

Predicting Prostate Cancer Recurrence via Maximizing the Concordance Index

Lian Yan
Aureon Biosciences, Inc.
28 Wells Ave., Building 3
Yonkers, NY 10701
lian.yan@aureon.com

David Verbel
Aureon Biosciences, Inc.
28 Wells Ave., Building 3
Yonkers, NY 10701
david.verbel@aureon.com

Olivier Saidi
Aureon Biosciences, Inc.
28 Wells Ave., Building 3
Yonkers, NY 10701
olivier.saidi@aureon.com

ABSTRACT

In order to effectively use machine learning algorithms, e.g., neural networks, for the analysis of survival data, the correct treatment of censored data is crucial. The concordance index (CI) is a typical metric for quantifying the predictive ability of a survival model. We propose a new algorithm that directly uses the CI as the objective function to train a model, which predicts whether an event will eventually occur or not. Directly optimizing the CI allows the model to make complete use of the information from both censored and non-censored observations. In particular, we approximate the CI via a differentiable function so that gradient-based methods can be used to train the model. We applied the new algorithm to predict the eventual recurrence of prostate cancer following radical prostatectomy. Compared with the traditional Cox proportional hazards model and several other algorithms based on neural networks and support vector machines, our algorithm achieves a significant improvement in being able to identify high-risk and low-risk groups of patients.

Categories and Subject Descriptors

H.4 [Database Management]: Database Applications - Data Mining; I.2.6 [Artificial Intelligence]: Learning; I.5.2 [Pattern Recognition]: Design Methodology - classifier design and evaluation

General Terms

Algorithms

Keywords

Survival analysis, concordance index, neural networks, prostate cancer recurrence, nomogram

1. INTRODUCTION

Most medical data which is used to train a prognostic predictive survival model consists of both censored and non-censored observations. Censorship indicates whether the outcome under observation, e.g., recurrence of prostate cancer, has occurred within a patient's follow-up visit time. If the recurrence of prostate cancer has not been observed at a patient's follow-up visit, this patient's data is censored. In our problem of predicting post-prostatectomy recurrence of prostate cancer, about 85% of the patients were censored. Censored observations provide incomplete information about the outcome, since the event may eventually occur after the follow-up visit. This must be taken into account by a prognostic model. Typically, traditional survival analysis, e.g., the Cox proportional hazards model, is used to deal with censored data. The survival model is used to determine the probability of the event occurring within a specific time. However, in general, the reliability of the Cox model deteriorates if the number of features is greater than the number of events divided by 10 or 20 [9]. In our application, it is expensive and difficult to collect the patient data, and the data set consists of only 130 patients, each of which is represented by a vector of 25 features. We have seen that a Cox model could not be successfully derived from this data set until we reduced the feature dimensionality.

Machine learning techniques, e.g., neural networks, have been applied to survival analysis. There are two types of approaches to using neural networks for survival data. The first type models the hazard or survival function as a neural network structure. In [2], the authors construct the survival curve by a hazard function modeled by a neural network, for which the i th output is the estimated hazard at the discretized time interval i . Biganzoli et al. [1] used the discretized time interval as an additional input to a neural network to model the survival probability. Other authors used several separately trained networks, each used to model the hazard function at a different time interval, e.g., [17]. Neural networks have been shown to be able to outperform traditional statistical models, probably due to neural networks' capacity to model nonlinearities [13]. However, these algorithms require a large number of samples in the training set in order to be successful [2].

Different from the traditional survival analysis, some applications' primary goal is to predict whether an event will *eventually* occur or not. Such applications are common in the medical diagnostic field, where the sample size is typically small, e.g., hundreds. In our application, we predict whether prostate cancer will eventually recur after a patient

