EVALUATION OF PROGNOSTIC MODELS: DISCRIMINATION AND CALIBRATION PERFORMANCE

A.F.G. Taktak*, A. Eleuteri*, S.P. Lake, A.C. Fisher

*Dept. Clinical Engineering, Royal Liverpool University Hospital, Liverpool, UK

Keywords: survival analysis, C index, goodness-of-fit, artificial data

Abstract

Prognostic models are developed to assist clinicians in making decisions regarding treatment and follow-up management. The accuracy of these models is often assessed either in terms of their discrimination performance or calibration but rarely both. In this paper, we describe a method for assessing both these aspects using the Harrell C index of discrimination and a Hosmer-Lemeshow type analysis for calibration. We show some illustrative examples to demonstrate the importance of assessing both discrimination and calibration. A tool is available to assess models online on the following URL: www.clineng-liverpool-nhs.com/AADP/Welcome.htm

1 Introduction

Numerous models have been developed to provide prognostic information in survival analysis. Such information is valuable both to clinicians and patients. It allows clinicians to choose appropriate treatment and plan follow-up as appropriate. Patients at high risk could be followed up more frequently than those at lower risk in order to channel valuable resources to those who need it most. For patients, obtaining information about their prognosis, although can be morbid, it is also extremely valuable in terms of planning their lives and providing care for their dependents.

There is a large number of methods described in the literature for developing prognostic models. Some of these models are based on linear statistical methods such as the Cox model and others are based on non-linear machine learning methods such as artificial neural networks. A number of studies have been carried out to compare these methods [B8], [B6], [B5]. The advantages and disadvantages of these types of analyses are also well understood and documented [B9], [B10].

Whatever model is used for studying survival, it is important to assess the performance of the model in two ways; its discrimination and calibration aspects. Discrimination is the ability of the model to correctly separate the subjects into different groups. Calibration is the degree of correspondence between the estimated probability produced by the model and the actual observed probability [B3]. It can be argued that discrimination performance is more important than calibration since calibration can be adjusted whereas a model that cannot discriminate between the different groups can not be put into practice. On the other hand, poor calibration can occur in highly discriminating models when the output is transformed monotonically [B2]. In a previous study, we performed double-blind comparison of five prognostic models in a multi-centre trial using a benchmark data set [B11]. In this paper, we describe the implementation of the assessment procedure using a web-based tool. We show two examples using this tool to illustrate the need to study both discrimination and calibration concurrently.

2 Background

2.1 Survival Analysis

Survival analysis is the study of time elapsed from some particular starting point to the occurrence of an event. The starting point of observation is usually a medical intervention such as first diagnosis of a given disease, a surgical intervention or the beginning of a treatment in a clinical study. The survival time is actually the time up to a certain event. Such event may be death, a relapse, or the development of a given disease.

Let T be the random variable denoting the survival time. The survival function *S (t),* is defined as the probability that an individual survives longer than t:

$$S(t) = P(T>t) \qquad (1)$$

S(t) is a nonincreasing function of time *t* with the following properties:

$$S(t) = \begin{cases} 1 & \text{for } t = 0 \\ 0 & \text{for } t = \propto \end{cases} \qquad (2)$$

meaning that the probability of surviving at least at time zero is 1 and that the probability of surviving at an infinite time is zero.

The function *S (t)* is also known as the cumulative survival rate. The graphic representation of *S (t)* is called the survival curve. The basic problem in survival analysis is to estimate from the sampled data one or more of these three functions, and to draw inferences about the survival pattern in the population.

In any observational study of survival, T is not known for all subjects entered in the study. Subjects ‘drop out’ from the study either because they are lost to follow-up (e.g. changing their address, dying of other causes, etc.) or because the event has not happened by the time the study has started. To deal with this problem, each subject in the study is only included up to the time they have been observed. If we assume that no subjects are lost to follow-up, the *follow-up time* (T_f) is the time from when the diagnosis was made up to either when the event has happened for deceased subjects or when the study has started for subjects who are still alive. The *Event Status Indicator* (*D*) is a binary variable representing the presence (=1) or absence of an event (=0). For example, if subject *i* died 4.5 years after he/she was diagnosed of the disease, then T_{fi} = 4.5 and D_i = 1. If another subject *j* was diagnosed 7.5 years before the study started and was still alive when the study has started, then T_{fj} = 7.5 and D_j = 0.

