Learning predictive models that use pattern discovery—A bootstrap evaluative approach applied in organ functioning sequences

Tudor Toma a, Robert-Jan Bosman b, Arno Siebes c, Niels Peek a, Ameen Abu-Hanna a,⁎

a Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands
b Department of Intensive Care, Onze Lieve Vrouwe Gasthuis, Oosterpark 9, 1091 AC Amsterdam, The Netherlands
c Department of Information and Computing Sciences, University of Utrecht, Padualaan 14, 3584 CH Utrecht, The Netherlands

A R T I C L E   I N F O

Article history:
Received 23 July 2009
Available online 21 March 2010

Keywords:
Prognostic methods
Temporal sequences
Pattern discovery
Intensive care
Evaluation
Accuracy
Calibration
Discrimination
Bootstrapping
.632 bootstrap

A B S T R A C T

An important problem in the Intensive Care is how to predict on a given day of stay the eventual hospital mortality for a specific patient. A recent approach to solve this problem suggested the use of frequent temporal sequences (FTSs) as predictors. Methods following this approach were evaluated in the past by inducing a model from a training set and validating the prognostic performance on an independent test set. Although this evaluative approach addresses the validity of the specific models induced in an experiment, it falls short of evaluating the inductive method itself. To achieve this, one must account for the inherent sources of variation in the experimental design. The main aim of this work is to demonstrate a procedure based on bootstrapping, specifically the .632 bootstrap procedure, for evaluating inductive methods that discover patterns, such as FTSs. A second aim is to apply this approach to find out whether a recently suggested inductive method that discovers FTSs of organ functioning status is superior over a traditional method that does not use temporal sequences when compared on each successive day of stay at the Intensive Care Unit. The use of bootstrapping with logistic regression using pre-specified covariates is known in the statistical literature. Using inductive methods of prognostic models based on temporal sequence discovery within the bootstrap procedure is however novel at least in predictive models in the Intensive Care. Our results of applying the bootstrap-based evaluative procedure demonstrate the superiority of the FTS-based inductive method over the traditional method in terms of discrimination as well as accuracy. In addition we illustrate the insights gained by the analyst into the discovered FTSs from the bootstrap samples.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Prediction of events can influence decisions. In the Intensive Care (IC), the predicted probabilities of mortality for the patients, based on their severity-of-illness at admission, are commonly used to benchmark the quality of care among various IC Units (ICUs). This is achieved by comparing a prognostic model’s predictions for the sample of patients of a given ICU with the actual proportion of death in that sample. The predictions are based on a model that has been retrospectively trained on a pooled sample of patients of a given ICU with the actual proportion of death in that sample. The predictions are based on a model that has been retrospectively trained on a pooled sample of patients of a given ICU with the actual proportion of death in that sample. The predictions are based on a model that has been retrospectively trained on a pooled sample of patients of a given ICU with the actual proportion of death in that sample.

When ICUs, about a decade ago, started collecting the daily SOFA scores [2] of their patients, various research groups tried to determine how to make prognostic use of this temporal information. A SOFA score on a given day is an integer quantifying the patient’s organ dysfunction. The higher the value of the score, the larger the organ function derangement. The SOFA score ranges between 0 and 24 and is calculated as the sum of six individual organ system dysfunction subscores (OD). Each subscore ranges between 0 and 4 and quantifies the degree of dysfunction in the following organ systems: respiratory (Resp), coagulation (Coag), hepatic (Hepa), circulatory (Circ), nervous system (Neuro), and renal (Ren).

For example if on day 2 score Resp = 2, score Coag = 0, score Hepa = 4, score Circ = 2, score Neuro = 1, and score Ren = 1 then the SOFA score on day 2 is 2 + 0 + 4 + 2 + 1 + 1 = 10. Table 1 illustrates the scoring scheme for the Coag and Hepa systems. The mapping between the physiological values and the scores is deterministic and is performed by the computer. The prognosis of the patient is dependent on his or her condition, which can change over time. Hence, the ability to provide probabilities of death (at discharge from the ICU or hospital) on a specific day.

When ICUs, about a decade ago, started collecting the daily SOFA scores [2] of their patients, various research groups tried to determine how to make prognostic use of this temporal information. A SOFA score on a given day is an integer quantifying the patient’s organ dysfunction. The higher the value of the score, the larger the organ function derangement. The SOFA score ranges between 0 and 24 and is calculated as the sum of six individual organ system dysfunction subscores (OD). Each subscore ranges between 0 and 4 and quantifies the degree of dysfunction in the following organ systems: respiratory (Resp), coagulation (Coag), hepatic (Hepa), circulatory (Circ), nervous system (Neuro), and renal (Ren).

