A Nearest Neighbor Approach to Predicting Survival Time with an Application in Chronic Respiratory Disease

Maurice Prijs[1], Linda Peelen[1], Paul Bresser[2], and Niels Peek[1]

[1] Dept. of Medical Informatics
[2] Dept. of Pulmonology
Academic Medical Center – Universiteit van Amsterdam
m.c.prijs@amc.uva.nl

Abstract. The care for patients with chronic and progressive diseases often requires that reliable estimates of their remaining lifetime are made. The predominant method for obtaining such individual prognoses is to analyze historical data using Cox regression, and apply the resulting model to data from new patients. However, the black-box nature of the Cox regression model makes it unattractive for clinical practice. Instead most physicians prefer to relate a new patient to the histories of similar, individual patients that were treated before. This paper presents a prognostic inference method that combines the k-nearest neighbor paradigm with Cox regression. It yields survival predictions for individual patients, based on small sets of similar patients from the past, and can be used to implement a prognostic case-retrieval system. To evaluate the method, it was applied to data from patients with idiopathic interstitial pneumonia, a progressive and lethal lung disease. Experiments pointed out that the method competes well with Cox regression. The best predictive performance was obtained with a neighborhood size of 20.

1 Introduction

In patients with a chronic progressive disease, medical decisions are often influenced by the patient's prognosis. Therefore reliable estimates of remaining lifetime, based on the current condition of the patient, are required. In communication with the patient, the expected prognosis is often expressed in terms of survival probabilities for various time intervals (e.g., 12, 24, and 60 months).

The predominant method for obtaining such survival probabilities is to analyze historical data using Cox proportional hazards (PH) regression analysis and to use the resulting predictive model to construct a survival curve for the individual patient. Although the Cox PH analysis has proven to be a valuable tool in epidemiology and healthcare research, its usefulness in clinical practice is limited by its 'black-box' nature: for the clinical user of a Cox model, it is unclear how the patient's characteristics are used in constructing the survival curve. Instead most physicians prefer to relate a new patient to the histories of similar, individual patients that were treated before. These case histories also

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 77–86, 2007.

provide valuable other information, for instance about mobility status at various time points and subjective experience of the progressing disease.

This paper presents a prognostic inference method which combines Cox regression with the nearest neighbor paradigm, and which can be used to implement a prognostic case-retrieval system. In brief, the method selects nearest neighbors of the new patient from a database of historical observations, using a distance measure based on the variables and their weights resulting from the Cox regression analysis. The survival outcomes of the neighbors are used to construct a nonparametric Kaplan-Meier survival curve for the individual patient. As part of a case-retrieval system, it can be used to collect useful other types of information about similar patients from the past.

This paper is organized as follows. Section 2 provides a brief review of survival analysis, Cox regression models, and introduces our method. Section 3 presents a case study in the field of chronic lung diseases, including a statistical evaluation of predictive performance. The paper is completed with a discussion and conclusions in Sect. 4.

2 Methods

2.1 Survival Analysis

Let $t_1, \ldots, t_n$ be observed survival times for n individuals, and let $\delta_1, \ldots, \delta_n$ be associated censoring indicators, where $\delta_i = 0$ means that individual i was still alive at time t_i and $\delta_i = 1$ means that t_i was the individual's time of death. The statistical basis for both the Cox regression model and our prognostic method is the nonparametric estimation method from Kaplan and Meier [1]. It is a nonparametric estimate of the probability $S(t)$ that a random individual from the given population will have a lifetime exceeding t. When the survival times are ordered such that $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(j)}$, where $t_{(j)}$ is the jth largest unique survival time, the Kaplan-Meier estimate is defined as:

$$\hat{S}_0(t) = P(T \geq t) = \prod_{j \mid t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \quad , \tag{1}$$

where r_j is the number of individuals at risk (i.e., alive and not censored) just before $t_{(j)}$, and d_j is the number of individuals that died at $t_{(j)}$.