2.2 Discrimination

Ideally, if a large dataset exists, testing the model is carried out on an unseen set of data that has not been used for training and validation. As mentioned earlier, for prognostic models, it is desirable to assess the model in terms of its discrimination and calibration aspects. The most appropriate method for assessing discrimination in survival analysis is Harrell’s C index [B4] [B1]. The C index is an extension of the Area Under the Receiver Operator Curve (AUROC) but it is more suited to survival analysis since it is threshold independent. It is calculated by looking at all pairs of samples which are *comparable* and calculating the probability of these pairs being *concordant*. To calculate first the probability of any given pair being comparable, at least one of the samples in the pair must have developed the event (e.g. death) and that the follow-up time period for this sample is less than that of the other sample. For example, let us assume the same subjects *i* and *j* above. The subjects are defined comparable by the event $\{T_{fi} < T_{fj} \mid D_i = 1\}$. A binary variable $\pi_{comp(i,j)}$ can be defined as an indicator of the event. The (unnormalised) *probability of comparison* for the dataset can therefore be calculated as:

$$\Pr_{comp} = \sum_i \sum_j \pi_{comp(i,j)} \qquad (3)$$

Next, we need to calculate the *probability of concordance*. In the above example, the elements of the pair {*i, j*} are said to be concordant if the probability of survival for subject *i* as predicted by the model under test (S_i) is greater at time *t* than that for subject *j* (S_j). A second binary variable $\pi_{conc(i,j)}$ can be defined which is set to 1 if:

a) the pair is comparable, and
b) the above condition is true

In other words, $\pi_{conc(i,j)}$ is an indicator for the event $\{S_{ti} > S_{tj} \mid \pi_{comp(i,j)} = 1 \;\; \& \;\; t \leq T_{fi}\}$. The (unnormalised) *probability of concordance* can then be calculated as:

$$\Pr_{conc} = \sum_i \sum_j \pi_{conc(i,j)} \qquad (4)$$

Finally, the C index is calculated as:

$$C = \frac{\Pr_{conc}}{\Pr_{comp}} \qquad (5)$$

2.3 Calibration

Next, we shall look at the assessment of the calibration performance of the model. For this, the Hosmer-Lemeshow statistics is an appropriate test [B7]. In this approach, the S_t values are first rank ordered and divided into *N* groups. The upper group contains subjects who are least likely to develop the event whereas the lowest contains those who are most likely to develop the event. Now if o_l denotes the observed survival in group *l* calculated by:

$$o_l = 1 - \frac{\text{Number of events in group } l}{\text{Number of samples in group } l} \qquad (6)$$

and e_l denotes the average estimated *S* value for group *l*, a goodness-of-fit measure can be obtained by comparing o_l and e_l graphically for *l* = 1,….,*N*. To quantify this analysis, a chi-square statistic can be derived by:

$$\chi^2 = \sum_{l=1}^{N} \frac{(o_l - e_l)^2}{e_l} \qquad (7)$$

with *N*-(*p*+1) degrees of freedom.

3. Method

3.1 Simulation data

In order to demonstrate the process, a set of artificial data was created based on real clinical data. First, a random variable X simulating a covariate was sampled 2000 times from a log normal probability distribution function X~LOGN(2.5,0.09). Next, a second variable depicting noise was created having a normal probability distribution function such that Noise~N(1,16). A prognostic index P was calculated as follows:

$$P=10*e^{(-0.05*X)} \qquad (8)$$

The actual survival time T was sampled from a gamma probability distribution function T~GAMMA(P,4).

A censoring variable C was created also using a gamma distribution function C~GAMMA(CT,1) where CT is a control variable which can be used to control the proportion of events in the dataset. The follow-up time T_f was calculated:

$$T_f = \text{Min}(C,T) \qquad (9)$$

The *Event Status Indicator* (*D*) was then calculated:

$$D = \begin{cases} 0 & \text{if } C < T \\ 1 & \text{if } C \geq T \end{cases} \qquad (10)$$

The Cox model was used to predict survival using (X+Noise) as a covariate. Discrimination and calibration assessment was carried out at different event ratios.

