Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Double-Blind Comparison of Survival Analysis Models Using a Bespoke Web System

A.F.G. Taktak
Dept. Clinical Engineering
Royal Liverpool University Hospital
Liverpool, L7 8XP, UK

C. Setzkorn
Dept. Clinical Engineering
Royal Liverpool University Hospital
Liverpool, L7 8XP, UK

B.E. Damato
St. Paul's Eye Unit
Royal Liverpool University Hospital
Liverpool, L7 8XP, UK

***Abstract*- The aim of this study was to carry out a comparison of different linear and non-linear models from different centres on a common dataset in a double-blind manner to eliminate bias. The dataset was shared over the internet using a secure bespoke environment called geoconda (www.geoconda.com). Models evaluated included: (1) Cox model, (2) Log Normal model, (3) Partial Logistic Spline, (4) Partial Logistic Artificial Neural Network and (5) Radial Basis Function Networks. Graphical analysis of the various models with the Kaplan-Meier values were carried out in 3 survival groups in the test set classified according to the TNM staging system. The discrimination value for each model was determined using the area under the ROC curve. Results showed that the Cox model tended towards optimism whereas the partial logistic Neural Networks showed slight pessimism.**

I. Introduction

Survival analysis is an important issue for clinicians since it allows them to review their practice and plan treatment effectively. They are also of equal importance to patients as it gives them the opportunity to make choices and plan care for their dependents. Traditionally, the use of Kaplan-Meier (KM) non-parametric model has been used as the main method for analyzing survival. This model provides graphical representation of the dataset. Group-specific survival is often analyzed by comparing the KM curves for these groups (such as males vs females) to investigate the importance of variables in predicting outcome. However, the KM model does not provide predictions for unseen combination of inputs nor dies it provide predictions beyond the date of the last event or last follow-up. Probability of survival figures derived from the KM model must be interpreted with care as it is dependent on the number of subjects at risk. For example, if a single event is observed in a population of 10 subjects at risk of developing the event, the probability of survival drops by 10% whereas if there were only 2 subjects at risk, then it will drop by 50%.

To overcome those limitations, a number of models have been proposed and utilized in studying survival and other outcomes from cancer. Some of these models are based on statistics such as logistic regression, Cox proportional hazards model, and log normal model. Others are based on artificial intelligence and machine learning such as artificial neural networks (ANNs). Although the latter offer a number of advantages over their statistical counterparts, they do have a number of limitations as is expected from any model. The benefit of using systems with ANNs in cancer has been reviewed [1]. A number of studies compared the performance of ANNs with statistical models [2] [3] [4] [5]. Others compared ANNs with basic clinical indicators such as the Tumour-Node-Metastasis (TNM) staging system [6] [7] [8]. Almost all of these studies have shown that ANNs match if not supersede the other methods.

A review paper looked at a number of comparison studies showing that in the majority of cases, equal performance was claimed [9]. The study however could not rule out the possibility of bias.

In this study, we aimed to carry out a comparison of different models from different centers on a common dataset in a double-blind manner to eliminate bias.

II. Methodology

Subjects

Patients were selected from the database of the Liverpool Ocular Oncology Centre for the time period 1984 – 2004. The dataset was split into training and test sets with a 3:2 ratio. The sets were stratified to include roughly equal proportion of events. This resulted in a training set containing 1734 patients with 490 events and the test set containing 1146 patients with 305 events. The median follow-up time was 5.31 years. The tenets of the Helsinki Declaration were followed and institutional ethical committee approval for a multi-center outcomes analysis was obtained.

The variables used in this study are summarized in Table 1. The outcome of interest was all-cause mortality [10].

Models Used

Five different models were incorporated in this study. To give each model the best chance of obtaining good results, the participants were given the freedom of choice on cross validation and dealing with missing data. A brief description of each model and how they were applied in this work is included below. For full description of these models, the reader is referred to the relevant references cited.

The Cox Model (Cox)

The Cox model is described in many textbooks (see for example [11]). The model used all variables shown in Table 1 except UH. It was trained and validated on the whole training set.