2.2 Cox Regression Models

The method of Kaplan and Meier estimates marginal (i.e., population-averaged) survival probabilities. To arrive at individual survival estimates, we need to use a method that takes information from these individuals into account. Let $\mathbf{X}$ be an $n \times p$ matrix of covariate patterns, where x_{ij} denotes the jth covariate and $\mathbf{x}_i$ the covariate pattern of individual i. Cox PH regression models consist of a nonparametric and a parametric component, that collaborate in the construction

of estimated survival curves for individuals, based on their covariate patterns [2]. The non-parametric component is the population-based Kaplan-Meier survival curve $\hat{S}_0$; the parametric part is a linear regression model that adjusts the marginal survival probabilities through an exponential link function. Formally, the survival probability $\hat{S}_i^{\text{Cox}}(t)$ for individual i at time point t is computed as

$$\hat{S}_i^{\text{Cox}}(t) = P(T > t \mid \mathbf{x}_i) = \hat{S}_0(t)^{\exp(\eta(\mathbf{x}_i))} \ , \tag{2}$$

where

$$\eta(\mathbf{x}_i) = \beta_1(x_{1i} - \bar{x}_1) + \cdots + \beta_p(x_{pi} - \bar{x}_p) \ . \tag{3}$$

Here, $\bar{x}_j$ is the average value of the jth covariate in the dataset. Thus, depending on the individual's deviance from average values, the baseline survival probabilities are increased, decreased, or remain unchanged. A fundamental assumption of the Cox PH regression model is that accurate survival functions for individuals (or subgroups of the population) are obtained through a proportional adjustment of the baseline survival function over time.[1]

The regression parameters $\beta_1, \ldots, \beta_p$ are estimated by optimizing a partial likelihood function that is based on Eq. 2. This type of estimation procedure is implemented in all major statistical software packages. Dedicated feature subset selection procedures exist (e.g., [3]) for eliminating irrelevant covariates and preventing the model from overfitting. For further details on the Cox regression model and its various extensions, we refer to [4].

2.3 *k*-Nearest Neighbor Survival Prediction

The *k-Nearest Neighbors* (k-NN) algorithm [5] solves classification and regression problems without explicitly building a model. Instead, when a prediction needs to be made for a particular individual, the algorithm selects a set of k similar instances from the data, and returns their average (regression) or dominant (classification) response. The k-NN algorithm resembles the retrieval step of case-based reasoning methods, a well-known decision support paradigm [6,7].

Classification and regression using the k-NN algorithm have several advantages and disadvantages. On the one hand, the algorithm makes few assumptions about the underlying regularities in the domain, and can therefore be used to approximate virtually every function. On the other hand, however, k-NN methods behave poorly in high-dimensional domains [8]. Furthermore, the algorithm is easily fooled by differences in scales on which the covariates are expressed [9].

The k-NN algorithm is usually applied for classification, smoothing, or binary regression; it has not been used in the context of survival analysis. In this paper we present a k-NN method for making individual survival predictions, which operates as follows. First, a Cox regression analysis with stepwise feature subset selection is conducted on the training dataset. Let η denote the linear predictor

[1] More precisely, the model assumes that such functions can be obtained through a proportional adjustment of the baseline *hazard* function – hence the name 'proportional hazards'.

of the resulting Cox model, as in (3). The value $\eta(\mathbf{x}_i)$ is called the *score* of individual i. We use the score difference

$$d(i,j) = |\eta(\mathbf{x}_i) - \eta(\mathbf{x}_j)| \tag{4}$$

to quantify the distance between individuals i and j.

Second, when making a prediction for an individual i, its k nearest neighbors are selected, and a survival curve is constructed from their survival times, using the method of Kaplan and Meier. Formally, let $\mathbf{x}_{[1]}, , \ldots, \mathbf{x}_{[k]}$, be the k individuals in $\mathbf{X}$ closest to individual i in the dataset $\mathbf{X}$, and let $t^*_{(1)} < \ldots < t^*_{(m)}$, $m \leq k$, be their unique and ordered survival times. Then, the k-NN estimated survival probability for individual i at time point t equals

$$\hat{S}_i^{k\mathrm{NN}}(t) = \prod_{j \mid t^*_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \ , \tag{5}$$

where r_j is the number of neighbors at risk (i.e., alive and not censored) just before $t^*_{(j)}$, and d_j is the number of neighbors that died at $t_{(j)}$.