undergoes prostatectomy. If recurrence could be predicted prior to actual recurrence, a follow-up therapy could be administered more effectively at an earlier stage. The model directly outputs a prognostic score for an individual patient. It seems now that the problem is a simpler, classical classification problem. Still, in order to effectively use machine learning algorithms in this scenario, the correct treatment of censored data is crucial. Simply omitting the censored observations [3] or treating them as non-recurring samples [18] both obviously bias the resulting model and should be avoided. Zupan et al. [20] used Kaplan-Meier estimates of event probability as target values during training for the patients who had short follow-up times and did not have an occurred event. This algorithm does take into account, to some extent, both follow-up time and censoring, but it still fails to make complete use of all the available information. For instance, it treats two recurred patients as the same regardless of their survival time. deSilva et al. [5] tried to train a neural network by incorporating both follow-up time and censorship into a modified objective function. The model regresses over the survival time, but the error term for censored samples in the objective function is an asymmetric squared error, which becomes zero when the model output is larger than the censoring time. We have also tried a similar algorithm based on support vector machines, where the loss function is asymmetric for censored samples and assigns a smaller penalty to outputs larger than the censoring time via different slack variable and margin. These algorithms do not output a probabilistic estimate. Instead, their outputs are proportional to the survival time, and a higher score is associated with a lower probability of occurring.

In this paper we propose a new algorithm, which utilizes both censored and non-censored observations in a more effective and elegant way. The algorithm directly uses the concordance index (CI) as the objective function to train a model, which predicts whether an event will eventually occur or not. The concordance index is a typical metric for quantifying the predictive ability of a survival model [10]. In our application, the CI estimates the probability that, of a pair of randomly chosen comparable patients, the patient with the higher prognostic score from the model will recur within a shorter time than the other patient. Here a pair of patients are comparable if one of the patients recurred and had a shorter follow-up time. Using the CI as the objective function during training allows the model to make complete use of the information from both censored and non-censored observations. We present the new algorithm in the next section, and the patient data in our application is described in Section 3. In Section 4, we compare the empirical results of the new algorithm, the traditional Cox model, and three other methods. Finally, we recalibrate the original outputs from the model to clinically meaningful scores, which are the probabilities of remaining free of recurrence in the next 7 years following the surgery.

2. THE NEW ALGORITHM

The typical objective functions used to train a neural network model, e.g., mean square error, cross entropy, and likelihood, cannot directly treat censored data. Here we propose an objective function that is an approximation to the concordance index and allows the model to make complete use of the information from both censored and non-censored observations. The proposed objective function is differentiable,

and thus can be applied to any parametric classifier. For any such classifier, one can optimize the objective function with respect to the parameters using gradient based methods.¹ In the results below, we apply the proposed objective function to a typical multilayer perceptron (MLP) with softmax outputs, with a single hidden layer and direct connection between the input and output layers.

2.1 The concordance index

The concordance index (CI) can be expressed in the form

$$CI = \frac{\sum_{(i,j) \in \Omega} I(\hat{t}_i, \hat{t}_j)}{|\Omega|}, \quad (1)$$

where

$$I(\hat{t}_i, \hat{t}_j) = \begin{cases} 1 & : \hat{t}_i > \hat{t}_j \\ 0 & : \text{otherwise} \end{cases}, \quad (2)$$

is based on pairwise comparisons between the prognostic scores \hat{t}_i and \hat{t}_j for patients i and j , respectively. Here Ω consists of all the pairs of patients $\{i, j\}$ who meet any of the following conditions:

1. Both patients i and j experienced recurrence and the recurrence time t_i of patient i is shorter than patient j 's recurrence time t_j .
2. Only patient i experienced recurrence and t_i is shorter than patient j 's follow-up visit time t_j .

The CI estimates the probability that a patient with the higher prognostic score from the model will recur within a shorter time than a patient with a lower score.

The CI is tightly associated with the area under the ROC curve (AUC), which is a common performance metric for classification [7]. The AUC is equivalent to the Wilcoxon-Mann-Whitney statistic in the form

$$U = \frac{\sum_{(i,j) \in \Theta} I(\hat{s}_i, \hat{s}_j)}{|\Theta|}, \quad (3)$$

where \hat{s}_i and \hat{s}_j are the classifier outputs for a positive sample i and a negative sample j , respectively. Note that Θ consists of only the pairs of positive sample i and negative sample j . Thus, the AUC considers only the pairwise comparisons between a pair of positive and negative samples, but ignores any comparison within the same class. This is correct for a typical classification problem, but, in survival analysis, the AUC cannot take into account the other critical information, which is the survival time. The CI encodes this important information by comparing the model outputs for the within-class pairs which meet the first condition above. Though the CI has long been used as a performance metric for survival analysis, it has never been used as an objective function to allow complete use of information from both censored and non-censored observations for training a prognostic model, e.g., neural network. The difficulty of using the CI as a training objective lies in that it is nondifferentiable and cannot be optimized by gradient-based methods.

