DOI 10.1007/s10916-008-9234-9

ORIGINAL PAPER

Prediction of Clinical Conditions after Coronary Bypass Surgery using Dynamic Data Analysis

K. Van Loon · F. Guiza · G. Meyfroidt · J.-M. Aerts · J. Ramon · H. Blockeel · M. Bruynooghe · G. Van den Berghe · D. Berckmans

Received: 28 August 2008 / Accepted: 14 November 2008 / Published online: 6 January 2009

Abstract This work studies the impact of using dynamic information as features in a machine learning algorithm for the prediction task of classifying critically ill patients in two classes according to the time they need to reach a stable state after coronary bypass surgery: less or more than 9 h. On the basis of five physiological variables (heart rate, systolic arterial blood pressure, systolic pulmonary pressure, blood temperature and oxygen saturation), different dynamic features were extracted, namely the means and standard deviations at different moments in time, coefficients of multivariate autoregressive models and cepstral coefficients. These sets of features served subsequently as inputs for a Gaussian process and the prediction results were compared with the case where only admission data was used for the classification. The dynamic features, especially the cepstral coefficients (aROC: 0.749, Brier score: 0.206), resulted in higher performances when compared to static admission data (aROC: 0.547, Brier score: 0.247). The differences in performance are shown to be significant. In all cases, the Gaussian process classifier outperformed to logistic regression.

K. Van Loon · J.-M. Aerts · D. Berckmans (✉)
Division Measure, Model & Manage Bioresponses,
Katholieke Universiteit Leuven,
Kasteelpark Arenberg 30,
B-3001 Leuven, Belgium
e-mail: daniel.berckmans@biw.kuleuven.be

F. Guiza · J. Ramon · H. Blockeel · M. Bruynooghe
Department of Computer Science,
Celestijnenlaan 200a,
B-3001 Leuven, Belgium

G. Meyfroidt · G. Van den Berghe
Department of Intensive Care Medicine,
University Hospital Gasthuisberg,
Herestraat 49,
B-3000 Leuven, Belgium

Keywords Gaussian processes · Time series analysis · Intensive care · Binary probabilistic classifier · Cepstral coefficients

Introduction

In cardiac surgery, optimal use of intensive care unit (ICU) and operating room (OR) capacity requires the prediction of future availability of ICU beds. On the level of the management of the department, a number of beds are reserved for cardiac surgery patients. In order to manage the planning of the intensive care unit and the operating theatre, it would be very helpful to have a system that provides an early alert if there is a high probability that a patient will be disconnected from ventilation during the next day. When the patients are still ventilated, they cannot be sent to a normal ward, the bed does not become available and the surgeon cannot operate on new patients. This medically relevant prediction task, concerning the time instant on which patients can be disconnected from mechanical ventilation, is well suited for our research, where we want to focus, in the first place, on the impact of using dynamic information in the prediction task.

Information on vital signs such as heart rate, blood pressure, oxygenation, etc. is routinely gathered in the ICU. Continuous evaluation of the values of these variables starting from the arrival in the ICU is important because the alterations are relevant to patient management [1]. For our analysis, we wanted to use the trends of the vital signals during the first hours of ICU stay to predict a short or prolonged length of stay from early on. Prognostic models in medicine can be useful for various tasks: from capacity planning to individual patient interventions. For an overview of uses and development approaches in the statistical and artificial intelligence field, we refer to the work of Abu-Hanna et al. and Ohno-Machado et al. [2, 3].

Minimal conditions used to start the weaning from mechanical ventilation in these critically ill patients are: hemodynamic and respiratory stability, absence of bleeding and normothermia. A lot of different physiological parameters related to these criteria are measured, monitored and stored in a typical ICU. From research it was shown that humans have difficulties with interpreting and handling more than seven variables at the same time. On top of that, the interpretation of the information can differ between clinicians and interpretation of temporal data seems to be the most important problem [4]. So, the need for decision support in the medical environment is very high [5]. Medical diagnostic decision support systems already have been an established component of medical technology [6]. A number of quantitative models, including logistic regression, neural networks and many others, have been used in this kind of systems to assist human decision-makers in several applications [6, 7], e.g. in epileptic detections [8].

Since all living organisms are characterized by the fact that they are complex individually different time-variant and dynamic (so called CITD systems) [9], it is expected that taking these characteristics into account will lead to better models of the physiological signals of intensive care patients. For example, Cappi et al. used repeated measurements of Acute Physiology and Chronic Health Evaluation (APACHE) instead of only one score on the day of admission [10], since this APACHE score is based on physiological measurements on a certain moment in time and does not consider the evolution of the signals in time. Chang et al. predicted deaths among ICU patients on the basis of trend analysis of daily measured APACHE II scores that were corrected for organ system failure. They applied this approach because they were convinced that the patho-physiological processes affecting ICU patients are dynamic and cannot be reflected by a single assessment of a static score on the day of admission [11–13]. Also Clermont et al. used repeated static scores in their micro simulation model to predict temporal patterns of multiple outcomes on the basis of demographic variables and the Sequential Organ Failure (SOFA) scores on admission [14]. The work of Toma et al. describes a method that captures the temporal evolution of organ functioning which is quantified by SOFA scores or Individual Organ System Failure (IOSF) scores and uses these patterns in a logistic regression modeling framework [15, 16].

Instead of using repeated static scores as described before to obtain dynamic information about a patient, it is also possible to extract dynamically relevant features from the commonly measured physiological data itself. Since a lot of time series of physiological variables are available in the ICU environment, these signals could be well suited as inputs for different modeling techniques and for cepstral coefficient analysis. These techniques can all be used to analyze individual patients whose health status varies with time. So far, univariate autoregressive analyses of physiological variables have been applied in several studies in the field of intensive care medicine [17, 18]. Akaike used a multivariate autoregressive method for the identification of a multivariate feedback system [19]. A lot of systems, e.g. in the human body, can be explained using this kind of systems [20]. His method has been applied in several medical applications [21–23] and helps to detect the relationships between all variables included in the model. The calculation of cepstral coefficients is another possibility to extract the significant features from time series. Curcie et al., for example, used this technique to identify individual heart rate patterns [24].

