Instance-based Prognosis in Intensive Care Using Severity-of-illness Scores

Clarence Tan^1**, Linda Peelen**^{1,2,*}**, and Niels Peek**^{1}

^{1} Department of Medical Informatics, Academic Medical Center
University of Amsterdam, The Netherlands

^{2} Dutch National Intensive Care Evaluation (NICE) Foundation

Abstract

This paper explores the use of instance-based reasoning (IBR) to estimate the probability of hospital death in patients admitted to the Intensive Care Unit (ICU). The predictions are based on severity-of-illness scores that indicate the state of the patient. We have implemented an instance-based reasoning algorithm as an alternative to logistic regression (LR) models to predict hospital mortality. The performance was measured and prospectively validated. Results show that instance-based reasoning is competitive to logistic regression.

1 Introduction

Clinical scoring systems are tools for assessing the states of patients and quantifying the severity of their condition [Wyatt, 1990]. They are used in many medical disciplines, including cardiology, oncology, and critical care, and can be used for a variety of clinical and management tasks such as comparative audit among practitioners, measuring the effects of treatment, and risk assessment and prognosis. In this paper, we focus on the application of scoring systems in prognosis with binary outcome variables.

In most scoring systems, patient-specific data is used to arrive at an integer value that represents the severity of a patient's illness. Because points are assigned to deviations from normal values, low values (close to zero) generally represent mild conditions, whereas higher values are associated with more serious conditions. When clinical scores are used in prognosis, a model has to be developed that converts these scores into patient-specific predictions. With a binary outcome variable, the model needs to convert scores into either predicted outcome classes or into probabilities. The predominant methodology for doing this is logistic regression (LR) analysis [Hosmer and Lemeshow, 2000], where the score is used as a linear covariate.

Although LR analysis has proven to be a powerful modeling methodology in the biomedical field, it is based on assumptions that are questionable for most clinical scoring systems. In particular, logistic regression assumes that there exists a fixed (usually linear) relationship between score and log odds of the outcome probability over the entire score range. In practice, however, most scoring systems were not designed to have this property, and the relationship between score and (log odds of the) outcome may vary over the score range, and may be highly nonlinear.

In this paper, we study the use of instance-based reasoning (IBR) as an alternative for LR analysis in scoring-based prognosis. IBR is a nonparametric prediction method that is based on the assumption that the prognosis of a new patient resembles those of past patients with similar characteristics. The IBR method employed is the weighted k-NN regression algorithm with an adaptive neighborhood size. The main advantage of instance-based reasoning is that it makes few assumptions regarding the relationship between predictors and outcome. Furthermore, being a 'lazy' learning method, it is less sensitive to population drift than eager (model-based) learning methods such as LR. The main disadvantage is that it requires relatively large datasets (compared to parametric methods), and does not work well in high-dimensional domains. Finally, when it is used to estimate probabilities, as in our application, these may be biased (structurally too high or too low), a phenomenon that does not occur in model-based methods.

The method was applied to data from two popular scoring systems for intensive care patients, the APACHE II [Knaus *et al.*, 1985] and SAPS II [Le Gall *et al.*, 1993] scores. The resulting mortality estimators were validated and compared with LR models internally (with cross-validation on the training dataset) and externally (on a prospectively collected dataset).

The paper is organized as follows. Section 2 reviews the two scoring systems that were employed; Section 3 provides details on the datasets, IBR prediction method, and validation procedure. Section 4 describes the results from our study and Section 5 finishes the paper with a discussion and conclusions.

2 APACHE II and SAPS II scoring systems

Various scoring systems have been developed for the field of intensive care medicine [Gunning and Rowan, 1999]. In this study, we have used the Acute Physiological And Chronic Health Evaluation (APACHE) II [Knaus *et al.*, 1985] and the Simplified Acute Physiology Score (SAPS) II [Le Gall *et al.*, 1993] scores. Both scores are assessed during the first 24h of a patient's ICU stay, and can be converted into an estimated probability of death by means of

[*] Corresponding author. E-mail: l.m.peelen@amc.uva.nl

an associated LR model. The APACHE II score has a minimum of 0 and a maximum of 71 points; it summarizes mainly physiological information, and the associated LR model employs information on the patient's diagnosis at admission (54 categories) and type of admission (6 categories) besides the score. The SAPS II score ranges from 0 to 163 points; it summarizes physiological, diagnostic, and admission-type information; the associated LR model only employs the score itself.