2.2 Maximizing the concordance index

A differentiable approximation to the step function in Eq. 2 has been proposed by one of us [19] to directly optimize the AUC during training and thus to improve classifier performance. In [19], several alternative approximations

¹We use the limited memory BFGS method in [16].

to Eq. 2 are discussed. The intuitive choice is the sigmoid function

$$S(\hat{t}_i, \hat{t}_j) = \frac{1}{1 + e^{-\beta(\hat{t}_i - \hat{t}_j)}}, \quad (4)$$

where $\beta > 0$. However, this is found to be less effective than the following function

$$R(\hat{t}_i, \hat{t}_j) = \begin{cases} (-(\hat{t}_i - \hat{t}_j - \gamma))^n & : \hat{t}_i - \hat{t}_j < \gamma \\ 0 & : \text{otherwise} \end{cases}, \quad (5)$$

where $0 < \gamma \leq 1$ and $n > 1$. $R(\hat{t}_i, \hat{t}_j)$ can be regarded as an approximation to $I(-\hat{t}_i, -\hat{t}_j)$. An example of $R(\hat{t}_i, \hat{t}_j)$ with $I(\hat{t}_i, \hat{t}_j)$ is shown in Figure 1. Thus, in order to maximize the CI in Eq. 1, we train a prognostic model by *minimizing* the objective

$$C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|}. \quad (6)$$

Empirically, we have found that a weighted version of C in the following form generally achieves better results:

$$C_w = \frac{\sum_{(i,j) \in \Omega} -(t_i - t_j) \cdot R(\hat{t}_i, \hat{t}_j)}{D}, \quad (7)$$

where

$$D = \sum_{(i,j) \in \Omega} -(t_i - t_j) \quad (8)$$

is the normalization factor. Here, each $R(\hat{t}_i, \hat{t}_j)$ is weighted by the difference between t_i and t_j . The process of minimizing C_w (or C) tries to move each pair of samples in Ω to satisfy $\hat{t}_i - \hat{t}_j > \gamma$ and thus to make $I(\hat{t}_i, \hat{t}_j) = 1$ in Eq. 1. When the difference between the outputs of a pair in Ω is larger than the margin γ , this pair of samples will stop contributing to the objective function. This mechanism effectively overcomes overfitting during training [19], and makes the optimization focus on only moving *more* pairs of samples in Ω to satisfy $\hat{t}_i - \hat{t}_j > \gamma$. Essentially, the influence of the training samples is adaptively adjusted according to the pairwise comparisons during training. Note that the positive margin γ in R is needed for better generalization performance.

3. PATIENT DATA

The prostate is a muscular, walnut-sized gland, located directly beneath the bladder and in front of the rectum that surrounds part of the urethra, the tube that transports urine and sperm out of the body. The prostate is part of the male reproductive system, and its main function is to produce seminal fluid, the solution that carries sperm. Prostate cancer (PCa) is a malignant tumor that usually begins in the outer-most part of the prostate and is the most common form of cancer found in American men. More than 180,000 men in the U.S. will be diagnosed with prostate cancer this year, and more than 30,000 will die of the disease. While the number of men diagnosed with prostate cancer remains high, survival rates have been steadily improving primarily due to early detection. Eighty-nine percent of the men diagnosed with the disease will survive at least five years, while 63% will survive 10 years or longer. The American Urological Association and the American Cancer Society recommend annual screening for men ages 50 to 70. The most effective screening tests available include a blood test for an enzyme

