

Patient survival estimation with multiple attributes: adaptation of Cox's regression to give an individual's point prediction

Ann E. Smith*, Sarabjot S. Anand*.

Abstract. In the field of medical prognosis, Cox's regression techniques have been traditionally used to discover "hazardous" attributes for survival in databases where there are multiple covariates and the additional complication of censored patients. These statistical techniques have not been used to provide point estimates which are required for the multiple evaluation of cases and comparison with other outputs. For example, neural networks (NNs) are able to give individual survival predictions in specific times (with errors) for each patient in a data set. To this end, an evaluation of predictive ability for individuals was sought by an adaptation of the Cox's regression output. A formula to transform the output from Cox's regression into a survival estimate for each individual patient was evolved.

The results thus obtained were compared with those of a neural network trained on the same data set. This may have wide applicability for the performance evaluation of other forms of Artificial Intelligence and their acceptance within the domain of medicine.

1 INTRODUCTION AND RATIONALE

Evidence-based medicine, with its attendant requirement for accountability in treatment strategy at the individual patient level, requires that we try to obtain a quantitative assessment of the prognosis for a patient. This prognosis is based on multiple relevant factors or attributes collected about that patient. Extremely precise predictions at the individual level are, of course, impossible because of unexplained variability of individual outcomes [1]. Cox's regression technique [2] has been the standard statistical tool utilised where "censored" patients exist. Censorship means that the event does not occur during the period of observation and the time of event is unknown, but these cases are incorporated into the analysis. Those whose event is unknown, or who are lost to the study (right censored) or new patients introduced into the study (left censored), add to the information on patients whose event time is known (uncensored), at each time interval. Cox's regression is used to derive the hazard ratio, and hazard regression coefficients, where there are multiple attributes associated with survival and a variable time to an event, e.g. death.

However, criticism has been made of traditional statistics in providing prognostic outcomes for individual patients [3]. Neural networks have entered this field [4] and have been shown to possess some advantages in overcoming the drawbacks. These include being able to give point estimates on multiple cases as a system output whilst not having to depend on assumptions of linearity, or the proportionality with time, of the hazard variables. The authors were carrying out research into A.I. methods of modelling the prognosis of colorectal cancer patients so a comparison and evaluation between neural networks and Cox's regression, was sought [5].

We have evolved a formula to calculate outputs of an individual point estimate for each patient, using Cox's regression as a basis for handling both uncensored and censored data, to give outcomes of survival times.

2 THE DATA SET

For purposes of illustration of the methodology, rather than any particularly good prognostic model, we used a local database of clinico-pathological attributes on 216 colorectal cancer patients, which contained details of both uncensored and censored patients. The uncensored patients had a time of death noted in intervals of 1 month, up to 60 months, and censored patients had only records of attendance at clinics after 60 months, which gave a minimum survival. The data collection instrument contained questions of patients demographic details as well as pathological co-variables, such as polarity; tubule configuration; tumour pattern; pathological type; lymphocytic infiltration; fibrosis; venous invasion; mitotic count; penetration; differentiation; Dukes stage; obstruction and site.

However, the database could be any large validated data set for any disease process that results in an "event".

3 COX'S REGRESSION

Application of Cox's regression, in SPSS [6] to the multiple covariates has produced a parsimonious model for the hazard, or death rate, $h(t)$, with significant variables of Dukes stage, patient age and fibrosis category, transforming

* Faculty of Informatics, University of Ulster at Jordanstown, Newtownabbey, Co. Antrim, N. Ireland. BT37 0QT
Email: {ae.smith,ss.anand}@ulst.ac.uk

a hazard baseline, where the covariates are set to zero, which changes with time:-

$$h(t) = [h_0(t)] * \exp \sum_{i=1}^n (B_i X_i) \quad (1)$$

where n is the number of explanatory variables, B_is are the partial regression co-efficients, X_i's the values of the covariates for each patient and h₀(t) is the hazard baseline. It is possible to obtain a survival baseline s₀(t) in discrete time intervals, accounting for censorship, and a survival function s(t):-

$$s(t) = s_0(t) \exp \sum_{i=1}^n (B_i X_i) \quad (2)$$

This equation gives a probability of survival at each time interval. Note that the exponential is as before, but is the power term rather than a multiplicative factor. In general, Cox's regression has more commonly been used for comparing hazards to survival in two populations, e.g. patients undergoing different treatment modalities. Cox's regression does not give an output of a direct point survival estimate for each patient, but a survival function in discrete time intervals of h(t). An exact 50% (or a median) chance of survival, is not possible as an output other than by manually reading off an individual survival curve at the approximate half-life. This is because the non-parametric estimates of s(t) are step functions. This does not lend itself to direct comparability with other systems where a time in months, years etc., as an **automated output** from a system where many cases are involved, is required.

