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

Citation for published version (APA):
Chapter 1

General Introduction

The availability of computer technology in hospitals leads to opportunities for improving the health-care process, especially in data-rich environments like the Intensive Care Unit (ICU). One direct application of computerized devices in the ICU is the continuous monitoring of the patients’ health status. These devices record substantial amounts of data over time about each patient. Such repeatedly measured data over time can be referred to as temporal (or longitudinal) data, to contrast it with static data, which is recorded only once. The medical staff is often unable to cope with the amount of collected temporal data of multiple variables. This calls upon automated tools to make more sense of these data to clinicians. Two particular challenges facing clinicians and clinical managers in the ICU are (1) obtaining a summary of the progression of the health status of their patients and (2) acquiring a daily updated prognosis of the survival status of their patients upon discharge from hospital. These tasks are currently not supported and it is the main objective of this thesis to investigate methods that can adequately perform these tasks. Our approach is based on summarizing health status progression over time by discovering frequently occurring patterns in the temporal data and developing prognostic models that use these patterns for prediction of patient survival probability at hospital discharge. In this thesis we focus on one type of data in the ICU: daily organ failure scores, but the analytic approach should be applicable to various medical and non-medical domains. Below
we introduce organ system measurement in the context of scoring systems in the ICU, and present our approach for developing prognostic models based on patterns of organ system scores.

ICU scoring

Much of the collected ICU data is in the form of integer scores. These scores are constructed from different data types (physiological, demographic) and are indicators of the patients health status. A score from the Simplified Acute Physiology Score (SAPS) \[1\] family or the Acute Physiology and Chronic Health Evaluation (APACHE) \[2\] family is calculated once per patients stay and indicates the general severity of illness of a patient upon admission to the ICU. These scores are also meant for making probabilistic predictions using the logistic regression framework. Other scores such as the Sequential Organ Failure Assessment score (SOFA score) \[3\] are calculated daily. The SOFA score measures the degree of organ dysfunction for each admitted patient and is calculated as the sum of six individual scores, one for each of the following organ systems: respiratory, coagulation, hepatic, circulatory, central nervous and renal. Unlike other scores, the SOFA score was not specifically developed to assist mortality prediction.

Organ dysfunction and the challenge of prognosis in ICU

Patients are admitted in the ICU due to their severe medical condition reflected by their level of organ failure. These patients require special medical attention to stabilize, support and restore their functions. Most people are discharged from the unit once their organs are functional again, with the recovery process continuing in other regular wards. An example clearly indicating the strong link between organ function and ICU is respiratory failure. Respiratory support is only provided in the operating rooms and the ICU and most patients are discharged as soon as they are able to breathe autonomously if no other critical factors impede it. Our hypothesis is that the daily quantification of the degree of organ dysfunction, as is done by the SOFA scoring scheme, holds prognostic value
beyond that available in static scoring schemes. Specifically, the SOFA scoring scheme maps for each of the 6 organ systems its degree of dysfunction to an integer between 0 and 4. A value of 0 indicates normal functioning and the higher the value the greater the dysfunction, there is also consensus that the values of 3 and 4 indicate organ failure. In consequence the SOFA score, the sum of the 6 organ scores, is positively proportional to the amount of general organ dysfunction.

Although the daily collection of the SOFA score provides the opportunity to describe the evolution of organ dysfunction over time, its predictive potential has not been fully exploited in prognostic research. As described in [4] the majority of existing research using SOFA for prediction relies on simple summaries like the mean SOFA value, total maximum score (TMS), the difference in SOFA score between different pre-specified days etc. [5, 6, 7, 8, 9, 10]. These measures abstract away much of the temporal information available in the SOFA score (and sub-scores) over time. We expect that this temporal information provides a positive effect on the quality of predictions and contributes to a better understanding of the phenomenon of organ (dys)functioning.

Extracting temporal information from data and using it for prediction calls upon suitable data representations for this information and adequate strategies for generating and evaluating predictive models. The rest of this chapter will introduce the essential prerequisites pertaining to Intensive Care, SOFA scores, patterns capturing organ dysfunction progression, prognostic models, and their evaluation. Thereafter we state our research questions and show how they map to the other chapters of this thesis.