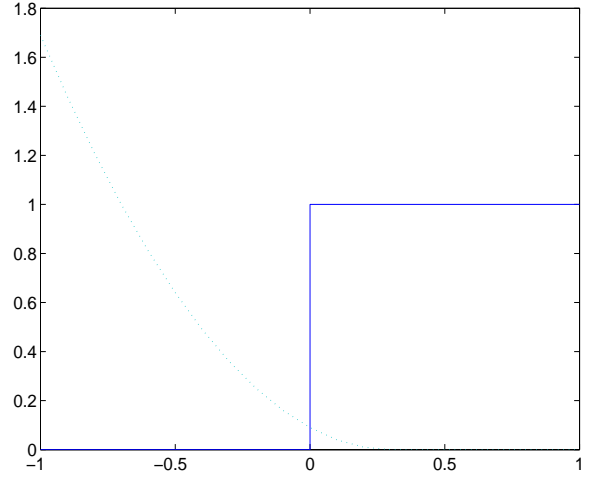


Figure 1: $R(\hat{t}_i, \hat{t}_j)$ compares with $I(\hat{t}_i, \hat{t}_j)$. The horizontal axis is $t_i - t_j$. $\gamma = 0.1$ and $n = 2$ in $R(\hat{t}_i, \hat{t}_j)$.

called prostate-specific antigen (PSA) which is produced by the prostate gland and the employment of a digital rectal exam (DRE). Elevated PSA levels ($> 4\text{ng/ml}$ or greater) may indicate prostate cancer. However, increases in PSA are also reported in benign conditions such as prostatitis and a pathologic enlargement of the prostate known as benign proliferative hyperplasia (BPH). The standard of care once PCa is suspected is to obtain a biopsy, typically a sextant (six-part) biopsy to assess presence or absence of disease.

The most common treatment for localized or confined PCa, in men under the age 70 who do not have other health complications is a radical prostatectomy, i.e., surgery to remove the prostate gland, seminal vesicles, vas deferens and some surrounding tissue. After surgery, the PSA levels in the blood should be reduced to 0.2ng/ml or less. If the PSA levels begin to rise at any time after treatment (also known as a biochemical recurrence BCR), a local or distant recurrence may be suspected, and will necessitate restaging the cancer, as well as a discussion of possible salvage therapies with the patient including radiation or hormonal therapy, experimental protocols or observation [8]. Thus, the ability to predict which patients will have a BCR would be very important to urologists and oncologists in managing the course of future treatment. A number of prostate cancer nomograms which combine clinical and/or pathologic factors to predict an individual patients probability of disease recurrence or survival have been published, e.g., [14] [6] [11]. The postoperative nomogram developed by Kattan et al. [14] is widely used by clinicians and allows a prediction of the probability of disease recurrence for patients who have received radical prostatectomy as treatment for prostate cancer. The postoperative nomogram used Cox proportional hazards regression analysis to model the clinical and pathologic data and disease follow-up for men treated with radical prostatectomy by a single surgeon. Prognostic variables included pretreatment serum prostate-specific antigen level, radical prostatectomy Gleason sum, prostatic capsular invasion, surgical margin status, seminal vesicle invasion, and lymph node status. Treatment failure was recorded when there was clinical evidence of disease recurrence, a rising serum prostate-specific antigen level, or initiation of adju-

vant therapy. Despite the widespread use of the postoperative nomogram and its reasonable predictive accuracy, better tools are needed to predict an individual patients probability of disease recurrence after radical prostatectomy.

Systems pathology or biology is a new discipline that is positioned to significantly impact biological discovery processes. This emerging approach attempts to facilitate discovery by systematic integration of technologies, gathering information at multiple levels (instead of only one) and examining complex interactions which results in a superior output of data and information, thereby enhancing our understanding of biological function and chemico-biological interactions [4]. The number of features generated by these technologies can be larger than standard survival methods can handle. Thus, the underlying hypothesis of this study is that an improved predictive model for disease recurrence after radical prostatectomy can be derived from a novel integrated or systems pathology approach, that will use neural networks to handle the expanded multidimensional sources of data input, including

- clinical and pathological variables (variables used in original nomogram plus additional clinical variables);
- molecular biomarker data derived from IHC analyses of tissue microarrays;
- results of machine vision image analysis which quantify histopathological features of H&E slides.