TABLE 1 VARIABLES INCLUDED IN THE STUDY

Name	Description	Type	Values
AGE	Age	Continuous	In years
SEX	Sex	Categorical	0 – Female 1 – Male
ANTMAR	Extraocular Extension	Binary	0 – No 1 - Yes
LUBD	Longest ultrasound basal dimension	Continuous	In mm
UH	Tumour height measured by ultrasonography	Continuous	In mm
EPI	Presence of epithelioid melanoma cells	Binary	0 – No 1- Yes

The Lognormal Model (Lognorm)

The lognormal model assumes that the logarithm of the survival time (S) is normally distributed [12]. The model used all variables shown in Table 1 except SEX. It was trained and validated on the whole training set.

Partial Logistic Splines (PLSPL)

The PLSPL model is a parametric generalized linear model for discrete hazard function based on partial logistic regression [13] [14]. The model used all variables shown in Table 1 except SEX and UH It was validated using a bootstrap method.

Partial Logistic Neural Networks Model with Auto-Relevance Determination (PLANN)

This model is an expanded version of the PLANN model [14] to includes a regularization framework based on Automatic Relevance Determination (ARD) [15]. The model used all variables shown in Table 1. It was validated using a 5-fold cross-validation method.

Radial Basis Function (RBF) Networks

Radial basis function networks were extracted to estimate probability of survival using a Gaussian basis function. A multi-objective evolutionary approach was used to select the best model [16]. The model used all variables shown in Table 1 except EPI. It was validated by randomly splitting the training set into training and validation subsets with a 2:1 ratio.

III. EVALUATION AND BENCHMARKING

The evaluation procedure was carried out in a double-blind manner to avoid bias. The dataset was shared over the internet using a bespoke secure environment called geoconda (www.geoconda.com) [17]. After the training period was complete, the test dataset was distributed without the outcome.

Next, the participants sent their predictions to a referee who anonymized the results and forwarded them to the assessor. The assessor carried out the analysis without knowing which result belonged to which model .

The discrimination value for each model was determined with receiver-operator characteristic (ROC) analysis on the hazard values. The area under the ROC curve (AUC) was calculated at time intervals 2, 3, 5, 7 and 10. Since the different models were based on a different number of samples due to the various methods in which they dealt with missing data, bias might be present in the analysis. In order to remove such bias, the AUC figures were also calculated for the largest common dataset.

The probability of survival S was also calculated from the hazard values. The models were compared on a case-by-case basis. In this comparison, each model was scored according to the number of times in which they produced the highest S figure for cases that were observed alive and the lowest S for cases that have died.

The calibration of the different models was also evaluated. The test set was split into 3 prognostic groups according to the tumour diameter using the TNM staging system [18]. Average survival curves in each group were generated by the different model and were compared with the KM curve for that group.

IV. RESULTS

The number of predictions made by each model, the mean probability of survival for alive subjects (S_0) and deceased ones (S_1) and the AUROC figures at each time interval were calculated.

On a case-by-case basis, the percentage of cases in the common set where predictions were best (i.e. highest S for alive patients and lowest S for deceased) were plotted for each model as shown in Fig. 1. If all models perform equally then the bars would have equal proportions for all models.

The largest common set where prediction was provided by all models contained 498 records. Results for this set are shown in Table 2.

These results seem to suggest that Cox is the best model for predicting survival and the worst for predicting death, i.e. the most optimistic model. The PLANN model on the other hand is generally the best model for predicting death (apart from time interval 10) and the worst for predicting survival making it the most pessimistic.

As described earlier, the test set was split into 3 prognostic groups according to the tumour diameter using the TNM staging in order to determine the calibration properties the models. These groups were labeled as 1, 2 and 3 representing good, medium and poor survival respectively. Calibration was analyzed graphically by comparing the average group survival curves for each model with the KM values as shown in Fig. 2. These results again suggest that the Cox model is rather optimistic in the good and medium survival groups whereas the PLSPL and the PLANN models are slightly pessimistic in the medium survival group.

6. Discussion

The study had several strengths. First, the various methods were carried out by different centers. Second, there were no restrictions imposed on the researchers performing the analyses thereby giving each method the best chance of obtaining maximum performance. Third, the comparison was carried out in a double-blind manner. Finally, the study was undertaken over the Internet thereby minimizing the costs associated with multi-center studies.