For example if on day 2 score Resp = 2, score Coag = 0, score Hepa = 4, score Circ = 2, score Neuro = 1, and score Ren = 1 then the SOFA score on day 2 is 2 + 0 + 4 + 2 + 1 + 1 = 10. Table 1 illustrates the scoring scheme for the Coag and Hepa systems. The mapping between the physiological values and the scores is deterministic and is performed by the computer. The prognosis of the patient is dependent on his or her condition, which can change over time. Hence, the ability to provide probabilities of death (at discharge from the ICU or hospital) on a specific day.
of stay using historical information known up to and including that day is clinically useful. This information could be used in several ways, e.g. to assess if the current treatment plan is effective in this patient or to be of assistance in discussing the expectations for this patient with the family. As described in the section on related work, and based on our systematic review [3] various approaches have been suggested to make use of the SOFA scores to provide mortality predictions. Some approaches, most notably in the clinical literature, rely on pre-specified derivations of the SOFA score to be calculated. Other approaches, particularly in the medical artificial intelligence literature rely on data-intensive techniques to search for useful features. The validity of these approaches have been demonstrated in the following way. First, the method was applied to a data sample to induce a model, or a set of models, for each day of prediction. Then, the model’s performance was measured on a test set1 whereupon statements were made about the validity or superiority of the method over other methods.

As discussed in the framework described by Dietterich [4] and in the statistical literature [5] one should distinguish between statements about a specific model (e.g. an instance of a logistic regression model) and statements about the methods for inducing them. The approaches described above, including recent work of ours, fall in the category of statements about models. They do provide indirect evidence to the validity of the respective inductive methods but they fall short of providing direct evidence on the inductive methods themselves because the variability in the training and test sets has not been accounted for. This is especially important when the given sample is relatively small because the measured performance could be biased due to the idiosyncrasies of the particular training and test sets at hand.

In this work we describe a design for the evaluation of, and comparison between, methods for inducing predictive models and address the question of whether a particular method we suggested recently for exploiting frequent temporal sequences (FTSs) in organ function status (category) [6] is superior over a traditional method that does not use temporal sequences. The FTS-based method will be referred to as TESIM (TEmporal Sequence-based Inductive Method) and the traditional method as TRIM (TRaditional Inductive Method). The approach in this paper is characterized by the following aspects. First, it uses a bootstrap procedure, specifically the .632 bootstrap, for accounting for the sources of variability in the training and test sets. Second, it specifically targets the prognostic role of the FTSs. Third, it gives explicit attention to aspects of both discrimination (by means of the area under the ROC curve) and calibration (by means of the Brier score which is an accuracy measure that includes a calibration component) of prognostic performance measures. Our results provide evidence to the superiority of TESIM applied in organ function status in comparison to a method based on recalibrating, for each day of stay, a traditional severity-of-illness model that does not use data about the preceding days. The .632 bootstrap method has been applied elsewhere for logistic regression models where the covariates have been pre-specified. In this work we apply pattern discovery within each bootstrap sample and use the patterns as covariates in a logistic regression model. This means that different patterns may emerge in the bootstrap samples. This results in an intensive computational effort (as patterns are discovered in each bootstrap sample), but provides inference about the method not otherwise available. It is important to note that we use bootstrapping for the evaluation and comparison between models, not for constructing the model itself. In this paper we consider patterns which are frequent temporal sequences (FTSs), but the idea is applicable to more general types of patterns.

The rest of the paper is organized as follows. The next section provides a background on the underlying SOFA-based prediction problem and how it was approached in the literature. Section 3 presents our comparison setup based on bootstrapping. The next section describes the two proposed methods for the induction of models for daily prediction of mortality. In Section 5 we use the bootstrap approach to compare these inductive methods on each of the first 7 days of ICU stay (except for the day of admission) using a real world dataset, and report on the results obtained. We examine the organ function FTSs and their prognostic properties by inspecting their frequencies and associated model coefficients in the bootstrap samples. Section 6 discusses and concludes this paper.

### 2. Problem statement and current approaches

The core of the prediction problem is stated in Table 2. To illustrate, assume one is interested in predicting the outcome (in-hospital mortality) for the ICU patients on their 5th day of stay. We will have a training set for patients who stayed at least 5 days including: the severity-of-illness at admission $\text{Sev}$; the SOFA score $(\text{SOFA})_d^i$; and its underlying individual organ system dysfunction scores $(\text{OD})_d^i$. The model trained on this set can be used to make predictions on (unseen) patients on their 5th day of stay and the expected performance in prediction can be computed using the AUC (Area under the ROC curve) and the Brier score (a loss function reflecting the inaccuracy of predictions).

