Predicting the probability of survival in intensive care unit patients from a small number of variables and training examples

Oscar Luaces[a], Francisco Taboada[b], Guillermo M. Albaiceta[b], Luis A. Domínguez[c], Pedro Enríquez[c], GRECIA Group[d], Antonio Bahamonde[a,*]

[a] *Artificial Intelligence Center, Universidad de Oviedo at Gijón. E33204 Gijón, Asturias, Spain*
[b] *Hospital Universitario Central de Asturias. Universidad de Oviedo. E33006 Oviedo, Asturias, Spain*
[c] *Hospital Universitario Río Hortega, E47010 Valladolid, Spain*
[d] *Grupo de Estudios y Análisis en Cuidados Intensivos*

Summary

Objective: Survival probability predictions in critically ill patients are mainly used to measure the efficacy of intensive care unit (ICU) treatment. The available models are functions induced from data on thousands of patients. Eventually, some of the variables used for these purposes are not part of the clinical routine, and may not be registered in some patients. In this paper, we propose a new method to build scoring functions able to make reliable predictions, though functions whose induction only requires records from a small set of patients described by a few variables.

Methods: We present a learning method based on the use of support vector machines (SVM), and a detailed study of its prediction performance, in different contexts, of groups of variables defined according to the source of information: monitoring devices, laboratory findings, and demographic and diagnostic features.

Results: We employed a data set collected in general ICUs at 10 units of hospitals in Spain, 6 of which include coronary patients, while the other 4 do not treat coronary diseases. The total number of patients considered in our study was 2501, 19.83% of whom did not survive. Using these data, we report a comparison between the SVM method proposed here with other approaches based on logistic regression (LR), including a second-level recalibration of release III of the acute physiology and chronic health evaluation (APACHE, a scoring system commonly used in ICUs) induced from the available data. The SVM method significantly outperforms them all from a statistical point of view. Comparison with the commercial version of APACHE III shows that the SVM scores are slightly better when working with data sets of more than 500 patients.

Conclusions: From a practical point of view, the implications of the research reported here may be helpful to address the construction of cheap and reliable prediction systems in accordance with the peculiarities of ICUs and kinds of patients.

Key words: Intensive care, mortality risk prediction, support vector machines, reduced number of variables and cases

1. Introduction

Survival probability predictions in critically ill patients are mainly used to measure the efficacy of intensive care unit (ICU) treatment. Risk stratification of patients allows comparison of the observed outcomes versus accepted standards provided by probability prediction functions. The importance of ICU assessment should be borne in mind, as it is estimated that critical care consumes 10% to 12% of all healthcare costs. Moreover, the average daily cost per patient in ICUs was about $3000 in the USA in 2001 [1]. On the other hand, the literature also shows that prognoses have constituted an important dimension of critical care, as patients and their families seek predictions about the duration and outcome of illness [2]. For a very thorough study of the uses of prognostic models in Medicine, see [3].

The available models for predicting outcomes in ICUs are usually scoring functions that estimate the probability of hospital mortality of critically ill adults [4]. This is the case of *acute physiology and chronic health evaluation* (APACHE) [5], *simplified acute physiology score* (SAPS) [6], and *mortality probability models* (MPM) [2]. These

*Corresponding author. Full address: Campus de Viesques. E33271 Gijón, Asturias. Spain. Email: antonio@aic.uniovi.es. Telephone: +34 985 18 21 22. Fax: +34 985 18 21 25

Email addresses: oluaces@aic.uniovi.es (Oscar Luaces), francisco.taboada@sespa.princast.es (Francisco Taboada), guillermo.muniz@sespa.princast.es (Guillermo M. Albaiceta), ldominguez@hurh.sacyl.es (Luis A. Domínguez), penriquez@hurh.sacyl.es (Pedro Enríquez)

predictors were induced from data on thousands of patients using logistic regression. The data required by these systems is gathered for each patient in a set of variables that can be split into 3 groups according to the source of information: monitoring devices, laboratory findings, and demographic and diagnostic features.

For instance, APACHE III inputs include age, 16 acute physiologic variables that use the worst values from the first 24 hours in the ICU (temperature, heart rate, blood pressure, respiratory rate, oxygenation, acid-base status, serum sodium, serum blood urea nitrogen, serum creatinine, serum albumin, serum bilirubin, serum glucose, white cell count, hematocrit, itemized Glasgow coma score, and urine output), preexisting functional limitations, major comorbidities, and treatment location immediately prior to ICU admission.

A major limitation to regular use of these predictors is the high cost of data collection. In fact, some of the variables required do not eventually form part of the clinical routine, and may not be recorded in some patients. In addition, other variables are not automatically included in a computerized database of patients. Therefore, careful records of data for every patient would need the assistance of specialized personnel for manual data entry, which would increase the costs of ICU cares [7, 8]. Notice that although the quality of data would theoretically be guaranteed, a number of studies have found significant interobserver variability in collecting data [4]. Furthermore, in order to be useful, these prediction functions need to be frequently updated. They must somehow capture the newest treatments and practices.

This paper provides a pair of tools that overcome these difficulties and make predictors more usable. First, we propose a new method to efficiently build effective scoring functions. In fact, it is acknowledged that one way to improve the calibration consists in building *local models*, as the accuracy of predictors is very sensitive to patient population changes [4], although the problem is actually how to obtain sufficient data records to build robust predictors. The method proposed here is able to learn functions that accurately predict the probability of survival of patients; however, the induction of these functions does not require records from thousands of patients.

We shall see that, using the appropriate method, building new predictors may be better than re-calibrating heavy predictors like APACHE. Moreover, we shall show that it is possible to build customized prediction systems in accordance with the peculiarities of ICUs and patients. These predictors could be reliable, and their construction and use could be economically feasible, given that they would require only a small number of variables. Notice that predictors built with a small number of variables could be used even for retrospective studies where the available data is reduced.

The second contribution of the paper is a study of the relevance of the variables involved in predictions. The aim is to gain insight into the factors that contribute to the actual prediction capabilities in different meaningful medical contexts. We shall discuss the role of groups of variables in units with and without coronary patients, and in patients aggregated according to their treatment location immediately prior to ICU admission. Surprisingly, we found that most of the prediction capability can be achieved using only a group of basic clinical variables. This group is made up of demographic and diagnostic data, adding simple tests or observations that are routinely recorded for ICU admissions. Additionally, we identify which groups of variables are more or less useful, depending on the context. For instance, the detection of multi-organ failures is related to laboratory findings more than monitoring, which would be more predictive in coronary patients.

In the next section we shall present a novel prediction method. It makes intensive use of the so-called support vector machines (SVM), a powerful family of algorithms for learning classification and regression tasks [9]. The method requires two stages: the first one was described in [10], while the second stage utilizes a *grid search* to optimize model parameters of SVM [11].

Throughout the paper, we use a data set collected in general ICUs at 10 units of hospitals in Spain, 6 of which include coronary patients, while the other 4 do not treat coronary diseases. The total number of patients considered in our study was 2501, 19.83% of whom did not survive. Using these data, we report a comparison between the SVM method with other approaches based on logistic regression (LR), including a second-level recalibration of APACHE induced from the available data.

2. Predicting probabilities with SVM

This paper addresses the task of predicting probabilities from the point of view of Machine Learning. In order to induce such predictions, we collect training sets with descriptions of patient states and their outputs codified by '+1' when the patient has survived, and '−1' otherwise. As there are two classes, the initial temptation is to tackle this learning task as a binary classification. However, we shall try to extract from the data all the useful knowledge represented by probabilities of survival.