For making classifications using many variables at the same time, several data mining techniques are available. It has been demonstrated that machine learning algorithms can analyze data from a collection of patients and can be trained to make predictions on new unseen patients. Machine learning algorithms have been used in a variety of medical applications [25] and have been shown to be specially valuable in data mining scenarios involving large databases and where the domain is poorly understood and therefore difficult to model by humans [26]. Intensive care is one of those domains that can benefit from the use of machine learning techniques [27]. In this field they have been used for prediction and classification tasks. For instance they have been used for classifying pressure-volume curves into different measurement methods for artificially ventilated patients suffering from the Adult Respiratory Distress Syndrome (ARDS) [28]. They have been shown to outperform to logistic regression in the task of classifying ICU patients with head injuries according to their outcome: good vs. poor Glasgow Coma Scores (GOS) and dead vs. alive [29]. In a different prediction task, Tong and colleagues successfully classified a neonatal ICU population according to ventilation duration, a study that extends their previous success with the same machine learning technique and classification task but on an adult ICU setting [30]. Giraldo and colleagues [31] classified respiratory patterns of patients on weaning trials into those that will succeed or fail to sustain spontaneous breathing. Gaussian processes (GP) have been applied to the problem of neonatal seizure detection from electroencephalograph (EEG) signals, where they are shown to outperform other modeling methods currently in clinical use for EEG analysis [32].

However, in the above cases no dynamic information about the patients is taken into account when applying the data mining approach although it is important and useful to capture and analyze the temporal aspects of the data as part of the knowledge discovery process [33]. Several attempts on temporal feature extraction for time series classification have been made [34–37]. According to Kadous et al. [38], abstracting temporal features is not a trivial task, especially not when it has to be done automatically. In this work, several automatically extracted representations of the dynamics of time series will be studied. Moreover, the classification results will be compared with the classification based on admission data only, while in the references cited above only monitoring data was considered, even though in clinical practice, static admission data plays an important role in the calculation of health evaluation scores such as APACHE scores.

The general objective of this study was to explore and quantify the prognostic value of dynamic information that was abstracted from time series data in various ways. More specifically, it was investigated whether the prediction of the timeframe in which the minimal clinical conditions to start weaning of the mechanical ventilation are reached, can be more accurately predicted by using dynamic information of the individual patients when compared to predictions on the basis of static admission data.

Materials and methods

Figure 1 gives an overview of the consecutive steps in our analyses. In this section the used signals, different types of time series analyses, the GP classifier and the prediction task are briefly explained.

Data generation

In the surgical ICU of the university hospitals of Leuven, 22 beds are reserved for cardiac surgery patients. We screened all patients admitted to the ICU after planned coronary bypass surgery, between February 2006 and December 2006 for this retrospective study. Ethics committee approval was obtained, and the need for informed consent was waived because of the retrospective nature of the study. We selected five physiological variables, routinely monitored in these patients (Philips Merlin monitor), to be used as inputs. Since we were focusing on the dynamics of the patients in this study, we took into account signals that were measured with the highest frequency (i.e. a sample interval of 1 min) in the Patient Data Management System (Metavision®, iMD-Soft®) and that were, on top of that, almost always measured and registered and showed enough variability. For an overview, see Table 1. Data of a total of 203 patients was used for analysis.

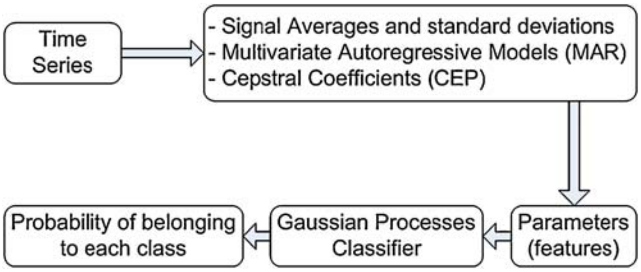
Time Series
- Signal Averages and standard deviations
- Multivariate Autoregressive Models (MAR)
- Cepstral Coefficients (CEP)
Parameters (features)
Gaussian Processes Classifier
Probability of belonging to each class

Fig. 1 Schematic overview of the analyses performed in this research

For these patients also admission data was used (see Table 2). For this, parameters from the Parsonnet score [39] and Euroscore [40] were selected, as far as they were available. Both scores have been shown to be predictive for ICU length of stay. The following seven variables were taken into account: age, sex, body mass index (BMI), normal lung function, diabetes, creatinine level, and NYHA class. The NYHA (New York Heart Association) classifies the extent of heart failure and ranges from I (no symptoms or limitations) to IV (severe limitations).

Modeling analysis

Abstraction of dynamic information

In order to quantify the dynamics of the patients' physiological variables, we used the mean and standard deviations of the signals (Avgstd), we applied multivariate autoregressive models (MAR) and calculated cepstral coefficients (CEP). The latter two are explained in more detail in this section.

Multivariate autoregressive models (MAR) A time series is a sequence of observations taken sequentially in time. Most time series consist of elements that are serially dependent. A common approach for analyzing this dependence is the AR model. In this type of model, a coefficient or a set of coefficients is estimated that describes the association between consecutive elements of the series [41]. The general equation of a multivariate autoregressive model (MAR) can be written as

$$Y(t) = \sum_{m=1}^{M} A(m)Y(t-m) + E(t) \qquad (1)$$

Every observation is made up of a linear combination of M prior observations (the order of the model) and a white noise term, which is a vector of mutually independent white noises. $Y(t) = [y_1(t), y_2(t), \ldots, y_K(t)]$ is the vector of simultaneously measured values at time t for K variables, in this case all variables of Table 1, and $E(t) = [e_1(t), e_2(t), \ldots, e_K(t)]$ is a prediction error vector. The

Table 1 Physiological variables

Var nr	Physiological variable	Unit	Sampling frequency
1	Arterial blood pressure, systolic	mmHg	1 / min
2	SpO_2 (oxygen saturation in arterial blood flow)	%	1 / min
3	Heart rate	bpm	1 / min
4	Blood temperature	°C	1 / min
5	Arterial pulmonary pressure, systolic	mmHg	1 / min

generation of the AR models was performed using the ARfit package for Matlab [42]. The matrices $A(m)$ are the MAR coefficients and are estimated using a stepwise least squares algorithm. In this study, the coefficients of matrix A are used as features in the data mining (cfr. Fig. 1) since they describe the dynamics of the considered system.