Our goal in the present study was to use clinical, histopathological, immunohistochemical (IHC), and bio-imaging data to predict prostate cancer BCR. In order to achieve this objective, a cohort of 539 patients who underwent radical prostatectomy at a single hospital in the US was studied. 16 clinical and histopathological features were collected, which include patient age, race, Gleason grade and score, and other pre- and post-operative parameters. In addition, high-density tissue microarrays (TMAs) were constructed from the patients' prostatectomy specimens. A single hematoxylin and eosin-stained (H&E) slide for each patient was used for image analysis, while the remaining sections made from the paraffin-embedded tissue blocks were used to conduct IHC studies of selected biomarkers in the laboratory. Data generated by the IHC studies included the number of cells which stained positive for a particular biomarker, if any, and the level of intensity at which the cell(s) stained positive for the biomarker. We obtained 43 IHC features from 12 biomarkers studied. Images of the H&E slides were captured via a light microscope at 20X magnification using a SPOT Insight QE Color Digital Camera (KAI2000). Using a proprietary image analysis system, pathologically meaningful objects were identified and various statistical features associated with these objects were generated. They include spectral-based characteristics (channel means, standard deviations, etc.), position, size, perimeter, shape (asymmetry, compactness, elliptical fit, etc.), and relationships to neighboring objects (contrast). In the end, 496 bio-imaging features were produced. In this study we restricted ourselves to those patients who had non-missing data for each of the above three domains (clinical and histopathological, IHC, and bio-imaging). Thus, the effective sample size consisted of only 130 patients. For these patients, the time from the surgery to the most recent follow-up visit ranged from 1 month to 133 months. Patients who had measurable prostate-specific

antigen (PSA) at this visit were considered to have recurrent prostate cancer. If a patient did not recur as of their last visit, or the patient outcome was unknown as of their most recent visit (e.g. due to loss to follow-up), then the patient outcome was considered censored, specifically right-censored. 20 patients experienced PSA recurrence among the 130 patients, while the remaining patients were censored. Thus, the available sample was very small and heavily censored. By consulting with our domain experts and using an in-house domain specific feature selection procedure, which combines greedy forward selection and backward elimination based on the relevant importance of feature groups given by domain experts, we were able to reduce the final feature set to 25 features.

4. EMPIRICAL COMPARISON

4.1 Algorithms

We compare the new algorithm, denoted as NN_{ci} , with four other algorithms over our data. The first algorithm is based on [20]. The Kaplan-Meier estimate of recurrence probability is used as the target value for the patients who had short follow-up times and did not experience recurrence. We will then refer to the MLP network trained by this algorithm as NN_{km} . The patients who had follow-up times longer than 7 years and remained disease free are assumed to be successfully cured and a target value of 0 is assigned. Those patients who had recurrence at the follow-up visit have a target value of 1. Like the new algorithm, this model should output a higher score for a higher risk patient. The second algorithm trains an MLP model to regress over the survival/censoring time [5]. It uses an asymmetric squared error function for the censored patients, which becomes zero when the model output is larger than the censoring time. We refer this model as NN_{ae} . Unlike NN_{ci} and NN_{km} , a higher risk patient should have a lower score, an estimate proportional to the survival time, from this model. We also implemented a support vector machine regressor with an asymmetric penalty function, which incurs a smaller penalty when the model output is larger than the target value (survival time) and a larger penalty when the output is smaller than the target value. We call this model SVR_c , which should output a higher score for a lower risk patient. The last algorithm we compared NN_{ci} with is the Cox proportional hazard model. The Cox model outputs a prognostic hazard score, which is a function of a linear combination of the covariates (input features) [15]. The higher the score is, the more risk the model predicts the patient would have.