In the following subsections, we discuss different options to learn probabilities from the point of view of SVM, including the approach presented in [10]. This approach stresses the idea that the severity of illness can be seen as a ranking computed from patients' records; probabilities are thus *just* a mapping from rankings onto a $[0, 1]$ scale. After providing a detailed explanation of this approach, we devote a subsection to discussing the readability of the models thus obtained. Finally, we make an experimental comparison of the SVM approaches versus other two learners based on LR.

2.1. The core idea of the method

The misclassification rate (or accuracy) is commonly used in Machine Learning to measure the performance of

predictors in classification tasks. However, this measure is not adequate when classes are very unbalanced and thus is rarely used in medical contexts like survival prediction. Instead, the area under a receiver operating characteristic (ROC) curve (AUC for short) is often used. This amount can be interpreted as the degree of coherence between a continuous output (such as the probability) and a binary classification. It should be stressed that this coherence is established in terms of orderings. For this purpose, continuous outputs or scores are used to rank available cases.

Within this context, Hanley and McNeil showed in [12] that the AUC is the probability of a correct ranking; in other words, it is the probability that a randomly chosen subject of class '+1' is (correctly) ranked with greater output than a randomly chosen subject of class '−1'. Therefore, the AUC coincides with the value of the Wilcoxon-Mann-Whitney statistic. This observation is crucial in the method presented in the following subsections. Notice that from this point of view, the AUC would not be a measure of the discriminant power of a classifier. If the threshold of probability to decide a class is 0.6 or 0.5, the rate of successful classifications will be different, but the AUC will remain invariable, since it only depends on the ranking induced by the probability. We shall return to this issue later, in Section 2.4, when presenting a formal definition of the AUC as a measure of the quality of a ranking.

In addition to probabilities, there are other possible ways to map cases of a classification task into continuous values. For instance, SVM learn hypotheses that return positive values for cases of one class, and negative values for the other class. Hence, in order to learn a probability distribution using SVM, it is necessary to transform these scores or continuous outputs into probabilities. But this is precisely what the method presented by Platt in [13] does. The core idea is to fit a sigmoid using a maximum likelihood procedure.

The novelty of the method reported in this paper is that, in order to fit a sigmoid to the SVM outputs using Platt's method, it is better to optimize the AUC than to minimize the error rate with a classification SVM. For this reason, in the next section we shall discuss how to optimize the area under the ROC curve (AUC) using a support vector method [14, 15], and how to then obtain probability predictions. The rationale behind this approach is that the quality of the sigmoid fit depends on the quality of the ranking of the scores. Thus, if most of the cases with a higher score than a given one of class y have a class greater than y, then the task of the sigmoid can be easily accomplished, and the performance of the final probability is improved.

As explained in the introduction, the method proposed in this paper comprises two stages. Once we have defined a parametrized learning algorithm, such as the SVM outlined above, we shall use a grid search to optimize the set of parameters involved in the learning process. This mechanism has been successfully used in many Machine Learning applications; in [11] the authors describe its use in learning mortality prediction models for percutaneous coronary interventions. We shall explain this important stage of the method later, in the experimental comparison.

2.2. The goodness of probability predictions

Let us now present the formal setting used in the remainder of the paper to describe both the learning algorithms and the measures of the goodness of probability estimations. As explained above, the probabilities may be considered as a ranking of the severity of illness; from this perspective, the AUC is a measure of the quality of probabilities. However, there is another, more direct measure of the quality of probabilities, the Brier score, which will be presented here.

Let $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ be a training set for a learning task in which a function (or hypothesis) is sought that is able to return outputs y_i from points $\boldsymbol{x}_i$ of an input space $\mathcal{X}$. An important issue when we are learning is to fix the way in which we are going to measure the quality of the result. Formally, given S, the aim of learning is to find a hypothesis h (from a given hypotheses space) that minimizes the average *loss* extended over the set of independently identically distributed (i.i.d.) test sets S', usually represented by $\Delta(h, S')$.

In the survival prediction task, training and test examples have no probability attached, they are labeled with $+1$ or -1. Therefore, we shall assume that the *true* class probability, $Pr^{true}(y = +1|\boldsymbol{x})$, is 1 when the class of $\boldsymbol{x}$, y, is $+1$, and 0 otherwise. In this context, there is basically one standard loss function: the average quadratic deviation. In symbols, the probability loss is given by

$$\Delta_{\mathrm{Pr}}(h, S') = \frac{1}{|S'|} \sum_{\boldsymbol{x}'_i \in S'} (h(\boldsymbol{x}'_i) - p_i)^2 \qquad (1)$$

where the hypothesis h returns the estimation of the probability $h(\boldsymbol{x}) = Pr(y = +1|\boldsymbol{x})$, and p_i stands for the observed probability of the i-th case, $p_i = Pr^{true}(y = +1|\boldsymbol{x}_i)$.

The measurement in Equation (1) is frequently used in medicine and meteorology, and is known as the *Brier* [16] index or *score*. If the number of possible outputs is greater than two, the estimated probabilities can be seen as a vector, and the Mean Square of the Euclidean (MSE) distance from predicted and observed probabilities is then used. It can be seen that, in the survival prediction task, the MSE is 2 times the Brier score.

2.3. Optimizing accuracy plus a sigmoidal transformation

The straightforward approach to the ICU problem is a binary classification SVM followed by a sigmoid estimated using Platt's method [13]. Thus, given the training set S, we can use a transformation ϕ defined from input points in $\mathcal{X}$ into a feature space $\mathcal{H}$, where classes should be mostly separable by means of a linear function. As is well known, $\mathcal{H}$ must have an inner product $\langle \bullet, \bullet \rangle$, and

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \qquad (2)$$

is called the *kernel* function of the transformation. We shall use the radial basis function (RBF) kernel, which is defined by

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}. \quad (3)$$

The work of the SVM consists in solving the following convex optimization problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},\xi} & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum\limits_{i=1}^{n}\xi_i, \\ \text{s.t.} & y_i(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad i = 1, \ldots, n. \end{array} \quad (4)$$

Then, the classification is accomplished by the hypothesis

$$\text{sign}(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b). \quad (5)$$

It can be seen that the kernel and the vector $\boldsymbol{\alpha} = (\alpha_i : i = 1, \ldots, n)$ of Lagrange multipliers define the implementation of function (5) computed from input space points $\boldsymbol{x}$ as follows:

$$\text{sign}(\langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + b) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right). \quad (6)$$

According to (4), the aim of this function is to maximize the margin (between classes) and to minimize the training loss. In fact, the sum of the so-called slack variables, $\sum_{i=1}^{n}\xi_i$, is an upper bound of misclassifications of (6) on the training set. It is acknowledged that the function (6) thus obtained has good classification accuracy on unseen cases.

In order to compute the probabilistic outputs, we get rid of the sign function and only consider the continuous outputs

$$f_{ac}(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + b = \sum_{i=1}^{n}\alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \quad (7)$$

Platt's method then fits a sigmoid to estimate probabilities:

$$h_{ac}(\boldsymbol{x}) = Pr(y = +1|\boldsymbol{x}) = \frac{1}{1 + e^{A_{ac}\cdot f_{ac}(\boldsymbol{x}) + B_{ac}}}. \quad (8)$$

Figure 1 depicts the fit of this sigmoid to the data set of all patients (2501) in all the available units. Since this figure is just drawn to illustrate the learning process, the train and test sets used are the same. Notice that the frequencies of f_{ac} values follow a bell-shape distribution, with most individuals having positive values, which means that they have a survival prediction.