4 THE TRANSFORMATION

The aim was to transform this survival function into a formula for a survival estimate, the time term \hat{t} , for comparison with other approaches, such as neural networks (NN) which give a point estimate as an output from the system, for each patient in the data set.

Cox' regression makes two main assumptions about the hazard function. Firstly, it assumes that the covariates are independent. Secondly, it assumes proportionality of the hazard covariates with time. The output is a step function in time, which gives the hazard ratio per time interval for all patients with these particular covariates. This does not predict survival (time to event) for individual patients. Thus, we propose a method for providing a point estimate of survival for individual patients given a particular threshold probability of survival. This method, rather than assuming a particular form for the baseline survival, fits a curve to the discrete data points of survival. In the simplest form, the baseline could be linear, however, non-linear baselines may also be investigated in the future. The important issue here is ensuring a "good" fit.

4.1 The hypothesis

If the discrete survival baseline, s₀(t), for uncensored patients, can be shown to be linear by curve-fitting, then it is possible to suggest that $s_0(t) = mt+1$, where **m** is the slope of the fitted line and **t** is the time variable, and the constant is the probability of survival when t = 0. Note that for S₀(t) to be a survival baseline, t ∈ [0, 1/|m|]

Assuming linearity of S₀(t) and using equation 2 with S(t) = p. In the general case where p is the probability of event, of the survival distribution, we get:-

$$p = (mt + 1) \exp \sum_{i=1}^n (B_i X_i) \quad (3)$$

Performing simple algebraic transformations, and taking the logarithm of both sides and extracting t, we now are able to provide a formula for the point estimate of t, as shown below.

$$\hat{t} = [\exp(\exp(-\sum_{i=1}^n B_i X_i) * \ln p) - 1] / m \quad (4)$$

Note that this formula (Point Cox) for \hat{t} is only valid if it can be demonstrated that the baseline is linear by empirical testing.

5 AN EVALUATION OF THE APPLICATION

In this section we used the cancer data set as a validation data set for the method proposed above. Firstly, we selected all uncensored patients, where we know the actual survival, in one month intervals up to 60 months, in order to do direct comparisons of time to event.

Here, the important result is to show linearity of the survival baseline and compare the individual patient survival from our method with those of a NN. The overall accuracy of the prognostic model is not important as it is not an accurate measure for the evaluation of the suitability of the covariates for modelling the survival. The overall accuracy is heavily dependent on the quality of the data. This is an evaluation of the comparability of the paradigms.

The first stage of evaluating the method proposed is to show that the survival baseline is linear. For validating the results obtained, 10-fold cross validation was used. Curve fitting to fit the survival baseline of all the patient data, in each of the cross-folds, to a linear curve gave a mean **R**² (goodness-of-fit measure of a linear model) of 0.989. Clearly, the cross-validation of this curve fit implies that the survival baseline for the cancer data set is sufficiently close to linearity. The slope of this was estimated as -5.30704E-05 for this model. The survival baseline for the data set as a whole is given in Figure 1.

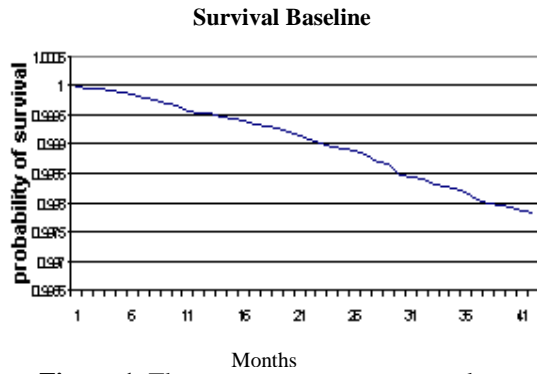


Figure 1. The survival baseline for the data

Note that this is the survival baseline, not the distribution of actual survival of patients. The model's survival baseline slope along with the individual patient's multiple covariates must be used for arriving at that individual patient's survival point estimate.