4.2 Results

The empirical results are based on leave-one-out cross validation. For all the algorithms, we conducted model selection based on cross validation over the training data for fold 1, and the same model setting was used for all the folds. All the neural networks based models have 5 hidden units after the model selection. For NN_{ci} , γ was chosen to be 0.01. We have found that the results in terms of the CI value are more sensitive to γ than the AUC metric in [19]. n is typically set as 3. The RBF kernel is used in SVR_c . To obtain a Cox model, we had to reduce the number of covariates to 23 since the data set is too small.² The performance is mea-

²We used the statistical package R in our experiments.

sured in two ways. The first measure is the Concordance Index, which evaluates the model’s general predictive accuracy by estimating the probability that a patient with the higher prognostic score will recur within a shorter time than a patient with a lower score. Table 1 shows the Concordance Index values for all the models. Not surprisingly, the new algorithm NN_{ci} achieved the largest CI value over the cross validation results.

NN_{ci}	NN_{km}	NN_{ae}	SVR_c	Cox
0.8178	0.5411	0.7375	0.6206	0.7037

Table 1: Concordance index for the cross validation results of the five algorithms.

To measure the model’s predictive accuracy, we can also specifically evaluate the model’s ability to identify the high-risk and low-risk patient groups. We have shown the survival curves for both the predicted high-risk and low-risk patients in Figures 2 to 5 for all the models. Survival curves are constructed for both high-risk and low-risk patients by Kaplan-Meier estimates based on the scores from each model. In each figure, we present the new algorithm with one of the four other algorithms. The high-risk and low-risk survival curves would be far apart if the model can successfully distinguish the two patient groups. For NN_{ci} , NN_{km} , and the Cox model, the high-risk group is defined as the patients in the highest quartile of the scores, and the low-risk group consists of the patients in the lowest quartile of the scores.³ However, for both NN_{ae} and SVR_c , the high-risk patients are in the lowest quartile of the scores, and the low-risk patients fall in the highest quartile of the scores. The log-rank test [12] is conducted for each pair of survival curves of high-risk and low-risk groups within each model. The p value indicates how significant the model can distinguish high-risk and low-risk patient groups.⁴ Table 2 summarizes the p values for all the methods. We can see that the p value (< 0.0001) for the new algorithm NN_{ci} is much smaller than all the other models, and this further demonstrates the new algorithm’s improvement in identifying high-risk and low-risk patient groups.

NN_{ci}	NN_{km}	NN_{ae}	SVR_c	Cox
< 0.0001	0.15	0.01	0.10	0.01

Table 2: p values of the log-rank test over the cross validation results of the five algorithms.

4.3 Recalibration

Note that the outputs from NN_{ci} are not well-calibrated probabilities. In order to make the score from the model have a clinically meaningful interpretation, we recalibrate

³We conveniently chose the quartiles as the high risk and low risk groups here. We have seen similar results when the highest 30% and lowest 30% scores are defined as high risk and low risk groups.

⁴The log-rank test’s null hypothesis is that all hazard rates between groups are equal for all time points. The null hypothesis is rejected if the risk groups differ at any one time, and thus the p value is very sensitive to any difference between the risk groups over the entire time spectrum.

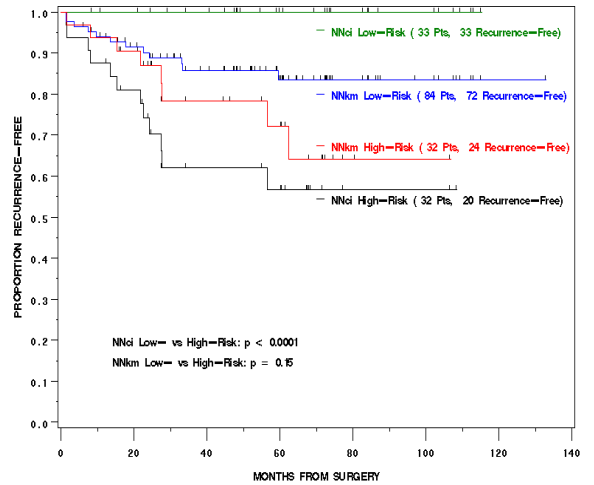


Figure 2: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and NN_{km} . Note that the low risk group of NN_{km} consists of 84 patients because of tied scores.