2.4. Optimizing the AUC instead of accuracy

When classification predictions are made comparing the values returned from patients' descriptions $\boldsymbol{x}$ by a rating function with a threshold, as in SVM classification (see Equations (5, 6)), then the performance of these predictions can be assessed using the AUC. According to its probabilistic interpretation, the complementary of this amount (1-AUC) can be used as a loss function. Thus, if g is a hypothesis, its loss evaluated on a test set S' is

$$\begin{array}{rcl} \Delta_{AUC}(g, S') & = & Pr(g(\boldsymbol{x}'_i) \leq g(\boldsymbol{x}'_j)|y'_i > y'_j) \\ & = & \frac{\sum_{i,j:y'_i > y'_j} 1_{g(\boldsymbol{x}'_i) \leq g(\boldsymbol{x}'_j)}}{\sum_{i,j} 1_{y'_i > y'_j}}. \end{array} \quad (9)$$

Let us stress that the explicit objective of the SVM presented in the preceding section is not to minimize Equation (9). Paper [17] provides a detailed statistical analysis of the difference between maximizing the AUC and minimizing the error rate in binary classification tasks.

In [18], Herbrich et al. presented a direct approach that solves a general ranking problem which is applicable to maximizing the AUC. The core idea is that if a hypothesis $f : \phi(\mathcal{X}) \to \mathbb{R}$ is linear and has to fulfill that $f(\phi(\boldsymbol{x}_i)) > f(\phi(\boldsymbol{x}_j))$, since $y_i > y_j$, then

$$f(\phi(\boldsymbol{x}_i)) > f(\phi(\boldsymbol{x}_j)) \Leftrightarrow f(\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)) > 0. \quad (10)$$

Notice that this statement converts ordering constraints into classification constraints (with one class), though now the input space is $\mathcal{X} \times \mathcal{X}$ and each pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is represented by the difference $\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)$. According to this approach, the aim is to find a hypothesis $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle$ such that $\boldsymbol{w}$ solves the following convex optimization problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},\xi} & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum\limits_{i,j:y_i > y_j}\xi_{i,j}, \\ \text{s.t.} & \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle - \langle \boldsymbol{w}, \phi(\boldsymbol{x}_j)\rangle \geq 1 - \xi_{i,j}, \\ & \xi_{i,j} \geq 0, \quad \forall i, j : y_i > y_j. \end{array} \quad (11)$$

For each $\boldsymbol{x}$ of the input space, the hypothesis thus found returns

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle = \sum_{y_i > y_j}\alpha_{i,j}(K(\boldsymbol{x}_i, \boldsymbol{x}) - K(\boldsymbol{x}_j, \boldsymbol{x})), \quad (12)$$

where $\alpha_{i,j}$ are once again the Lagrange multipliers computed by the optimizer.

Unfortunately, this approach leads to dealing with one constraint for each element of the data set

$$\bar{S} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j; +1) : y_i = +1 > y_j = -1\}, \quad (13)$$

whose size is the number of positive (class +1) examples times the number of negatives, #pos×#neg, i.e. $\mathcal{O}(n^2)$ when the size of S is only n. This means that some applications become intractable, although the approach (or a simplified version of it) has been successfully used on other occasions [19, 20].

To mitigate the difficulties caused by the size of data sets, Herbrich's approach cannot be directly reformulated

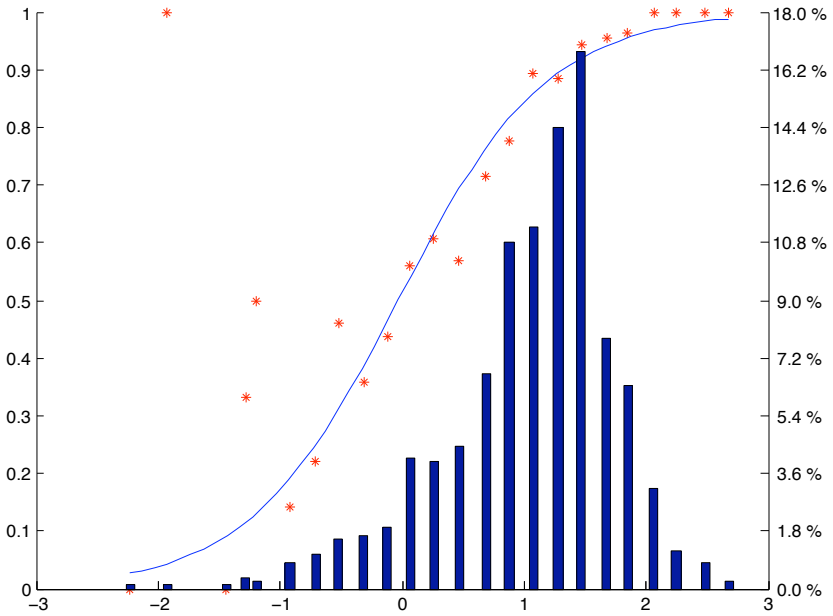

Figure 1: The fit of the sigmoid to the data set of all patients (2501). The horizontal axis represents the outputs of an SVM. Each '*' mark is the average posterior probability for all examples falling into a bin of width 0.2. The sigmoidal function is the estimation computed by Platt's method [13] (the output values are labeled on the left vertical side), while the bell-shaped function is the histogram for $Pr(f(\boldsymbol{x}))$ for all the examples. Frequencies are labeled on the right

in a straightforward way as an optimization problem with a *small* number of constraints. The main problem is that the loss function (1-AUC) (see Equation (9)) cannot be expressed as a sum of disagreements or errors produced by each input $\boldsymbol{x}_i$.

Following a different procedure, Joachims recently proposed a multivariate approach in [14, 15] to solve this problem with a convex optimization problem that converges using only a few constraints.

The optimization problem is:

$$\begin{aligned} \min_{\boldsymbol{w},\xi} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\xi, \\ \text{s.t.} \quad & \langle \boldsymbol{w}, \sum_{y_i > y_j} (1 - y'_{i,j})(\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j))\rangle \\ & \geq \Delta_{AUC}((1,\ldots,1)(y'_{i,j})) - \xi, \\ & \forall y'_{i,j} \in \{+1,-1\}^{\#pos\cdot\#neg} - \{(1,\ldots,1)\}. \end{aligned} \tag{14}$$

Despite the enormous potential number of constraints, the algorithm proposed in [14, 15] converges in polynomial time. Moreover, it only requires a small set of constraints. However, the most interesting result is that the solution $\boldsymbol{w}$ of problem (14) is also the same as that of the optimization problem (11). Additionally, the slack variables in both cases are related by

$$\xi = 2 \sum_{y_i > y_j} \xi_{i,j}. \tag{15}$$

Finally, the multivariate SVM returns a function f_{AUC} of the form

$$f_{AUC}(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle. \tag{16}$$

Then Platt's method can fit a sigmoid to transform the output of f_{AUC} into a probability.

$$h_{AUC}(\boldsymbol{x}) = Pr(y = +1|\boldsymbol{x}) = \frac{1}{1 + e^{A_{AUC}\cdot f_{AUC}(\boldsymbol{x}) + B_{AUC}}} \tag{17}$$

2.5. Regression is a baseline approach

Considering that probabilities are real numbers, regression algorithms must constitute an initial attempt to learn them. However, the regression approach ignores the fact that the aim is to obtain a probability; therefore, this approach should only be considered as a baseline.