1.1 Prerequisites

Intensive Care

Intensive Care is a special form of care delivered in ICUs in which patients with life threatening health conditions are admitted. Patients are continuously monitored and treated by physicians (intensivists) and nurses who are assisted by various computer-
based devices. These devices are used to measure and to influence the vital functioning of organ systems of patients. For example there are devices to measure the heart rate and others to determine the tidal volume required by patients on respiratory support. The introduction of computerized data acquisition facilitates the collection and storage of historical patient data. A system designed for the collection and storage of electronic patient records is the Patient Data Management System (PDMS), which is currently in use in various ICUs. Aside from information manually entered by clinicians, this system also often records information originating from several bedside monitoring and treatment devices it is coupled with.

Intensive care is one of the most expensive forms of care, especially due to the presence of a relatively large number of care providers at all times. Patients are admitted to the ICU in the belief that it will be beneficial to them and decisions about treatment and discharge are influenced, implicitly or explicitly, by their survival chances as estimated by clinicians. This emphasizes the need for understanding the relationship between the patients condition over time and probability of patient outcome such as survival or re-admission. This relationship can be captured in the form of a prognostic model that takes as input patient information and delivers as output a probability of the adverse event under consideration (death or re-admission). Much of the patient information is already available in the PDMS, which includes static data collected at admission as well as temporal data collected daily such as the Sequential Organ Failure Assessment score.

**Sequential Organ Failure Assessment score**

The data gathered in the ICU consists of demographic (e.g. age, gender), physiology (e.g. heart rate, body temperature), laboratory (e.g. measurements of substances in the blood), and outcome data (e.g. mortality, length of stay, re-admission). In 1996 the SOFA score was introduced and is currently widely used to quantify and record the degree of organ dysfunction. Recall that SOFA is the sum of 6 individual scores, one for each of the following organ systems: respiratory, coagulation, hepatic, circulatory, central nervous,
and renal systems. It ranges from 0, indicating normal functioning, to 24, representing the worst possible dysfunction of organs.

<table>
<thead>
<tr>
<th>SOFA Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Respiration</strong></td>
</tr>
<tr>
<td><strong>Coagulation</strong></td>
</tr>
<tr>
<td><strong>Hepatic</strong></td>
</tr>
<tr>
<td><strong>Circulatory</strong></td>
</tr>
<tr>
<td><strong>Central nervous system</strong></td>
</tr>
<tr>
<td><strong>Renal</strong></td>
</tr>
<tr>
<td>or</td>
</tr>
</tbody>
</table>

Figure 1.1: SOFA scoring scheme. “Dopa” is Dopamine, “Dobu” is Dobutamine, “Norepi” is Norepinephrine.

Figure 1.1 illustrates the SOFA scoring scheme for the six underlying organ systems. Scores of 3 and 4 are considered to indicate organ system failure. Associated with each organ system are one or more clinical parameters whose values determine the score. For example, when Bilirubin is < 1.2 mg/dl then the Hepatic score becomes 0, but when Bilirubin is ≥ 6 and < 12 mg/dl then the score becomes 3. For each day those measured values of the clinical parameters indicating the worst dysfunction are used in the calculation of the score. The scores are often generated automatically by a computer and are validated by the medical staff.

There are various options for using the SOFA score in prediction. First, one should decide on how to capture organ (dys)function development over time. For example, this
can be captured quantitatively (like average change per day) or qualitatively (like indicating whether there is increase or not in the score). Second, one should decide on whether to use the individual sub-scores in the model or the general SOFA score. Third one should decide on which other covariates should be included in the prediction model. For example, multiple organ failure (MOF) consists of the leading cause of death in critically ill patients, and it is hence interesting to investigate whether the temporal information provides added predictive value beyond the number of failing organs.

**Patterns of organ function**

A temporal pattern represents a temporal association among different measurements from one or multiple variables over time, such as sequences of daily SOFA scores. An example of a complex association is the increase in one sub-score followed by a decrease in another. One can consider patterns that are either predefined by a human analyst or discovered from data. Pre-defined patterns may benefit from the pre-existing knowledge about the relevance of the pattern for the prediction task at hand. However, this approach may miss unknown relevant patterns. Discovered patterns, in contrast, can reveal new relevant temporal associations. However, one should state constraints on the form and/or frequency of these patterns lest the complexity of the search becomes prohibitive or the number of discovered potentially relevant patterns becomes too large. A common practice in pattern discovery is to retain only patterns having a relative frequency exceeding a predefined minimum frequency (called support).

