Real-time prediction of an anesthetic monitor index using machine learning

Olivier Caelen[a], Olivier Cailloux[b], Djamal Ghoundiwal[b], Abhilash Alexander Miranda[a], Luc Barvais[b], Gianluca Bontempi[a]

[a]*Machine Learning Group, Département d'Informatique, Faculté des Sciences, Université Libre de Bruxelles, Bruxelles, Belgium*
http://dev.ulb.ac.be/mlg/
[b]*Service d'Anesthésiologie-Réanimation, Faculté de Médecine, Université Libre de Bruxelles, Bruxelles, Belgium*

Abstract

An anesthesiologist may control the level of consciousness of a patient undergoing surgery by appropriately dosing hypnotic drugs. The information provided by the monitoring devices may be utilized in order to accomplish this task. One such monitor provides a dimensionless quantity derived from the electroencephalogram called bispectral index (BIS), which could quantify the level of awareness of the patient. This article discusses the use of machine learning techniques to implement a predictive model of the BIS based on the variation of the hypnotic drugs. Such a model learned from a database of recorded operations can aid real-time decision making during the course of an operation. In order to deal with inter-individual variability, the proposed model takes into account patient physiology as well as the reactions of the patient during the early phases of the operation. Two models of the bispectral index behavior are assessed and compared in this work: a linear predictor and a local learning predictor. These prediction models were software implemented and their accuracies were assessed by a computerized cross-validation study and were tested in real situations.

Key words: anesthesia, bispectral index, machine learning, local modeling

1. Introduction

During surgery, the anesthesiologist controls the depth of anesthesia by administrating three types of drugs: hypnotics to cause and maintain loss of consciousness, analgesics to inhibit pain, and very often muscle relaxants to block muscle reactions. In this paper, the drugs considered are propofol as hypnotic and remifentanil as analgesic. Nowadays, anesthesiologists may take advantage of devices which monitor unconsciousness in real-time in order to choose the appropriate dose of hypnotics. Typically, such monitors are connected via electrodes to the patient's forehead and display a signal that has been derived from the electro-encephalographic activity of the patient. The value of the signal gives the anesthesiologist an indication of the level of unconsciousness of the patient. A commonly used monitor is the bispectral index (BIS) commercialized by Aspect Medical Systems [39]. The BIS monitor provides a single dimensionless number, the BIS value, which ranges from 0 to 100. A BIS value of 0 equals EEG silence, while near 100 is the expected value for a fully awake adult, and between 40 and 60 indicates a level for general anesthesia recommended by the manufacturer. Figure 1 shows a typical temporal pattern of the BIS signal during a surgical operation. The BIS signal is close to 100 at the beginning of the operation when the patient is still conscious and falls to about 50 after the induction phase when the pa-

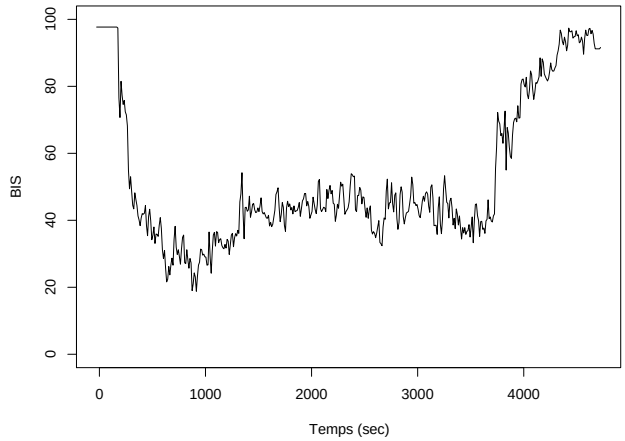

Figure 1: An example of BIS signal evolution.

tient loses conscience. Then, it is typically controlled around values in the 40-60 range until the end of the operation when the anesthesiologists stops the delivery of hypnotics and the patient awakes. The BIS monitor allows the anesthesiologist to detect situations of excessively high or low hypnosis and consequently to adapt the titration of the agents in order to avoid unsafe states.

Note that remifentanil is often used together with propofol because of its known impact on the interaction between propofol and the BIS index [16].

Despite the real-time insight that BIS provides about the depth of anesthesia, it remains difficult for the anesthesiologist, especially if inexperienced, to predict how the BIS signal could vary after a change in the administered anesthetic agents. This is generally due to the high variability of the reactions of the patients to drugs even within classes of patients having similar physiology, which is referred to in the literature as the *inter-individual variability problem* [36]. This means that, two patients treated with a similar dose of drug at the same phase of the surgery could manifest a very different BIS evolution, these differences being impossible to explain solely by age, sex, weight, or other apparent differences.

In this article we analyze a historical database of surgeries in order to learn the short and medium term influence of hypnotic drugs on the level of consciousness of the patient. Specifically, this is achieved by applying *machine learning* methods to predict the bispectral index. Also, this article introduces an original technique to deal with the inter-individual variability.

Two learning techniques are considered: a linear predictive model [31] and a local learning predictive model based on the lazy learning algorithm [7]. Because of the high number of descriptive variables we take into account, the learning procedure is preceded by a forward feature selection procedure to reduce the input dimensionality.

Two validation strategies are adopted to assess the quality of the predictive models: (i) a simulated cross-validation procedure which takes advantage of the large number of available historical surgeries and (ii) a preliminary assessment of the accuracy of the predictions returned by the prototype software BisPrediction. The software was run five times in the operating room under the supervision of the anesthesiologists coauthoring this paper and the results as well as a discussion on their impact on the decision making process are reported in Section 9.2.

The paper is organised as follows: The next section presents a review of related studies. Section 3 gives some information about surgical anesthetics concerning the learning problem studied in this paper and is formally defined in Section 4. This section also introduces the two learning techniques used in this paper. Section 5 presents the feature selection procedure adopted to improve the prediction accuracy of the learning machines. For safety purposes, Section 6 proposes a mechanism which detects the use of the model in a region of the learning dataset where the density of the samples is too low. BisPrediction is a software-tool which implements the predictive model and is presented in Section 7. Section 8 presents the learning datasets used in Section 9 for the experiments. Section 10 gives the conclusions of this paper.

2. Literature review

The link between drug administration and level of conscience has often been studied from a con-

trol theory perspective. Some authors [3, 2, 22, 29] propose proportional-integral-differential algorithms for automatic control of BIS by propofol anesthesia. Liu et al. [24, 25] compare a proportional-differential controller to a manual TCI administration.

While PID controllers require no a priori model of the BIS behavior, model-based approaches [30, 42, 18] integrate patient models to predict the BIS response to some drug input. Struys et al. [41] propose a simulation methodology to test automated controllers and apply it to compare two propofol controllers [42, 3] where BIS is the controlled variable. De Smet et al. introduce a Bayesian closed-loop model which is first assessed in a simulation setting [11] and then tested in a clinical environment [12]. However, unlike the data driven approach presented in this paper, the BIS prediction models adopted in state-of-the-art model-based controllers rely on conventional pharmacokinetic-pharmacodynamic (PKPD) models whose parameters are tuned to specific patients.