To rewrite the learning task as a regression problem, all training examples of class -1 are labeled as 0. Moreover, in order to maintain the uniformity of approach with preceding subsections, we consider the regression based on support vectors and subsequently use the so-called support vector regression (SVR). Although there are least squares SVR, we use the standard version; i.e. a learner of a function

$$f_{Re}(\boldsymbol{x}) = \sum_{i=1}^{n} (\alpha_i^- - \alpha_i^+) K(\boldsymbol{x}_i, \boldsymbol{x}) + b^*, \tag{18}$$

where K is once again the RBF (3) kernel, and α_i are the Lagrange multipliers of the solution to the convex opti-

mization problem:

$$\begin{aligned} \min_{\boldsymbol{w},\xi} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-), \\ \text{s.t.} \quad & (\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b) - y_i \leq \epsilon + \xi_i^+, \\ & y_i - (\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b) \leq \epsilon + \xi_i^-, \\ & \xi_i^+, \xi_i^- \geq 0, \quad i = 1, \ldots, n. \end{aligned} \tag{19}$$

However, given that nothing forces f_{Re} (18) outputs to be in $[0, 1]$, we set the hypothesis output to 1 whenever f_{Re} returns values above 1, and to 0 for f_{Re} values below 0. Finally, in symbols, we have the hypothesis

$$h_{Re}(\boldsymbol{x}) = \max\{0, \min\{1, f_{Re}(\boldsymbol{x})\}\}. \tag{20}$$

2.6. Readability of models

After reading the formulas presented in the preceding sections, it is clear that the predictors obtained by training the SVM algorithms are black-boxes that hide the way in which variables operate. This is also the case of models obtained using artificial neural networks, like in [21] for instance.

We think that the achievements in performance, measured by objective tests, are the main merit of predictors; nevertheless, users sometimes want to *read* and *understand* the model. An option to mitigate this problem might be to use more readable knowledge representations to describe the models. This is the case of [22], where the authors show how a classification tree can be used for prognosis in ICU. In [23], the predictor is a set of decision rules. Notice that, in addition to the benefits of readability, these explicit languages of knowledge representation exploit an economy of variables. Thus, in [23], the authors stress the fact that the prediction models obtained by a rule learner require the collecting of less data than other conventional predictors.

To address the interpretability of our models, we explored another path. We shall discuss the interactions of meaningful sets of variables, explicitly computing the differences in prediction performance achieved by SVM learners with and without those sets of variables. Moreover, we shall compute performance differences in several medical contexts. The aim is to gain insight into the prediction mechanism of the models proposed in this paper.

Nonetheless it is worth mentioning that readability is not always present in popular ICU predictors. It is noticeable that APACHE III is not at all readable. As far as it is possible to say without unveiling the commercial secret of the APACHE III formula, predictions are computed in two stages. First the acute physiology score *(APS)* is determined: this is an integer score of the severity of patients defined as the sum of rules of parabolic functions associated with raw data from laboratory findings and monitoring devices. Then, a 2^{nd}-degree polynomial is fitted using logistic regression; the monomials are built using APS and the other demographic and diagnostic variables.

2.7. Experimental evaluation of learning approaches

In order to evaluate the learning approaches presented in previous sections, we compared them experimentally. Let us recall that the SVM approaches are: SVM followed by Platt's fit of a sigmoid, the accuracy optimizer described in Subsection 2.3, which will be represented simply by SVM; the multivariate version, aimed at optimizing the AUC[1] (Subsection 2.4), SVM_{AUC} for short; and finally the regression approach, SVR (Subsection 2.5).

The second group of learners used in these experiments employ logistic regression. The most representative member of this group is the commercial system APACHE III; we used the customization described in [24], which was developed to improve its performance in Spain. First of all, we should point out that comparison of SVM learners versus APACHE III is unfair, since APACHE III was trained with a cohort of 17440 patients from 40 different hospitals in the USA [5]; the Spanish version used records of 10929 patients from 86 ICUs; while the available data sets in our experiments only included 2501 patients. Nevertheless, the comparison is useful to test whether or not the scores achieved by SVM methods are good enough to be considered for future learning tasks.

However, it may also be argued that SVM learners have an advantage over APACHE III, since they have been trained in our data set, while APACHE only sees the data for testing. For this reason, we built a second-level recalibration of the APACHE III using the same data sets used for SVM learners (see Table 1): the coefficients of APACHE III formula were induced on each learning task. Notice that this version could be called a *local APACHE III*. In the experiments, we shall refer to this learner as LR_{APS}. Additionally, to test whether possible differences between SVM and logistic regression families arise from the use of APS, we shall use a logistic regression algorithm applied to the same raw data used by SVM learners; this algorithm will be represented by *LR*. The implementation of logistic regression used in all cases is described in [25][2].

To estimate the performance of the algorithms described in the preceding section, we used data collected from ICUs at 10 different Spanish hospitals, 6 of which include coronary patients. It is acknowledged among the medical community that coronary diseases generally have a lower mortality risk than other critical illnesses. Hence, from a learning perspective, it makes sense to differentiate between ICUs with and without coronary patients.

The data were organized in 13 different training sets, one for each single unit, two collecting the data from non coronary/coronary ICUs respectively, and the last one containing all the data; see Table 1. Each patient in these data sets was described by the same set of variables used by APACHE III. However, to be handled both by SVM

[1] Software available at (Accessed: September 2, 2008) http://svmlight.joachims.org/svm_struct.html

[2] Software available at (Accessed: September 2, 2008) http://www.csie.ntu.edu.tw/∼cjlin/liblinear

	Data sets												
Codes	1	2	3	4	5	6	7	8	9	10	NC	C	ALL
# patients	108	189	194	194	195	239	269	297	337	479	919	1582	2501

Table 1: Size (number of patients) and codes of data sets used in the experiments. The first 10 sets correspond to the intensive care units separately. The 11-th data set (NC) is formed by non coronary patients: those from units coded 2, 3, 6 and 8. Units including coronary patients (1, 4, 5, 7, 9, 10) are grouped in the data set coded by C. Finally, the 13-th data set is the set of ALL patients. Notice that data sets are sorted by number of patients