The main weakness of the study was the handling of the missing data. Although participants were given the freedom in the way they dealt with missing data, this resulted in downgrading the performance of some models. It was observed, for example, that survival of patients with missing values was higher than those without missing data, which indicated that the data were not missing at random. In this case, imputing the missing data from the rest of the dataset was unsatisfactory. Attributing a missing category was also unsatisfactory as it implicitly assumed that a specific status was associated with a lack of information.

This study provides a template for evaluation and benchmarking various analysis methods aiming to avoid the evaluation bias. The slight differences in the model performances were insignificant and there was generally good agreement in all the results. Comparison with the KM figures for the different prognostic groups showed a slight tendency towards optimism for the Cox model in the excellent and good prognosis groups and a slight tendency towards pessimism for the PLANN model in the medium prognosis group. However, it must be emphasized here that these conclusions cannot be generalized to other datasets.

Overall, there was no indication in this study that a particular model is superior to the others. The study therefore gives confidence in the use of ANN models in survival analysis since they are not constrained by rigid assumptions regarding the functional dependence of the outcome on the investigated covariates and time. They do not therefore rely on the availability of prior knowledge as is the case with linear models. Such favorable property also allows the exploration of unknown covariate effects when accounting for them. However, careful consideration has to be made in the design of ANN models and the handling of the data.

ACKNOWLEDGMENT

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] Lisboa, P. J. and Taktak, A. F. G., "The use of artificial neural networks in decision support in cancer: A Systematic Review," *Neural Netw.*, Jan.2006.

[2] 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, Dec.2003.

[3] 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.*, vol. :1-8. pp. 1-8, Oct.2005.

[4] 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, Jan.1997.

[5] Ravdin, P. M., Clark, G. M., Hilsenbeck, S. G., Owens, M. A., Vendely, P., Pandian, M. R., and McGuire, W. L., "A demonstration that breast cancer recurrence can be predicted by neural network analysis," *Breast Cancer Res.Treat.*, vol. 21, no. 1, pp. 47-53, 1992.

[6] De Laurentiis, M., De Placido, S., Bianco, A. R., Clark, G. M., and Ravdin, P. M., "A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients," *Clin.Cancer Res.*, vol. 5, no. 12, pp. 4133-4139, Dec.1999.

[7] Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E., Jr., Marks, J. R., Winchester, D. P., and Bostwick, D. G., "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer*, vol. 79, no. 4, pp. 857-862, Feb.1997.

[8] Sato, F., Shimada, Y., Selaru, F. M., Shibata, D., Maeda, M., Watanabe, G., Mori, Y., Stass, S. A., Imamura, M., and Meltzer, S. J., "Prediction of survival in patients with esophageal carcinoma using artificial neural networks," *Cancer*, vol. 103, no. 8, pp. 1596-1605, Apr.2005.

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

[10] Kroll, S., Char, D. H., Quivey, J., and Castro, J., "A comparison of cause-specific melanoma mortality and all-cause mortality in survival analyses after radiation treatment for uveal melanoma," *Ophthalmology*, vol. 105, no. 11, pp. 2035-2045, Nov.1998.

[11] Armitage, P., *Statistical methods in medical research* Oxford: Blackwell Scientific, 2000.

[12] van Belle, G., Heagerty, P. J., Fisher, L. D., and Lumley, T. S., *Biostatistics : A Methodology For the Health Sciences* New Jersey: Wiley, 2004.

[13] Efron, B., "Logistic regression, survival analysis, and the Kaplan Meier curve," *J.Amer.Statist.Assoc*, vol. 83 pp. 414-425, Jan.1988.

[14] Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E., "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Stat.Med.*, vol. 17, no. 10, pp. 1169-1186, May1998.

[15] Lisboa, P. J., Wong, H., Harris, P., and Swindell, R., "A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer," *Artif.Intell.Med.*, vol. 28, no. 1, pp. 1-25, May2003.

[16] Setzkorn, C., Taktak, A. F. G., and Damato, B., "On The Use Of Multi-Objective Evolutionary Algorithms For Survival Analysis," *BioSystems*, Jan.2006.

[17] Setzkorn, C., Taktak, A. F. G., and Damto, B., "Geoconda: A Web Environment for Multi-Centre Research," in Taktak, A. F. G. and Fisher, A. C. (eds.) *Outcome Prediction in Cancer* Amsterdam: Elsevier, 2006.

[18] *TNM Classification of Malignant Tumours*, 5 ed. New York: Wiley, 1997, pp. 158-161.