An important difference between the APACHE II and SAPS II scoring systems is that the former is based on knowledge from practitioners, whereas the latter is based on data analysis. The APACHE II scoring system was designed during a consensus meeting with experienced intensive care clinicians; the associated prognostic model is based on LR analysis of a multicenter dataset of ICU admissions. The SAPS II scoring system, in contrast, was obtained by scaling the coefficients that were derived by multiple LR analysis on a large multicenter dataset.

Both scoring systems consider patients who have undergone cardiac surgery as special cases. These patients usually stay for observation at the ICU and leave for further recovery at the nursing ward once their condition is stable. We can compute scores for these patients, but the associated probability estimates from the LR models are believed to be unreliable [Knaus *et al.*, 1985; Le Gall *et al.*, 1993].

3 Data and methods

3.1 Data

The Dutch National Intensive Care Evaluation (NICE) register [NICE, 2005] provided two datasets containing information on ICU admissions. The first dataset describes 1559 ICU admissions from 7 Dutch hospitals between January 2003 and August 2003 and was used as a training set. In this dataset the hospital mortality is 14.8%.

During our study a second dataset was provided consisting of 1868 ICU admissions from August 2003 to June 2004. It was used to validate the IBR estimators that were developed on the first set. The hospital mortality in this dataset is 16.3%. The difference in mortality between the two datasets is not significant ($\chi^2 = 1.38$; $p = 0.24$). Both sets contain all variables required to compute the APACHE II and SAPS II scores, the scores themselves, and the associated probabilities of death estimated by the APACHE II and SAPS II LR models.

Using these data in total eight IBR estimators were developed using different (combinations of) predictive features. Two univariate IBR estimators were developed, one for the APACHE II score, and one for SAPS II score. Because the APACHE II score does not include information on the patient's diagnosis and type of ICU admission, also three multivariate estimators were developed for the APACHE II score in combination with diagnosis category, admission type, and both. Finally, a multivariate IBR estimator was developed on the basis of the two scores together.

As discussed in Section 2, predictions from the APACHE II and SAPS II LR models are believed to be unreliable for cardiac surgery patients and therefore should not be used. In the IBR estimators described above we have neglected this exclusion criterion and make predictions for all ICU patients in the same manner. Therefore we refer to these IBR estimators as *single method estimators*.

To take the exclusion criterion for cardiac surgery ICU admissions into account, we developed two more estimators, called the *dual method estimators*. Here we use the clinical scores (APACHE II and SAPS II respectively) to arrive at predictions for the patients who did not undergo cardiac surgery, and four alternative features for patients who arrive at the ICU after cardiac surgery. The four alternative features are minimum temperature, minimum systolic blood pressure, minimum bicarbonate, and maximum creatinine (all during the first 24h of ICU stay); they have been shown to be important predictors of mortality in cardiac surgery patients [Verduijn, 2002].

All IBR estimators were constructed with an extension of the *weighted k-NN regression algorithm*.

3.2 Prediction method

In weighted k-NN regression, predictions are obtained by computing a weighted average of the outcomes of the k training instances that are most similar to query instance $\mathbf{x}_q$. In the case of a binary outcome Y, we have

$$\hat{p}(Y = 1|\mathbf{x}_q) \quad = \quad \frac{\sum_{i=1}^{k} K_\lambda(\mathbf{x}_q, \mathbf{x}_{[i]}) \cdot y_{[i]}}{\sum_{i=1}^{k} K_\lambda(\mathbf{x}_q, \mathbf{x}_{[i]})}, \qquad (1)$$

where $\mathbf{x}_{[1]}, \ldots, \mathbf{x}_{[k]}$ are the k training instances most similar to $\mathbf{x}_q$, and $K_\lambda(\mathbf{x}_q, \mathbf{x}_{[j]})$ is the weight assigned to training instance $\mathbf{x}_{[j]}$. This is called the Nadaraya-Watson kernel-weighted average [Hastie *et al.*, 2001, Ch. 6]. In our application, $\hat{p}(Y = 1|\mathbf{x}_q)$ is the patient's estimated probability of hospital death, given the feature-value vector $\mathbf{x}_q$ (e.g. APACHE II score and diagnosis category).