3.2 Algorithm implementation

The above algorithm was implemented in MATLAB® (The MathWorks). The program examines the data in pairs and for each pair (i,j) it carries out a test of comparability and a test of concordance as described above. The program calculates the total number of comparable pairs *r*, the number of comparable and concordant pairs *w* and the number of comparable and disconcordant pairs *v*. Under the assumption that the above algorithm provides a Gaussian estimator for the C index, the chi-square standard error of the mean can be calculated as follows:

$$s.e. = \frac{\sqrt{\left(\sum_{ij} r^2 \times \sum_{ij} w\right)^2 - 2\left(\sum_{ij} r \times \sum_{ij} w \times \sum_{ij} rw\right) + \left(\sum_{ij} r \times \sum_{ij} w^2\right)^2}}{(\sum_{i} r)^2} \qquad (11)$$

3.3 Analysis web tool:

We have developed a simple-to-use and robust method to employ a MATLAB function to carry out these analysis across the Internet. The tool is accessed from a Client/User web-page using the generic Simple Object Access Protocol (SOAP) interface to multiple concurrent MATLAB Automation Services on a single host (single processor) machine. MATLAB Automation Services offers remote access to an instance of MATLAB, with commands being submitted from local programs such as MS Visual Basic or MS Excel effectively as command line entries. This method lends itself to the rapid deployment of teaching materials, research works and turn-key applications which are feature of modern signal and image processing in medicine. It is not necessary for the Developer, Host Administrator or the Client/User to have any specialised knowledge relating to web-interfacing or server management: the MatSOAP method is constructed as a ready-to-go software tool. The interface layer, inherent between the MATLAB command line and the User/Client, complies entirely with the legal restrictions on the MATLAB licences.

The MatSOAP arrangement can be summarised as a Client Web-Page which *submits*, over a TCPIP network, 2 pieces of information, *i.e.* the MATLAB function name and input variables as a single CSV string, to the MatSOAP Web-Page. MATLAB executes the function and returns output as a single string to the Client web-page.

In this application, the user first uploads the datafile onto the server (Filename). They then specify the time point at which the analysis is to be carried out (τ) and the number of groups for the goodness-of-fit test (NumGrps). The input string to the MatSOAP tool therefore has the following format:

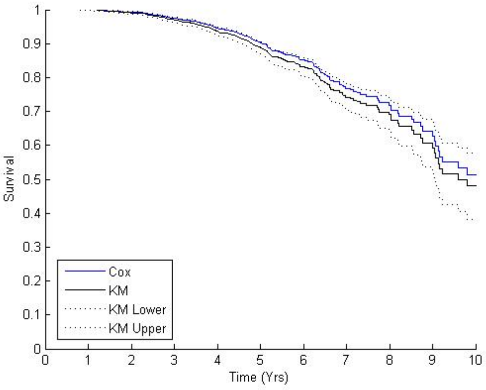
Input String = (<Filename>,<τ>,<NumGrps>)

The MATLAB function decodes the input, carries out the analysis and passes the output in a single CSV string which is processed at the client's machine using JavaScript. The output string contains information about the value of the C index (C) with the confidence intervals (C_H, C_L), the χ^2 statistic for the goodness-of-fit test with the p value and the name of the JPEG calibration image on the server (CalImg). The output string therefore has the following format:

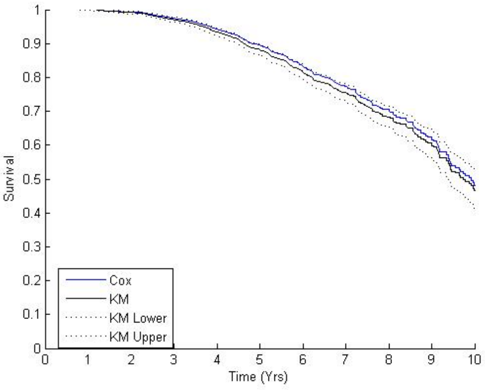
Output String = (<C>,< C_H, C_L >,< χ^2>,<Pval>,<CalImg>)

4. Results

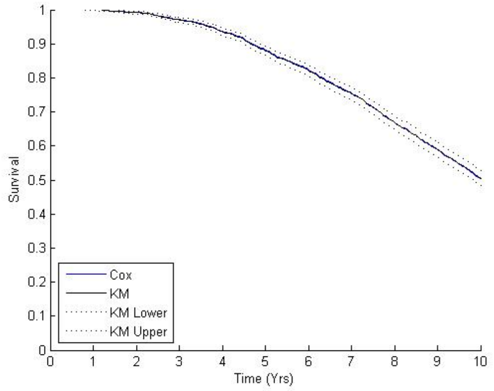
Two values for mean parameter for CT were chosen. These were 6 and 9 corresponding roughly to 10% and 20% event proportions (R) respectively at 10 years. The Cox model provided a good estimate of survival in each case as can be seen in figure 1. The model was also evaluated with no censorship which resulted in R = 40% for a 10-year follow-up. Figure 2 shows that variable (X+Noise) was a good discriminator in the model.