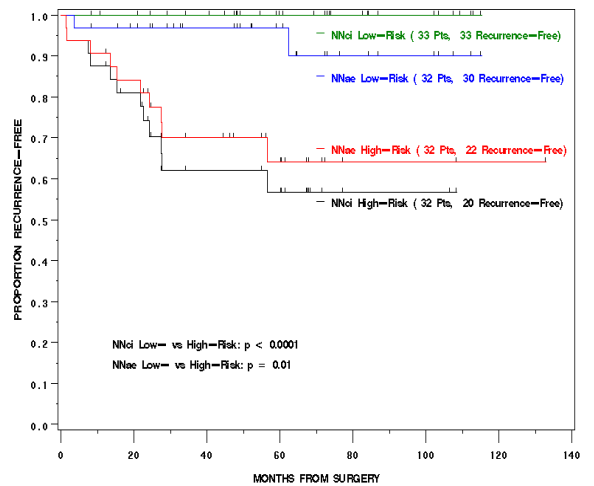


Figure 3: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and NN_{ae} .

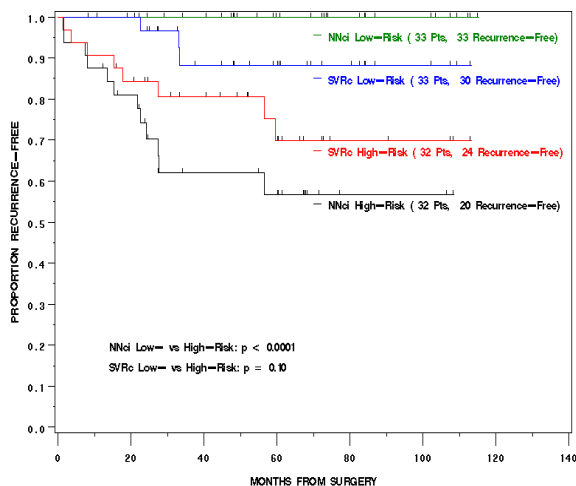


Figure 4: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and SVR_c .

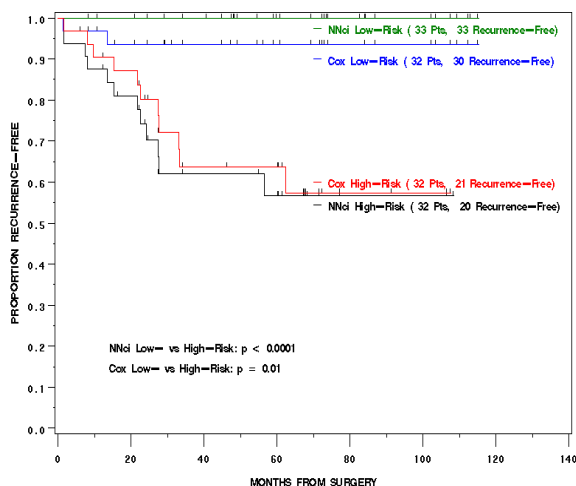


Figure 5: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and the Cox model.

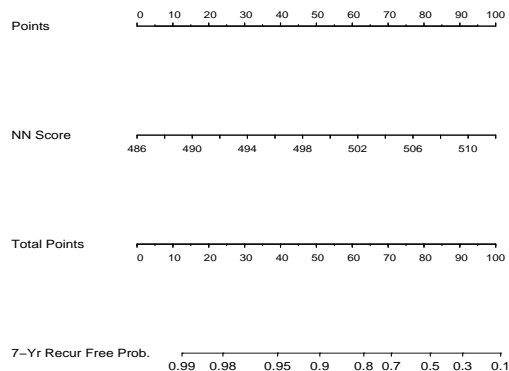


Figure 6: A nomogram based on the score from the NN_{ci} model.

the scores to probabilities of remaining free of recurrence in the next 7 years following the surgery. We estimated the probability by the cumulative hazard function, also known as the Nelson-Aalen estimator, which incorporates both the baseline hazard rate and the hazard function, estimated via partial likelihood maximization using the Newton-Raphson method [15]. Based on these probability estimates, we can generate a nomogram in Figure 6.⁵ To use the nomogram, one would draw a straight line up to the Points axis from the patient's NN score to determine how many points towards recurrence the patient receives for his NN Score. In a nomogram with more than one feature, this process would be repeated for each feature, and the points added together would equal the Total Points. In our example, with a single feature, which is the NN score, the Points and Total Points axes are identical. Once the total points are known, a straight line would be drawn down from the Total Points axis to the corresponding probability of the patient remaining recurrence-free for 7 years following the surgery, assuming the patient does not die of another cause first.