Machine learning approaches are however not new in anesthesia. Greenhow et al. [14] use fuzzy logic and Bayesian reasoning with about 400 rules derived from expert anesthetists to provide decision support for dosing of inhaled volatile anesthetics. Nebot [32] extends this work by proposing fuzzy inductive reasoning to control depth of anesthesia as defined by the heart rate, the systolic arterial pressure and respiratory rate. Other papers use neural networks to assess the depth of anesthesia (using systolic arterial pressure, heart rate and respiration rate as inputs) [43] and analyse the resultant networks using principal components analysis and canonical discriminant variates [23]. Some studies use fuzzy-logic and neural networks to build a controller of depth of anesthesia based on the auditory evoked potentials (an other monitor of depth of anesthesia) [17, 5]. Shieh et al. [38] help choosing the concentration of inhaled volatile anesthetics using a hierarchical architecture and self-organizing fuzzy logic for reasoning. The rules are given by expert anesthetists or derived from machine-learning techniques. Some authors [34, 27] use fuzzy modeling to build a comprehensive patient model which predicts the effects of the propofol and remifentanil on the depth of anesthesia, but this depth is measured from the heart rate, the systolic arterial pressure and several auditory evoked potential features rather than on the BIS as in our study. Several surveys [21, 1, 26, 13, 35] cite other related studies using fuzzy logic theory, control and expert systems.

Unlike the approach presented in this paper, most of existing studies [3, 2, 22, 24, 25, 30, 42, 18, 41, 11, 12, 17, 32, 5, 38] either model the BIS reactions to a single input (propofol or another hypnotic drug) or rely on very little experimental evidence. This is the case of of Nunes et al. [33] who tune a Takagi-Sugeno-Kang fuzzy model to predict the future BIS on only two patients.

3. The data collection system

This paper presents and assesses a BIS predictive model whose learning was made possible by the availability of a large database of historical anesthesia sessions. The sessions were recorded by the *Infusion Toolbox* (ITB) software [10]. This software, implemented in C and Smalltalk[1], has been used for several years by the anesthesiologists of the Erasme Hospital, Brussels to achieve Total IntraVenous Anesthesia (TIVA), which will be discussed shortly in the following.

General anesthesia could be considered as having three components [28]: amnesia, analgesia and possibly muscle relaxation. Total intravenous anesthesia (TIVA) achieves these components by administering a combination of exclusively intravenous anesthetic drugs, by infusing them separately and allowing titration of each to the specific dose required to meet the specific needs of the case.

Understanding pharmacokinetics is capital to perform TIVA. The most classical way to describe the decline of blood concentrations after a bolus dose or end of infusion is an open tricompartmental model [6] (see Figure 2).

After a bolus of an intravenous anesthetic drug, there is a rapid, initial distribution phase which

[1] http://www.smalltalk.org/

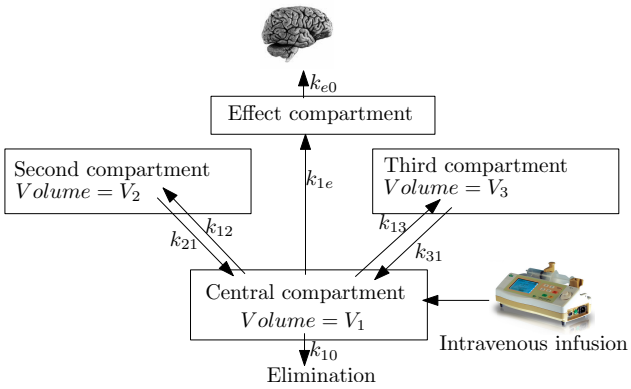

Figure 2: Compartment modeling seeks to estimate the contraction evolution with time of a drug in an effect site (e.g. the brain). The values of the transfer rates $(k_{10}, k_{21}, k_{12}, k_{31}, \ldots)$ and volumes (V_1, V_2, V_3) are defined by the model.

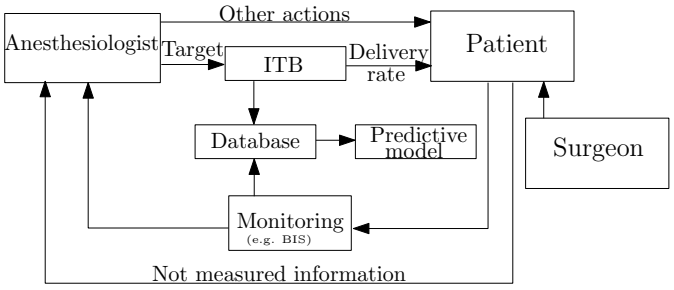

Figure 3: The ITB software and the anesthesia procedure. ITB accomplishes two main tasks: (i) servo-control of the drugs' delivery rate on the basis of the targets fixed by the anesthetist and (ii) monitoring and storing the patient's signals and the anesthetist's actions in a database.

represents distribution from central plasmatic compartment to highly perfused organs such as the brain (also called effect site). This is followed by a slower, second phase representing redistribution to less well perfused tissues such as muscles and fat (second and third compartment). Significant metabolism occurs during the second phase. Recovery from anesthesia is due to extensive redistribution from the brain and to metabolic clearance (liver/renal/plasmatic).

This helps to calculate the ideal loading dose and infusion rate to maintain a certain concentration of a drug at an effect site. The best way to achieve this goal is a computer controlled infusion called TCI (Target Controlled Infusion) [15]. Instead of setting an infusion rate as a flow, the anesthetist sets and adjusts the target brain concentration as required on clinical grounds. Then, a microprocessor computes the flow of drug needed to achieve and maintain the desired brain concentration, according to a pharmacokinetic model, using a microprocessor-controlled syringe pump. Such models exist for propofol (hypnotic agent) and remifentanil (analgesic agent).

The ITB software monitors the patient state and acts as a servo-controller on the multiple intravenous drug infusions (Figure 3). Before and during the operation, ITB stores statistics and monitoring information like: (i) basic details regarding the doctor, the patient and his general state, (ii) the type of surgery, (iii) the evolution of the hemodynamic and physiological parameters (e.g. the BIS) of the patient, (iv) the evolution of the concentration levels of the drugs as chosen by the anesthetist.

Figure 4 shows an example of the evolution of the BIS index during a short period of time of 14 minutes. In this example, the patient is a 70 year-old man of 76 kg, 174 cm and his lean body mass (lbm) equals 59.18. We fix the zero reference for time when the target of propofol is modified for the first time. Note that at 443 seconds, the anesthetist modifies the target of propofol from $0.5\mu g/ml$ to $2\mu g/ml$. The propofol target's modification is made 412 seconds since the previous modification and the BIS drops, in general, as a consequence in the subsequent minutes.

4. Learning of BIS predictive models

This section discusses the learning procedure to estimate the BIS prediction model from data collected by the ITB system. We intend to solve the following: predict the short or medium term (about 10 minutes) temporal evolution of the BIS signal upon change of target of propofol by the anesthesiologist.

We adopt a well known forecasting strategy called *direct prediction* [40]. We wish to predict the BIS signal at regular intervals of Δ units of time from the time t when the target of propofol was changed. In doing so, the problem is decomposed into a set of distinct prediction tasks, one for each horizon $t+h$, where $h \in \{\Delta, 2\Delta, \ldots, H\Delta\}$.