3 Case Study in Idiopathic Interstitial Pneumonia

Idiopathic interstitial pneumonia (IIP) comprises a group of disorders of unknown etiology which are characterized by a variable pattern of inflammation and/or fibrosis of the pulmonary interstitium, causing shortness of breath on exertion and dry cough. It is a rare disease, with a median survival generally being reported as 2 to 4 years, but there is substantial heterogeneity in survival time among patients [10]. A reliable prediction of survival probabilities for each new patient is of great value for physicians caring for these patients. Decisions on referral for lung transplantation are based on this prognosis, which can have far-reaching consequences for the patient. For this reason IIP is a typical area of disease in which the method we propose in this paper could have added value as compared to the standard Cox model.

3.1 Data

For this evaluation study we used prospectively collected data of patients with IIP, collected between November 1993 and December 2005 at the Department of Pulmonology of the Academic Medical Center in Amsterdam, The Netherlands. Data collection was performed as part of a local protocol for examination and treatment for IIP patients eligible for lung transplantation. The variables used in this study are histopathologic pattern, age, sex, and eight pulmonary function testing (PFT) variables. All of these variables have been identified as predictors of IIP survival in several clinical studies (see e.g., [11]). The dataset we used in this study contained information on 103 patients. Median survival of the patients in the dataset was 47.9 months, and the 5-year survival rate was 35.7%.

3.2 Evaluation of Predictions

The survival probabilities $\hat{S}_i^{\text{Cox}}(t)$ and $\hat{S}_i^{k\text{NN}}(t)$, derived from both the Cox regression model and the k-NN algorithm, provide estimates for survival for each individual i at each observed time point t. The accuracy of these predictions is assessed by comparing the observed individual vital status at each time point $Y_i(t)$ with the predicted survival probabilities $\hat{S}_i(t)$. The means of the squared differences are a measure of prediction error, ranging from 0 to 1, also known as the quadratic or Brier score [12]. The lower the Brier score, the more accurate the prediction. In order to compensate for the loss of information due to censoring, we used an adjusted version of the Brier score $BS_i^c(t)$ as proposed by Graf et al [13], which uses a reweighing of the individual contributions based on their probability of being censored.

With $BS_i^c(t)$ the Brier score for each time point and for each individual is calculated. We used three types of cumulative scores for the comparison of the accuracy of the predictions of both methods. First, BS_i^c describes the cumulative prediction error per patient. Second, $BS^c(t)$ cumulates over individuals for each time point. Third, by cumulating over all individuals and all time points the total cumulative prediction error BS^c is calculated. Another measure of accuracy we used is the area under the receiver-operating characteristic curve (AUC) [14] for specific time points. It represents the probability that a patient who lived at the given time point is assigned a higher survival probability than a patient who then had died. An AUC of 0.5 indicates that the predictions are not better than chance, and is generally found for nondiscriminative models such as the Kaplan-Meier estimate.

3.3 Design

In this section the design of our experiment is outlined. The goal was to measure the performance of both methods in making individual survival predictions. Our strategy to achieve this is summarized in Fig.1. In order to reduce the risk of overfitting, performance was measured in a 10-fold cross-validation setting. Figure 1 represents the operation for a single individual in a single fold.

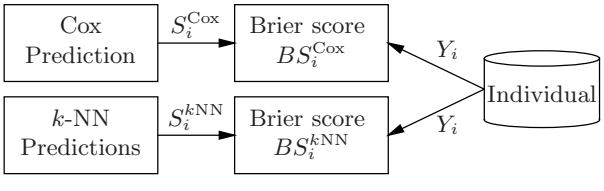

Fig. 1. Measuring the performance of the Cox regression model and the k-NN algorithm. The performance of both methods is measured by means of the Brier score, which compares the estimated probability of survival S_i to the observed vital status Y_i of the individual for each time point.

Calculating Cox- and *k*-NN Predictions. The Cox predictions and the k-NN predictions were obtained according to the procedures described in Sect. 2. For each of the folds in the cross-validation the following procedure was applied.

Based on the data in the training set a standard univariate Cox regression analysis was used to identify prognostic variables related to survival. Significant variables (p-value<0.05) were entered into a multivariate Cox model. Feature subset selection was performed using a backward stepwise selection method based on exact Akaike's Information Criterion (AIC) [3]. The model with the lowest AIC was considered the final model. From this model, the baseline hazard function and the linear predictor function were extracted.