There are many variations on the problem stated in Table 2. For example some work uses only the SOFA scores without using the severity-of-illness scores at admission, as presented in [7]. There is work that develops a set of models targeting a range of days [6–8] whereas another approach develops a model for a specific given day but a different outcome, e.g. day 3 in [9] where outcome was time to survival within 180 days after admission. In another variation no specific day of stay is provided for making the prediction but an attempt is made to find the prognostic merit of a pre-specified variable for the whole stay (for example the maximum SOFA score during the whole stay) [10]. Still, other work

### Table 1

<table>
<thead>
<tr>
<th>SOFA score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coagulation Platelets $\times 10^9/\mu L$</td>
<td>$&gt;150$</td>
<td>(100–150]</td>
<td>(50–100]</td>
<td>(20–50]</td>
<td>$\leq 20$</td>
</tr>
<tr>
<td>Hepatic Bilirubin mg/dL</td>
<td>$&lt;1.2$</td>
<td>(1.2–2.0]</td>
<td>(2–6.0]</td>
<td>(6–12]</td>
<td>$&gt;12.0$</td>
</tr>
</tbody>
</table>

1 In 5 out of 16 studies in the clinical literature reviewed in Minne et al. [3] there was no report on using an independent test set for the validation.
focusses on predicting outcomes at multiple future intervals [11]. Finally, some work investigated the use of variables other than the SOFA, most notably adverse events [12].

The prediction problem, regardless of its specific formulation, entails the following issues: how to represent the temporal data up to a certain day and how to use this representation in prediction. Two main approaches were devised for representing the temporal data. The first, exemplified by [10,13,14], applies pre-specified abstractions such as the mean, maximum or minimum (up to day d) or a pre-specified variable such as the difference in SOFA scores between day d and day 0. Logistic regression, the most dominant model type, and neural networks [12] have been applied in this approach. In the other approach a data-driven search was applied to identify relevant covariates to be used in the prognostic model. The covariates commonly indicate frequent sequences observed over the earlier days of ICU stay. Predictive models, aside from logistic regression, using such covariates include Naïve Bayes [11] and support-vector-machines [15].

In all the above approaches in which a model was developed to predict mortality, performance evaluation consisted of one of the following ways: the model was tested on the same training set (as probably was the case in the five clinical studies identified in [3]), or the split-sample approach was used [12,8,11,6]. Both ways fall short of validating the model inductive method itself.

3. Comparing inductive methods

We are interested in answering the following question: “Does Method i for (pattern-based) model induction perform, on average, better than Method j when both are trained on an arbitrary but same training set, S, of a given size K?” In the taxonomy of statistical questions presented in [4] this question corresponds to comparing predictive methods (in contrast to specific models trained on S). In practice, because datasets are not abundant, it corresponds to the more concrete question of choosing between methods when the size of the sample data is relatively small. To answer this question one must take into account the following sources of variation: variation due to the random selection of the training set and the test set (from the population); due to internal randomness in the learning algorithm (e.g. initialization of internal parameters in neural networks); and due to noise in the data (e.g. random classification error). In our application it is sufficient to account for the variations in the training and test sets because the methods we intend to compare—described below—do not apply internal parameter initialization, and classification errors in the data are very unlikely as the survival status of the patient is unambiguous and the datasets undergo routine quality checks. In other applications, accounting for these latter two sources of variation would require more insight in the processes determining the internal parameters in the algorithms and those responsible for the errors in the labeling.

Accounting for variation in the training and test data sets implies generating a multitude of training and test sets. For the training set, and because data is limited, one will need to rely on some form of resampling. Because one is interested in training on a set with size K (and not less), bootstrapping [16] seems to be more appropriate than cross-validation [17]. Although computationally expensive, by directly simulating the sampling distribution of a statistic (e.g. median or a model’s coefficient), bootstrapping yields nearly unbiased estimates of the statistic along with confidence intervals. In bootstrapping a number (usually hundreds) of resamples (each called a bootstrap sample) are drawn from the original dataset of size K. A bootstrap sample consists of K equally-likely draws with replacement from the original dataset. Some observations may appear multiple times, whereas others may not appear at all. To illustrate, the sets \{a, c\} and \{b, b, b, c\} are two possible bootstrap samples of the set \{a, b, c, d\}. The inductive method is then applied to each bootstrap sample. When evaluating each model learned on a bootstrap sample one may want to obtain its performance on only observations not already selected in that bootstrap sample (called out-of-bag, or simply OOB). Alternatively, as we do in this paper and as explained below, one may use both the performance on the original dataset as well as on the OOB observations in that iteration.