Cepstral coefficients (CEP) Cepstrum analysis is a nonlinear signal processing technique with a variety of applications in areas such as speech and image processing. The cepstrum is defined as the inverse Fourier transform of the short-time logarithmic amplitude spectrum [43, 44]. More detailed, the real cepstrum for a sequence x is given by the sequence y:

$$y = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left| X\left(e^{j\omega}\right) \right| e^{j\omega t} d\omega \tag{2}$$

where $X\left(e^{j\omega}\right)$ is the Fourier transform of y.

The difference between the cepstral coefficients of different time series can be used as a similarity measure between these time series. Cepstral coefficients decay rapidly to zero, so only the first few coefficients are needed to capture most of the dynamic information in the time series. An example of the cepstrum of the heart rate signal of one patient is shown in Fig. 2. Because of the good clustering results of Kalpakis et al. [43] on the basis of cepstral coefficients, it is interesting to use these coefficients as input features in the data mining analysis as an alternative summary of the dynamics of the signals. Moreover, other techniques based on frequency information, such as the calculation of wavelet coefficients [34], have been applied for the summarization of data. Given the good results of Zhang et al. it is worthwhile to explore and use frequency information (such as cepstrum coefficients) in the classification task.

Gaussian processes for classification

Gaussian processes [45], a type of kernel method, are a machine learning technique that has been successfully used to model and forecast real dynamic systems because of their flexible modeling abilities and their high predictive performances. They allow for multi-dimensional inputs and they assign a confidence value to their predictions. The main advantage of using a GP classifier over other kernel method classifiers is that it produces an output with a clear probabilistic interpretation [46].

In probabilistic binary classification the task is to determine for an unlabeled test input vector x_* the probability of belonging to the class $C: \pi_c(x_*) = p(t_* = 1|x_*)$ when a training set $\{X, t\}$ is given. The training set is comprised of N training input vectors $X = \{x_1, x_2, \ldots, x_N\}$ and their corresponding N binary class labels $t = \{t_1, t_2, \ldots, t_N\}$ such that t_i=+1 if x_i belongs to a given class C and t_i=–1 if x_i does not belong to the class. The probability that x_* does not belong to the class can then be computed as $p(t_* = -1|x_*) = 1 - \pi_c(x_*)$. In the remainder of this text the input vectors X will be referred to as *examples*.

Table 2 The population description table

	Class 1 (Criteria met<=9 h)	Class 2 (Criteria met>9 h)
Number of patients	102	101
Age (mean ± std)	66±11	66±10
Sex (male/female) %	66% / 34%	84% / 16%
BMI (mean±std)	27.7±4.6	27.7±3,8
Normal lung function (%)	90%	84%
Diabetes (%)	65%	68%
Creatinine (mg/dL) (mean ± std)	1,15±0,45	1,18±0,40
NYHA class (I/II/III/IV) (%)	55% / 28% / 12% / 5%	61% / 19% / 19% / 1%

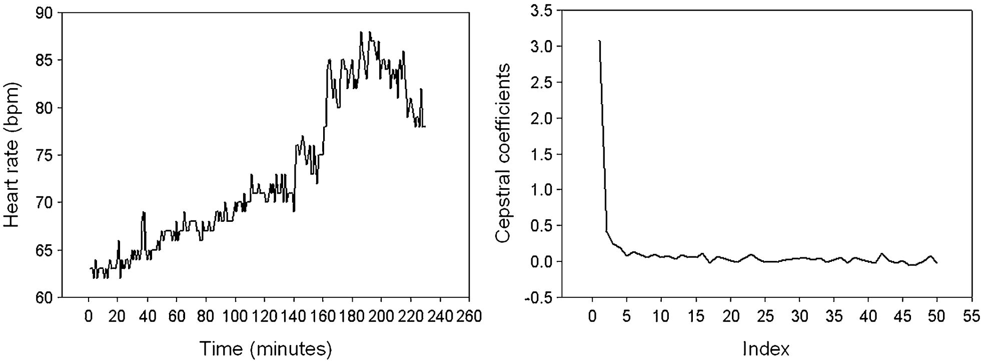

Fig. 2 A heart rate signal in beats per minute of 230 samples. Right: The corresponding cepstrum truncated at 50 cepstral coefficients

In GP binary classification [47], a GP over a function *f*(*x*) is defined and then transformed through a logistic or *squashing* function $\sigma(.\ )$ so that its outputs lie in the [0,1] interval, and can be thus interpreted as probabilities: $\pi_c(x) := p(t = +1|x) = \sigma(f(x))$. Conditioning the predictive distribution on the training data allows for a probabilistic prediction on a test input example [48].

A GP is a distribution over functions and is a natural generalization of a Gaussian Distribution, the latter of which has a vector as mean and as covariance a matrix. The GP over a function is accordingly specified by a mean function and a covariance function. The covariance function is given by a positive semi-definite kernel function $k(x_i, x_j)$. The covariance function determines the properties of the function distribution in the GP, for example it can impose smoothness so that nearby inputs x_i, x_j have similar values $f(x_i)$, $f(x_j)$, with high probability.

Learning from data in the GP case means to modify the function distribution by conditioning it on the observed data. This modified or posterior function distribution has a mean function that coincides with the target values when evaluated on the training examples.

Figure 3 shows a GP learned from the one-dimensional training data depicted with crosses. The shaded area corresponds to the 95% confidence region learned for the function distribution; it can be seen that the uncertainty of the prediction grows in regions where there are few training points. Figure 4 shows a cut-section of the predicted distribution for the test input at -6, which has a mean predicted value of 1.92. Also shown (dashed line) is the predicted distribution before training, which has a mean predicted value of 0 and is very broad to reflect the uncertainty associated with this prediction. Once learning has occurred, the predictions become more certain because data has been seen in the vicinity of the test point, and the predictions must be consistent with these observations.

