

BOOTSTRAP INVESTIGATION OF THE STABILITY OF A COX REGRESSION MODEL

DOUGLAS G. ALTMAN

*Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, Lincoln's Inn Fields,
London WC2A 3PX, U.K.*

AND

PER KRAGH ANDERSEN

Statistical Research Unit, Blegdamsvej 3, 2200 Copenhagen N, Denmark

SUMMARY

We describe a bootstrap investigation of the stability of a Cox proportional hazards regression model resulting from the analysis of a clinical trial of azathioprine versus placebo in patients with primary biliary cirrhosis. We have considered stability to refer both to the choice of variables included in the model and, more importantly, to the predictive ability of the model. In stepwise Cox regression analyses of 100 bootstrap samples using 17 candidate variables, the most frequently selected variables were those selected in the original analysis, and no other important variable was identified. Thus there was no reason to doubt the model obtained in the original analysis. For each patient in the trial, bootstrap confidence intervals were constructed for the estimated probability of surviving two years. It is shown graphically that these intervals are markedly wider than those obtained from the original model.

KEY WORDS Cox proportional hazards regression model Bootstrap Model selection Prediction
Primary biliary cirrhosis

INTRODUCTION

It is often necessary to estimate a regression relation between an outcome variable and a set of explanatory variables while at the same time selecting the variables for inclusion. The model is usually selected and evaluated using the same set of data.

It is well known that when variable selection and model fitting are carried out on the same set of data, the estimated predictive ability of the model will be too high. With regression the estimated coefficients will be biased away from zero by the criteria for inclusion of variables in the model.^{1, 2}

It is also well known that the choice of variables for inclusion in the regression model will vary across repeated samples or sub samples. One reason for this is the interrelationship (correlation) between the explanatory variables. However, it does not necessarily follow that models with different selections of explanatory variables are very different in their ability to predict the outcome variable.

There has been relatively little exploration of the practical implications of these effects. Miller² suggested several techniques for investigating the bias in regression coefficients, but not all are suitable for exploring the stability of a regression model. We have considered stability to refer both

to the choice of variables included in the model and, more importantly, to the predictive ability of the model.

One suitable method is data splitting, in which a sample is randomly split into two subsamples which are used independently for model selection and evaluation.³ Miller² rightly observes that sample sizes are usually insufficient to make this approach practical. Another is the jack-knife method in which the same analysis is performed on a large number of subsets of $n-r$ out of n observations, where r/n is small; Miller suggests $r/n = 0.1$. The closely related bootstrap method (see Diaconis and Efron⁴ for a readable introduction; Efron and Gong⁵ and Efron and Tibshirani⁶ for a more detailed account; and Efron⁷ for theoretical development) may be considered an elaboration of the jack-knife.^{8, 9}

Chen and George¹⁰ used the bootstrap to investigate the choice of variables in a Cox regression model, but they did not consider the prognostic implications for individual patients. Gong¹¹ presented a bootstrap analysis investigating variation in the selection of variables in a stepwise logistic regression model.

In this paper we describe a bootstrap investigation of the stability of the Cox proportional hazards regression model arising from the analysis of a clinical trial of patients with primary biliary cirrhosis (PBC). This trial is of particular interest as it showed a statistically significant ($P < 0.05$) benefit of active treatment only after adjustment for imbalance in a covariate.^{12, 13}

DATA

The data come from a multinational double-blind randomized clinical trial of azathioprine versus placebo in 248 patients with PBC. Recruitment of patients covered six years and follow-up a further six years. Details of this trial together with the results of Cox proportional hazards regression analysis were given by Christensen *et al.*¹² A final model comprising six variables was obtained, partly using forward stepwise selection from a set of 25 variables found to have significant marginal associations ($P < 0.05$) with survival. No variable was found to interact with therapy at $P = 0.05$. The final model was refitted to the data from the 216 patients who had complete data for those six variables. This model fitted to the original data we will call M_0 . The six variables in M_0 are shown in Table I. The transformations of age and serum bilirubin were chosen after careful examination of log cumulative hazard functions for several strata. The transformation of age is referred to as $\exp(\text{age})$ in the rest of this paper.

Of the 216 cases, 103 had died and 113 had censored survival times. As well as the six variables in M_0 a further 11 binary variables, which had shown strong marginal effects on survival, were included in the bootstrap investigation. One other such variable, the presence or absence of ascites, had to be excluded as there were too few patients with ascites to obtain reliable results, not because it was prognostically unimportant. The data for these 17 explanatory variables from the 216 patients, together with the survival time and a death/censoring indicator, were used in the bootstrap investigations.