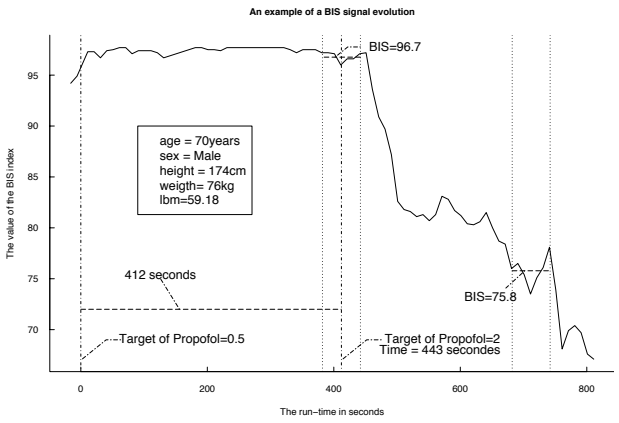

Figure 4: The evolution of the BIS index during a short period of time. At time $t = 443$ seconds, the anesthetist changes the setting of propofol from $0.5\mu/ml$ to $2\mu/ml$.

Once these predictors are built, the set of predictions for the intervals are assembled to obtain the discrete-time values of the BIS signal in the interval $[t + \Delta, t + H\Delta]$.

In this work, we estimate a NARMAX (Nonlinear AutoRegressive Moving Average with eXternal input) model f^h for each prediction task [37] according to

$$B(t+h) = f^h(B(t), x(t)) + \epsilon^h(t), \quad h = \{\Delta, 2\Delta, \ldots, H\Delta\} \quad (1)$$

where t denotes an instant of the operation at which the propofol target is changed, $t = 0$ corresponds to the time of the first modification of the propofol target, $\epsilon^h(t)$ is an error term of the model, $B(t)$ is the value of the BIS at time t and $x(t)$ is the value at time t of the vector x which lumps the set of variables described in Table 1. Note that in the setting of our problem, we generate a set of models by specifying $H = 20$ and $\Delta = 30$ seconds.

4.1. The set of variables

Table 1 reports for each input variable its notation, definition, whether the variable is static (i.e. constant during the operation), and whether this variable is a session specific one. We mean by *initialization* the phase when session specific variables are introduced and recorded to help the model deal with the issue of inter-individual variability. It is a 10 minutes period (including the paptient induction) starting with the first modification of the propofol target. The variables recorded during the initialization are essentially related to the reaction of the patient to the surgical and anesthetic stimuli of the early phases of the operation. The rationale is that due to inter-individual variability, the evolution of the patient's level of awareness during the early phases of the operation can help the system to better understand the patient.

Let us now explain some of the input variables of our model as given in Table 1. The variable $p(t)$ denotes the target of propofol ($\mu g/ml$) before the modification at time t while $\Delta p(t)$ denotes the target modification. In other words, the new target of propofol after the modification amounts to $p(t) + \Delta p(t)$. The variable $r(t)$ measures the target concentration of remifentanil (ng/ml) when the target of propofol is modified (i.e. time t) while T_r is the time elapsed since the last modification of the remifentanil. The fifteen remaining variables are static variables. Among them the first five are conventional descriptive variables of the patient profile. The latter ten variables are session-specific and are recorded during the initialization phase. The quantities $p80$, $p70$ and $p60$ denote the target concentration of propofol during the initialization phase when the BIS equals 80, 70, and 60, respectively. The variables μ_r and μ_p contain the value of the average target of remifentanil and propofol during the first 10 minutes. The quantity $\max_p$ is the maximum target of propofol used by the doctor during the initialization phase and $tMax_p$ is the instant at which the target of propofol reaches its maximum value $\max_p$. $\min_B$ is the minimum reached by the BIS and $tMin_B$ is the instant at which the BIS attains its minimum value $\min_B$. Finally, R is the ratio between $\max_p$ and $\min_B$. A large R means that the induction was aggressive i.e. the maximum target of propofol is high and the BIS decreases to a small value.

Variable	Signification	Static	Session specific
$\Delta p(t)$	The modification's magnitude of the propofol target concentration ($\mu g/ml$) at time t.		
$p(t)$	The target concentration of propofol before the target modification.		
$\Delta timeP$	The time (sec) elapsed between the last modification of the target concentration of propofol and t.		
t	The instant (sec) when the target concentration of propofol is modified.		
$r(t)$	The target concentration of remifentanil (ng/ml) at time t.		
$T_r(t)$	The time (sec) elapsed between the last modification of the target concentration of remifentanil and t.		
a	The age of the patient.	✓	
s	The sex of the patient.	✓	
he	The height of the patient (cm).	✓	
w	The weight of the patient (kg).	✓	
lbm	The Lean Body Mass (lbm) of the patient equals $\begin{cases} 1.1 \cdot weight - 128\frac{weight^2}{height^2} & \text{if } sex = \text{man} \\ 1.07 \cdot weight - 148\frac{weight^2}{height^2} & \text{if } sex = \text{woman} \end{cases}$	✓	
$p80$	The target of propofol ($\mu g/ml$) when the BIS equals 80.	✓	✓
$p70$	The target of propofol ($\mu g/ml$) when the BIS equals 70.	✓	✓
$p60$	The target of propofol ($\mu g/ml$) when the BIS equals 60.	✓	✓
μ_r	The average target concentration of remifentanil (ng/ml).	✓	✓
μ_p	The average target concentration of propofol ($\mu g/ml$).	✓	✓
max_p	The maximum reached by the target of propofol ($\mu g/ml$).	✓	✓
$tMax_p$	The instant (sec) when max_p is reached.	✓	✓
min_B	The minimum reached by the BIS.	✓	✓
$tMin_B$	The instant (sec) when min_B is reached.	✓	✓
R	The ratio between max_p and min_B.	✓	✓

Table 1: This table describes the variables of the vector $x(t)$ in equation (1).

4.2. The learning procedure

Two learning techniques are used to implement the set of prediction models according to (1). The first one is a conventional linear technique [31]. The second is a local modeling technique, called lazy learning [7], which has been proved to be successful in many problems of nonlinear modeling [9] and in two international competitions on data analysis and time series prediction [8].

In local modeling, the value of an unknown mapping is estimated focusing on the region surrounding the point where the estimation is required. The procedure essentially consists of these steps: (*i*) for each query point $q(t)$, select a set of neighbors and weigh their relevance according to some relevance criterion (e.g., the distance) (*ii*) choose a local regression function h in a restricted family of parametric functions (*iii*) compute the regression value $h(q)$. Doing so, the approach requires to keep in the memory the set of observations for each prediction, unlike a global modeling approach (e.g., linear regression) which discards it. At the same time, local modeling requires only simple approximators (e.g., constant and/or linear) to model the dataset in a neighborhood of the query point. Moreover, the method is intrinsically adaptive, since the availability of new measurements simply requires the updating of the observations' set.

Lazy learning is a particular instance of local modeling which provides an automatic way of selecting the optimal number of neighbors for each query point. The idea consists in starting from a minimum number of neighbors and recursively adding neighbors until the predicted performance of the corresponding local approximation significantly decays or until a maximum number of examples is reached. This procedure allows the detection of a linearity region around the query point. For more details on local modeling methods and the distinctive features of lazy learning, we refer the reader to [7].

5. Feature selection

Techniques which estimate non-linear dependencies from multidimensional data are vulnerable to ill-conditioning and overfitting. Having recourse to feature selection techniques is a typical solution to such situations which at the same time provides an useful insight to the anesthetist about which variables play an important role on the evolution of the patient physiology.

Examples of feature selection approaches are filtering [20] and wrapper techniques [19]. In this paper, we use a well-known wrapper technique called *sequential forward selection* [4] where a leave-one-out cross-validation procedure is used to assess the robustness of the input feature set.

The sequential forward selection begins by considering each of the variables individually and by selecting the most accurate. At each successive stage of the algorithm, one additional input variable is added to the set, again chosen on the basis of the best accuracy.