	SVM_{AUC}	SVM	SVR	LR	LR_{APS}	APACHE III
1	0.786 ± 0.037	0.788 ± 0.056	0.751 ± 0.063	0.729 ± 0.054	0.825 ± 0.046	0.829 ± 0.041
2	0.775 ± 0.032	0.759 ± 0.024	0.725 ± 0.042	0.778 ± 0.032	0.751 ± 0.038	0.764 ± 0.054
3	0.746 ± 0.043	0.745 ± 0.038	0.687 ± 0.032	0.731 ± 0.045	0.684 ± 0.035	0.768 ± 0.046
4	0.794 ± 0.045	0.782 ± 0.051	0.794 ± 0.032	0.779 ± 0.045	0.698 ± 0.065	0.837 ± 0.041
5	0.851 ± 0.048	0.848 ± 0.045	0.836 ± 0.044	0.833 ± 0.024	0.766 ± 0.048	0.880 ± 0.036
6	0.768 ± 0.032	0.743 ± 0.031	0.736 ± 0.033	0.760 ± 0.022	0.762 ± 0.024	0.775 ± 0.035
7	0.824 ± 0.036	0.786 ± 0.044	0.822 ± 0.044	0.802 ± 0.048	0.831 ± 0.032	0.891 ± 0.039
8	0.846 ± 0.010	0.843 ± 0.024	0.824 ± 0.025	0.848 ± 0.024	0.829 ± 0.016	0.868 ± 0.021
9	0.792 ± 0.015	0.802 ± 0.024	0.817 ± 0.015	0.815 ± 0.018	0.802 ± 0.027	0.813 ± 0.024
10	0.782 ± 0.025	0.721 ± 0.021	0.729 ± 0.049	0.763 ± 0.045	0.775 ± 0.027	0.779 ± 0.024
NC	0.809 ± 0.020	0.799 ± 0.017	0.784 ± 0.025	0.798 ± 0.018	0.789 ± 0.020	0.807 ± 0.016
C	0.824 ± 0.021	0.806 ± 0.024	0.813 ± 0.020	0.816 ± 0.018	0.812 ± 0.020	0.826 ± 0.013
ALL	0.824 ± 0.007	0.819 ± 0.009	0.822 ± 0.009	0.820 ± 0.008	0.809 ± 0.009	0.823 ± 0.015

Table 2: AUC estimated by a 10-fold cross-validation for the learners described in the text and, for the commercial system APACHE III, computed using the data available as the test set. The first column reports the code of the data set described in Table 1

	SVM_{AUC}	SVM	SVR	LR	LR_{APS}	APACHE III
1	0.173 ± 0.013	0.168 ± 0.016	0.188 ± 0.024	0.186 ± 0.018	0.140 ± 0.017	0.147 ± 0.014
2	0.182 ± 0.008	0.190 ± 0.007	0.222 ± 0.009	0.191 ± 0.012	0.190 ± 0.011	0.170 ± 0.021
3	0.171 ± 0.009	0.182 ± 0.011	0.209 ± 0.006	0.191 ± 0.015	0.193 ± 0.012	0.159 ± 0.017
4	0.107 ± 0.007	0.110 ± 0.006	0.118 ± 0.005	0.120 ± 0.008	0.129 ± 0.013	0.096 ± 0.012
5	0.104 ± 0.007	0.105 ± 0.009	0.131 ± 0.010	0.121 ± 0.009	0.135 ± 0.008	0.108 ± 0.012
6	0.156 ± 0.008	0.156 ± 0.008	0.178 ± 0.008	0.158 ± 0.009	0.158 ± 0.007	0.146 ± 0.009
7	0.100 ± 0.006	0.109 ± 0.006	0.114 ± 0.004	0.106 ± 0.010	0.101 ± 0.011	0.085 ± 0.011
8	0.125 ± 0.005	0.121 ± 0.007	0.129 ± 0.006	0.125 ± 0.009	0.130 ± 0.007	0.113 ± 0.005
9	0.113 ± 0.002	0.112 ± 0.005	0.122 ± 0.003	0.116 ± 0.006	0.123 ± 0.007	0.107 ± 0.006
10	0.110 ± 0.005	0.121 ± 0.004	0.120 ± 0.005	0.118 ± 0.008	0.118 ± 0.005	0.122 ± 0.007
NC	0.146 ± 0.005	0.146 ± 0.005	0.167 ± 0.005	0.149 ± 0.006	0.151 ± 0.006	0.143 ± 0.006
C	0.108 ± 0.004	0.113 ± 0.006	0.126 ± 0.003	0.110 ± 0.005	0.110 ± 0.005	0.109 ± 0.005
ALL	0.121 ± 0.002	0.122 ± 0.003	0.133 ± 0.003	0.123 ± 0.002	0.124 ± 0.002	0.122 ± 0.005

Table 3: Brier scores estimated by a 10-fold cross-validation for the learners described in the text and, for the commercial system APACHE III, computed using the data available as the test set. The first column reports the code of the data set described in Table 1

and logistic regression algorithms, we codified each discrete variable using as many new binary variables (with values 0 and 1) as the number of possible values of the original variable, setting only the variable corresponding to the discrete value actually taken by the original variable to '1'. There is one exception: the three components of the Glasgow coma score (eye, verbal and motor responses) were used in their original numeric form. The idea is to take advantage of the ordered relation between their values and the severity of the illness of the patient.

Performance estimations were made using a 10-fold stratified cross-validation on each of the data sets, for all the algorithms except APACHE III. As it was already trained with a different data set, we used the available data just to test its predictions. Additionally, the data was standardized according to the mean and deviation observed on each training fold.

It is important to recall that the AUC achieved by the Spanish version of APACHE III in our experiments, 82.27% (in percentage) is similar to the amount reported by Rivera-Fernández et al. in [24]: 81.82%. This fact supports the representativeness of the sample of critically ill patients considered in the experiments described here.

Grid search

As is usual when dealing with SVM, the parameter setting stage is very important. Thus, in order to carry out the experiments described in this section, we wrapped each algorithm with a grid search procedure [11]. Thus, every time an algorithm was trained on a data set T, an internal two-fold cross validation (repeated 3 times) was performed to estimate the best combination of parameter values for T.

The values examined for the regularization parameter C (see optimization problems in Section 2) ranged from 10^{-3} to 10^2, while the values for the RBF kernel parameter σ ranged from 10^{-3} to 10^0, varying the exponents in steps of 1 in both cases. The grid search mechanism selected the combination of parameters that minimized a given loss function in the internal cross validation; this loss function was the Brier score for all learning algorithms except for the SVR, which used the mean average deviation (MAD), commonly used in regression tasks.

The implementation used for logistic regression [25] also requires a regularization parameter, C. To estimate the best value for C we used the same search as that used for SVM with the same range of possible values.

Scores and comparisons

Table 2 show the results obtained in AUC in the experimental setting described above. Focusing on the results obtained by the three support vector algorithms, we can observe that, in general, the best performance (highest AUC) is achieved by multivariate SVM_{AUC}. In the logistic regression family, the commercial version of APACHE III outperforms all the other learners considered, both SVMs and local logistic regressions. Apart from this, LR performs better than LR_{APS}, the local APACHE III, the majority of times: 9 out of 13. However, SVM_{AUC} is superior to LR in 10 out of 13 data sets. Figure 2 depicts the differences in AUC multiplied by 100 for ease of reading. The same conclusions can be achieved if we look at Brier scores; see Table 3.

Following Demšar in [26], the recommended test for statistically comparing two classifiers by measuring the AUC (or other performance measurements like the Brier score) on multiple data sets is the Wilcoxon signed ranks test. Using this test, the differences of the commercial system APACHE III with all the other learners are statistically significant according to the test with a threshold $p < 0.05$. Additionally, we also found statistically significant differences between SVM_{AUC} and the rest of learners ($p < 0.05$); that is, with learners that used the same training sets as SVM_{AUC}. The differences between LR and LR_{APS} are not significant. The same significant differences appear in the Brier scores reported in Table 3.

Let us stress that, although the optimization problem posed to SVR is precisely the minimization of the distance between true and predicted probabilities, a large amount of data is required to tie the SVM_{AUC} Brier scores. The underlying reason explaining this behavior may be that the hypothesis space used by SVR is not adequate to induce probability distributions from a reduced set of training data, even with an RBF kernel.