(a)



(b)



(c)

Figure 1 Survival function in the dataset showing good agreement between the Cox model and the Kaplan-Meier estimate for events proportions of (a) 0.1, (b) 0.2 and (c) 0.4

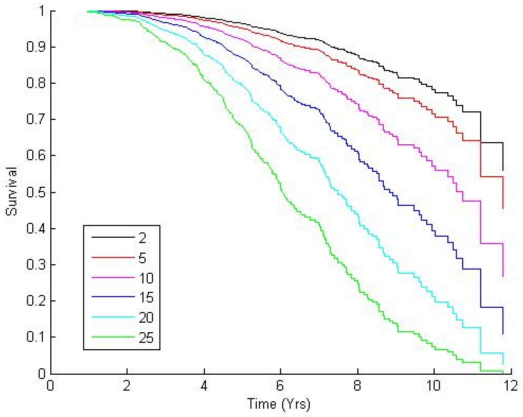


Figure 2 Survival curves showing worsening survival with increased value of the covariate

The C index with the 95% confidence intervals was calculated for different values of R at time points τ = 3, 5 and 10 years. The results are shown in table 1.

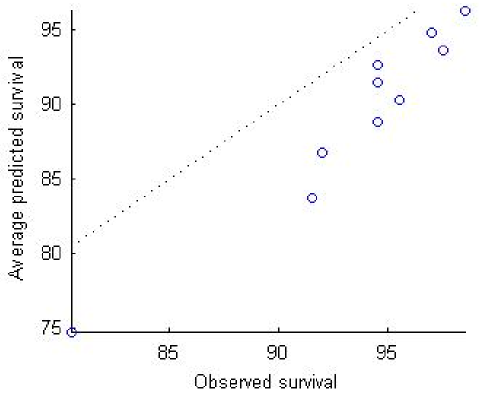
The χ^2 statistics and the corresponding p values are shown in table 2 for the same time intervals and the same proportion of events values. Calibration curves are shown in figures 3-6.

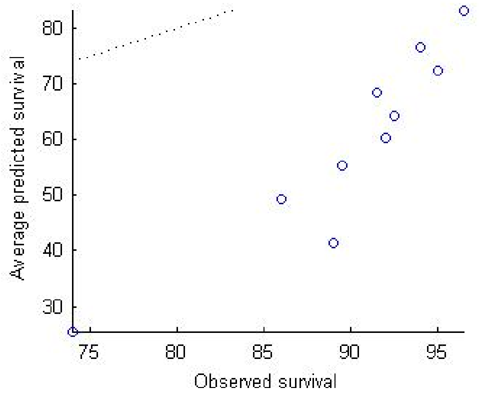
R	τ	C index	95% CI
0.1	3	0.7454	(0.6164 – 0.8744)
	5	0.7074	(0.6142 – 0.8006)
	10	0.7012	(0.6174 – 0.7854)
0.2	3	0.7656	(0.6456 – 0.8856)
	5	0.7172	(0.6406 – 0.7938)
	10	0.6809	(0.6189 – 0.7429)
0.4	3	0.7535	(0.6391 – 0.8679)
	5	0.6975	(0.6293 – 0.7657)
	10	0.6118	(0.5752 – 0.6484)

Table 1: The C index and 95% confidence intervals at different time points (τ) for different proportions of events (R)

R	τ	χ^2	P
0.1	3	0.0622	1
	5	2.5144	0.9805
	10	244.2917	0
0.2	3	0.1742	1
	5	1.0490	0.9993
	10	285.6117	0
0.4	3	0.3956	1
	5	3.9817	0.9126
	10	4.2722	0.8926

Table 1: The x2 statistics for the goodness-of-fit at different time points (τ) for different proportions of events (R)





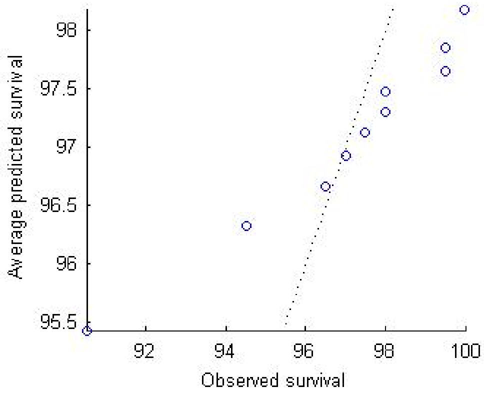


Figure 3 Calibration curves at τ = 3 (top), 5 (middle) and 10 (bottom) for events proportion of 0.1. The dotted line represents the line of equality