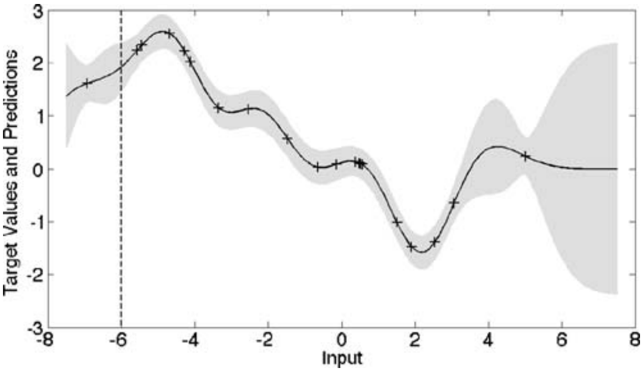

Fig. 3 Gaussian Process learned from the one-dimensional training data depicted with crosses. The *shaded area* corresponds to the 95% confidence region learned for the function distribution, and *bold line* indicates the mean predictions. The *dashed line* indicates a test point more thoroughly studied in Fig. 4

Given that the GP is defined by its covariance function, and that the covariance or kernel function is defined by a set of parameters (referred to as hyper parameters), then training the GP amounts to finding the values of the hyper parameters such that the probability of the data given these hyper parameters is maximized.

Because of the inclusion of the logistic function $\sigma(.\ )$ required for classification, the inference of the predictive or posterior distribution requires the solution to integrals which are analytically intractable, a problem that is solved either by resorting to Monte Carlo sampling or analytical approximations to the integrals. In this study we follow the latter approach through the use of expectation propagation [49].

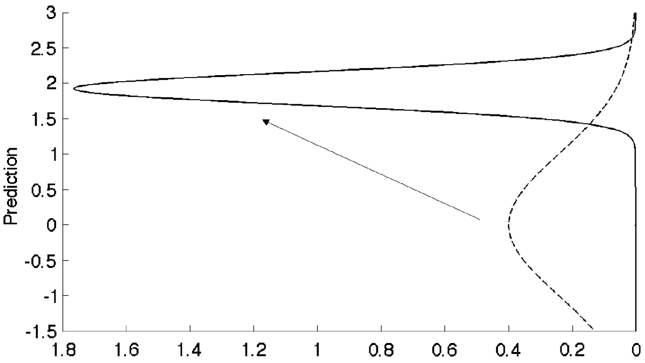

Fig. 4 Predicted distribution for the test input at -6, with a mean predicted value of 1.92. *Dashed line* is the predicted distribution before training, which is very broad and has a mean predicted value of 0

The covariance function used in this study is the so called rational quadratic with ARD (automatic relevance determination) defined as follows:

$$k(x_i, x_j) = \sigma_f^2 \left[1 + \frac{(x_i - x_j)^T M (x_i - x_j)}{2\alpha} \right]^{-\alpha} \tag{3}$$

Recall that each example x corresponds to a vector obtained from the different time series models. In this equation M=*diag*$(l)^{-2}$, and the $l_1, l_2, \ldots, l_D$ parameters in the diagonal matrix are characteristic length-scales for each dimension of the input examples. The σ_f is the signal variance of the process, which controls its magnitude and α is the shape parameter. Learning or training the GP amounts to finding the values for the parameters $\theta = \{\sigma_f, \alpha, l_1, l_2, \ldots, l_D\}$ (which are iteratively updated according to the expectation propagation algorithm) so as to maximize the likelihood of the class labels given the training data [46]. The values of the parameters of the diagonal matrix M determine the relevance of the corresponding input dimension. If after training, a length-scale has a very large value, the covariance will become almost independent of that input dimension. This ARD covariance function has been found in other works to successfully remove uninformative input dimensions [50]. Also, the increase in degrees of freedom of the ARD covariance function given by the increase in hyper parameters allows for more complex mappings between the inputs and the targets to be found.

It has been shown [46] that in the limit $\alpha \rightarrow \infty$, the covariance function of Eq. (3) converges to the squared exponential covariance function, one of the most frequently used covariance function in kernel methods. The rational quadratic covariance function can thus be seen as an infinite sum of squared exponential covariance functions with different characteristic length-scales. A detailed description of commonly used covariance functions can be found in the work of Rasmussen et al. [46].

Protocol

Prediction task

Can we predict the time frame in which the patients fulfill the criteria for stability that will lead to weaning from mechanical ventilation? In our ICU, cardiac surgery patients are weaned off the ventilator using a protocol. In this protocol, the following criteria have to be met before sedation can be switched off: hemodynamic stability (dobutamine ≤5 μg/kg/min, levophed ≤0,2 μg/kg/min and lactate <2 mmol/L), respiratory stability (the oxygen saturation in arterial blood flow (PaO_2) ≥75 mmHg, the fraction of inspired oxygen concentration (FiO_2) ≤0.5, the positive end-expiratory pressure (PEEP) ≤8 mbar), temperature stability (blood temperature >36°C, peripheral temperature >30°C) and blood loss stability (sum of blood loss of all drains <100 ml/h).

To enable future comparisons with predictions performed by intensivists, the considered task was restated as follows: Predict the probability that the patient will begin to satisfy the stability criteria within each of the following time frames (classes): class 1: earlier than 9 h after admission; class 2: later than 9 h after admission. This 9 h threshold was chosen such that the resulting classes contained roughly same amount of patients. In class 1 there was a total of 102 patients and class 2 contained 101 patients. These classes also conform to an intuitive classification into patients that recover quickly and those that require prolonged ICU stays.

Preprocessing

Before doing any analysis, the signals were normalized: the mean was put to zero and the standard deviation to one. Furthermore, the recorded time series contained a limited number of missing values or artifacts, usually due to sensor disconnections. The missing data points were calculated using linear interpolation. In order to remove these artifacts, a peak-shaving algorithm was applied. This algorithm consisted of three major parts. In the first step, the trend of the original time series was calculated. Secondly, an upper and lower bound were computed as the trend plus and minus four times the standard deviation of the trend respectively. In the third step, values of the original signal that did not lie in between the lower and upper bound were replaced by linearly interpolated values calculated from the previous and next value that lay in between the two borders.