From Figure 2, it can be observed that the differences mainly appear in the first data sets, i.e. the smallest. SVR and LR performance was particularly poor in these cases. Bearing in mind that the rows of Tables 2 and 3 are in ascending order of size of the data sets, the trend indicates that performance in all learners could be improved if more training cases were available. In fact, when the data set included all available patients' records, the results obtained were similar to or better than those yielded by APACHE III. The exceptions are LR_{APS} in AUC, and SVR in Brier score.

On the other hand, it can also be observed that, for all learners, correct survival predictions seem to be slightly harder to obtain for ICUs without coronary patients (data set with code NC, including units 2, 3, 6 and 8) than for ICUs including coronary patients.

3. The role of variables involved in predictions

In this section we attempt to gain insight into the role played in predictions by the variables used to describe patients' records. Let us recall that we recorded the set of variables employed by APACHE III, the golden standard in the field. In addition to demographic data and a brief clinical history, these variables include 16 acute physiological records that use the worst values from the first 24 hours in the ICU. In order to study the prediction capabilities of these variables, we divided the whole set into 3 groups

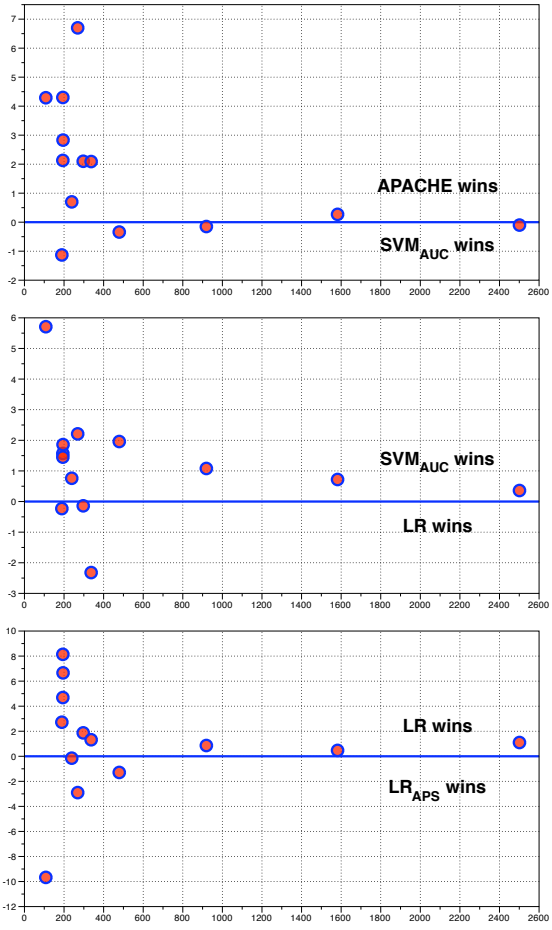

Figure 2: Differences in $(AUC * 100)$ scores achieved by APACHE, SVM_{AUC}, LR, and LR_{APS}, see Table 2. The horizontal axis represents the number of patients used for training, while the vertical axis represents the differences between couples of those predictors. In all cases, the differences decrease as the number of training patients increase. Note the scores of SVM_{AUC} when the number of patients is greater than 500; recall that APACHE III was trained with thousands of cases: 17 thousand in the original version, and almost 11 thousand in the Spanish version

according to the source of information of these variables, see Table 4.

We labelled the first group of variables with the tag *clinical*. In this group, we collect demographic and diagnostic data, adding simple tests or observations. Let us emphasize that the recording of these data can be done at no cost. The second group of variables, *monitoring*, is formed by data supplied by monitoring devices. Finally, the third group of variables comes from *laboratory* analyses.

In the following subsections, we shall measure the weight of these groups of variables when the predictions are sought in different contexts. We shall discuss contexts defined by different kinds of ICUs and contexts characterized by the treatment location of patients immediately prior to ICU admission.

To measure the weight of a group of variables, we compare the scores in AUC (and in Brier score) achieved by a learning method (estimated by a 10-fold cross validation) when using all the variables and when using only those of the group to be measured. We use two learning methods for this study: the best option of support vectors' family, SVM_{AUC}, and the logistic regression that uses raw data, LR. Notice that APACHE III or its local recalibration could not be used here, as the APS would have no sense if it were computed with a limited number of variables.

Group	Variables
Clinical	age, sex, mechanical ventilation, pre-existing comorbidities, major diagnostic category, type of patient (scheduled or urgent surgery, trauma, medical) location prior to ICU (other hospital, ward, scheduled or urgent surgery), Glasgow coma score
Monitoring	temperature, blood pressure, heart and respiratory rate, urinary output
Laboratory	gas exchange (PaO_2, $PaCO_2$, pH, FiO_2), white cell count, hematocrit, serum: sodium, blood urea nitrogen, creatinine, albumin, bilirubin, glucose

Table 4: The division of variables used to record the state of ICU patients into 3 groups according to their source of information

3.1. Groups of intensive care units

From a medical point of view, the most obvious division between ICUs can probably be stated in terms of those that include coronary patients or not. Let us recall that there are important differences in mortality risk and treatments applied to patients in both kinds of ICUs. Therefore, we decided to consider whether there are also differences in the hypotheses that predict the probabilities of survival.

Table 5 shows the scores achieved by SVM_{AUC} in these kinds of units considering different groups of variables defined in Table 4. To contrast the results, we included three data sets: coronary, non-coronary, and all units. On the other hand, the groups of variables used were: all variables, clinical, and clinical plus each one of the other two groups: laboratory and monitoring. We included clinical variables as they somehow constitute the basic information about a patient that it is routinely recorded.

First of all, we observe that the basic clinical variables provide surprisingly good results when we deal with the data set of all patients. The differences in AUC with the whole set of variables are just around 2 points down, while in the Brier score the gap is around 0.5 when the units of these scores are multiplied by 10^2.

	AUC			Brier Score		
	All ICUs	*Coronary*	*Non-coronary*	*All ICUs*	*Coronary*	*Non-coronary*
All	0.824 ± 0.007	$\mathbf{0.824} \pm 0.021$	$\mathbf{0.809} \pm 0.020$	0.121 ± 0.002	$\mathbf{0.108} \pm 0.004$	$\mathbf{0.146} \pm 0.005$
C+M	0.821 ± 0.008	$\mathbf{0.823} \pm 0.022$	0.799 ± 0.021	0.122 ± 0.001	$\mathbf{0.108} \pm 0.004$	0.150 ± 0.005
C+L	0.818 ± 0.008	0.809 ± 0.020	$\mathbf{0.805} \pm 0.018$	0.123 ± 0.002	0.112 ± 0.004	$\mathbf{0.149} \pm 0.004$
C	0.805 ± 0.010	0.794 ± 0.022	0.789 ± 0.021	0.126 ± 0.002	0.113 ± 0.004	0.158 ± 0.004

Table 5: Performance of survival predictions of SVM_{AUC} by groups of ICUs and groups of variables using 10-fold cross validation. Small differences (in bold) both in the AUC and Brier score indicate that monitoring has more relevancy than laboratory data in units with coronary patients; in units without these patients, the opposite situation occurs

