<td>88-99.25 (83)</td>
</tr>
<tr>
<td>SAPS II predicted mortality, %</td>
<td>(82.5,89.1)</td>
<td>(97,97.8)</td>
<td>(82,86.4)</td>
<td>(97.4,97.2)</td>
</tr>
<tr>
<td>Recalibrated SAPS II predicted mortality, %</td>
<td>72.6-91.2</td>
<td>89.3-93.1</td>
<td>72.6-90.8</td>
<td>87.7-92.8</td>
</tr>
<tr>
<td>Age, yrs</td>
<td>81-85 (83)</td>
<td>81-85 (82)</td>
<td>81-85 (83)</td>
<td>81-85 (82)</td>
</tr>
<tr>
<td>Male, %</td>
<td>49.2</td>
<td>47.5</td>
<td>46.1</td>
<td>43.4</td>
</tr>
<tr>
<td>Admisation type, %</td>
<td>73.8</td>
<td>85.2</td>
<td>73.7</td>
<td>82.9</td>
</tr>
<tr>
<td>Medical</td>
<td>12.6</td>
<td>13.1</td>
<td>22.4</td>
<td>13.9</td>
</tr>
<tr>
<td>Surgical unscheduled</td>
<td>0</td>
<td>1.5</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>Surgical scheduled</td>
<td>3.15 (15)</td>
<td>3.6 (3)</td>
<td>3.15 (15)</td>
<td>3.4 (3)</td>
</tr>
<tr>
<td>GCS 24 hrs after admission</td>
<td>29.5</td>
<td>34.4</td>
<td>23.7</td>
<td>38.8</td>
</tr>
<tr>
<td>CPR, %</td>
<td>0.2-1.3</td>
<td>0.5-1.3</td>
<td>0.2-0.9</td>
<td>0.2-1.4</td>
</tr>
<tr>
<td>ICU LOS</td>
<td>0.5-1.0</td>
<td>0.5-1.0</td>
<td>0.5-1.0</td>
<td>0.5-1.0</td>
</tr>
<tr>
<td>ICU mortality</td>
<td>85.5</td>
<td>80.3</td>
<td>84.2</td>
<td>82.9</td>
</tr>
<tr>
<td>Hospital LOS</td>
<td>1-4 (2)</td>
<td>1.3-6 (2)</td>
<td>1.3-7.7</td>
<td>1.3-7.2</td>
</tr>
<tr>
<td>Overlap, %</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Intersection PPV (Hospital mortality), %</td>
<td>91.3</td>
<td>92.9</td>
<td>91.6</td>
<td>93.4</td>
</tr>
<tr>
<td>Intersection, Patients at risk, %</td>
<td>1.0</td>
<td>1.3</td>
<td>3.6</td>
<td>6.4</td>
</tr>
</tbody>
</table>
2.5. Discussion

Using PRIM, we found and validated subgroups of patients at a very high risk to die before hospital discharge within the population of very elderly ICU patients. The subgroups are described by conjunctions of simple conditions based on data which are routinely collected for virtually all ICU patients during the first 24 hours after admission. Almost 10% of elderly ICU patients were identified as having a risk greater than 85% to die before hospital discharge and, in an independent sample of patients, the positive predictive value of this prediction was 87%. Our subgroups had a similar positive predictive value as the SAPS II model after recalibration for Dutch very elderly ICU patients. However, a major advantage of the PRIM generated rules is that they are easy to interpret and, more importantly, they describe homogenous populations in terms of patient characteristics, which can be beneficial in therapeutic efficacy research and are likely to be more intuitive for decision makers. As an example, consider a subgroup of what are high-risk patients according to the SAPS II model. The makeup of this group can be very diverse (compared to one derived using PRIM) because the total SAPS II score is composed of many small sub-scores for different risk related factors. It is therefore hard for a decision maker to get insight into the general cause for patients being in this subgroup, whereas with PRIM subgroups, a subgroup consists of a number of conditions that are linked with "AND" and in that sense all patients within the subgroup are "alike".

In comparing the characterization of the PRIM to the SAPS groups, the following differences clearly stand out. The SAPS II scores of the PRIM groups are markedly lower than those assigned to the SAPS II groups. This means that the SAPS II mean predicted probabilities assigned to the PRIM groups will be lower than the observed mortality mean. This is even more pronounced for the recalibrated SAPS II model. For example while the mean probability (which is equal to PPV) found in PRIM group A is 91.8%, the mean predicted probability according to the recalibrated SAPS II is only 89.1%. This is evidence that PRIM is arriving at "bumps" at different regions of the feature space than those found by the models based on SAPS II, or for similar scoring systems in general. The make-up of patients in the PRIM and SAPS groups are different: the SAPS groups generated by accumulating the patients at most risk will tend to first exhaust all patients with the feature associated with the maximum penalty (this is GCS of value below 6, contributing 26 points to the SAPS II score). This is easily seen in SAPS groups A and B (note that the worst value of GCS is 3). PRIM can discover patients corresponding to subranges of features that are not penalized heavily enough by SAPS. One way to capitalize on our observations on the differences between PRIM and SAPS is adding dummy variables in the SAPS model corresponding the the PRIM groups.

The most important prognostic factors in our model were GCS, admission type, blood pressure, urine production and acidosis. Interestingly, in another prognostic model based on recursive partitioning [21], aiming at predicting the likelihood of survival for all elderly ICU patients, similar risk factors were found, although there were differences in the cut-off values, and the risk was not required to be as high as in our study. Few other
models have been published that predict mortality specifically in elderly ICU patients. However, they were either specialized for pneumonia patients only and not validated in an independent patient population [22], or used data on functional status prior to ICU admission that are not available in our data set [23].

Our study has some limitations. First, PRIM requires some user-interaction and is not exhaustive in its search for subgroups and other adequate subgroup definitions are likely to exist. Second, the developmental and validation sets were randomly selected samples from the same population. This kind of validation eliminates the effects of changes in population and treatment over time. We cannot exclude that our model will be less accurate in identifying high risk patients in the future if therapeutic options may be improved. It should also be noted that our dataset was obtained solely from ICUs in the Netherlands. Third, we compared our high-risk subgroups to those derived from the SAPS II model. SAPS II was developed for patients of all ages. We recalibrated the SAPS II model for an elderly Dutch population and only included patients fulfilling the SAPS II inclusion criteria, however, a completely new model based on logistic regression specifically developed for elderly patients is likely to have higher predictive accuracy than SAPS II.

Although this study was part of a research project on prognosis and preferences of elderly ICU patients aged 80 years and older, the methodology used in this paper can be used for other patient groups. A model such as PRIM might also be used to find regions in the data where logistic regression models perform poorly by finding regions where the difference between the predicted probability and the outcome is high.

2.6. Conclusion

In sum, we successfully identified non-parametric descriptions of subgroups with very high probability of death in the very elderly ICU population. These descriptions are comparable in performance to SAPS II, but require less information, are easier to understand, and result in groups of relatively homogenous patients. Future research will focus on comparing PRIM to other subgroup discovery algorithms and using the same approach on other patient populations.

2.7. Acknowledgements

This work is performed within the ICT Breakthrough Project "KSYOS Health Management Research" and the I-Catcher project (number 634.000.020), which are funded respectively by the grants scheme for technological co-operation of the Dutch Ministry of Economic Affairs, and the Netherlands Organization for Scientific Research (NWO).
2.8. References