Three important choices have to be made when weighted k-NN regression is applied: 1. How do we find similar training instances (*choice of distance metric*)?, 2. How are distances transformed into weights (*kernel function*)?, and 3. How many neighbors are used to make predictions (*neighborhood size*)? Each of these questions is addressed below.

Distance metric In the univariate IBR estimators we have used the *score difference* to quantify the distance between instances. In the multivariate estimators, local distance metrics were constructed for each of the predictive features. For non-numeric features (diagnosis category and ICU admission type), these local metrics were defined by distance matrices based on the hierarchical relations between feature values; we refer to [Tan, 2005] for details. In the prediction phase, local distances were normalized and then combined using the *Manhattan metric* (i.e. taking the unweighted sum of all normalized local distances).

Kernel function The kernel is a function that assigns a nonzero weight to all instances within the neighborhood of k nearest training instances, and zero weight to all other instances. We have used two kernel functions in our experiments, the *uniform kernel* and the *Epanechnikov kernel*

[Silverman, 1986]. The uniform kernel assigns unit weight to all k nearest neighbors, thus treating them as equally important. The Epanechnikov kernel, in contrast, is a nonlinear function that approaches 1 at small distances to the query instance, and 0 at the boundaries of the neighborhood:

$$K_\lambda(\mathbf{x}_q, \mathbf{x}_{[i]}) = \begin{cases} \frac{3}{4}(1-t^2) & \text{if } |t| \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $t = d(\mathbf{x}_q, \mathbf{x}_{[i]})/d(\mathbf{x}_q, \mathbf{x}_{[k]})$ is the normalized distance between neighbor $\mathbf{x}_{[i]}$ and the query instance $\mathbf{x}_q$.

Neighborhood size Most algorithms for k-NN classification and regression (e.g. those implemented in WEKA [Witten and Frank, 2001]) choose a fixed number of neighbors to make all predictions. However, usually the values of predictive features are not uniformly distributed over their theoretical range. As a result the width of the neighborhood that is necessary to obtain the k nearest neighbors varies with the sparsity of the data in the neighborhood of the query instance. However, when the neighbors are weighed according to their distance to the query instance, a single close neighbor yields the same amount of weight as multiple distant neighbors together. A better option is therefore to let the neighborhood width depend on the total weight of the neighbors rather than the number of neighbors [Hastie *et al.*, 2001]. This implies that the neighborhood width varies with the position of the query instance in the instance space and is locally adapted to the sparsity of the data.

In our application, a *target total weight* (`ttw`) of the instances in the neighborhood was established during the learning phase. The value of `ttw` is constant over the feature space, but needs to be optimized for the predictive feature(s) and the type of kernel function that are used to predict mortality. To find the optimal value for `ttw`, the following method was employed. For each IBR estimator, both kernel types and `ttw` values of 5, 10, 20, 50, 100, 200 and 500 were employed in a jackknife cross-validation procedure. In each run of the procedure, the estimator's accuracy was determined. Based on the results, the kernel type and `ttw` value were chosen.

Within this procedure, predictive accuracy was measured by the R^2 statistic [Ash and Shwartz, 1999]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{p}(Y=1 \mid \mathbf{x}_i) - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}, \quad (3)$$

where N is the size of the training dataset and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the mean outcome value. The R^2 statistic is inversely proportional to the mean squared error and the Brier score.

Figure 1 shows an example for the APACHE II score. The best performance is obtained with the Epanechnikov kernel and `ttw` values of 50 and 100. Because larger `ttw` values correspond to simpler models, we choose the Epanechnikov kernel with a `ttw` value of 100.

This procedure of selecting the optimal settings has been applied for all IBR estimators.

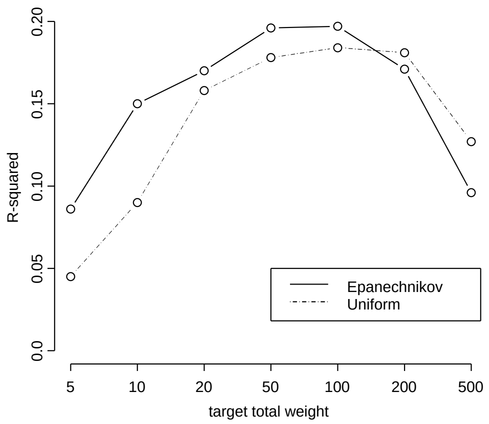


Figure 1: R^2 performance statistic for the APACHE II IBR estimator, plotted against the `ttw`, for both kernel types.