In total, the inclusion of the missing values and the removal of artifacts affected 1.9% of all data points.

Time series models

Data from each patient, collected during the first 4 h ICU stay, were used to generate the different time-series models, the parameters of which were used as the features of the examples. One of the two possible class labels was assigned to each example. Figure 5 shows data from one patient used to generate a training example, and how the appropriate class label was assigned.

On the one hand, sufficient data points should be taken into account in the modeling process. On the other hand, the sooner after admission of the patient a reliable prediction can be made about the extubation time, the better. Therefore, an interval duration of 4 h was chosen for

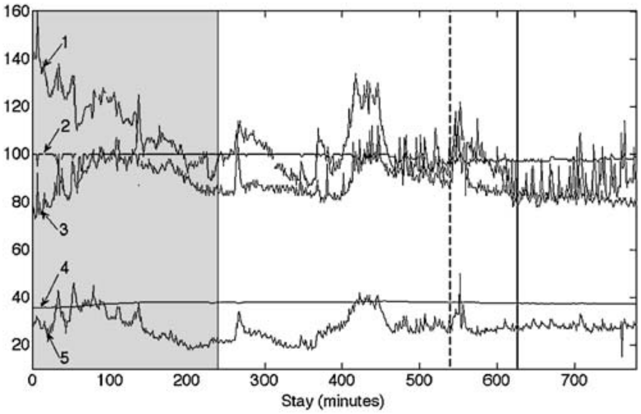

Fig. 5 The *gray area* corresponds to 4-hour interval of data used to generate the example. The signals are numbered according to Table 1 (1: arterial blood pressure, 2: SpO2, 3: heart rate, 4: blood temperature, 5: arterial pulmonary pressure). The *dashed vertical line* depicts the 9-hour class-boundary and the *solid vertical line* indicates the moment when the patient satisfies the stability criteria (minute 627). The example generated from this data is labeled as belonging to Class 2 (Stability criteria met after 9 h)

Table 3 The aROC's and Brier scores for all experiments

aROC / Brier score	LOGREG	GP
Admission (7)	0.543	0.547
	0.249	0.247
Avgstd (20)	0.628	0.713
	0.241	0.214
MAR (25)	0.591	0.708
	0.250	0.219
CEP_4 (20)	0.542	0.707
	0.250	0.218
CEP_5 (25)	0.542	0.749
	0.247	0.206

The first column gives the logistic regression results, the second column the results of the Gaussian process classifier

our analysis. Shorter time intervals led to non-stable MAR models. The different time-series analysis techniques described above, were applied to each of the 4-hour intervals of data for each patient in order to generate the examples used as inputs for the GP classifier. In order to avoid over-fitting, the dimension of the examples should be kept low enough. According to traditional rules of thumb, 5 to 10 observations are required for each parameter to be estimated [51]. This leads to maximum number of parameters between 18 and 36 in our 10-fold cross-validation schema (explained below) for 203 patients. The types of examples (input vectors) used to train the GP classifiers in our experiments are explained below. They were designed in a way that the rule above is not violated.

Signal Average and standard deviation (Avgstd) Each example is a 20 dimensional vector containing four values for each of the five physiological variables of Table 1. For two intervals of 2 h the mean and the standard deviation were calculated for each signal.

MAR coefficients All five variables of Table 1 were used as input of a first order MAR model. The first order was chosen in order to keep all models as simple and compact as possible. Moreover, higher order models would lead to examples of high dimensions and in that case there is a higher chance for over-fitting (cfr. supra). So, matrix A of Eq. 1 was a 5×5 matrix of which all 25 parameters were put in a 25 dimensional vector that served as input example of the GP.

Cepstral coefficients (CEP) Each example contained the four (CEP_4) or five (CEP_5) first cepstral coefficients of all variables in Table 1, i.e. the four or five first numbers of the sequence y of Eq. 2. This resulted in a 20 or 25 dimensional vector respectively.

Gaussian processes

A binary probabilistic classifier was learned for class 1, such that for each patient a probability p of belonging to the class was obtained, the probability of belonging to class 2 could readily be determined as $1-p$. Training examples for each classifier were labeled positive (t=+1) if the moment when the patient became stable started within the corresponding time interval and were labeled negative (t= −1) otherwise.

All examples generated for all patients from one type of time series model and their corresponding class labels were collected in one dataset. The dataset was randomly split into 10-folds, 1 fold was removed and used as test set, while the data from the remaining folds was used as training for the classifier. Once the classifier has been trained, the predicted probability of belonging to class 1 was determined for each example in the test set. The described process was repeated for each of the 10 folds so that a probability of belonging to each class was assigned to each of the N patients. In other words, a 10-fold cross-

Table 4 The statistical significance of the differences in performance between the GP classifier and the logistic regression shown for the Brier scores as well as the aROC's

INPUT	Brier score	aROC
Admission	No	No
Avgstd	Yes	Yes
MAR	Yes	Yes
CEP_4	Yes	Yes
CEP_5	Yes	Yes

Table 5 The statistical significance of the differences between the different GP classifiers

Brier score / aROC	Admission	Avgstd	MAR	CEP_4	CEP_5
Admission	–	Yes/yes	Yes/yes	Yes/yes	Yes/yes
Avgstd		–	No/yes	No/no	Yes/yes
MAR			–	No/no	Yes/yes
CEP_4				–	Yes/yes

validation was performed. The obtained probabilities allowed for the computation of an aROC (area under the receiver operating characteristic curve) for each classifier. If a hard-classification is required, each patient would be assigned to the class for which it had the highest probability. To evaluate the calibration of the predicted probabilities the Brier Score [52] was also computed.