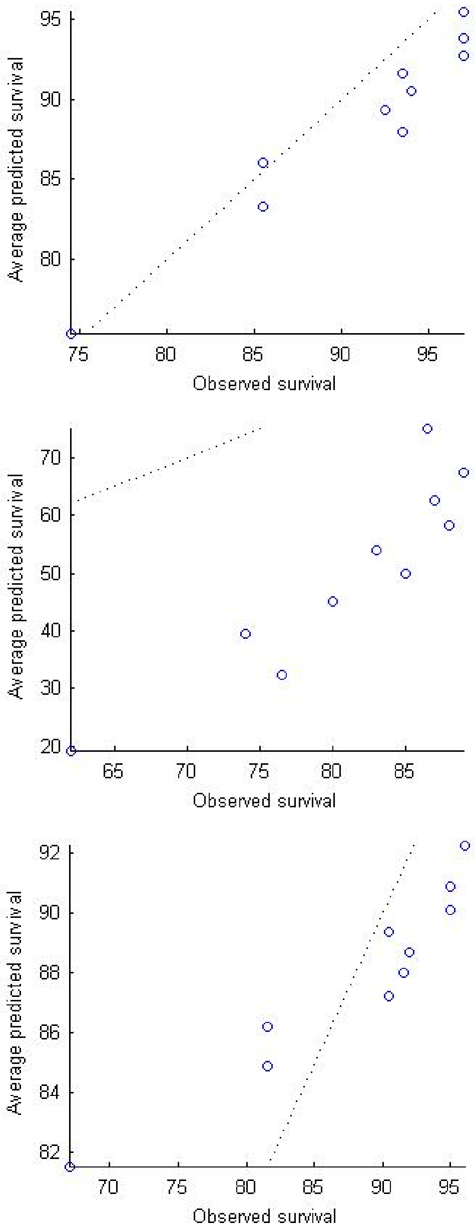


Figure 4 Calibration curves at τ = 5 (top), 5 (middle) and 10 (bottom) for events proportion of 0.2. The dotted line represents the line of equality

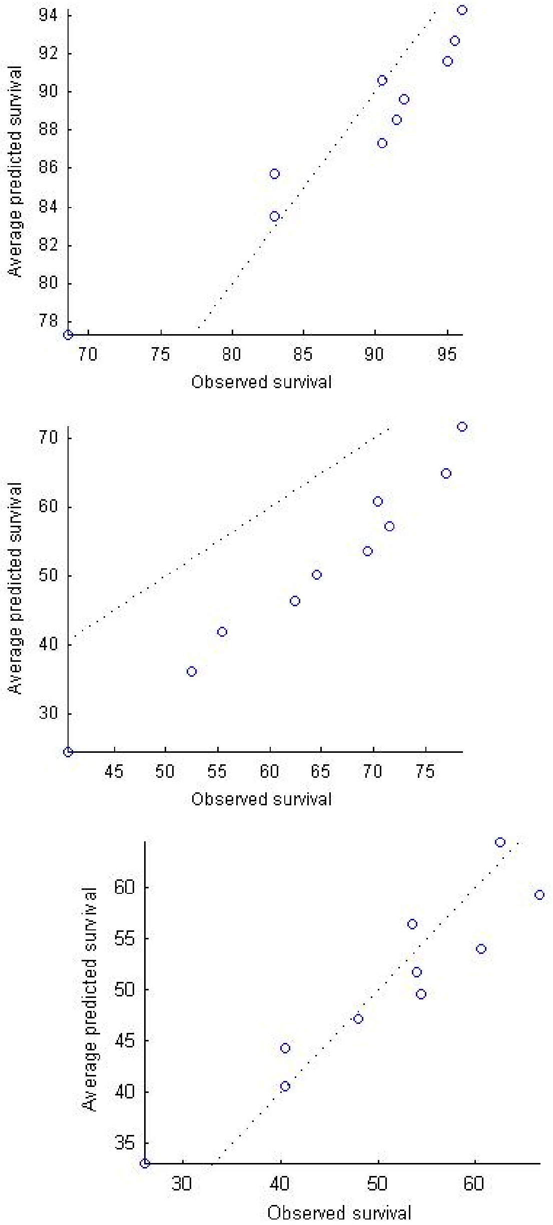


Figure 5 Calibration curves at τ = 10 (top), 5 (middle) and 10 (bottom) for events proportion of 0.4. The dotted line represents the line of equality