3.3 Validation

The IBR estimators were internally and prospectively validated. In the internal validation the performance of the estimator was measured by jackknife cross-validation on the training data. The estimator was also validated on prospectively collected data, using the second data set provided by the NICE register. Three different procedures were used in this prospective validation.

The first prospective validation procedure, the *settings validation*, aims to check whether the settings for kernel type and target total weight that were optimized on the training data, yield comparable performance on the prospective test dataset. To this end, we only use these settings, but not the training data for prediction; instead jackknife cross-validation is applied on the test set. Because the test set is larger than the training set, we expect the measured performance to be equally good or better if the chosen settings are valid.

The second prospective validation procedure is called the *plain prospective validation*. This procedure aims to investigate how well the algorithm generalizes to prospective data. To this end, predictions are made for all instances in the test set, while the training set serves as the instance base. We use the settings for kernel type and `ttw` value that were found on the training set.

One interesting property of IBR is the fact that it is a *lazy learning* method: generalization over examples in the instance base takes place no sooner than at the time of making predictions. The third prospective validation procedure, called *incremental prospective validation*, takes advantage of this property by incrementally adding instances from the test set and using them for future predictions. To this end, records in the test set were ordered by ICU admission date, and evaluated in that order. When evaluating a given record with admission date d, the instance base consists of all records from both the training set and test set with discharge date prior to d. For the first record from the test set this procedure yields the same prediction as in the second validation procedure. But for later records, the number of possibly similar instances is much larger, and there-

Estimator method	Predictive feature(s)	Kernel type	ttw	Relative bias	AUC ± S.D.
Single	APACHE II	Epan	100	-3.85	0.792 ± 0.033
Single	SAPS II	Unif	50	-1.65	0.860 ± 0.030
Single	APACHE II, SAPS II	Epan	20	4.50	0.854 ± 0.031
Single	APACHE II, admission type	Epan	20	0.04	0.828 ± 0.029
Single	APACHE II, diagnosis category	Unif	20	-4.22	0.831 ± 0.029
Single	APACHE II, adm. type, diag. category	Unif	20	-6.70	0.818 ± 0.029
Dual	APACHE II or alternative features	Epan	100	-11.90	0.818 ± 0.033
Dual	SAPS II or alternative features	Epan	50	-1.07	0.854 ± 0.030
LR model	APACHE II	-	-	-	0.796 ± 0.033
LR model	SAPS II	-	-	-	0.867 ± 0.027

Table 1: Results from the internal validation (jackknife cross-validation on the training set, 1559 ICU admissions). The predictive bias, averaged over all cases, is expressed as a percentage of the hospital mortality (14.8%). The alternative features for the dual method estimator are minimum temperature, minimum systolic blood pressure, minimum bicarbonate, and maximum creatinine values during the first 24h of ICU stay.

fore the predictions may be more accurate. Furthermore, in this way the prediction method accommodates to changes in the population characteristics (*drift*), a phenomenon that frequently occurs in medical applications.

In each validation procedure we computed the area under the ROC curve (AUC) for all IBR estimators. The AUC quantifies a estimator's ability to discriminate between patients who survive and those who die. An AUC value of 0.5 indicates that the estimator does not predict better than chance, while an AUC value of 1 indicates perfect discrimination. For the APACHE II and SAPS II scoring systems an AUC of > 0.80 is considered to be good.

4 Results

4.1 Internal validation

Table 1 shows the results from the internal validation. When regarding the AUCs, we see that the SAPS II IBR estimator is superior to the APACHE II IBR estimator (0.860 vs. 0.792). The LR model of SAPS II is better than that of APACHE II (0.867 vs. 0.796), and the SAPS II IBR estimator. The multivariate IBR estimator that uses both scores yields a slightly worse performance than SAPS II alone (0.860 vs 0.854) but these differences have not been tested for significance.