Table 2 gives the input variables selected by the sequential forward selection algorithm for each prediction problem. Note that some variables are selected more often than others and are consequently expected to bring higher information on the evolution of the BIS index. For instance, the variables $B(t)$, $\Delta p(t)$ and $p(t)$ are always selected. Except for the predictive model $f^{h=540}$, the variable min_B is also always present. The variables $p70$, R, $\Delta timeP$, and t are also selected frequently. We decided to take these eight input variables as the most informative and we included them in the final predictive models

$$\widehat{B}(t+h) = \widehat{f}^h(B(t), \Delta p(t), p(t), min_B, p70, R, \Delta timeP, t, \alpha_N). \quad (2)$$

where α_N is a vector containing the parameters of the model.

To predict the BIS, the forward selection procedure confirms the importance of taking into account the current BIS index, the magnitude of the drug modification and the old target of propofol. Most of the static variables (age, sex, lbm, etc.) are already integrated in the pharmacokinetic model and this could explain why these vari-

h	Variables selected by the sequential forward selection algorithm
30sec	$B(t) + \Delta p(t) + p(t) + min_B + r(t) + p70 + \Delta timeP + R + w + tMin_B + p60 + he + lbm$
60sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + p80 + \Delta timeP + t + r(t)$
90sec	$B(t) + \Delta p(t) + p(t) + min_B + t + \Delta timeP + w + p70 + R$
120sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + \Delta timeP + t + R + max_prop$
150sec	$B(t) + \Delta p(t) + p(t) + min_B + tMin_B + p70 + R + t + \Delta timeP + he$
180sec	$B(t) + \Delta p(t) + p(t) + min_B + \mu_r + \Delta timeP + tMin_B + p70 + R + p80 + lbm + age$
210sec	$B(t) + \Delta p(t) + p(t) + min_B + \mu_p + R + t + \Delta timeP + r(t) + he + max_prop + p60$
240sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + R + t + \Delta timeP$
270sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + R + he + t + \Delta timeP + r(t)$
300sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + p80 + \Delta timeP + tMax_prop + t + R + max_prop$
330sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + p80 + \mu_r + \Delta timeP + t + r(t) + R + max_prop + p60$
360sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + R + max_prop + t + r(t) + \Delta timeP + \mu_r$
390sec	$B(t) + \Delta p(t) + p(t) + min_B + r(t) + p70 + R + p80$
420sec	$B(t) + \Delta p(t) + p(t) + min_B + p70 + p80 + \mu_r$
450sec	$B(t) + \Delta p(t) + p(t) + min_B + max_prop + p80 + R + \mu_r + p70$
480sec	$B(t) + \Delta p(t) + p(t) + min_B + p60 + R + t + \Delta timeP + \mu_r + p80 + max_prop$
510sec	$B(t) + \Delta p(t) + p(t) + t + min_B + p70 + R + max_prop + \Delta timeP + \mu_r + p80$
540sec	$B(t) + \Delta p(t) + p(t) + t + p70 + R + p60 + p80 + \mu_r$
570sec	$B(t) + \Delta p(t) + p(t) + p70 + p80 + min_B + \mu_p + t + R + \Delta timeP + w + p60 + \mu_r + age$
600sec	$B(t) + \Delta p(t) + p(t) + p70 + p80 + min_B + t + \Delta timeP + \mu_p + sexe + age + R + p60 + \mu_r$

Table 2: For each learning set, this table gives the input variables selected by the sequential forward selection algorithm.

ables are rejected by the forward selection. Note that input variables min_B, $p70$, and R selected by the feature selection algorithm and are specific to the patient and measures the patient's reaction as they are recorded during the first ten minutes of the intervention. We expect that this subset of variables carries within the model sufficient information to reduce the inter-individual variability problem.

6. Domain of clinical validity

An important issue in using prediction models for safety-critical applications is to determine whether they are used in their validity domain in a clinical setting. Since our final goal is to use the BIS prediction models to support decision making during surgery, we need the definition of a mechanism which detects when the tool is used outside its validity region. For that reason, we implemented an outlier detection algorithm which relies on a multivariate estimator of the density of the inputs. A warning message is then issued to the user when the model is asked to operate in a point where the density of the historical dataset is too low. This density is estimated based on $B(t)$, $\Delta p(t)$ and $p(t)$, which were consistently selected by the feature selection procedure described in Section 5.

We adopted a parametric outlier detection procedure which relies on a conventional multivariate Gaussian estimation and the Mahalanobis distance. Let $\widehat{\mu} = \left[\widehat{\mu}_{B(t)}, \widehat{\mu}_{\Delta p(t)}, \widehat{\mu}_{p(t)}\right]$ be the sample mean vector computed with the training set and $\widehat{\Sigma}$ the sample covariance matrix.

Let $X_q = \left[q_{B(t)}, q_{\Delta p(t)}, q_{p(t)}\right]$ be the query point, that is the vector containing the BIS index, the drug's modification and the old propofol target at time t.

The Mahalanobis distance

$$d_M = \sqrt{(X_q - \widehat{\mu})^T \, \widehat{\Sigma}^{-1} \, (X_q - \widehat{\mu})}.$$

can be used to detect whether the query is an outlier or equivalently whether the model is requested to work in an unsafe region.

In our application, the anesthetist is warned that the model is running out of its validity region when the Mahalanobis distance d_M is higher than the 90% quantile of the Mahalanobis distance computed on the samples of the training set.

7. The BisPrediction tool

One of the contributions of this paper is the implementation of a decision support software tool called BisPrediction. It implements $H = 20$ linear predictive models $\widehat{f}^h$. The estimated coefficients of these models are reported in Appendix A. BisPrediction interacts with ITB to obtain the input variables each time the anesthesiologist intends to perform a propofol target variation.

Figures 5 and 6 contain screenshots of BisPrediction in manual and automatic modes. The first two panels from the left are the usual ITB panels. The first panel is used by the anesthetist to control the propofol and the second one is dedicated to remifentanil, while the third panel is BisPrediction.

BisPrediction receives all the information regarding the current session by communicating with ITB via a server which is running in the background. This communication is completely transparent to the anesthetist and allows to have a very simple Graphical User Interface (GUI) which minimizes the interaction between BisPrediction and the user.

The two tabs of the BisPrediction interface are used to choose its two main working modes: the *manual mode* (Figure 5) and the *automatic mode* (Figure 6). There is a third tab which displays all the internal variables used by the models for debugging purposes.

In the manual mode, the anesthetist chooses the target of propofol, clicks on GO and a graphic with the estimation of the BIS index is displayed. This allows the anesthetist to estimate the effect of a drug modification before it is applied.