Subsequently, predictions were made for each patient in the test set. The survival probability function based on Cox regression, $\hat{S}_i^{\text{Cox}}(t)$, was determined by calculating the score $\eta(\mathbf{x}_i)$ for each of the individuals in the test set based on their covariate pattern $\mathbf{x}_i$ and using this in (2). To estimate the survival probability function based on k-NN regression, the score $\eta(\mathbf{x}_i)$ was calculated for each of the individuals in the test set *and* in the training set. Subsequently, $\hat{S}_i^{k\text{NN}}(t)$ for each individual in the test set was calculated by determining the k nearest neighbors from the training set (using the distance function in (4)), and using their survival times to construct a Kaplan-Meier estimate. This procedure was repeated for different values of k, resulting in multiple predictions of $\hat{S}_i^{k\text{NN}}(t)$.

Calculating the Prediction Errors. The error in the predictions obtained by both methods is estimated for each individual by means of the Brier scores $BS_i^{c^{\text{Cox}}}(t)$ and $BS_i^{c^{k\text{NN}}}(t)$. We calculated the mean, standard deviation, and median of all BS_i^cs. To evaluate and display the predictive accuracy of the different methods, error curves were plotted using $BS^{c^{\text{Cox}}}(t)$ and $BS^{c^{k\text{NN}}}(t)$. In interpreting these error curves we focused on the first part of the curves, as the confidence intervals (not shown in the figure for clarity) rapidly increased for higher values of t. To compare the performance of the Cox model and the k-NN algorithm with that of a model that does not include information on the covariate pattern, we used the performance of the Kaplan-Meier estimate (1) as a benchmark value.

3.4 Results

All results are calculated over the 10 cross-validation folds. The mean, standard deviation and median of the cumulative Brier scores are shown in Table 1. The large standard deviations suggest an unbalanced distribution of BS_i^c. Therefore the median is used for comparison. We first compared the performance of the k-NN predictor for different values of k. The lowest median BS_i^c is found for k=20.

In addition, we plotted the prediction error curves over time for the survival predictions based on various neighborhood sizes, as shown in Fig.2. The prediction error for k=5 is markedly higher over the entire time period than for the other values of k. The differences between 20, 25 and 40 as values of k are less distinct. For approximately the first 50 months, the survival predictions based on k=25 seem to have the lowest prediction error. In the period between 50 and

Table 1. Cumulative Brier scores for the Kaplan-Meier (KM, $\hat{S}_0(t)$), Cox regression ($\hat{S}_i^{\text{Cox}}(t)$), and k-NN method ($\hat{S}_i^{k\text{NN}}(t)$) with different numbers of neighbors k

	KM	Cox	k-NN									
			k=5	k=10	k=15	k=20	k=25	k=30	k=35	k=40	k=50	k=70
Mean	19.9	22.4	30.1	21.3	21.0	20.4	20.4	20.0	19.8	19.1	19.2	19.6
s.d.	21.0	25.1	38.9	24.2	23.0	21.9	21.5	20.9	20.3	19.5	19.7	20.6
Median	18.7	14.8	16.2	13.7	12.6	12.2	13.8	13.6	14.3	15.4	16.4	18.4

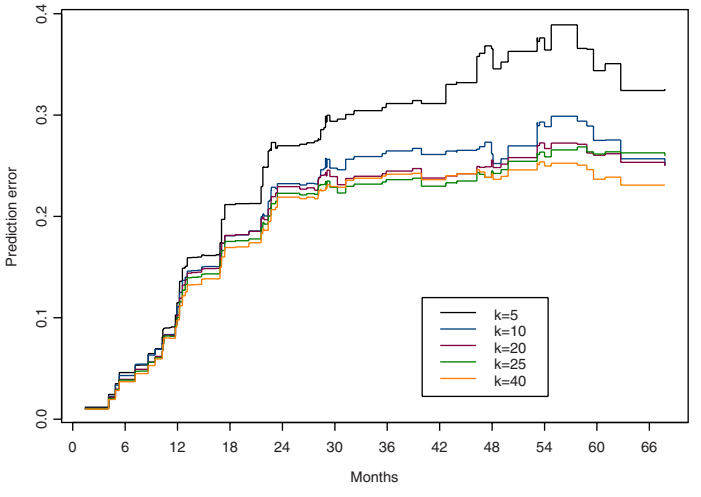

Fig. 2. Estimated prediction error curves for the k-NN algorithm with different numbers of neighbors k for the first 72 months

70 months, this is true for k=40, whereas the prediction error curves of the different values of k after 70 months intersect multiple times due to the increased confidence interval. Based on the median BS_i^c and the error curves, we choose to use 20 as the value of k in the comparison of the different prediction methods.