The APACHE II LR model employs information on the patient's diagnosis and type of admission besides the score, so employing this information with the IBR estimator should lead to better results as well. This is done by combining the APACHE II score and the admission type and/or diagnosis category in the IBR estimator. We see in Table 1 that adding either APACHE II admission type or diagnosis category leads to a increase in performance compared to that of the APACHE II alone in the IBR estimator. The performance is slightly worse when both the admission type and diagnosis category are used.

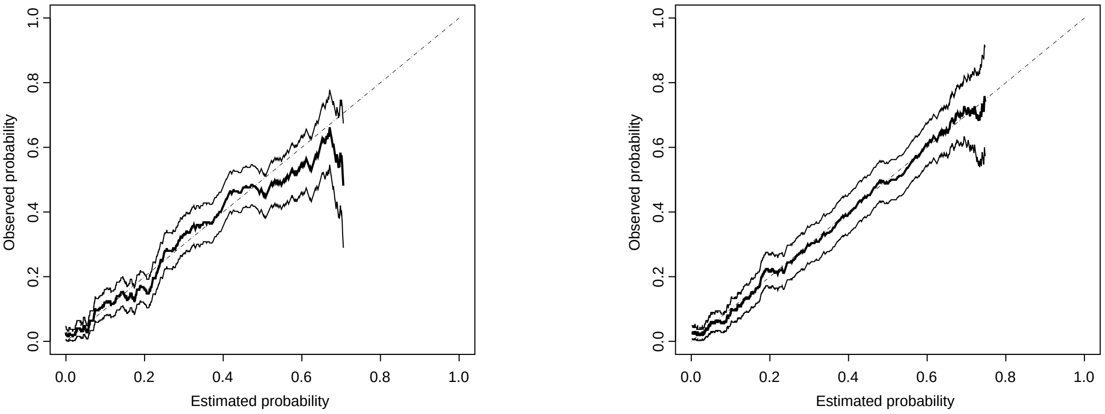
Since predictions for cardiac surgery patients by the APACHE II and SAPS II LR models are believed to be unreliable, the predictions by the single method IBR estimator may be unreliable as well. The dual method estimator attempts to improve performance by using alternative features for these patients. The desired effect is however only obtained for the APACHE II score and not for SAPS.

Table 1 also shows that the uniform kernel and Epanechnikov kernel were almost equally often selected by the optimization algorithm. So, the uniform kernel (i.e., equal weight for all instances in the neighborhood) may perform equally well or better than the Epanechnikov kernel in practical circumstances, even though the Epanechnikov kernel appears to be superior from a theoretical point of view.

Interestingly, the optimization algorithm has chosen `ttw` values (i.e., effective neighborhood sizes) that are relatively large compared to the values that are usually reported in the literature (less than 20 neighbors is common). Presumably, the explanation is that in our application, the neighboring outcomes are used to estimate the probability of death instead of the dominant class, and therefore a larger neighborhood size is required. Note that lower `ttw` values are selected for the multivariate estimators, due to sparsity in the multidimensional feature space of these estimators.

Because k-NN regression does not optimize a global likelihood formula, its predictions may show structural deviations from the observed outcome; we refer to this phenomenon as *predictive bias*. In Table 1 we have listed the predictive bias of each of the estimators, expressed as a percentage of the observed outcome. The APACHE II IBR estimator (first row), for instance, predicts a total of 221.1 deaths, whereas 230 out of 1559 patients actually died; the estimator thus underestimates mortality with 3.85%. The dual method estimator for APACHE II (seventh row) has a serious negative bias of -11.9% (202.6 deaths predicted).

Figure 2 shows smoothing plots of observed versus predicted probabilities for the APACHE II and SAPS II IBR estimators. The plots illustrate well the superior fit of the SAPS II estimator to the data: its plot is far more smooth and extends further into the upper region of the probability interval. The APACHE II plot, in contrast, is rather bumpy and the estimator appears to perform very poorly for patients with a high score. So, the APACHE II score appears to contain 'errors' that are difficult to repair, even for a highly adaptive method such as k-NN regression.



(a)

(b)

Figure 2: Observed vs. predicted probabilities in the APACHE II (a) and SAPS II (b) IBR estimators. The observed probabilities (on the y-axis) are obtained by loess smoothing on the observed outcome values (0 and 1), and are surrounded by 95% confidence intervals.