In the automatic mode, no interaction is required via the GUI. Here, whenever the level of concentration of propofol is modified on ITB (the first panel starting from the left in Figure 6), a sig-

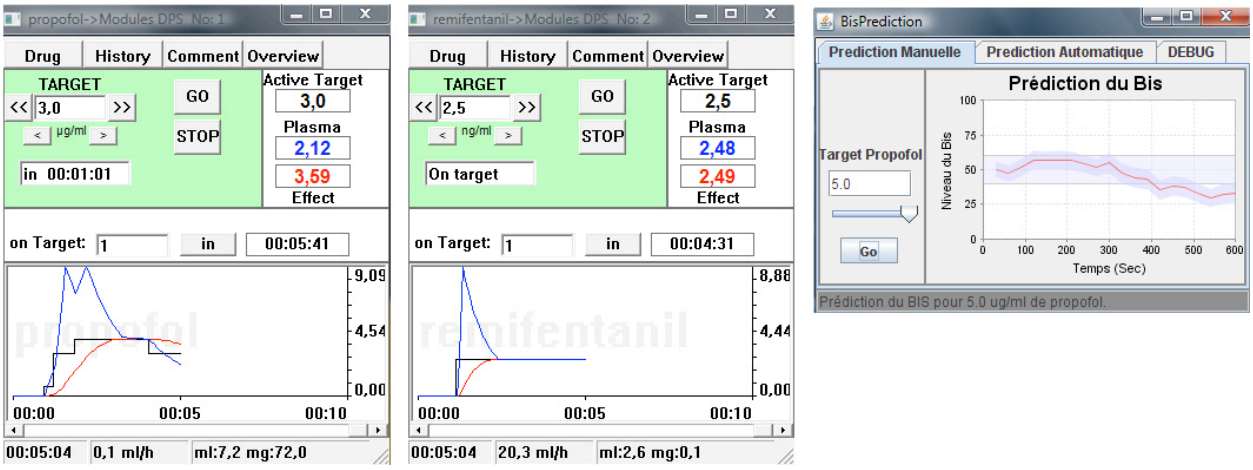

Figure 5: A screenshots of ITB and BisPrediction (manual mode)

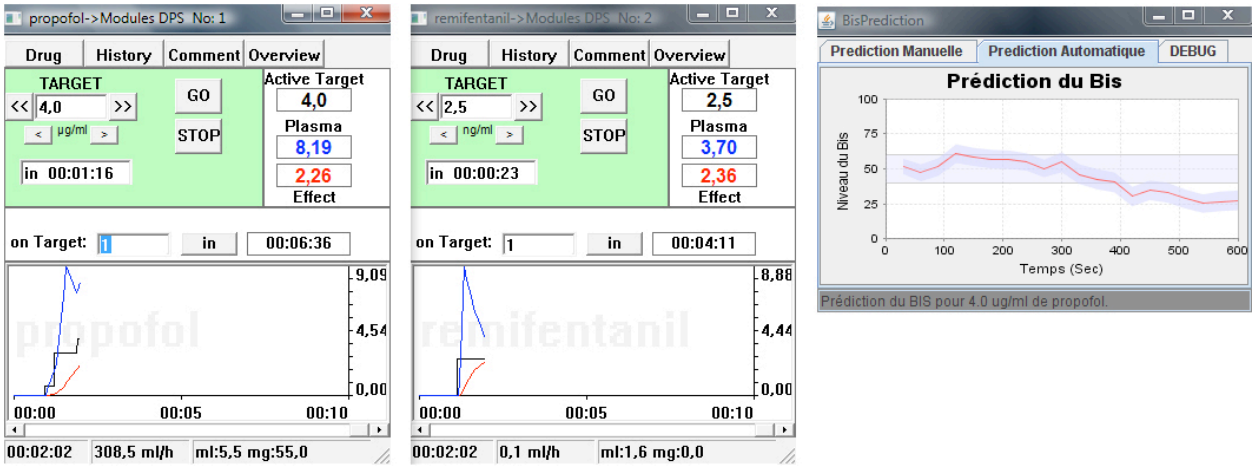

Figure 6: A screenshots of ITB and BisPrediction (automatic mode)

nal requesting a plot of the estimated future BIS index is automatically sent from ITB to BisPrediction, via the server running in the background.

BisPrediction is implemented with the Java / Swing[2] language and the JFreeChart[3] library is used to display the curves of the BIS evolution and Java Native Interface technology is used to interact with the server.

8. The learning set

By means of the ITB software, we collected a huge MySQL[4] database containing the recordings of 1069 interventions.

In order to learn the predictive model f^h, the database is used to create 20 learning sets D_N^h, $h = \{30, 60, 90, 120, \ldots, 570, 600\}$. Note that a sample belongs to the dataset D_N^h if the target concentration of propofol is modified at time t (with $t > 600$ sec, i.e. after the initialization phase) and if no other modifications of the target of propofol appears in the time interval $[t, t + h]$. Table 3 gives the number of samples for each learning set thus obtained.

We used the statistical language R[5] for the statistical analysis and the package RODBC[6] for connecting R with the MySQL database.

[2]http://java.sun.com/
[3]http://www.jfree.org/jfreechart/
[4]http://www.mysql.com/
[5]http://www.r-project.org/
[6]http://cran.r-project.org/web/packages/RODBC

	$D_N^{h=30}$	$D_N^{h=60}$	$D_N^{h=90}$	$D_N^{h=120}$	$D_N^{h=150}$
$N =$	1786	1656	1542	1465	1390
	$D_N^{h=180}$	$D_N^{h=210}$	$D_N^{h=240}$	$D_N^{h=270}$	$D_N^{h=300}$
$N =$	1333	1270	1244	1195	1148
	$D_N^{h=330}$	$D_N^{h=360}$	$D_N^{h=390}$	$D_N^{h=420}$	$D_N^{h=450}$
$N =$	1101	1060	1017	986	957
	$D_N^{h=480}$	$D_N^{h=510}$	$D_N^{h=540}$	$D_N^{h=570}$	$D_N^{h=600}$
$N =$	917	890	860	824	801

Table 3: This table shows for each of the twenty learning sets, the number of samples available.

9. Experiments and results

9.1. Assessment using cross validation

The validation of the models is an important aspect in machine learning. Three different cross-validation criteria are used for the validation and two of them rely on the notion of *leave-one-session-out error*. Let $S(i)$ be the set of all the samples belonging to the session from which the sample i originates. Let $\widehat{B}_{(-S(i))}(t+h)$ be the prediction for the sample i returned by a model trained on all the samples except the ones in the set $S(i)$. The leave-one-session-out error made on the sample i is then given by

$$\widehat{E}_i^{loo}(h) = \widehat{B}_{(-S(i))}(t+h) - B_i(t+h).$$

The first criterion which is used to assess the quality of the prediction is the *normalized mean squared error* (NMSE)

$$NMSE = \frac{\sum_{i=1}^{N}\left(\widehat{E}_i^{loo}(h)\right)^2}{\sum_{i=1}^{N}\left(\widehat{\mu}_b - B_i(t+h)\right)^2}$$

where $\widehat{\mu}_b = \frac{1}{N}\sum_{i=1}^{N} B_i(t+h)$ is the average of the future BIS index. The NMSE measure is commonly used in the time series prediction community to assess how the predictor behaves with respect to the simplest one (i.e. the sample average). Note that if $NMSE > 1$, we may interpret that the prediction error is worse than the error obtained had the average of the future BIS been used.

The second criterion is the *mean of the absolute errors* (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{E}_i^{loo}(h)\right|.$$

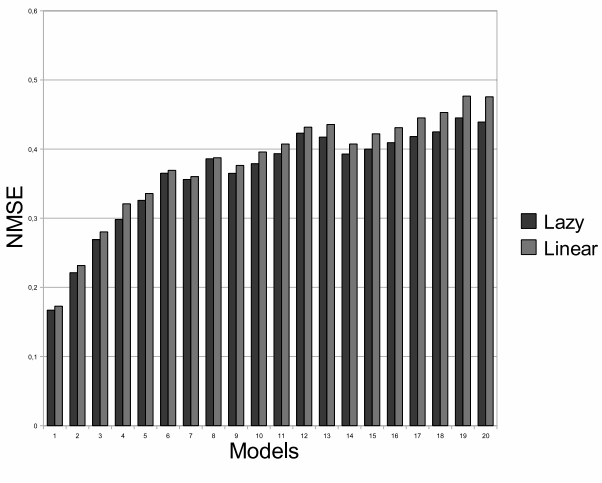

Figure 7: NMSE (*normalized mean squared error*) of the $H = 20$ predictive models (the lower the better).