Figure 3 visualizes the differences between the estimated prediction error over time for the two methods. For purposes of comparison, we have also depicted the performance of the Kaplan-Meier estimate (population average survival). For predictions of survival up to approximately two years, the curves yield a similar pattern, with a slightly better performance of the k-NN algorithm. In the period between approximately 30 and 50 months the k-NN algorithm yields the lowest estimated prediction errors.

We calculated the median of the cumulative Brier scores up to t=12, 24 and 48 months and the AUC at the same time points. As shown in Table 2, the k-NN algorithm yields the lowest median prediction error for each period. The AUC

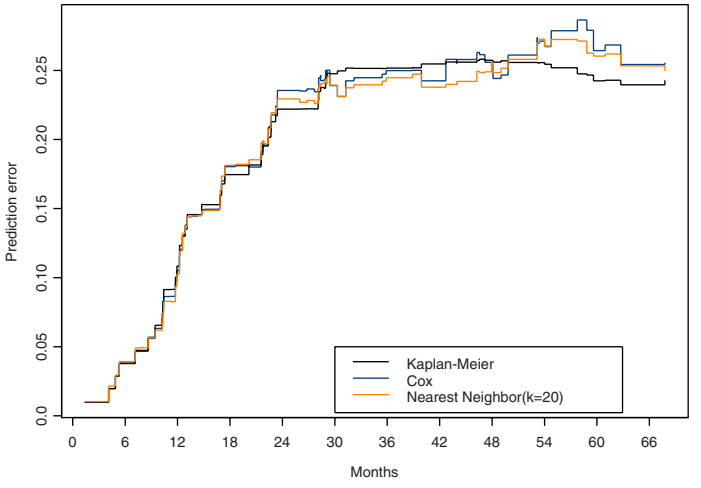

Fig. 3. Estimated prediction error curves for the Cox regression model and the k-NN algorithm with k=20 neighbors for the first 72 months. The error curve of the Kaplan-Meier estimate serves as a benchmark.

Table 2. Median cumulative Brier score per individual (BS^c) and area under the receiver-operating characteristic curve (AUC) for the Kaplan-Meier estimate, the Cox regression model, and the k-NN algorithm for t=12, 24 and 48 months

	12 mo.		24 mo.		60 mo.	
	BS^c	AUC	BS^c	AUC	BS^c	AUC
KM	0.04	0.49	0.77	0.53	11.47	0.46
Cox	0.03	0.61	0.73	0.61	11.73	0.62
20-NN	0.01	0.62	0.63	0.61	8.57	0.59

of the k-NN algorithm is almost equal to that of the Cox regression model at t=12 and 24 months, but lower at t=60 months.

4 Discussion and Conclusions

In this paper we proposed a prognostic inference method that combines the k-NN paradigm with Cox regression. It yields probabilistic survival predictions for individual patients that are based on small sets of similar patients that were seen before by the doctor. The intended application of the method is a prognostic case-retrieval system, but other types of application are conceivable.

From a theoretical point of view, our prediction method is more flexible than the Cox regression model because it does not need the proportional hazards assumption. In some situations covariate patterns are associated with survival curves that are fundamentally different in shape from the population-based

survival curve. In the IIP domain, for instance, it is known that differences in histopathologic pattern do not only influence the steepness of the survival curve, but also its shape [15]. However, as all k-NN algorithms, our method relies on statistical estimates from small samples (neighborhoods), and this may cause high variation (and therefore low reliability) of predictions.

In a case study on IIP we compared the predictive performance of this method with the commonly used Cox regression model. On the given dataset, our method has the best predictive performance with k=20. With this number of neighbors, k-NN survival curves were largely similar to those obtained with the Cox model, and the median cumulative Brier score was even slightly lower, indicating superior performance in most cases. However, both methods also performed only slightly better than the population-based Kaplan-Meier survival estimate (in terms of the cumulative Brier score), and had but moderate discriminative abilities (as measured by the AUC). These findings impel to further investigations on survival prediction in this domain.