3.1. Implementation of strategy

Our strategy is based on three main design choices regarding resampling, performance estimation, and evaluation. First, we choose for the .632 bootstrap method [17,18] which accounts for the variability in both the training and test sets. In our experiments we rely on 300 bootstrap samples as this allows to obtain valid experimental results in a reasonable amount of computational time\(^2\) [17]. Second, we will look at daily predictions. In our experiments we cover day 2 (patterns are less meaningful at day 1) till day 7 (the great majority of patients would have by now left the ICU). Third, we rely on the discrimination as well as accuracy measures. In our experiments we rely on the AUC (discrimination) as well as the Brier score (accuracy) for the predictive performance measures. The AUC indicates how well a prognostic model can discriminate between survivors and non-survivors. The Brier score is defined as \(\frac{1}{N} \sum_{i=1}^{N} (P(Y_i = 1 | x_i) - y_i)^2\), where \(N\) denotes the number of patients, \(x_i\) the covariate values for patient \(i\), and \(y_i\) the actual outcome for this patient (0: survival, 1: non-survival). The Brier score describes how close the predictions are to the real outcome, and as such it includes a calibration aspect.

Algorithm 1.

The .632 bootstrap for evaluating inductive methods on day d.

---

2 We noted that increasing this number did not qualitatively change the results and hence the source of variation introduced by the bootstrap procedure itself can be neglected.
all steps performed during the bootstrap procedure for estimating the predictive performance. For comparing methods, the same set-up is used but one considers differences between performance measures in each bootstrap sample, as we describe below.

We can see in the last steps of the algorithm that two elements determine the predictive performance of a method. The first element is simply the performance of its models on the original data-set when trained on the bootstrap samples (the models are denoted by Mod.). To illustrate, for the Brier score this is \( \rho_{\text{Brier}}(\text{Mod.}, \text{PATd}) \) which is denoted as \( \text{Brier}^\text{PATd} \). The second element consists of performance on the observations when they were in the OOB sample. To illustrate, for each observation \( \text{PATd}_i \in \text{PATd} \) we calculate the mean Brier score on \( \text{PATd}_i \) when \( \text{PATd}_i \) was in an OOB (observations may vary in the number of times they appear in an OOB over the 300 randomization trials). Then the mean of these average Brier scores, denoted \( \text{Brier}^\text{OOB} \), is calculated. Taking the weighted average of the two elements with weight 0.368 for \( \text{Brier}^\text{PATd} \) and 0.632 for \( \text{Brier}^\text{OOB} \) corrects the biased (optimistic) performance as measured by \( \text{Brier}^\text{PATd} \) and represents the .632 performance estimate. The factor .632 arises because it is the probability for an observation to appear in a bootstrap sample. For the AUC, \( \text{AUC}^\text{PATd} \) is calculated in the same manner as for the Brier score. However, because we cannot obtain the AUC on individual observations, \( \text{AUC}^\text{OOB} \) is calculated in a slightly different manner: for each of the 300 iterations we calculate the AUC on the OOB in that iteration as one set \( \text{AUC}(\text{Mod.}, \text{OOB})_i \), and then we take the mean over all iterations to obtain \( \text{AUC}^\text{OOB} \).

Aside from evaluating predictive performance of a method one may readily compare the performance of the two methods on any given day. One simply takes the differences between the performance measures (Brier and AUC) obtained by the two methods in each bootstrap sample. For each bootstrap sample \( b \), \( \text{AUC}^\text{ooB} = 0.632 \cdot \text{AUC}(\text{Mod.}, \text{PATd}) + 0.632 \cdot \text{AUC}(\text{Mod.}, \text{OOB}) \) and \( \text{Brier}^\text{ooB} = 0.368 \cdot \text{Brier}(\text{Mod.}, \text{PATd}) + 0.632 \cdot \text{Brier}(\text{Mod.}, \text{OOB}) \). Note that because we consider each bootstrap sample separately we also calculate the Brier per iteration on the OOB sample as a whole, just like with the AUC. Because we now have 300 points of the bootstrap distribution of the differences between each of these performance measures we can declare statistical difference between the inductive methods in various ways. In this paper we resort to the bootstrap percentile method at the 0.05 level: The 2.5 and 97.5 percentiles of the bootstrap distribution of each statistic define its 95% confidence interval. The difference is declared significant at the 0.05 level if the 0 is not included in this interval.

We defined the performance difference as the estimated performance of TESIM (the FTS-based inductive method) minus the estimated performance of TRIM (the traditional method). A positive difference in the AUC value and a negative difference in a Brier score are indicative of better performance for TESIM over TRIM in discrimination and accuracy, respectively.

4. The inductive methods

The two methods compared in this study, and presented in [6], predict on a daily basis the probability of hospital discharge mortality. For each day of prediction during the ICU stay each method generates one prognostic model.