This measure has the same dimension as the BIS index variable and gives an indication about the average magnitude of the errors made by the decision support system.

The last criterion is the percentage P of cases where the change of the predicted future BIS has the same direction as the real future BIS

$$P = \frac{100}{N}\sum_{i=1}^{N} I\left[\left(\widehat{B}_{(-S(i))}(t+h) - B_i(t)\right)\cdot (B_i(t+h) - B_i(t))\right]$$

where $I[A] = \begin{cases} 1 & \text{if } A \geq 0, \\ 0 & \text{if } A < 0. \end{cases}$

This criterion measures how often the changes of BIS predicted by the decision support tool agree with the real ones.

Figures 7, 8 and 9 shows the cross-validated values of NMSE, MAE and P for the 20 prediction models. The results show that, in most of the cases, lazy learning outperforms the linear model and supports the argument that a non-linear relationship probably holds between the target of propofol and the BIS index. Table 4 gives the p-values (paired t-test) of the normalized mean squared error in comparing the lazy learning model with the linear model. Although the lazy learning is always better than linear model in having a lower NMSE, according to Table 4, this difference in performance is significant for $h \leq 120$ and

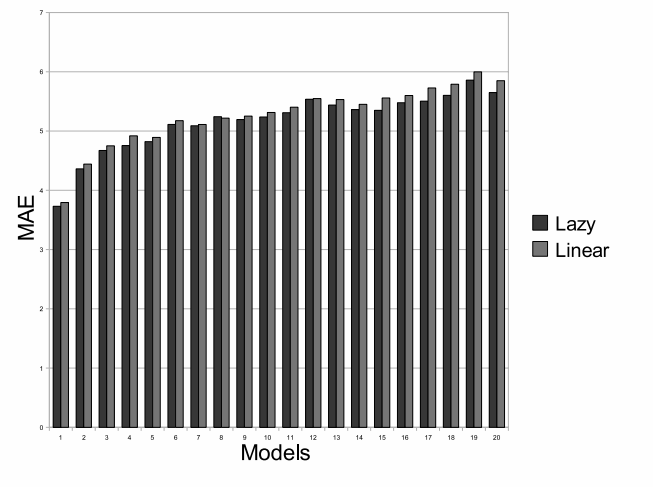

Figure 8: MAE (*mean of the absolute errors*) of the $H = 20$ predictive models (the lower the better).

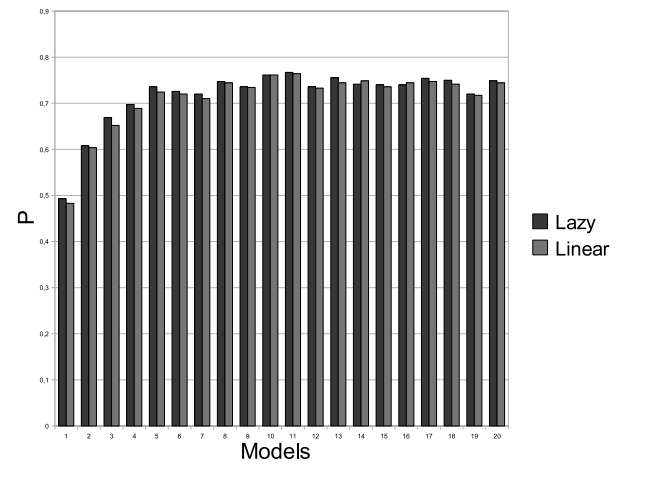

Figure 9: Evolution of the P for the twenty models (the higher the better).

Figure 10: An exemple of an equipment cart used at the Erasme Hospital.

$h \geq 450$. We are yet to incorporate the lazy learning in BisPrediction.

9.2. Validation in real settings

The ITB and BisPrediction softwares were installed in an equipment cart of an operation room at the Erasme Hospital to carry out preliminary validation. This equipment cart (Figure 10) holds three syringe pumps of hypnotic, analgesic and muscle relaxant drugs together with a BIS monitor.

The BisPrediction functionality, in terms of 10 minutes ahead predictive information provided to the anesthesiologist, was assessed during five surgical operations.

For one of these operations (29 year-old woman undergoing a gynaecological surgery), Figure 11 shows the real BIS evolution together with the predictive output of BisPrediction after each propofol target modification (vertical lines). The light grey area represents the initialization phase while shaded areas represent intervals of confidence. Black

Models	$\widehat{f}^{30}$	$\widehat{f}^{60}$	$\widehat{f}^{90}$	$\widehat{f}^{120}$	$\widehat{f}^{150}$	$\widehat{f}^{180}$	$\widehat{f}^{210}$	$\widehat{f}^{240}$	$\widehat{f}^{270}$	$\widehat{f}^{300}$
p-value	**<0.01**	**<0.01**	**<0.01**	**<0.01**	0.047	0.52	0.51	0.79	0.051	0.012
Models	$\widehat{f}^{330}$	$\widehat{f}^{360}$	$\widehat{f}^{390}$	$\widehat{f}^{420}$	$\widehat{f}^{450}$	$\widehat{f}^{480}$	$\widehat{f}^{510}$	$\widehat{f}^{540}$	$\widehat{f}^{570}$	$\widehat{f}^{600}$
p-value	0.061	0.29	0.0172	0.049	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**

Table 4: The p-values (paired t-test) of the NMSE (*normalized mean squared error*).

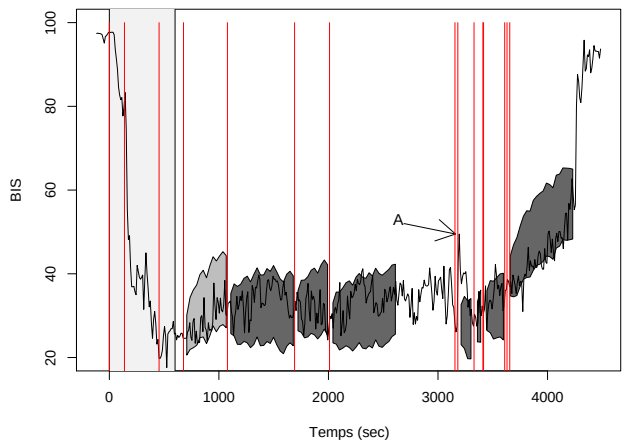

Figure 11: The Bis evolution during surgery. Vertical red lines are propofol target modifications. Black area represents prediction of Bis during ten minutes after propofol modifications. Grey area occurs when the Mahalanobis distance is too high.

shaded areas characterize predictions made for regimes inside the validity region (Section 6) while grey regions denote predictions made in a configuration out of the validity region. For instance, because of a very low BIS (under 30), the first prediction happens to be outside the validity of our model defined by the Mahalanobis distance. Note also that the strong BIS increase (marked by the A character in Figure 11) is due to coughing of the patient. Because of the fortuity of this event, our model was enable to return a reliable prediction.