5. Discussion

The results showed that the C index decreased with increasing time. We have also calculated the area under the ROC curve figures for these results as an alternative measure of discrimination and found similar results. Calibration results also showed that the results were better numerically at $\tau = 3$ and 5 than those at $\tau = 10$. The visual plot provides a good clue in determining the reason for the degradation of the performance at longer time points. Although the trends of the calibration points are plausible, it is shifted from the line of equality probably due to censoring. For example, the markers in figure 4 corresponding to $\tau = 10$ are closer to the line of equality than those shown in figure 3 for the same time period. The corresponding graph in figure 5 where there was no left censoring shows that the points lie on either side of the line of equality. It is very important when assessing the performance of any prognostic model to determine the median follow-up time to give an indication about the amount of censorship in the data. In the datasets used in this study, these figures were 5.2, 7.4 and 10.1 for R = 0.1, 0.2 and 0.4 respectively. These figures show that the performance of the model is determined by the data itself and the cause of the degradation in the discrimination performance is probably due to the data rather than the model.

6. Conclusion

In this paper we have shown that a true assessment of any prognostic model should be made in terms of both discrimination and calibration. A large number of studies have assessed the performance of the models they use in terms of accuracy or the C index of discrimination only. However, neither of these measures are sufficient indicators of the performance by themselves. Both measures have been shown to be dependent on the sample size and the proportion of events, which can lead to misleading results.

The illustrative examples given in this paper also showed that graphical representation of the performance is very important in the assessment and can compliment numerical analysis. On the other hand, if one looks at the graphical assessment alone, it will be difficult to decide at which point the evidence in front of them is strong enough. We therefore conclude that survival models should be assessed in terms of discrimination and calibration both graphically and numerically.

Acknowledgements

This project is funded by the Biopattern Network of Excellence FP6/2002/IST/1; proposal N° IST-2002-508803; Project full title: Computational Intelligence for Biopattern Analysis in Support of eHealthcare; URL: www.biopattern.org

References

[1] Antolini, L., Boracchi, P., and Biganzoli, E., "A time-dependent discrimination index for survival data," *Stat.Med.*, **vol. 24, no. 24**, pp. 3927-3944, 2005.

[2] Chiu, J. S., Hu, T. M., Li, Y. C., and Hsu, C. Y., "Choroidal melanoma prognosis," *Ophthalmology.*, **vol. 113, no. 8**, pp. 1474-1475, 2006.

[3] Dreiseitl, S. and Ohno-Machado, L., "Logistic regression and artificial neural network classification models: a methodology review," *J.Biomed.Inform.*, **vol. 35, no. 5-6**, pp. 352-359, 2002.

[4] Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A., "Evaluating the yield of medical tests," *JAMA.*, **vol. 247, no. 18**, pp. 2543-2546, 1982.

[5] Jerez, J. M., Franco, L., Alba, E., Llombart-Cussac, A., Lluch, A., Ribelles, N., Munarriz, B., and Martin, M., "Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks," *Breast Cancer Res.Treat.*, pp. 1-8, 2005.

[6] Kattan, M. W., "Comparison of Cox regression with other methods for determining prediction models and nomograms," *J.Urol.*, **vol. 170, no. 6 Pt 2**, pp. S6-S9, 2003.

[7] Lemeshow, S. and Hosmer, D. W., Jr., "A review of goodness of fit statistics for use in the development of logistic regression models," *Am.J.Epidemiol.*, vol. 115, no. 1, pp. 92-106, 1982.

[8] Ohno-Machado, L., "A comparison of Cox proportional hazards and artificial neural network models for medical prognosis," *Comput.Biol.Med.*, **vol. 27, no. 1**, pp. 55-65, 1997.

[9] Sargent, D. J., "Comparison of artificial neural networks with other statistical approaches: results from medical data sets," *Cancer*, **vol. 91, no. 8** pp. 1636, 2001.

[10] Schwarzer, G., Vach, W., and Schumacher, M., "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Stat.Med.*, **vol. 19, no. 4**, pp. 541-561, 2000.

[11] Taktak, A. F. G., Antolini, L., Aung, M., Boracchi, P., Campbell, I., Damato, B., Ifeachor, E. C., Lama, N., Lisboa, P. J., Setzkorn, C., Stalbovskaya, V., and Biganzoli, E., "Double-blind evaluation and benchmarking of survival models in a multi-centre study," *Comput.Biol.Med.*, In Press.