**Prognostic Models in Intensive Care**

Prognostic models in the Intensive Care often predict the probability of mortality (or other events like re-admissions) for a given patient. Mortality may refer to hospital mortality, ICU mortality, or mortality within some time interval (e.g. 30 days after admission). Prognostic models are developed on the premise that they can be useful for their intended purpose. The traditional use of prognostic models in the ICU is benchmarking in which
quality of care is compared among a set of ICUs. In this use the model should adjust for differences in patient case-mix. Predictions of prognostic models can also be used for risk stratification in clinical trials (e.g. to enroll or exclude high or low risk patients). Another potential use of prognostic models, which we are mainly concerned with here, is informing treatment decisions in individual patients. Existing models from the SAPS and APACHE families are mainly used for benchmarking purposes and for this reason they solely rely on data known at the time of ICU admission. Like these models, we use hospital mortality as our outcome and logistic regression as the form of our prognostic model. However, we include discovered patterns as possible predictors of outcome.

Model evaluation

In order to assess the credibility of a prognostic model it is imperative to evaluate its performance (often referred to as validating the model). The evaluation should be performed on a dataset other than the one used for developing the model otherwise the measured performance may be inflated. Relevant measures of predictive performance include discrimination (the ability to differentiate between those with different outcomes) and calibration (the ability to provide predicted probabilities that are equal to the relative frequencies of events). In this thesis we use the Area Under the Receiver Operating Characteristics Curve (AUC) as a measure of discrimination, and the Brier score, which possess both elements of discrimination and calibration.

Prognostic methods versus prognostic models

When evaluating or comparing predictive performance, one can distinguish between prognostic models and prognostic methods. A prognostic model is the result of applying an inductive predictive method. A method concerns the algorithm used to learn the model from a given training data. Comparing models originating from two different methods does not provide sufficient evidence for making conclusions about the quality of the corresponding inductive methods. In order to compare methods, one should account for the inherent sources of variation caused by e.g. the random splits that a method pre-
scribes for obtaining the data for training and for testing a model. An important modality for accounting for the possible sources of variation is the use of resampling techniques. In essence, the variability provided by these resamples from the given sample resembles the variability provided by drawing independent samples from the population. There are several resampling techniques available such as cross-validation and bootstrapping. In this thesis we make use of a bootstrap-based technique [11] in which a given inductive method, or in case of comparison two, are applied to each resample. Various statistics, such as (difference in) performance measures or the set of frequent patterns, are then obtained from each sample. This provides a distribution, the bootstrap distribution, of these statistics. This distribution can then be used for further inference (such as testing whether the differences in performance between two inductive methods are statistically significant).

1.2 Thesis Research Questions and organization

The contents of this thesis address the following Research Questions:

Research Questions

Question 1: How can sequential organ function data be integrated, in a symbolic representation, into models for predicting hospital mortality of ICU patients?

Question 2: How can one systematically evaluate methods for prediction, correcting for sources of variation in the model construction process?

Question 3: What is the increase in predictive performance when patterns of sequential organ failure assessment score are incorporated into prognostic models, and how do these models compare to the predictive abilities of ICU physicians?
Thesis structure

Chapter 2: Discovery and inclusion of SOFA score episodes in mortality prediction

Context: The existing predictive models in the ICU rely solely on data describing each patients severity of illness at arrival in the unit. The role of the daily SOFA score has not yet been well exploited in outcome prediction. If we can generate daily predictions for the eventual hospital survival status then the utility of predictive models for counseling patients and supporting clinicians in making informed decisions about withholding or withdrawing treatment may increase.

Addressed research questions: This chapter addresses question 1 and question 3.

Approach: A retrospective single-centre study in which we use the collected data to discover frequent sequential patterns (in the chapter sequential patterns are referred to as episodes) of SOFA scores. These patterns are subsequently used as potential predictors together with admission information to develop logistic regression models for predicting, at each day of stay, the hospital mortality risk. We investigate the predictive performance of the newly devised prognostic models and compare them to the traditional SAPS-II models, which are first re-calibrated to the same dataset used to fit the models with the frequent patterns. For example for day 5 of stay both models are fit on the same sample of patients staying at least 5 days.

Chapter 3: Discovery and integration of univariate patterns from daily individual organ-failure scores for Intensive Care mortality prediction

Context: The SOFA score is an aggregate score, and relies on six individual organ failure scores. The prognostic value of these individual scores has not been investigated before, although models enriched with patterns of individual scores may boost the prognostic accuracy of the models.
Addressed research questions: This chapter addresses question 1 and question 3.