METHODS

The Cox model

The Cox proportional hazards regression model¹⁴ has the form

$$\lambda(t) = \lambda_0(t) \exp(\sum \beta_k z_k) \quad (t > 0)$$

where $\lambda(t)$ is the hazard (instantaneous risk of death) at time t , $\lambda_0(t)$ is the 'baseline' hazard (when the covariates z_k are all zero), and the β_k ($k = 1, \dots, p$) are regression coefficients. For any set of covariate values from either a specific or a hypothetical patient, a prognostic index (PI) may be calculated which is defined by

$$PI = \sum_{k=1}^p \hat{\beta}_k z_k,$$

where $\hat{\beta}_1, \dots, \hat{\beta}_k$ are the maximum partial likelihood estimates¹⁵ of the regression coefficients. Denoting by $\hat{\Lambda}_0(t)$ the estimated cumulative baseline hazard at time t the survival function can be estimated by

$$\hat{S}(t) = \exp[-\hat{\Lambda}(t)],$$

where

$$\hat{\Lambda}(t) = \hat{\Lambda}_0(t) \exp(PI)$$

is the estimated cumulative hazard at time t . The log cumulative hazard, given by

$$\log[\hat{\Lambda}(t)] = \log[\hat{\Lambda}_0(t)] + PI,$$

is thus linearly related to PI , and can be used to construct a confidence interval for $\Lambda(t)$ and thus for $S(t)$.

Bootstrap samples

The basic idea of the bootstrap technique in its simplest form is that if independent identically distributed (i.i.d.) observations X_1, \dots, X_n are available, then the variability of estimated characteristics of the distribution of the X s can be assessed by studying the variability of the estimate across a large number B of *bootstrap samples*. The bootstrap samples are obtained by taking samples of size n from the original data X_1, \dots, X_n using random sampling *with replacement*. The theoretical justifications for the analysis of such samples are discussed by Efron and Gong⁵ and Efron.⁷⁻⁹ In essence, the set of bootstrap samples is treated as B trials and the variability between the results from identical analyses of all the samples is investigated.

Hjort¹⁶ studied several bootstrap sampling schemes for the Cox regression model, including that used by Chen and George.¹⁰ In the present study, 100 samples were generated as follows. The j th patient's data X_j comprised an observed survival time, a death/censoring indicator and a set of 17 explanatory variables. Each bootstrap sample was obtained by randomly sampling 216 times from the data sets $\{X_j; j = 1, \dots, 216\}$ *with replacement*, using the random number generator of Wichmann and Hill.¹⁷ The method is justified if the sets of explanatory variables are i.i.d. and if survival and censoring times are conditionally independent given the explanatory variables, since X_1, \dots, X_{216} will then be i.i.d. if, for example, the censoring intensity also has a Cox-type form.

Fixed covariates

A Cox regression model with the same six variables that were included in M_0 was fitted to each bootstrap sample. These models are denoted M_{fix}^i ($i = 1, \dots, 100$) to indicate that the choice of variables was fixed in advance. If $\hat{\beta}_{ik}^*$ is the estimated regression coefficient for the k th variable in model M_{fix}^i fitted to the i th bootstrap sample (where $*$ indicates a bootstrap estimate), then bootstrap estimates of the regression coefficient and its standard error are simply the mean and standard deviation of $\hat{\beta}_{ik}^*$ ($i = 1, \dots, 100$), say $\hat{\beta}_k^*$ and $\text{SE}(\hat{\beta}_k^*)$. Ninety per cent confidence limits ('percentile limits') for β_k were obtained empirically from the cumulative distribution function of the $\hat{\beta}_{ik}^*$ s.^{7, 9}

Stepwise regression

Each of the 100 bootstrap samples was also analysed by stepwise selection as for the original data set, except that the number of potential explanatory variables was only 17. Forward stepwise regression on the original data set for these 17 variables does produce M_0 , although in practice M_0 was obtained by a more complex route, notably including investigation of the appropriate scaling for continuous variables. It was necessary to standardize the analysis of the 100 bootstrap samples, so forward stepwise Cox regression was performed using the program BMDP2L.¹⁸ Variables were entered into the model if significant at the 5 per cent level based on a maximum partial likelihood ratio test. At each step, variables in the model could be dropped if they were no longer significant at the 5 per cent level. The resulting model for the i th bootstrap sample we call M_{step}^i ($i = 1, \dots, 100$). There were five patients with missing values for at least one of the 11 variables not in M_0 , the number obviously varying among the bootstrap samples. The models M_{step}^i were thus based on analysis of slightly fewer than 216 patients.

Comparing models

It is natural to compare the three models M_0 , M_{fix}^i and M_{step}^i on each bootstrap sample. It is not possible to compare the goodness of fit of these models directly, because there are different numbers of cases (due to differential replication of missing values) and different numbers of variables in the models. However, it is possible to consider the results of each analysis in relation to some directly relevant information for individual patients.