4.1 The traditional inductive method, TRIM

TRIM is based on the current ICU method for mortality prediction. It applies logistic regression (see Appendix A) with input variables (covariates) whose values are obtained within 24 h of a patient’s admission. These covariates quantify the severity-of-illness of the patient at admission. Exemplary to this family of models is the SAPS model which is the object of study in this paper. The SAPS model uses the score SAPS and the \( \log(\text{SAPS}+1) \) as its sole covariates. The probability of hospital mortality is described by:

\[
P(\text{died} = 1 | \text{SAPS}) = \frac{\exp(b_0 + b_1 \times \text{SAPS} + b_2 \times \log(\text{SAPS}+1))}{1 + \exp(b_0 + b_1 \times \text{SAPS} + b_2 \times \log(\text{SAPS}+1))}
\]

(1)

where the term \( b_0 + b_1 \times \text{SAPS} + b_2 \times \log(\text{SAPS}+1) \) is called the linear predictor (LP). Unlike the original SAPS-II method, which is applied only once for all patients, TRIM develops various models one for each ICU day of stay. Each model for a day \( d \) of prediction is trained on data of patients staying at least \( d \) days in the unit. The TRIM models differ only in the estimates of the coefficients. Recall that in this comparison TRIM is used as the reference method against which the TESIM method is to be compared.

4.2 The FTS-based inductive method, TESIM

The inductive method based on the discovery of frequent temporal sequences was developed in our previous work [6]. The main focus of the current paper is not improving the pattern-based method itself. Instead, the focus is on the evaluation of such inductive methods. We hence only briefly explain TESIM here as applied to organ function sequences. TESIM generates daily prognostic models based on a variable selection strategy for predicting hospital mortality. Just like in TRIM, a prognostic model for each day of stay uses the admission information (SAPS and \( \log(\text{SAPS}+1) \)) but now also in combination with abstractions from the temporal data collected during patient stay, up to and including the day of prediction. This information includes the six daily individual organ system scores available from each patient. The scores are transformed from a quantitative to a qualitative representation based on their median values over all days of all the patients from the data set (and hence will not be obtained by bootstrapping on individual days). The median, rather than a pre-defined cut-off point on the organ score, was chosen because the distribution may be quite skewed leaving very little observations at one side of the pre-defined cut-off point.

To illustrate, if the sequence of the coagulation system’s functioning in the first 5 days of a patient’s stay was 3–4–4–2–0, and the median value of this score is 1 for all patients, then the qualitative representation of this sequence will be Coag{c2, c2, c2, c1} where all values for this organ system that are \( \geq 1 \) fall in category c2. The qualitative representation of the data serves the purpose of discovering frequent univariate patterns of organ function categories. For each day of prediction \( d \), frequent temporal sequences (our patterns) are discovered in the training set from the patients staying at least \( d \) days in the unit. The discovery procedure is based on an adaptation of the A-priori algorithm which was previously developed in [19]. We consider a specific type of FTs, those aligned to the day of prediction and consisting of qualitative values coming from consecutive days of stay. This restriction reflects the belief that more recent events, particularly on the last day, are more relevant than events happening earlier in time. An FT is frequent if the number of patients in which it occurs is larger than a pre-defined threshold value (e.g. 5% of the sample). For example the length-2 respiratory FT Resp{c1, c2} occurs as aligned to day 3 within the following patient sequence showing the first 5 days of the respiratory organ dysfunction status: c2–c1–c2–c2–c1. This is because this pattern agrees with the patient’s sequence at day 3 and day 2. In contrast, the pattern Resp{c2, c2} does not align on day 3 as a consecutive pattern (it does however for day 4). By virtue of being a list of events ordered in time of their occurrence, our patterns are temporal sequences. In particular the event is an organ with a dysfunction score above the median; the time of
occurrence is a day; the sequence consists of consecutive days; and the last event occurs at the day of prediction. Our patterns, however, are simple and do not bear other important temporal aspects. In particular, only univariate patterns are allowed (so a pattern cannot express that the liver failed after the renal system); only sequential patterns without explicit trends are allowed; and the temporal relationship we allow indicates that an event occurs one day (and no other time interval) after another. This means that the patterns discovered in TESIM constitute a restricted form of serial episodes as defined in [20] where a general partial order between the events is allowed. In the context of our experiments, the term pattern in the sequel refers to a frequent temporal sequence (FTS), but should be applicable to other forms of patterns in general.