However, the majority of the predictions performed during the five surgical case studies were deemed to be sufficiently close to the real BIS. As a consequence, the overall evaluation of the performance was positive and the tool was judged as promising by our partner anesthesiologists to provide useful information during the surgery.

Useful suggestions were also made for future works mainly related to the improvement of the performance around the final phases of the surgery. In machine learning terms this request can be easily adressed by building a specific database to deal with the end of the operation and the awakening phase.

10. Conclusions and Future Work

This paper proposes the use of machine learning techniques to build a predictive model supporting the activities of anesthetists during surgical operations via a decision support tool. The cross-validation results (Section 9.1) as well as the preliminary tests in real conditions (Section 9.2) are encouraging. It is our belief that such a tool can reduce risks in daily medical practice by helping the anesthesiologist to chose the best hypnotic dose in advance. The performances of the predictive models as well as its successful implementation and use in real operations as detailed in this article are strong indications that this area of research is worth exploring in more detail.

We propose the following ways to extend this work. Three of the eight features selected by the feature selection procedure are created during the initialization phase ($minB$, $p70$ and R). This is an argument in favor of the use of this kind of features to deal with the inter-individual variability problem. To improve the prediction accuracy of the model, research must be extended in this direction to explore new features created during the initialization phase.

Currently, BisPrediction was assessed in only five experimental conditions. In order to analyze the impact on the daily practice of this kind of decision support system in anesthesia, a protocol should be defined to test BisPrediction on a larger number of real cases.

Finally, the whole work is focused on the impact of Propofol on the BIS. The same method

presented in this paper could be generalized to the prediction of other clinical variables (for instance blood pressure and heart rate).

A. The coefficients of the linear models

This appendix gives the linear version of the twenty predictive models $\widehat{f}^h$ implemented is the BisPrediction software. The linear predictive models are defined as

$$\begin{aligned}\widehat{f}^h =& a_1 \cdot B(t) + a_2 \cdot \Delta p(t) + a_3 \cdot p(t) + \\ & a_4 \cdot min_B + a_5 \cdot p70 + a_6 \cdot R + \\ & a_7 \cdot \Delta timeP + a_8 \cdot t + a_0 \end{aligned} \quad (3)$$

where the values of the coefficients a, depending on h, are described is Table 5.

References

[1] M. F. Abbod, D. G. von Keyserlingk, D. A. Linkens, and M. Mahfouf. Survey of utilisation of fuzzy technology in medicine and healthcare. *Fuzzy Sets and Systems*, 120(2):331–349, 2001.

[2] A. R. Absalom and G. N. C. Kenny. Closed-loop control of propofol anaesthesia using bispectral index (tm): performance assessment in patients receiving computer-controlled propofol and manually controlled remifentanil infusions for minor surgery. *British Journal of Anaesthesia*, 90(6):737–741, 2003.

[3] A. R. Absalom, N. Sutcliffe, and G. N. Kenny. Closed-loop control of anesthesia using bispectral index - performance assessment in patients undergoing major orthopedic surgey under combined general and regional anesthesia. *Anesthesiology*, 96(1):67–73, 2002.

[4] D. W. Aha and R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. *Artificial Intelligence and Statistics*, 5, 1996.

[5] R. Allen and D. Smith. Neuro-fuzzy closed-loop control of depth of anaesthesia. *Artificial Intelligence in Medicine*, 21(1-3):185–191, 2001.

[6] J. M. Bailey, W. M. Haddad, and T. Hayakawa. Closed-loop control in clinical pharmacology: Paradigms, benefits and challenges. In *Proceedings of the 2004 American Control Conference*, pages 2268–2277, 2004.

[7] G. Bontempi. *Local Learning Techniques for Modeling, Prediction and Control*. PhD thesis, IRIDIA-Université Libre de Bruxelles, 1999.

[8] G. Bontempi, M. Birattari, and H. Bersini. Lazy learners at work: the lazy learning toolbox. In *Proceeding of the 7th European Congress on Intelligent Techniques and Soft Computing EUFIT '99*, 1999.

[9] G. Bontempi, M. Birattari, and H. Bersini. Lazy learning for modeling and control design. *International Journal of Control*, 72(7/8):643–658, 1999.

[10] F. Cantraine and E. Coussaert. The first object oriented monitor for intravenous anesthesia. *PubMed - indexed for MEDLINE*, 16:3–10, 2000.

[11] T. De Smet, Mmrf Struys, S. Greenwald, E. P. Mortier, and S. L. Shafer. Estimation of optimal modeling weights for a bayesian-based closed-loop system for propofol administration using the bispectral index as a controlled variable: A simulation study. *Anesthesia and Analgesia*, 105(6):1629–1638, 2007.

[12] T. De Smet, Mmrf Struys, M. M. Neckebroek, K. Van den Hauwe, S. Bonte, and E. P. Mortier. The accuracy and clinical feasibility of a new bayesian-based closed-loop control system for propofol administration using the bispectral index as a controlled variable. *Anesthesia and Analgesia*, 107(4):1200–1210, 2008.

[13] A. M. Fields, K. M. Fields, and J. W. Cannon. Closed-loop systems for drug delivery. *Current Opinion in Anesthesiology*, 21(4):446–451, 2008.

[14] S. G. Greenhow, D. A. Linkens, and A. J. Asbury. Development of an expert system adviser for anesthetic control. *Computer Methods and Programs in Biomedicine*, 37(3):215–229, 1992.

[15] F. Guarracino, F. Lapolla, C. Cariello, A. Danella, L. Doroni, R. Baldassarri, and A. Boldrini ans M.L. Volpe. Target controlled infusion: Tci. *Minerva Anestesiol*, 71(6):335–342, 2005.

[16] B. Guignard, C. Menigaux, X. Dupont, D. Fletcher, and M. Chauvin. The effect of remifentanil on the bispectral index change and hemodynamic responses after orotracheal intubation. *Anesth Analg*, 90:161–168, 2000.

[17] J. W. Huang, Y. Y. Lu, A. Nayak, and J. Rob. Depth of anesthesia estimation and control. *Ieee Transactions on Biomedical Engineering*, 46(1):71–81, 1999.

[18] C. A. Ionescu, R. De Keyser, B. C. Torrico, T. De Smet, M. M. Struys, and J. E. Normey-Rico. Robust predictive control strategy applied for propofol dosing using bis as a controlled variable during anesthesia. *Ieee Transactions on Biomedical Engineering*, 55(9):2161–2170, 2008.

[19] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[20] D. Koller and M. Sahami. Toward optimal feature selection. *International Conference on Machine Learning*, pages 284–292, 1996.

[21] M. Krol and D. L. Reich. Expert systems in anesthesiology. *Drugs of Today*, 34(7):593–601, 1998.