	AUC			Brier Score		
	All ICUs	*Coronary*	*Non-coronary*	*All ICUs*	*Coronary*	*Non-coronary*
All	0.820 ± 0.008	$\mathbf{0.816} \pm 0.018$	$\mathbf{0.798} \pm 0.018$	0.123 ± 0.002	$\mathbf{0.110} \pm 0.005$	$\mathbf{0.149} \pm 0.006$
C+M	0.819 ± 0.009	$\mathbf{0.824} \pm 0.018$	0.788 ± 0.022	0.123 ± 0.002	$\mathbf{0.109} \pm 0.005$	0.152 ± 0.007
C+L	0.811 ± 0.008	0.799 ± 0.022	$\mathbf{0.803} \pm 0.016$	0.125 ± 0.002	0.113 ± 0.004	$\mathbf{0.147} \pm 0.005$
C	0.804 ± 0.011	0.797 ± 0.021	0.787 ± 0.020	0.126 ± 0.002	0.113 ± 0.004	0.153 ± 0.006

Table 6: Performance of survival predictions of LR by groups of ICUs and groups of variables using 10-fold cross validation. Small differences (in bold) both in the AUC and Brier score indicate that monitoring has more relevancy than laboratory data in units with coronary patients; in units without these patients, the opposite situation occurs

When we add monitoring or laboratory variables to the clinical records, we almost reach maximum predictive capacity. In the data set of patients from all units, the differences are inappreciable. However, in ICUs with coronary patients, monitoring is more useful for a prediction task than laboratory variables. In contrast, the opposite situation is true in units without coronary patients. These results are consistent when we measure the performance with AUCs or Brier scores.

Table 6 repeats the same situation, though in this case measured with LR scores. Notice that the LR scores are generally worse than those achieved by SVM_{AUC} in Table 5.

3.2. Groups of patients

The second context in which we studied the differences in the weight of variable groups arose from considering the treatment location of patients immediately prior to ICU admission. Table 7 reports the number and percentages of patients for each location in the whole data set of 2501 patients and the percentage of deaths. Notice that survival percentages are dramatically different. We accordingly used the situations with higher death rates in order to further our knowledge of the weight of variables.

Thus, we now consider 3 contexts to study the performance of the groups of variables as we did in the experiments reported in the preceding subsection. Table 8 gathers the results so obtained using SVM_{AUC}, while Table 9 reports the scores obtained by LR.

We observe that in the case of patients coming from a different hospital, no matter whether they come from an ICU, ward, or any other location in the other hospital, laboratory data are more predictive than monitoring, at least measured in AUC with SVM_{AUC}. If we learn using LR, however, these observations could not be induced from the obtained Brier score, see Table 9. One possible reason for this behavior is that the number of patients in this situation is too small, only 179; in these cases (see the scores in Table 3), the performance of LR decreases dramatically.

On the other hand, for patients who come from a ward in the same hospital, monitoring devices are more useful to predict their survival probabilities than laboratory data.

The third situation considered is that of patients admitted to an ICU subsequent to urgent surgery. The scores of SVM_{AUC} indicate that laboratory findings would be more useful than monitoring. Once again, however, the reduced number of data means that the LR scores do not present any appreciable difference, since in this case monitoring and laboratory variables have the same weight.

3.3. Discussion

We have described three prediction contexts in which laboratory findings are more useful than monitoring in predicting survival: units without coronary patients, and patients coming from other hospitals or from urgent surgery. In these cases, the risk of death of patients is usually related to multi-organ (respiratory, renal or hepatic) failure. The medical way of controlling the evolution of these diseases is by means of laboratory findings, which explains the results obtained.

On the other hand, monitoring is more useful than laboratory findings for patients coming from a ward in the same hospital as the ICU under consideration, or for units

	Other hospital	Ward	Scheduled surgery	Urgent surgery	Urgencies	Totals
% patients	7.16%	20.03%	15.11%	9.08%	48.58%	100%
# patients	179	501	378	227	1216	2501
% deaths	**27.93**%	**36.53**%	7.94%	**25.11**%	14.49%	19.83%

Table 7: Distribution of patients according to the treatment location immediately prior to ICU admission. In some cases (in bold), the percentage of deaths is significantly high

	AUC			Brier Score		
	Other Hospital	*Wards*	*Urg. Surgery*	*Other Hosp.*	*Wards*	*Urg. Surgery*
All	$\mathbf{0.760} \pm 0.037$	$\mathbf{0.780} \pm 0.018$	$\mathbf{0.764} \pm 0.023$	$\mathbf{0.175} \pm 0.009$	$\mathbf{0.184} \pm 0.005$	$\mathbf{0.162} \pm 0.005$
C+M	0.726 ± 0.034	$\mathbf{0.775} \pm 0.017$	0.757 ± 0.017	0.180 ± 0.009	$\mathbf{0.187} \pm 0.005$	0.163 ± 0.004
C+L	$\mathbf{0.746} \pm 0.041$	0.764 ± 0.015	$\mathbf{0.772} \pm 0.029$	$\mathbf{0.179} \pm 0.010$	0.190 ± 0.005	$\mathbf{0.160} \pm 0.007$
C	0.719 ± 0.044	$\mathbf{0.774} \pm 0.014$	0.700 ± 0.025	0.185 ± 0.008	$\mathbf{0.185} \pm 0.004$	0.173 ± 0.006

Table 8: Performance of survival predictions of SVM_{AUC} by groups of patients and groups of variables using 10-fold cross validation. Small differences (in bold) both in the AUC and Brier score indicate that laboratory data are more relevant than monitoring in patients from other hospitals or from urgent surgery, but monitoring variables are more relevant than laboratory data in patients from wards

with coronary patients. In these cases, survival is mostly threatened by cardiovascular complications, which are controlled by means of monitoring devices.

4. Conclusions

We have presented a reliable learning method for estimating the probability of hospital survival of critically ill patients. The method consists of a grid search [11] applied to the algorithm presented in [10], a variant of the standard procedure using SVM and Platt's method [13] to fit a sigmoid. Instead of using an SVM devised to optimize classification accuracy, we propose the use of a learner that optimizes the Area Under the ROC Curve (AUC). This can be done using a multivariate SVM described in [14, 15].

We experimentally compared the results obtained by this method with other approaches including logistic regression, and with a commercial scoring system trained with thousands of cases, APACHE III [5, 24]. One of the learners compared is a second-level recalibration of APACHE III, LR_{APS}. In the reported experiments, we used real data from 10 ICUs at hospitals in Spain that contain records from 2501 patients. The medical description of each patient includes monitoring variables, clinical analysis, and demographic and diagnostic features.

The method presented here, SVM_{AUC}, outperforms the other methods trained with the same data: the standard SVM approach, logistic regression with raw data, and the local APACHE LR_{APS}. Comparison with the commercial system APACHE III reveals similar scores (in fact slightly better for SVM_{AUC}) when the number of patients available for training SVM_{AUC} is higher than 500.

In addition, we have identified a number of medical contexts in which the weights of monitoring and laboratory variables have meaningful differences. These results have clear medical explanations. Furthermore, we have established that most of the prediction capability of models learned from our data can be achieved by means of a group of basic clinical variables. This group is made up of demographic and diagnostic data, adding simple tests or observations that are routinely recorded for ICU admissions.

From a practical point of view, the implication of the research reported here can be helpful to address the construction of cheap and reliable prediction systems in accordance with the peculiarities of ICUs and kinds of patients.