To evaluate whether there was statistical significance between the differences in performance of the classifiers two approaches were followed. Regarding the Brier scores, a non-parametric bootstrap method [53] was used to generate a bootstrap distribution of 1,000 samples of mean differences, from which a 95% confidence interval could be determined based on the 2.5% and 97.5% quartiles. If the confidence interval did not include 0, then a statistically significant difference at the 0.05 level was declared. Regarding the aROC scores the non-parametric method described in DeLong et al. [54] was implemented to determine significance at the 0.05 level.

Results and discussion

Table 3 gives the obtained aROCs as well as the Brier scores for each experiment. The left column of 3 contains to the results of the corresponding GP probabilistic binary classifier with the covariance function of Eq. 3. The right column contains the results obtained when using a logistic regression (LOGREG) model [55], included here as a baseline for performance.

The main goal of this research was to investigate the prognostic value of dynamic information abstracted in various ways when predicting how much time a critically ill patient needs to reach a stable state after coronary bypass surgery. Five physiological variables were considered, not including demographic or historical patient information. A separate model on the basis of admission data was developed for comparison purposes.

Table 4 shows that the increase in performance for all GP models versus the LOGREG models was found to be significant, except for the model based on admission data for which the difference in performance was not statistically significant. So, although logistic regression techniques are commonly used in medical applications, other classifiers might lead to better results. This was, among others, also concluded by Sakai et al. who found that artificial neural networks have a higher level of accuracy than logistic regression models for the diagnosis of acute appendicitis [56]. In another study about assessing the posttraumatic cerebral hemodynamia in minor head injured patients, Erol et al. obtained better classification results with multi-layer perceptron neural networks than with logistic regressions [57].

The statistical significances of the GP are shown in Table 5. From this it is clear that all dynamic models perform better than the model purely based on admission information, with respect to both the Brier score and aROC. In Table 4, the GP with 5 cepstral coefficients (CEP_5) had the best performance (lowest Brier score and highest aROC). From Table 5 it can be seen that the difference in performance is shown to be significant. This agrees with our assumptions that it is a promising approach towards feature extraction for time-series prediction tasks. Only the first five cepstral coefficients seem to contain enough information to result in a good classification, which is consistent with the findings of Kapalkis et al. [43]. The poor performance of the models based on static information alone can be attributed to the similarity of these parameters for the two classes in our particular population (see Table 2).

There is no statistically significant difference in performance between CEP_4 and MAR or between CEP_4 and

Table 6 The statistical significance of the differences between the different logistic regressions

Brier score /aROC	Admission	Avgstd	MAR	CEP_4	CEP_5
Admission	–	No/yes	No/yes	No/no	Yes/no
Avgstd		–	No/yes	No/yes	No/yes
MAR			–	No/yes	No/yes
CEP_4				–	No/no

Avgstd. With respect to the Brier score there is no statistically significant difference between MAR and Avgstd, but there is one regarding aROC values (in favor of Avgstd).

Table 6 gives the statistical significances for the logistic regression models. From this table can be concluded that there is no significant statistical difference between any of the two models with respect to the Brier score. When considering the aROC's, Avgstd has the best performance with statistical significance.

A possibility to improve the results is to combine the parameters of several dynamic analysis techniques in one input example for the GP classifier, or to combine static admission data and dynamic information of the first hours in the ICU.

To improve on the generalization capabilities of the classifiers it would also be of use to increase the number of patients used during training. When more patients are included, the models can be trained on more features what possibly results in better performances while over-fitting is still avoided. This increase both in the number of physiological variables and patients will however require more complex implementations of the algorithms presented such that they are able to cope with the data increase while still remaining computationally tractable. Possible variants of the GP classifier include the use of sparse methods, aggregation, dimensionality reduction techniques and the inclusion of more specialized kernels that better incorporate the available prior knowledge.

To our knowledge, the work of Verduijn et al. [35] is most closely related to our study. They compared two temporal abstraction procedures, one that resulted in symbolic descriptions of the data and one that resulted in numerical mate features. These procedures were applied to monitoring data from the ICU for the estimation of the risk of prolonged mechanical ventilation after cardiac surgery. The defined the outcome as "mechanical ventilation longer than 24 h" and used high frequently measured physiological data as well as laboratory values of the first 12 h in the ICU. The main conclusion of their work was that induction of numerical meta features is preferable to extraction of symbolic meta features using existing clinical concepts. These results compliment our own findings, in which for a particular population, extracted dynamic features can be used as predictors that outperform more typically used clinical concepts such as static admission data.

Conclusion

In this study, the use of dynamic information, obtained from physiological signals in various ways, was investigated for the prediction task about the future stability of ICU patients, resulting in weaning of mechanical ventilators. For every patient a probability of belonging to each of two classes was assigned. Each class was defined according to the time needed to reach a stable state after coronary bypass surgery: less or more than 9 h. For this prediction, dynamic data from the first 4 h of the patient's ICU stay were included and results were compared to a model built upon admission data only. The main conclusion of this work is that it is preferable to use dynamic information of the first few hours after admission in the ICU above using only static admission data for the considered prediction task. All models based on dynamic information preformed better with respect to aROC's and Brier scores and the differences were found to be significant. When compared to logistic regression, the Gaussian process classifier results in better performances in all cases.