Having confirmed linearity of the survival baseline, the Point Cox method was then applied to each patient's attributes. Individual predictions for survival of each patient were then generated. Table 1 gives some summary statistics of predicted survivals and actual survivals for those within the uncensored range for which direct comparisons were possible. The actual survivals have a distribution (in months) of mean 18.4, median 15.0, s.d. 13.4, range of 57.0 and skewness of 0.785. The neural net used in this study is part of Clementine Data Mining tool [7]. The learning algorithm used was a variant of the back propagation network, that attempts to optimise the topology of the neural net by starting with a large network and iteratively pruning out nodes in input and output layers.

Table 1 Summary statistics for predicted survivals for the uncensored patients, as output by the two methodologies

	N	Mean	Median	S.D.	Range	Skewness
Point Cox	105	22.8	20.3	13.8	55.6	0.945
NN	105	19.1	17.0	10.1	49.0	0.733

A Wilcoxon's signed rank test comparing directly the Point Cox and the neural network outputs indicated no really significant differences between the two approaches ($p=0.052$).

A histogram of the actual survivals recorded, the estimates from the Point Cox method and a neural network are given in Figure 2 for illustration and clarification of the above.

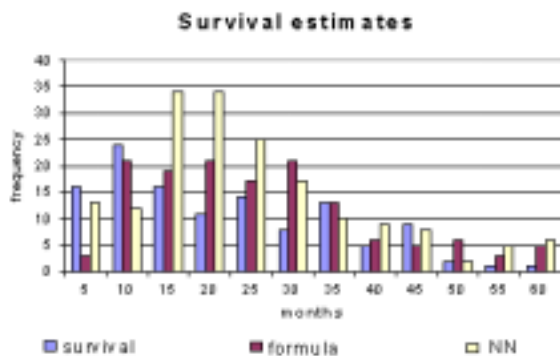


Figure 2. Histogram of survival estimation for the point Cox method, NN output and the actual survivals (uncensored patients)

This particular clinico-pathological model had regression coefficients of Dukes stage (0.67840), Age (0.03396), Fibrosis category (0.21160). When command syntax was written within SPSS and the Point Cox method was applied to these regression co-efficients and patient covariate values, a survival estimation for each patient was generated. As an example, the distribution output from Cox's for a 54 year old patient whose pathological attributes had the highest grade of Dukes staging (6) and the highest Fibrosis category (6) is shown in Figure 3. The half-life of this is about 13 months, the same value produced from the Point Cox method for the point estimate \hat{t} .

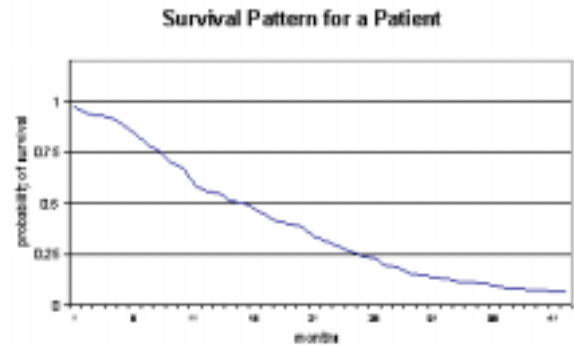


Figure 3. Survival distribution with PointCox method for a patient with covariates as described in text.

The overall results of applying the Point Cox method, within the colorectal cancer data set are given in the first row in Table 2 :-

Table 2 Comparing the actual survivals with both Point Cox method and a NN, for uncensored patients.

Model	Mean survival-predicted	95% C.I.	Wilcoxon
Point Cox	15.9 months	(7.0, 24.8) months	$p = 0.021$
NN	14.3 months	(7.2, 21.4) months	$p = 0.020$

The second column is the mean absolute error between the predicted survival and the actual survival in months for those uncensored patients for whom an actual survival is recorded. The mean actual survival is 18.4 months. It indicates the magnitude of the error in prediction. The Wilcoxon signed ranks test gives a measure of the probability that there is no difference between the observed and predicted survival distributions for the paired data. The results from this test indicate that there are no significant differences in both the Point Cox method and the neural net, when compared with actual survivals.

The errors of the estimates are large, as indicated by the confidence interval (C.I.) but this model only contained limited attributes and no indication of the treatment that the patient received, e.g. surgery, radiotherapy or chemotherapy, was included. We are aware that this particular model is not sufficient in itself for any prediction of survival in a patient, but we would expect the inclusion of information on such factors, if available, to improve the model and reduce the errors.