[22] K. Leslie, A. Absalom, and G. N. C. Kenny. Closed loop control of sedation for colonoscopy using the bispectral index. *Anaesthesia*, 57(7):693–697, 2002. Anaesthesia.

h	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_0
30sec	0.904	−1.467	−0.83	0.013	0.447	−5.772	-1.14×10^{-4}	-4.10×10^{-7}	5.078
60sec	0.845	−2.717	−1.12	0.020	0.512	−5.177	-2.23×10^{-4}	9.61×10^{-5}	7.087
90sec	0.808	−4.284	−1.23	0.021	0.543	−9.927	-2.93×10^{-4}	1.55×10^{-4}	8.714
120sec	0.767	−5.281	−1.74	0.024	0.913	−11.231	-3.16×10^{-4}	9.87×10^{-5}	11.103
150sec	0.749	−6.086	−1.70	0.024	0.816	−13.997	-2.31×10^{-4}	1.26×10^{-4}	11.849
180sec	0.738	−6.208	−1.60	0.022	0.762	−15.215	-1.95×10^{-4}	5.79×10^{-5}	13.064
210sec	0.736	−6.501	−1.80	0.022	0.871	−12.952	-2.92×10^{-4}	1.52×10^{-4}	12.508
240sec	0.729	−7.380	−2.07	0.028	1.048	−14.063	-2.61×10^{-4}	1.24×10^{-4}	12.969
270sec	0.721	−7.745	−1.86	0.018	1.328	−19.085	-1.97×10^{-4}	9.88×10^{-4}	12.889
300sec	0.717	−8.343	−2.32	0.038	1.337	−16.643	-3.18×10^{-4}	9.34×10^{-4}	13.584
330sec	0.705	−9.018	−2.21	0.033	1.114	−11.787	-2.56×10^{-4}	1.10×10^{-4}	13.914
360sec	0.701	−9.371	−2.11	0.023	1.132	−13.220	-1.94×10^{-4}	1.07×10^{-4}	14.347
390sec	0.688	−9.652	−2.27	0.031	0.994	−16.703	-6.10×10^{-5}	1.75×10^{-5}	15.589
420sec	0.698	−9.690	−2.33	0.042	0.893	−9.448	-2.12×10^{-4}	5.25×10^{-5}	14.843
450sec	0.667	−9.302	−2.28	0.042	0.953	−8.704	-1.85×10^{-4}	8.56×10^{-5}	15.640
480sec	0.667	−9.703	−2.34	0.041	1.056	−13.438	-3.02×10^{-4}	1.85×10^{-4}	15.535
510sec	0.648	−10.508	−2.76	0.021	1.534	−23.173	-3.38×10^{-4}	2.36×10^{-4}	17.385
540sec	0.645	−10.185	−2.69	0.018	1.415	−18.751	-1.72×10^{-4}	1.82×10^{-4}	17.933
570sec	0.645	−10.047	−3.17	0.028	1.830	−16.700	-1.39×10^{-4}	1.67×10^{-4}	17.428
600sec	0.642	−10.832	−3.51	0.031	1.877	−16.833	-1.71×10^{-4}	1.97×10^{-4}	17.739

Table 5: The parameters of the twenty linear models (3).

[23] D. A. Linkens and L. Vefghi. Recognition of patient anaesthetic levels: neural network systems, principal components analysis, and canonical discriminant variates. *Artificial Intelligence in Medicine*, 11(2):155–173, 1997. Artif. Intell. Med.

[24] N. Liu, T. Chazot, A. Genty, A. Landais, A. Restoux, K. McGee, P. A. Laloe, B. Trillat, L. Barvais, and M. Fischler. Titration of propofol for anesthetic induction and maintenance guided by the bispectral index: Closed-loop versus manual control - a prospective, randomized, multicenter study. *Anesthesiology*, 104(4):686–695, 2006.

[25] N. Liu, T. Chazot, B. Trillat, R. Pirracchio, J. D. Law-Koune, L. Barvais, and M. Fischler. Feasibility of closed-loop titration of propofol guided by the bispectral index for general anaesthesia induction: a prospective randomized study. *European Journal of Anaesthesiology*, 23(6):465–469, 2006.

[26] M. Mahfouf, M. F. Abbod, and D. A. Linkens. A survey of fuzzy logic monitoring and control utilisation in medicine. *Artificial Intelligence in Medicine*, 21(1-3):27–42, 2001. Artif. Intell. Med.

[27] M. Mahfouf, C. S. Nunes, D. A. Linkens, and J. E. Peacock. Modelling and multivariable control in anaesthesia using neural-fuzzy paradigms - part ii. closed-loop control of simultaneous administration of propofol and remifentanil. *Artificial Intelligence in Medicine*, 35(3):207–213, 2005.

[28] G. Morgan, M. Mikhail, and M. Murray. *Clinical Anesthesiology*. McGrawHill, 4 edition, 2006.

[29] A. Morley, J. Derrick, P. Mainland, B. B. Lee, and T. G. Short. Closed loop control of anaesthesia: an assessment of the bispectral index as the target of control. *Anaesthesia*, 55(10):953–959, 2000.

[30] E. Mortier, M. Struys, T. De Smet, L. Versichelen, and G. Rolly. Closed loop controlled administration of propofol using bispectral analysis. *Anaesthesia*, 53(8):749–754, 1998.

[31] R. H. Myers. *Classical and Modern Regression with Applications*. PWS-KENT, Boston, MA, 1990.

[32] A. Nebot, F. E. Cellier, and D. A. Linkens. Synthesis of an anaesthetic agent administration system using fuzzy inductive reasoning. *Artificial Intelligence in Medicine*, 8(2):147–166, 1996.

[33] C. S. Nunes, D. A. Ferreira, T. F. Mendonca, P. Amorim, and L. M. Antunes. Neuro-fuzzy techniques to model pharmacodynamic interactions between anesthetic drugs on the bispectral index: A preliminary study. *Journal of Intelligent & Fuzzy Systems*, 16(1):15–22, 2005.

[34] C. S. Nunes, M. Mahfouf, D. A. Linkens, and J. E. Peacock. Modelling and multivariable control in anaesthesia using neural-fuzzy paradigms - part i. classification of depth of anaesthesia and development of a patient model. *Artificial Intelligence in Medicine*, 35(3):195–206, 2005.

[35] R. Pauldine, G. Beck, J. Salinas, and D. W. Kaczka. Closed-loop strategies for patient care systems. *Journal of Trauma-Injury Infection and Critical Care*, 64(4):S289–S294, 2008.

[36] M. Huzmezan C.R. Ries S. Bibian, G.A. Dumont. Patient variability and uncertainty quantification in anaesthesia: Part i - modeling and identification. In *Proceedings of the 6th IFAC Symposium on Modeling and Control in Biomedical Systems*, 2006.

[37] T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989.

[38] J. S. Shieh, D. A. Linkens, and A. J. Asbury. A hierarchical system of on-line advisory for monitoring and controlling the depth of anaesthesia using self-organizing fuzzy. *Engineering Applications of Artificial Intelligence*, 18(3):307–316, 2005.

[39] J.C. Sigl and N.G. Chamoun. An introduction to bispectral analysis for the electroencephalogram. *Clin Monitor*, 10:392–404, 1994.

[40] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse. Methodology for long-term prediction of time series. *Neurocomputing*, 2007.

[41] Mmrf Struys, T. De Smet, S. Greenwald, A. R. Absalom, S. Binge, and E. P. Mortier. Performance evaluation of two published closed-loop control systems using bispectral index monitoring - a simulation study. *Anesthesiology*, 100(3):640–647, 2004.

[42] Mmrf Struys, T. De Smet, L. F. M. Versichelen, S. Van de Velde, R. Van den Broecke, and E. P. Mortier. Comparison of closed-loop controlled administration of propofol using bispectral index as the controlled variable versus "standard practice" controlled administration. *Anesthesiology*, 95(1):6–17, 2001.

[43] L. Vefghi and D. A. Linkens. Dynamic monitoring and control of patient anaesthetic and dose levels: time-delay, moving-average neural networks, and principal components analysis. *Computer Methods and Programs in Biomedicine*, 59(2):91–106, 1999.