Once FTS discovery is complete, each pattern $P_{att}$ is represented as an indicator variable $I(P_{att})$ in a logistic regression model where $I(P_{att}) = 1$ for patients in which FTS occurs and 0 otherwise. Next, a variable selection procedure based on the Akaike’s Information Criterion (AIC) [5] is applied to retain the best predictive variables. A model using frequent patterns will be described by a linear predictor with the following generic form: $\hat{P}(SAPS; P_{att}) = \beta_0 + \beta_1 \text{SAPS} + \beta_2 \log(\text{SAPS} + 1) + \sum_{i=1}^{K} \alpha_i I(P_{att})$, where $P_{att}, i = \{1, 2, \ldots, K\}$ represent the model’s selected frequent temporal sequences. The essential difference between a traditional model and a TESIM model for day $d$ is that TRIM uses only the SAPS information. However, the models in TRIM are refitted specifically for the population of patients that stayed at least $d$ days in the ICU. This means that the populations used for both kinds of models are exactly the same, hence controlling for the effect of the days themselves, but TESIM is allowed to use, in addition, the frequent sequences as dummy variables in the logistic regression model.

5. Case study and results

5.1. Data

The data available for analysis included 9103 admissions. It was collected between July 1998 and February 2007 from the adult ICU in the OLVG teaching hospital in Amsterdam. The data included: SAPS-II scores (short SAPS) upon admission, all daily SOFA scores, all daily individual organ system scores and in-hospital survival status. To allow for a fair comparison to the SAPS model alone (used in TRIM), we used the established SAPS exclusion criteria to eliminate records of patients admitted after cardiac surgery (5291 cases). Missing values in the individual organ system scores (and hence also for SOFA), were imputed using the maximum of the “left” and “right” adjacent individual organ scores, or the only adjacent value itself when the values for the first or last days are missing. Cases with 2 or more consecutive missing values were discarded. In total 2928 admissions were retained for analysis.

Table 3 depicts characteristics of survivors and non-survivors in the sample. The mortality rate was about 25% with higher organ dysfunction in the non-survivors compared to the survivors for all organ systems.

5.2. Results

5.2.1. Categorization

The obtained rules for categorizing the individual SOFA scores were:

- Respiratory: category c1 for values in $\{0, 1, 2, 3\}$ or category c2 otherwise.
- Coagulation: category c1 for the value 0 or category c2 otherwise.
- Hepatic: category c1 for the value 0 or category c2 otherwise.
- Circulatory: category c1 for values in $\{0, 1, 2\}$ or category c2 otherwise.
- Neurologic: category c1 for the value 0 or category c2 otherwise.
- Renal: category c1 for values in $\{0, 1\}$ or category c2 otherwise.

5.2.2. Performance comparison

In Table 4 the results of comparing the two methods per day of prediction are depicted covering the days 2–7 of ICU stay. A $\Delta$Brier $< 0$ indicates a better accuracy by the Brier score (and hence also calibration) of TESIM. By the same token a $\Delta$AUC $> 0$ indicates better discrimination between survivors and non-survivors for TESIM. The difference in the Brier scores ranged between $-10 \times 10^{-3}$ and $-5 \times 10^{-3}$ with confidence intervals strictly in the domain of negative real numbers. Because the corresponding 95% CIs do not include the value 0, which corresponds to the null hypothesis of no difference, all $\Delta$Briers are statistically significant at the 0.05 level. The TESIM models had also higher discrimination performance with differences in the AUC ranging from 0.016 to 0.059 and were always significant at the 0.05 level with the exception of day 6.

5.2.3. Insight

An important advantage of resampling methods is the insight obtained by inspecting the variability of an estimate over the (bootstrap) samples. One may focus on organ systems, days, patterns, and model coefficients. Below we provide a series of illustrative examples of such estimates. Consider how often an organ system is selected in a model, in other words when the model includes a pattern of that organ system.

Table 3

Descriptive statistics of the patient sample.

<table>
<thead>
<tr>
<th></th>
<th>Survivors</th>
<th>Non-survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>2200</td>
<td>728</td>
</tr>
<tr>
<td>Admission type (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>55</td>
<td>77</td>
</tr>
<tr>
<td>Urgent</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>Planned</td>
<td>24</td>
<td>9</td>
</tr>
<tr>
<td>Mean age ± SD (years)</td>
<td>61 ± 17</td>
<td>68 ± 14</td>
</tr>
<tr>
<td>Male (%)</td>
<td>60</td>
<td>62</td>
</tr>
<tr>
<td>Median LOS</td>
<td>1.7</td>
<td>3</td>
</tr>
<tr>
<td>Mean SAPS</td>
<td>39</td>
<td>61</td>
</tr>
<tr>
<td>Mean SOFA</td>
<td>7.3</td>
<td>9.6</td>
</tr>
<tr>
<td>Mean # org. fail.</td>
<td>4.4</td>
<td>9.8</td>
</tr>
<tr>
<td>Mean Resp</td>
<td>3.5</td>
<td>3.7</td>
</tr>
<tr>
<td>Mean Coag</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Mean Hepa</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Mean Cardio</td>
<td>1.6</td>
<td>1.9</td>
</tr>
<tr>
<td>Mean Neuro</td>
<td>0.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Mean Ren</td>
<td>0.8</td>
<td>1.4</td>
</tr>
</tbody>
</table>
The frequency in which an organ system’s FTS was included in the TESIM prognostic models for days 2–7 of prediction in the 300 bootstrap samples.