References

1. Rivera-Fernandez, R., Nap, R., Vazquez-Mata, G., and Miranda, D. R., Analysis of Physiologic Alterations in Intensive Care Unit Patients and Their Relationship With Mortality. *J. Crit. Care*. 22:120–128, 2007. doi:10.1016/j.jcrc.2006.09.005.
2. Abu-Hanna, A., and Lucas, P. J. F., Prognostic Models in Medicine-Ai and Statistical Approaches. *Methods Inf. Med*. 40:1–5, 2001.
3. Ohno-Machado, L., Resnic, F. S., and Matheny, N. E., Prognosis in Critical Care. *Annu. Rev. Biomed. Eng*. 8:567–599, 2006. doi:10.1146/annurev.bioeng.8.061505.095842.
4. Mcclish, D. K., and Powell, S. H., How Well Can Physicians Estimate Mortality in a Medical Intensive-Care Unit. *Med. Decis. Making*. 9:125–132, 1989. doi:10.1177/0272989X8900900207.
5. Stacey, M., and Mcgregor, C., Temporal Abstraction in Intelligent Clinical Data Analysis: A Survey. *Artif. Intell. Med*. 39:1–24, 2007. doi:10.1016/j.artmed.2006.08.002.
6. Ubeyli, E. D., Combining Neural Network Models for Automated Diagnostic Systems. *J. Med. Syst*. 30:483–488, 2006. doi:10.1007/s10916-006-9034-z.
7. Lee, N. K., and Lee, C. W., A Neural Network Application to Classification of Health Status of Hiv/Aids Patients. *J. Med. Syst*. 21:87–97, 1997. doi:10.1023/A:1022890223449.
8. Srinivasan, V., Eswaran, C., and Sriraam, N., Artificial Neural Network Based Epileptic Detection Using Time-Domain and Frequency-Domain Features. *J. Med. Syst*. 29:647–660, 2005. doi:10.1007/s10916-005-6133-1.
9. Quanten, S., De Valck, E., Mairesse, O., Cluydts, R., and Berckmans, D., Individual and Time-Varying Model Between Sleep and Thermoregulation. *J. Sleep Res*. 15:243–244, 2006. doi:10.1111/j.1365-2869.2006.00519.x.
10. Cappi, S. B., Sakr, Y., and Vincent, J. L., Daily Evaluation of Organ Function During Renal Replacement Therapy in Intensive Care Unit Patients With Acute Renal Failure. *J. Crit. Care*. 21:179–183, 2006. doi:10.1016/j.jcrc.2005.07.003.
11. Chang, R. W. S., Jacobs, S., and Lee, B., Predicting Outcome Among Intensive-Care Unit Patients Using Computerized Trend Analysis of Daily Apache-Ii Scores Corrected for Organ System Failure. *Intensive Care Med*. 14:558–566, 1988. doi:10.1007/BF00263530.
12. Chang, R. W. S., Jacobs, S., Lee, B., and Pace, N., Predicting Deaths Among Intensive-Care Unit Patients. *Crit. Care Med*. 16:34–42, 1988.