Acknowledgments

The research reported here is supported in part under grant TIN2005-08288 from the MEC (Ministerio de Educación y Ciencia, Spain). The authors acknowledge the work of the Grecia Group (*Grupo de Estudios y Análisis en Cuidados Intensivos*) in the collection of data. This group is formed by: Hospital Universitario Río Hortega, Valladolid, Spain (Luis A. Domínguez, Jesús Blanco, Pedro Enríquez), Hospital General Yagüe, Burgos, Spain (Martín De Frutos), Hospital Clínico Universitario, Salamanca, Spain (Víctor Sagredo), Hospital Río Carrión, Palencia, Spain (Juan J López-Messa, Ana Domínguez), Complejo Hospitalario de León, León, Spain (Demetrio Carriedo, Javier Collado), Hospital Central de Asturias, Oviedo, Spain (Francisco Taboada, Guillermo M. Albaiceta, José Antonio Gonzalo), Hospital General de Soria, Soria, Spain (Ángel García-Labattut), Hospital Clínico Universitario, Valladolid, Spain (Francisco Gandía, Felipe Bobillo), and Hospital San Agustín, Avilés, Spain (Manuel Valledor).

	AUC			Brier Score		
	Other Hospital	*Wards*	*Urg. Surgery*	*Other Hosp.*	*Wards*	*Urg. Surgery*
All	$\mathbf{0.721} \pm 0.028$	$\mathbf{0.772} \pm 0.016$	0.762 ± 0.027	0.200 ± 0.011	$\mathbf{0.188} \pm 0.007$	0.160 ± 0.007
C+M	0.706 ± 0.032	$\mathbf{0.788} \pm 0.016$	$\mathbf{0.746} \pm 0.019$	0.192 ± 0.013	$\mathbf{0.180} \pm 0.006$	$\mathbf{0.166} \pm 0.005$
C+L	$\mathbf{0.735} \pm 0.032$	0.761 ± 0.013	$\mathbf{0.747} \pm 0.025$	0.191 ± 0.011	0.190 ± 0.005	$\mathbf{0.165} \pm 0.007$
C	0.710 ± 0.042	0.779 ± 0.015	0.703 ± 0.027	0.186 ± 0.013	0.182 ± 0.005	0.176 ± 0.007

Table 9: Performance of survival predictions of LR by groups of patients and groups of variables using 10-fold cross validation. Differences (in bold) indicate noteworthy relevancies. In most cases, these coincide with those found with SVM_{AUC}; see caption in Table 8 and Subsection 3.2

References

[1] P. Provonost, D. Angus, Economics of end-life-care in the intensive care unit, Critical Care Med 29 (Suppl) (2001) 46–51.

[2] S. Lemeshow, D. Teres, J. Klar, J. S. Avrunin, S. H. Gehlbach, J. Rapoport, Mortality probability models (MPM II) based on an international cohort of intensive care unit patients, Journal of the American Medical Association 270 (20) (1993) 2478–2486.

[3] A. Abu-Hanna, P. Lucas, Prognostic models in medicine - AI and Statistical Approaches, Methods of Information in Medicine 40 (2001) 1–5.

[4] L. Ohno-Machado, F. Resnic, M. Matheny, Prognosis in Critical Care, Annu. Rev. Biomed. Eng. 8 (2006) 567–599.

[5] W. Knaus, E. Draper, D. Wagner, J. Zimmerman, M. Bergner, P. Bastos, C. Sirio, D. Murphy, T. Lotring, A. Damiano, The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults, Chest 100 (1991) 1619–1636.

[6] J. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, D. Villers, A simplified acute physiology score for ICU patients, Critical Care Medicine 12 (1984) 975–977.

[7] M. Render, D. Welsh, M. Kollef, J. Lott III, S. Hui, M. Weinberger, J. Tsevat, R. Hayward, T. Hofer, Automated computerized intensive care unit severity of illness measure in the Department of Veterans Affairs: Preliminary results, Critical Care Medicine 28 (10) (2000) 3540 – 3546.

[8] M. Render, J. Deddens, R. Freyberg, P. Almenoff, A. Connors Jr, D. Wagner, T. Hofer, Veterans Affairs intensive care unit risk adjustment model: Validation, updating, recalibration, Critical Care Medicine 36 (4) (2008) 1031–1042.

[9] V. Vapnik, Statistical Learning Theory, John Wiley, New York, 1998.

[10] O. Luaces, J. R. Quevedo, F. Taboada, G. M. Albaiceta, A. Bahamonde, Prediction of probability of survival in critically ill patients optimizing the Area Under the ROC Curve, in Manuel M. Veloso (Ed): Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '07), AAAI Press, Menlo Park, California, USA, 2007, pp. 956–961.

[11] M. Matheny, F. Resnic, N. Arora, L. Ohno-Machado, Effects of SVM parameter optimization on discrimination and calibration for post-procedural PCI mortality, Journal of Biomedical Informatics 40 (6) (2007) 688–697.

[12] J. Hanley, B. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.

[13] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, Massachusetts, USA, 2000, pp. 61–74.

[14] T. Joachims, A support vector method for multivariate performance measures, in: Luc de Raedt and Stefan Wrobel (Eds.): Proceedings of the International Conference on Machine Learning (ICML '05), Bonn, Germany, 2005, pp. 377–384.

[15] T. Joachims, Training linear SVMs in linear time, in: Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad (Eds.): Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM Press, New York, USA, 2006, pp. 217–226.

[16] G. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Rev 78 (1950) 1–3.

[17] C. Cortes, M. Mohri, AUC optimization vs. error rate minimization, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, Massachusetts, USA, 2004.

[18] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, Massachusetts, USA, 2000, pp. 115–132.

[19] T. Joachims, Optimizing search engines using clickthrough data, in: Osmar R. Zaïane, Randy Goebel, David Hand, Daniel Keim, and Raymond Ng (Eds.): Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM Press, New York, USA, 2002, pp. 133–142.

[20] A. Bahamonde, G. F. Bayón, J. Díez, J. R. Quevedo, O. Luaces, J. J. del Coz, J. Alonso, F. Goyache, Feature subset selection for learning preferences: A case study, in: R. Greiner, D. Schuurmans (Eds.), Proceedings of the International Conference on Machine Learning (ICML '04), Banff, Alberta (Canada), 2004, pp. 49–56.

[21] Á. Silva, P. Cortez, M. Santos, L. Gomes, J. Neves, Mortality assessment in intensive care units via adverse events using artificial neural networks, Artificial Intelligence in Medicine 36 (3) (2006) 223–234.

[22] S. de Rooij, A. Abu-Hanna, M. Levi, E. de Jonge, Identification of high-risk subgroups in very elderly intensive care unit patients, Critical Care 11 (2) (2007) R33.

[23] B. Nannings, A. Abu-Hanna, E. de Jonge, Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients, International Journal of Medical Informatics 77 (4) (2008) 272–279.

[24] R. Rivera-Fernández, G. Vázquez-Mata, M. Bravo, E. Aguayo-Hoyos, J. Zimmerman, D. Wagner, W. Knaus, The APACHE III prognostic system: customized mortality predictions for Spanish ICU patients, Intensive Care Medicine 24 (6) (1998) 574–581.

[25] C. J. Lin, R. C. Weng, S. S. Keerthi, Trust region newton method for logistic regression, Journal of Machine Learning Research 9 (Apr) (2008) 627–650.

[26] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, The Journal of Machine Learning Research 7 (2006) 1–30.