<table>
<thead>
<tr>
<th>Prediction day</th>
<th>Brier score TESIM</th>
<th>Brier score TRIM</th>
<th>ΔBrier 10⁻¹</th>
<th>AUC TESIM</th>
<th>AUC TRIM</th>
<th>ΔAUC 10⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.159</td>
<td>0.165</td>
<td>-6 (-9, -2)</td>
<td>0.742</td>
<td>0.726</td>
<td>16 (5, 24)</td>
</tr>
<tr>
<td>3</td>
<td>0.140</td>
<td>0.150</td>
<td>-10 (-13, -5)</td>
<td>0.734</td>
<td>0.702</td>
<td>32 (18, 45)</td>
</tr>
<tr>
<td>4</td>
<td>0.125</td>
<td>0.134</td>
<td>-9 (-13, -5)</td>
<td>0.710</td>
<td>0.663</td>
<td>46 (25, 62)</td>
</tr>
<tr>
<td>5</td>
<td>0.116</td>
<td>0.123</td>
<td>-7 (-10, -2)</td>
<td>0.688</td>
<td>0.651</td>
<td>36 (6, 58)</td>
</tr>
<tr>
<td>6</td>
<td>0.110</td>
<td>0.115</td>
<td>-5 (-8, -1)</td>
<td>0.670</td>
<td>0.636</td>
<td>33 (-3, 57)</td>
</tr>
<tr>
<td>7</td>
<td>0.101</td>
<td>0.110</td>
<td>-8 (-12, -4)</td>
<td>0.678</td>
<td>0.618</td>
<td>59 (17, 94)</td>
</tr>
</tbody>
</table>

6. Discussion

6.1. Interpretation of results

The covariate coefficients in the models for day 6, presented in Fig. 4, concord with our expectations. All patterns including only category 1 or showing “recovery” in organ functioning (Circ[c2, c1, c1, c1, c1, c1]) were indeed always associated with a negative coefficient as would be expected. By the same token, all patterns with positive coefficients included a category 2. The pattern Hepa[c2, c2, c2, c2] requires more investigation: it is associated with a negative coefficient and yet it depicts dysfunction in the patient’s hepatic organ system. It is unclear whether the unfavorable odds-ratio is due to the categorization not being sensitive enough to pick on the difference between moderate dysfunction and complete organ failure, or because the other patterns (not shown here), co-selected in the models, play a role in this phenomenon.

Table 4 shows that the TESIM models consistently outperformed the traditional models, both in aspects of calibration and discrimination and almost always with statistically significant difference (except for day 6 for the AUC difference). Note also that the pattern-based TESIM models are the result of a straightforward automatic variable selection procedure that facilitates performing
the bootstrap experiment without any additional fine tuning. This can sometimes lead to unstable models: a slight change in the data may cause a large change in the regression coefficients. We suggested improvements in constructing pattern-based models in [6].

6.2. Strengths and limitations of the proposed approach

The main strength of this work is the synthesis between pattern discovery, logistic regression, and resampling methods to assess and compare among inductive methods. In addition, we use both discrimination (AUC) and accuracy measures (the Brier score, which also includes a calibration measure) when evaluating models. Both of our measures generalize to multiple classes. In [21] the definition of the AUC was extended by averaging pairwise comparisons. The original definition of the Brier score was already expressed in terms of multiple classes: \[
BS = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} (P_{ic} - y_{ic})^2
\]
where \(P_{ic}\) is the predicted probability of class \(c\) (from \(C\) classes) for instance \(i\) [22]. This approach is novel in prognostic research in the ICU at least. We use the bootstrap procedure which, although computationally intensive, overcomes limitations of cross-validation [17] such as the subjective choice of the number of folds to be used and the fact that not all data are used in training.

One limitation of this study is that we do not prospectively validate the models to inspect performance on data collected at a later period in time than when models have been fit. But our approach allows to focus on the added value of methods without additional confounders.

Another limitation of our approach is that we did not take into account the correlation of results between days. We hence have shown that the pattern-based method performed better on any given day of a series of days but we did not address an aggregate summary measure that takes the correlation between days into account. This forms an interesting future work.