4.2 Prospective validation

Table 2 shows the results of all prospective validations. For each prospective validation, the mean predictive bias and AUC are displayed.

In the settings validation, the IBR estimators are applied to new data with the kernel type and target total weight settings that were optimized on the training set. For all estimators, the performance is equally good or better on the test set (explained by the fact that this set is somewhat larger than the training set). We conclude that the settings that optimized on the training set generalize well to new data.

Also in the plain prospective validation, where instances from the training set are used to make predictions on the test set, the performance is similar to the internal performance on the training set. So, we can use the IBR estimators to make predictions for future, unseen cases. The predictive bias, however, increases.

In the incremental prospective validation, the performance of estimators based on the APACHE II score further increases. Apparently, these estimators take advantage of the increasing size of the instance base. This does not hold for the estimators based on SAPS II. Furthermore, the predictive bias now reduces. An explanation for the latter fact is that the feature space becomes more densely populated since instances are added. A denser population means that the neighborhood does not have to expand as much as with a sparse population. This is especially advantageous near the boundaries, where the predictive bias is usually larger.

5 Discussion and conclusion

We have used IBR to predict hospital mortality for patients admitted to the ICU. Comparing our study to other applications of IBR in medicine, we note that in most studies IBR is used for classification (e.g. [Schmidt and Gierl, 2005; Lopez and Plaza, 1997]) and only sparsely for prediction. Anand et al. [Anand *et al.*, 2001] use k-NN in a hybrid system to predict time to survival in cancer patients, Gottrup et al. [Gottrup *et al.*, 2005] predict infarcted regions of the brain after cerebral stroke, based on MRI scans. Often IBR is used as part of a larger system, e.g. as in [Montani *et al.*, 2000].

From our experiments we conclude that IBR can be used to make reliable prognoses from clinical scores, and is competitive to LR in this task. For the APACHE II score, IBR prediction even outperforms the LR model. The applied method has been shown to generalize well to future patients, especially when new patients are added to the instance base to compensate for drift in the population characteristics.

When comparing the performance of the APACHE II and SAPS II scores in the IBR algorithm, we see that the SAPS II score performs better than the APACHE II score. The SAPS II scoring system was developed by scaling the coefficients that were obtained with multiple LR analysis. In contrast, the APACHE II scoring system is based on expert knowledge and the associated prognostic model was obtained from a LR analysis. This may be the reason that the IBR estimator does not perform better than the SAPS II LR model. These different approaches (expert knowledge vs. multiple LR analysis) to the development of a scoring system appears to be an important factor in the performance of IBR compared to a LR model. We think that this difference may also be apparent in other medical domains.

In the multivariate IBR estimators, we have used the Manhattan distance metric. Euclidean distance or other more sophisticated metrics may lead to better results. Similarly, it may be beneficial to weigh the predictive features, instead of treating them as equally important. However, Kohavi et al. [Kohavi *et al.*, 1997] found that weighing features rapidly leads to overfitting. Furthermore, we note that these adjustments only affect the multivariate estimators, whereas very good results were obtained already with our univariate estimators.

In the multivariate experiments, the combination of APACHE II and SAPS II scores performed worse than the

Feature(s)	Settings validation		Plain prospective validation		Incr. prospective validation	
	Bias	AUC ± S.D.	Bias	AUC ± S.D.	Bias	AUC ± S.D.
APACHE II	-3.83	0.821 ± 0.026	-11.22	0.784 ± 0.029	-4.37	0.809 ± 0.027
SAPS II	-0.73	0.865 ± 0.024	-2.23	0.867 ± 0.024	0.26	0.867 ± 0.024
APACHE II, SAPS II	-3.30	0.869 ± 0.022	-3.02	0.861 ± 0.025	0.83	0.863 ± 0.024
APACHE II, admission type	-2.81	0.843 ± 0.024	-8.57	0.832 ± 0.025	-4.74	0.839 ± 0.024
APACHE II, diagnosis category	-3.01	0.840 ± 0.023	-7.60	0.829 ± 0.025	-4.88	0.835 ± 0.024
APACHE II, adm. type, diag. category	-13.78	0.826 ± 0.024	-7.88	0.828 ± 0.024	-5.64	0.831 ± 0.024
Dual method APACHE II	-12.12	0.818 ± 0.024	-16.95	0.812 ± 0.027	-11.82	0.834 ± 0.025
Dual method SAPS II	-1.52	0.863 ± 0.024	-1.15	0.870 ± 0.024	1.68	0.872 ± 0.023
APACHE II *LR model*	-	-	-	0.804 ± 0.027	-	0.804 ± 0.027
SAPS II *LR model*	-	-	-	0.877 ± 0.022	-	0.877 ± 0.022

Table 2: Results from the prospective validations (1868 ICU admissions). The hospital mortality in this dataset is 16.3% .