13. Chang, R. W. S., Individual Outcome Prediction Models for Intensive-Care Units. *Lancet*. 2:143–146, 1989. doi:10.1016/S0140-6736(89)90193-1.
14. Clermont, G., Kaplan, V., Moreno, R., Vincent, J. L., Linde-Zwirble, W. T., Van Hout, B. et al, Dynamic Microsimulation to Model Multiple Outcomes in Cohorts of Critically Ill Patients. *Intensive Care Med*. 30:2237–2244, 2004. doi:10.1007/s00134-004-2456-5.
15. Toma, T., Abu-Hanna, A., and Bosman, R. J., Discovery and Integration of Univariate Patterns From Daily Individual Organ-Failure Scores for Intensive Care Mortality Prediction. *Artif. Intell. Med*. 43:47–60, 2008. doi:10.1016/j.artmed.2008.01.002.
16. Toma, T., Abu-Hanna, A., and Bosman, R. J., Discovery and Inclusion of Sofa Score Episodes in Mortality Prediction. *J. Biomed. Inform*. 40:649–660, 2007. doi:10.1016/j.jbi.2007.03.007.
17. Imhoff, M., Bauer, M., Gather, U., and Lohlein, D., Statistical Pattern Detection in Univariate Time Series of Intensive Care on-Line Monitoring Data. *Intensive Care Med*. 24:121305–1314, 1998. doi:10.1007/s001340050767.
18. Lambert, C. R., Raymenants, E., and Pepine, C. J., Time-Series Analysis of Long-Term Ambulatory Myocardial-Ischemia—Effects of Beta-Adrenergic and Calcium-Channel Blockade. *Am. Heart J*. 129:677–684, 1995. doi:10.1016/0002-8703(95)90315-1.
19. Akaike, H., On Use of a Linear Model for Identification of Feedback Systems. *Ann. I. Stat. Math*. 20:425—438, 1968.
20. Jones, R. W., *Principles of biological regulation: an introduction to feedback systems*. Academic Press, Inc., New York, p. 359, 1973.
21. Wada, T., Akaike, H., Yamada, Y., and Udagawa, E., Application of Multivariate Autoregressive Modeling for Analysis of Immunological Networks in Man. *Comput. Math. Appl*. 15:713–722, 1988. doi:10.1016/0898-1221(88)90125-3.
22. Wada, T., Sato, S., and Matsuo, N., Application of Multivariate Autoregressive Modeling for Analyzing Chloride Potassium Bicarbonate Relationship in the Body. *Med. Biol. Eng. Comput*. 31:S99–S107, 1993. doi:10.1007/BF02446657.
23. Wada, T., Yamada, H., Inoue, H., Iso, T., Udagawa, E., and Kuroda, S., Clinical Usefulness of Multivariate Autoregressive: (Ar) Modeling as a Tool for Analyzing Lymphocyte-T Subset Fluctuations. *Math. Comput. Model*. 14:610–613, 1990. doi:10.1016/0895-7177(90)90254-K.
24. Curcie, D. J., and Craelius, W., Recognition of Individual Heart Rate Patterns With Cepstral Vectors. *Biol. Cybern*. 77:103–109, 1997. doi:10.1007/s004220050371.
25. Lavrač, N., Selected techniques for data mining in medicine. *Artif. Intell. Med*. 16:3–23, 1999. doi:10.1016/S0933-3657(98)00062-1.
26. Mitchell, T., *Machine Learning*. McGraw-Hill, New York, 1997.
27. Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M. et al, Mining data from intensive care patients. *Adv. Eng. Inform*. 21:3243–256, 2007. doi:10.1016/j.aei.2006.12.002.
28. Ganzert, S., Guttmann, J., Kersting, K., Kuhlen, R., Putensen, C., Sydow, M. et al, Analysis of respiratory pressure-volume curves in intensive care medicine using inductive machine learning. *Artif. Intell. Med*. 26:1–269–86, 2002. doi:10.1016/S0933-3657(02)00053-2.
29. Andrews, P., Sleeman, D., McQuatt, A., Corruble, V., Jones, P.A., Howells, T., et al., Decision tree analysis of data from a neurological intensive care unit. *Proceedings of the international conference on Artificial Intelligence in Medicine*. 1999.
30. Tong, Y., Frize, M., and Walker, R., Extending ventilator duration estimations approach from adult to neonatal intensive care patients using artificial neural networks. *IEEE T. Inf. Technol. B*. 6:2188–191, 2002. doi:10.1109/TITB.2002.1006305.
31. Giraldo, B., Garde, A., Arizmendi, C., Jané, R., Benito, S., Diaz, I., Ballesteros, D., Support Vector Machine Classification Applied on Weaning Trials Patients. *Proceedings of the 28th IEEE EMBS Annual International Conference*. 5587–5590, 2006.
32. Faul, S., Gregorcic, G., Boylan, G., Marnane, W., Lightbody, G., and Connolly, S., Gaussian Process Modeling of EEG for the Detection of Neonatal Seizures. *IEEE T. Bio-med. Eng*. 54:122151–2162, 2007.
33. Roddick, J. F., and Spiliopoulou, M., A survey of temporal knowledge discovery paradigms and methods. *IEEE Trans. Knowl. Data Eng*. 14:750–767, 2002. doi:10.1109/TKDE.2002.1019212.
34. Zhang, H., Ho, T. B., Lin, M. -S., and Liang, X., Feature extraction for time series classification using discriminating wavelet coefficients. In: Wang, J., Yi, Z., Zurada, J. M., Lu, B. -L., and Yin, H. (Eds.), *Proceedings of the third international symposium on neural networks*Springer, Berlin, pp. 1394–1399, 2006.
35. Verduijn, M., Sacchi, L., Peek, N., Bellazzi, R., de Jonge, E., and de Mol, B. A., Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data. *Artif. Intell. Med*. 41:11–12, 2007. doi:10.1016/j.artmed.2007.06.003.
36. Bellazzi, R., Larizza, C., and Riva, A., Temporal abstractions for interpreting diabetic patients monitoring data. *Intell. Data Anal*. 2:1–15, 1998. doi:10.1016/S1088-467X(98)00020-1.
37. Guler, N. F., and Kocer, S., Use of Support Vector Machines and Neural Network in Diagnosis of Neuromuscular Disorders. *J. Med. Syst*. 29:271–284, 2005. doi:10.1007/s10916-005-5187-4.
38. Kadous, M. W., and Sammut, C., Classification of multivariate time series and structured data using constructive induction. *Mach. Learn*. 58:179–216, 2005. doi:10.1007/s10994-005-5826-5.
39. Lawrence, D. R., Valencia, O., Smith, E. E. J., Murday, A., and Treasure, T., Parsonnet Score Is a Good Predictor of the Duration of Intensive Care Unit Stay Following Cardiac Surgery. *Heart*. 83:429–432, 2002. doi:10.1136/heart.83.4.429.
40. Nashef, S. A. M., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., and Salamon, R., European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardio-thorac*. 16:9–13, 1999.
41. Box, G. E., Jenkins, G. M., and Reinsel, G. C., *Time series analysis: forecasting and control*. Prentice-Hall International, New Jersey, 1994.
42. Schneider, T., and Neumaier, A., Algorithm 808: ARfit—A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans. Math. Softw*. 27:58–65, 2001. doi:10.1145/382043.382316.
43. Kapalkis, K., Gada, D., and Puttagunta, V., Distance measures for effective clustering of ARIMA time-series. *Proceedings IEEE International Conference on Data Mining* (ICDM 2001):273–280, 2001.
44. Rangayyan, R. M., *Biomedical Signal Analysis: a Case Study Approach*. Wiley Interscience, New York, 2002.
45. Seeger, M., Gaussian Processes for Machine Learning. *Int. J. Neural Syst*. 14:269–106, 2004. doi:10.1142/S0129065704001899.
46. Rasmussen, C. E., and Williams, C., *Gaussian Processes for Machine Learning*. MIT, Cambridge, 2006.
47. Williams, C. K. I., and Barber, D., Bayesian Classification with Gaussian Processes. *IEEE T. Pattern*. 20:121342–1351, 1998. doi:10.1109/34.735807.
48. Bishop, C. M., *Pattern Recognition and Machine Learning*. Springer, 2006.
49. Minka, T. P., *A Family of Algorithms for Approximate Bayesian Inference*. PhD Thesis, Masachusetts Institute of Technology, 2001.
50. William, C. K. I., and Rasmussen, C. E., Gaussian Processes for Regression. *Proc. Adv. Neural Inf. Process. Syst*. 9:514–520, 1996.
51. Schwarzer, G., Vach, W., and Schumacher, M., On the misuses of artificial neural networks for prognostic and diagnostic classification

in oncology. *Stat. Med*. 19:4541–561, 2000. doi:10.1002/(SICI)1097-0258(20000229)19:4<541::AID-SIM355>3.0.CO;2-V.

52. Hand, D. J., *Construction and assessment of classification rules*. Wiley & Sons, Chichester, England, 1997.
53. Efron, B., and Tibshirani, R. J., *An introduction to the bootstrap, vol. 57 of monographs on statistics and applied probability*. Chapman and Hall CRC, Boca Raton, 1993.
54. DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L., Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 44:3837–845, 1988. doi:10.2307/2531595.
55. Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning, Data Mining, Ingerence and Prediction*. Springer, 2001.
56. Sakai, S., Kobayashi, K., Toyabe, S. I., Mandai, N., Kanda, T., and Akazawa, K., Comparison of the Levels of Accuracy of an Artificial Neural Network Model and a Logistic Regression Model for the Diagnosis of Acute Appendicitis. *J. Med. Syst*. 31:357–364, 2007. doi:10.1007/s10916-007-9077-9.
57. Erol, F. S., Uysal, H., Ergun, U., Barisci, N., Serhathoglu, S., and Hardalac, F., Prediction of Minor Head Injured Patients Using Logistic Regression and Mlp Neural Network. *J. Med. Syst*. 29:205–215, 2005. doi:10.1007/s10916-005-5181-x.