5. CONCLUSIONS

We have proposed an algorithm which allows machine learning techniques to make complete use of information from survival data. The algorithm directly maximizes a differentiable approximation to the concordance index. We have used this algorithm to train a neural network model for prediction of prostate cancer recurrence, and achieved significant improvement in being able to identify high-risk and low-risk groups of patients. The improvement is primarily due to the new algorithm's more complete use of censored data and its ability of overcoming overfitting. The Nelson-Aalen estimator is proposed to transform the original score from the model to a clinically meaningful probability, but we are still exploring and evaluating other post-processing methods for more reliable and interpretable prognostic scores. The proposed differentiable approximation to the step function appears to be of more general use. One

⁵For ease of use, we multiply the original scores from the model by 1000 in the nomogram.

possible extension is to use the approximation in a more general problem of learning ranks, where an objective function based on the step function can be formed.

6. ACKNOWLEDGMENTS

US patents have been filed based on some materials in this paper. The authors thank the helpful discussion with Junshui Ma from Ohio State University.

7. REFERENCES

- [1] E. Biganzoli, P. Boracchi, L. Mariani, and et al. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat Med*, 1998.
- [2] S. F. Brown, A. J. Branford, and W. Moran. On the use of artificial neural networks for the analysis of survival data. *IEEE Trans. on Neural Networks*, 8(5):1071–1077, 1997.
- [3] H. B. Burke, P. H. Goodman, D. B. Rosen, and et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 97(4):857–862, 1997.
- [4] E. Davidov, J. Holland, E. Marple, and S. Naylor. Advancing drug discovery through systems biology. *Drug Discov Today*, 8:175–183, 2003.
- [5] C. J. S. deSilva, P. L. Choong, and Y. Attikiouzel. Artificial neural networks and breast cancer prognosis. *Australian Comput. J.*, 26:78–81, 1994.
- [6] M. Graefen, P. I. Karakiewicz, I. Cagiannos, and et al. Validation study of the accuracy of a postoperative nomogram for recurrence after radical prostatectomy for localized prostate cancer. *Journal of Clin Oncol*, 20:951–956, 2002.
- [7] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. John Wiley & Sons, New York, 1966.
- [8] H. Gronberg. Prostate cancer epidemiology. *Lancet*, 361:859–864, 2003.
- [9] F. E. Harrell. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York, 2001.
- [10] F. E. Harrell, R. M. Califf, D. B. Pryor, and et al. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.
- [11] L. Hood. Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev*, 124:9–16, 2003.
- [12] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York, 1980.
- [13] M. W. Kattan, K. R. Hess, and J. R. Beck. Experiments to determine whether recursive partitioning or an artificial neural network overcomes theoretical limitation of cox proportional hazards regression. *Comput Biomed Res*, 31(5):363–373, 1998.
- [14] M. W. Kattan, T. M. Wheeler, and P. T. Scardino. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *Journal of Clin Oncol*, 17:1499–1507, 1999.
- [15] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 1997.
- [16] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.
- [17] L. Ohno-Machado and M. A. Musen. Modular neural networks for medical prognosis: Quantifying the benefits of combining neural networks for survival prediction. *Connection Science*, 9:71–86, 1997.
- [18] P. Snow, D. S. Smith, and W. J. Catalona. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. *J. Urology*, 152(5):1923–1926, 1997.
- [19] L. Yan, R. Dodier, M. Mozer, and R. Wolnienicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proc. of 20th Int'l Conf. Machine Learning*, pages 848–855, 2003.
- [20] B. Zupan, J. Demsar, M. W. Kattan, and et al. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, 20:59–75, 2000.