In the case of the censored patients, the observed survival values are a minimal estimate since they are only attendances recorded at a clinic. The degree of overestimate however, in this situation is valid, but cannot be used for assessment of the performance, as it is not possible to get an exact measure of

the error. The underestimate can be stated in minimal errors, for example, if a censored patient has a minimum survival recorded of 100 months and the predicted survival is 45 months, it can be assessed as an error in prediction of at least 55 months.

Table 3 gives a comparison of the minimal errors given by the Point Cox estimate and a neural network to see if the differences in the paradigms applied to censored patients are significant.

Table 3 T test of results when applying the Point Cox method compared with a neural network, for censored patients.

Model	% Underestimate	Mean error	T value	d.f.	Sig
Point Cox	63.2%	46.8	0.333	9	0.749
NN	93.5%	45.0			

This comparison indicates no significant difference between the two paradigms.

6 DISCUSSION

This paper presents an original technique, for deriving point estimates from Cox' Regression, aimed at enabling the evaluation of other methods e.g. regression trees or NN against the standard technique used in the medical domain. The authors believe that the presented Point Cox method can be used as a simple approach and incorporated into Cox's Regression to obtain point survival estimates for individual patients in any disease process which results in an event, and contains censored patients. This possibly has wide applicability for the prognosis of new patients and can certainly be used for direct comparisons with, and evaluation of, other applications and systems, such as emerging intelligent systems, where many cases and many attributes demand an automated output.

Alternative parametric methods may be employed, eg. the Breslow [8] approach, or other methods of smoothing of survival curves to give a median estimate, Collett [9]. However, we believe that the current method is more easily integrated into programs and understandable to non-statisticians, when using the Cox's model for multivariate analysis.

We believe that the empirically defined linear function allows a practicable simulation of the continuous function that exists in reality with patients dying. Once a model has been derived, from any similar database, with co-efficients for the attributes and slope of the baseline defined, the significant attribute values for each new or current patient can be inserted into the Point Cox to obtain a point estimate. The attributes are defined in the model set up by the user and can be attributes shown to be hazardous for any disease process which results in an "event", such as death. By changing the probability of survival from 0.5 in the formula, representing 50% chance, statements can also be made as to 20%, 80% etc. chance that the patient will live to a certain time. These statements are, of course constrained by the inherent limitations of the model of attributes applied, plus the uncertainty of patient variability. However, this paradigm may be useful for the planning of treatment profiles.

This Point Cox, derived from Cox' regression, is novel in its attempt to overcome the failure of Cox's regression so far to provide point estimates as a direct output. It, however, depends heavily on the assumptions of linearity of the survival baseline produced for both uncensored and censored data up to

censoring time, using the regression co-efficients for the attributes hazardous to survival. However, it may be possible to include non-linear functions if curve fitting can show that the survival baseline can be approximated adequately by a specific function.

The analysis of more survival data sets for observing the extent to which linear baselines can be expected, so that this current constraint can be lifted, is left as a future research goal.

ACKNOWLEDGEMENTS

We thank Dr. Peter Hamilton and others at the Dept. of Pathology, Royal Victoria Hospital, Belfast, U.K. for the use of the database on colorectal cancer patients.

We also thank Ms Adele Marshall for expert help on survival analysis.

We acknowledge the Grant, in the form of a Fellowship, from the Medical Research Council, which has enabled continuation of this work.

REFERENCES

- [1] M Buyse, P Piedbois. Comment. *Lancet*, **350 (9085)**, 1175-1176, (1997)
- [2] DR Cox. Regression models and life tables. *J R Stat Soc*, **34**, 187-220 (1972)
- [3] L Bottaci, PJ Drew, JE Hartley, MB Hadfield, R Farouk, PWR Lee, et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet*, **350 (9076)**, 469-472, (1997)
- [4] D Farragi, R Simon. A neural network model for survival data. *Stats Med*, **350**, 72-82, (1995)
- [5] S S Anand, A E Smith, P W Hamilton, J S Anand, J G Hughes, P Bartel. An evaluation of Intelligent prognostic systems for colorectal cancer. *J Art Int Med*, **15(2)**, 193-213, (1999).
- [6] SPSS for Windows, Release **9.0**, SPSS inc. Chicago. 1996
- [7] Clementine Data Mining, *User Manual*. Integral Solutions Ltd. 1997
- [8] P Breslow. Covariance analysis of censored survival data. *Biometrika* **57**, 579-594, (1974)
- [9] D Collett. *Modelling survival data in medical research*. Texts in Statistical Science. Chapman & Hill, London. 1994