Approach: A single-centre study in which we develop logistic regression models using a variable selection strategy from covariates including SAPS-II score, a summary of the SOFA score, and patterns from sequences of individual organ dysfunction scores. The patterns are obtained by a data-driven discovery algorithm performed for every day of prediction and for each of the 6 organ systems on temporal data from the day of admission up to and including the prediction day. Model performance is assessed in terms of the AUC and the Brier score and is used to establish the added value of temporal patterns to the quality of predictions compared to existing prognostic approaches.

Chapter 4: Frequent single and multiple organ failure in the Intensive Care: prevalence, persistence and prognostic value over days of stay

Context: Multiple organ failure (MOF) occurs when at least 2 individual organ systems fail (SOFA ≥ 3) at the same day. MOF increases mortality risk. The prevalence of frequent single and multiple organ failure over the course of time has not yet been investigated. In addition it is unclear whether the type of failing organ(s) improves predictive performance compared to models using only the number of organ failures on each day.

Addressed research questions: This chapter addresses question 1 and question 3.

Approach: Single and multiple organ failure patterns are discovered from all available single centre patient data. The patterns are used together with admission information as candidates for inclusion in daily logistic regression models for predicting hospital mortality up to day 7 of stay. The association of the MOF patterns to hospital mortality is investigated. The predictive performance (AUC and Brier score) of models including MOF patterns were compared to that of models using only the daily number of failures. A bootstrap based technique was used for the selection of the most predictive patterns,
the candidates for model inclusion, as well as for the evaluation of the models prognostic
performance.

Chapter 5: Learning Predictive Models that use Pattern Discovery – A Bootstrap Evaluative
Approach Applied in Organ Functioning Sequences

Context: In previous chapters we applied pattern-based inductive methods to data
and obtained daily models for predicting hospital mortality. The performance of these
models was then compared to that of the static models using only admission information.
Although this evaluation provides some evidence to the superiority of the pattern-based
models on the static models, it does not directly demonstrate the superiority of the
inductive method generating them.

Addressed research questions: This chapter addresses question 2.

Approach: We propose a setup for addressing the problem of prognostic method
evaluation. The approach is based on resampling, in particular a variation of bootstrap
called .632, for estimating the discrimination and calibration of pattern-based methods.
The setup for evaluation implies repeating all steps of an inductive method under inves-
tigation on the bootstrap samples. The evaluation of a methods accuracy is based on
the combined performance of its models on the original dataset and on the data not used
for training (“out-of-bag” cases) at each bootstrap iteration. The evaluation method
can also be used for comparing prognostic methods where the difference in performance
measures is calculated per bootstrap iteration. A case study is presented for comparing
the inductive method based on frequent temporal sequences of organ functioning to a
traditional inductive method that does not use patterns. The case study also illustrates
the insights gained by the analyst into the discovered frequent patterns from the bootstrap
samples.
Chapter 6: Assessing and Combining Repeated Prognosis of Physicians and Temporal Models in the Intensive Care

Context: The prognostic models generated so far were compared to the static models, exemplified in this work by SAPS-II. This gives an indication of the benefits of using patterns in mathematical models for prediction. However, physicians subjective mortality estimates remain the only source of temporal prognosis in use in the ICU. They have access to a larger variety of data to support their predictions and quickly react to changing situations but the temporal models are more consistent in making predictions. It is intriguing to investigate how the predictive performance of the mathematical models and the physicians' compare.

Addressed research questions: This chapter addresses question 3.

Approach: A prospective study where we compare the quality of repeated predictions on each day of stay, which were obtained for the same patients of an ICU, by temporal models and by physicians. Specifically we developed three types of temporal models. In addition to models using SAPS-II as baseline and the SOFA-patterns as temporal information, we considered a second type which enriches the patterns-based models with predictors representing SOFA summary measures such as the mean, maximum and delta SOFA (differences in the score). In the third model type we add to the covariates of the second model also subjective information consisting of summaries from the physicians' predictions such as mean, maximum and delta in expert opinion. Calibration and discrimination for all three temporal models and the experts estimates were assessed. To put the results in perspective we also fitted a classification tree model on the same data.
Chapter 7: Discussion and Conclusions

In this last chapter we summarize our findings per research question, and reflect on the strengths, weaknesses and significance of our work. We then delineate some important future work.

Bibliography