SAPS II score alone. A possible explanation is found in the fact that the distance metric regards these two scores on two independent axes, perpendicular to each other. This is not correct, because both scores indicate the severity of illness; they are collinear. In future experiments, we have planned to use local regression models [Cleveland, 1979], which is expected to adjust for this phenomenon.

Acknowledgement The authors wish to thank Ameen Abu-Hanna and Rob Bosman for valuable discussions on the work described. Niels Peek receives a grant from the Netherlands Organisation of Scientific Research (NWO) under project number 634.000.020.

References

[Anand *et al.*, 2001] S.S. Anand, P.W. Hamilton, and J.G. Hughes et al. On prognostic models, artificial intelligence and censored observations. *Methods Inf Med*, 40:18–24, 2001.

[Ash and Shwartz, 1999] A. Ash and M. Shwartz. R^2: a useful measure of model performance when predicting a dichotomous outcome. *Stat Med*, 18:375–84, 1999.

[Cleveland, 1979] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*, 74:829–836, 1979.

[Gottrup *et al.*, 2005] C. Gottrup, K. Thompson, and P. Locht et al. Applying instance-based techniques to prediction of final outcome in acute stroke. *Artif Intell Med*, 33:223–236, 2005.

[Gunning and Rowan, 1999] K. Gunning and K. Rowan. ABC of intensive care: Outcome data and scoring systems. *BMJ*, 319:241–4, 1999.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, New York, 2001.

[Hosmer and Lemeshow, 2000] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 2nd edition, 2000.

[Knaus *et al.*, 1985] W.A. Knaus, E.A. Draper, and D.P. Wagner et al. APACHE II: a severity of disease classification system. *Crit Care Med*, 13:818–29, 1985.

[Kohavi *et al.*, 1997] R. Kohavi, P. Langley, and Y. Yun. The utility of feature weighting in nearest-neighbor algorithms. In M. van Someren and G. Widmer, editors, *Proc. 9th Europ. Conf. on Machine Learning (ECML-97)*. Springer, Berlin, 1997.

[Le Gall *et al.*, 1993] J. Le Gall, S. Lemeshow, and F. Saulnier. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA*, 270:2957–63, 1993.

[Lopez and Plaza, 1997] B. Lopez and E. Plaza. Case-based learning of plans and goal states in medical domains. *Artif Intell Med*, 9:29–60, 1997.

[Montani *et al.*, 2000] S. Montani, R. Bellazzi, and L. Portiginale et al. Diabetic patients management exploiting case-based reasoning techniques. *Comput Methods Programs Biomed*, 62:205–218, 2000.

[NICE, 2005] NICE, 2005. National Intensive Care Evaluation (NICE) register. http://www.stichting-nice.nl, also in English, accessed June 20th, 2005.

[Schmidt and Gierl, 2005] R. Schmidt and L. Gierl. A prognostic model for temporal courses that combines temporal abstraction and case-based reasoning. *Int J Med Inform*, 74:307–315, 2005.

[Silverman, 1986] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[Tan, 2005] C.H.K. Tan. Instance-based prognosis in intensive care using severity-of-illness scores. Master's thesis, Faculty of Medicine, University of Amsterdam, 2005. Accesible through: http://dare.uva.nl/scriptie/159502.

[Verduijn, 2002] M. Verduijn. Prognostic tree models in cardiac surgery. Identifying interactions between risk factors in a process-oriented approach. Master's thesis, Faculty of Medicine, University of Amsterdam, 2002.

[Witten and Frank, 2001] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques with JAVA implementation*. Morgan Kauffmann, San Francisco, 5th edition, 2001.

[Wyatt, 1990] J. Wyatt. Construction of clinical scoring systems. *BMJ*, 300:538–9, 1990.