In the literature, k-NN methods have been used for a variety of prediction tasks, but rarely in the context of survival analysis. Hamilton et al. [16] used a k-NN algorithm to predict survival time of patients with colorectal cancer. Their algorithm returns the median survival time of four closest neighbors, based on a distance measure for three predefined variables. Anand et al. [17] enhanced several basic distance metrics by statistical techniques and used a framework based on Dempster-Shafer's Theory of Evidence to make survival predictions. Although the idea of embedding case retrieval in a multimodal reasoning task in general is not new [18], the idea of combining Cox regression with k-NN regression to predict survival, as proposed in this study, is novel.

There are several limitations to our study. In particular, the design of our evaluation method in the case study is somewhat biased: ten values of k were mutually compared in a cross-validation design, and the best performer (k=20) was compared to the Cox regression model within the same cross-validation loop. This design puts the Cox regression model at a slight handicap. We nevertheless believe that the two methods are competitive on the given dataset. Furthermore, the predictions from both methods could probably be improved by including additional information, e.g., on trends in physiological-respiratory variables [15,19].

The number of k=20 neighbors is quite large for a prognostic case-retrieval system, the intended application of our method. Clinical visits are restricted in time, and discussing 20 historical cases is not feasible for doctor and patient, even if they fancied doing so. Unfortunately, smaller numbers of k (e.g., 5 and 10) produce markedly worse predictions, and should therefore be avoided. We suggest that further research investigates how the information from some 20 cases can be conveniently presented to users of a system. Additional possibilities are to weigh cases according to their distance, and restrict the maximum distance to the query instance, as many search engines on the internet do.

We conclude that our k-NN method can compete in performance with a standard Cox regression model in predicting individual survival for IIP patients, with moderately-sized neighborhoods. It provides the starting point for a prognostic case-retrieval system.

Acknowledgements. We would like to thank Judith Blumenthal for her work on the IIP database.

References

1. Kaplan, E., Meier, P.: Nonparametric estimation from incomplete observations. JASA 53, 457–481 (1958)
2. Cox, D.R.: Regression models and life tables. J Royal Stat Soc Series B 34, 187–220 (1972)
3. Akaike, H.: A new look at the statistical model identification. IEEE Trans Auto Control AC-19, 716–723 (1974)
4. Therneau, T., Grambsch, P.: Modeling Survival Data: Extending the Cox Model. Springer, New York (2000)
5. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans Inform Theory 13, 21–27 (1967)
6. Schmidt, R., Montani, S., Bellazzi, R., Portinale, L., Gierl, L.: Cased-Based Reasoning for medical knowledge-based systems. Int J Med Inform 64, 355–367 (2001)
7. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: Whats next? Artif Intell Med 36, 127–135 (2006)
8. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
9. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. In: Data Mining, Inference, and Prediction, Springer, New York (2001)
10. American Thoracic Society/European Respiratory Society: International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. Am J Respir Crit Care Med 165(2), 277–304 (2002)
11. Perez, A., Rogers, R.M., Dauber, J.H.: The prognosis of idiopathic pulmonary fibrosis. Am J Respir Cell Mol Biol 129(Suppl. 3), 19–26 (2003)
12. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Rev 78, 1–3 (1950)
13. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18, 2529–2545 (1999)
14. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1), 29–36 (1982)
15. Latsi, P.I., du Bois, R.M., Nicholson, A.G., et al.: Fibrotic idiopathic interstitial pneumonia: the prognostic value of longitudinal functional trends. Am J Respir Crit Care Med 168(5), 531–537 (2003)
16. Hamilton, P.W., Bartels, P.H., Anderson, N., Thompson, D., Montironi, R., Sloan, J.M.: Case-based prediction of survival in colorectal cancer patients. Anal Quant Cytol Histol 21(4), 283–291 (1999)
17. Anand, S.S., Hamilton, P.W., Hughes, J.G., Bell, D.A.: On prognostic models, artificial intelligence and censored observations. Meth Inf Med 40(1), 18–24 (2001)
18. Aha, D., Daniels, J.J. (eds.): Case Based Reasoning Integrations: Papers from the 1998 Workshop (Technical Report, WS-98-15). AAAI Press, Menlo Park, CA (1998)
19. Collard, H.R., King Jr, T.E., Bartelson, B.B., et al.: Changes in clinical and physiologic variables predict survival in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 168(5), 538–542 (2003)