6.3. Patterns of organ function categories in prediction

Bootstrapping allows for a more comprehensive insight into the patterns of organ dysfunction and their predictiveness. Fig. 1 showed the dominance of the neurological system in the selected patterns. One may also observe a clear increase in the frequency of selecting the circulatory system in days 6 and 7 of prediction. Also the renal organ system seems to have more prognostic value in the later days of prediction whereas the respiratory system exhibits variations in the number of selections over the 6 days of prediction. The coagulation system was often selected on days 2, 3 and 4 and less in days 5–7. The hepatic system seems to hold relatively small predictive information and is infrequently selected on most days. The patterns Neuro(c2) and Neuro(c2, c2), indicating organ derangement in the neuronal system were the most often selected patterns.
In general, short patterns with the same category such as \( \{c_1, c_1\} \) or \( \{c_2, c_2\} \) were favored possibly because our categorization approach is not sensitive to the small day to day variations in the patients' scores and because shorter patterns are more frequent than longer ones. When considering day 6 of prediction, the circulatory pattern \( \text{Circ}(c_2) \) is shown to be the most frequent (Fig. 3). However, the dominant organ system remains the neurological system which has 5 patterns in the 10 most frequently selected patterns. The only pattern showing improvement in the functioning of an organ system that is often predictive in combination with other covariates is \( \text{Circ}(c_2, c_1, c_1, c_1, c_1, c_1) \), the other patterns indicated a sequence of the same organ function category. Fig. 4 shows the models' coefficients for each of the 10 most frequent patterns from the 300 bootstrap iterations. This provides insight into the variance of these coefficients over the bootstrap samples. All the patterns showing increased dysfunction with other covariates is \( \text{Circ}(c_2, c_1, c_1, c_1, c_1, c_1, c_1) \), the functioning of an organ system that is often predictive in combination with other covariates is \( \text{Circ}(c_2) \) is shown to be the most frequent than longer ones. When considering day 6 of prediction, the patients' scores and because shorter patterns are more frequent in terms of accuracy (with a calibration aspect) as well as discrimination over the traditional one. We also showed how inspecting patterns' frequencies and model coefficient estimates over the bootstrap samples provides insight in organ system dysfunction and their association with the outcome (survival status).

### 6.4. Conclusions

This work introduced a design and experiment for evaluating and comparing inductive methods of prognostic models based on the principles of accounting for the relevant sources of variation. Using this design we compared a prognostic method using patterns, in the form of frequent temporal sequences of organ dysfunction, to a method that merely refits models for each day of prediction. The pattern-based method was shown to be superior in terms of accuracy (with a calibration aspect) as well as discrimination over the traditional one. We also showed how inspecting patterns' frequencies and model coefficient estimates over the bootstrap samples provides insight in organ system dysfunction and their association with the outcome (survival status).

### Acknowledgments

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

### Appendix A. Logistic regression

A logistic regression model [25] specifies the conditional probability of a binary outcome variable \( Y \), given the values of the covariate vector \( \mathbf{x} = (x_1, \ldots, x_m) \) : \( p(Y = 1 \mid \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m}} \). For \( m \) covariates the natural logarithm of the odds (logit) is equal to the linear predictor \( \log\left(\frac{p(Y = 1 \mid \mathbf{x})}{1 - p(Y = 1 \mid \mathbf{x})}\right) = LP(\mathbf{x}) = \beta_0 + \sum_{i=1}^{m} \beta_i \cdot x_i \) where \( \beta_i, i = 1, \ldots, m \), denote the coefficients of the \( m \) covariates. A coefficient (\( \beta_i \)) can be interpreted in terms of an odds-ratio. Suppose the linear predictor is \( \beta_0 + \beta_1 \cdot \text{SAPS} + \beta_2 \cdot \text{Patt} \) where \( \text{Patt} = 1 \) for patients having some specific pattern and 0 for patients not having the pattern. The odds of dying for those having the pattern, \( \text{odds}(\text{Patt} = 1) = P(Y = 1 \mid \text{Patt} = 1)/P(Y = 0 \mid \text{Patt} = 1) \) and for those not having the pattern, \( \text{odds}(\text{Patt} = 0) = P(Y = 1 \mid \text{Patt} = 0)/P(Y = 0 \mid \text{Patt} = 0) \). The quantity \( e^{\beta_2} \) is equal to the odds-ratio \( \text{odds}(\text{Patt} = 1)/\text{odds}(\text{Patt} = 0) \). A positive coefficient corresponds to an odds-ratio >1 and indicates that, when adjusting for all other
variables (here SAPS), the odds of the event is higher for those with the pattern compared to those without it.

References