Christensen *et al.*¹² considered PI in relation to the estimated median survival time and the probability of surviving two, five or eight years. In the present analysis the survival function at 24 months, $\hat{S}(24)$, was calculated for each of the 216 patients in the original data set using M_0 and M_{fix}^i and M_{step}^i for all 100 bootstrap samples. Confidence intervals for $\hat{S}(24)$ for each patient were compared using (a) the estimated asymptotic variance given by Altman and Andersen¹⁹ based on that of Tsiatis;²⁰ (b) 90 per cent percentile bootstrap confidence limits based on M_{fix}^i ; and (c) the same method for M_{step}^i . When estimates based on M_{step}^i were calculated for all 216 original covariate patterns it was necessary to replace a few missing values. For continuous variables the mean value was used, and binary variables were set to the more common value. The widths of the three sets of confidence intervals were compared on the log cumulative hazard scale where they are symmetric (see Methods and Appendix in Christensen *et al.*¹²); we denote these by CI_0 , CI_{fix} and CI_{step} , respectively.

All subsets regression

For comparison with the results of forward stepwise model selection, 'all possible subsets' regression was carried out. The same 17 variables were used, for which 211 patients had complete data. The method was that of Kuk²¹ for censored survival data, using the program BMDP9R²² on a matrix derived from the covariance matrix of the parameter estimates for the full model. The best model was taken as that with the lowest value of Mallows's C_p .²³

RESULTS

Models with fixed covariates

Table I shows the regression coefficients and their asymptotic standard errors and 90 per cent confidence intervals for the model M_0 applied to the original data. Table II shows comparable

Table I. Clinical trial of azathioprine and placebo in primary biliary cirrhosis: variables in final Cox regression model M_0 in order of entry in stepwise analysis

Step	Variable (k)	Scoring	$\hat{\beta}_k$	Regression coefficient 90% confidence interval		
				(SE)	5	95
1	\log_{10} (serum bilirubin)	$\mu\text{mol/litre}$	2.51	(0.32)	1.98	3.04
2	$\exp(\text{age})$	$(\text{age in years} - 20)/10$	0.0069	(0.0016)	0.0043	0.0095
3	Cirrhosis	absent: 0 present: 1	0.88	(0.22)	0.52	1.24
4	Therapy	azathioprine: 0 placebo: 1	0.52	(0.21)	0.17	0.87
5	Serum albumin	g/litre	-0.050	(0.018)	-0.080	-0.020
6	Central cholestasis	absent: 0 present: 1	0.68	(0.27)	0.24	1.12

 Table II. Bootstrap estimates of regression coefficients and standard errors based on M_{fix}^i ($i = 1, \dots, 100$), and bootstrap median and 90 per cent percentile confidence limits

Variable	Regression coefficient		Median and 90% confidence limits		
	$\hat{\beta}^*$	(SE)	5	50	95
\log_{10} (serum bilirubin)	2.58	(0.27)	2.13	2.57	3.06
$\exp(\text{age})$	0.0073	(0.0019)	0.0047	0.0073	0.0106
Cirrhosis	0.93	(0.26)	0.51	0.91	1.40
Therapy	0.51	(0.21)	0.19	0.51	0.81
Albumin	-0.052	(0.017)	-0.078	-0.052	-0.020
Central cholestasis	0.71	(0.25)	0.34	0.72	1.14

bootstrap estimates derived from the models M_{fix}^i from the 100 bootstrap samples, with the median and 90 per cent bootstrap confidence limits. It can be seen that there is close agreement between M_0 and the bootstrap results, both for the estimated regression coefficients and their standard errors. The median and mean of the $\hat{\beta}_{ik}^*$ were very similar.

For each variable the distribution of the 100 estimated regression coefficients was examined using the Shapiro-Wilk W statistic.²⁴ After excluding one outlier for $\exp(\text{age})$, none of the distributions was significantly different from normal ($P > 0.3$).

Models derived from stepwise regression

Table III shows how often each of the 17 variables appeared in the 100 models M_{step}^i . $\text{Log}(\text{serum bilirubin})$ was included in all 100 models, and was always the first variable selected. $\text{Exp}(\text{age})$ was included in all but three of the bootstrap models. The other four variables in M_0 were present in

Table III. Frequency of selection of variables in stepwise regression on 100 bootstrap samples

Variable	Selection frequency
\log_{10} (serum bilirubin)	100
exp(age)	97
Cirrhosis	74
Therapy	68
Albumin	68
Central cholestasis	62
Pigmentation	41
Gastrointestinal haemorrhage	28
Fibrosis	19
Peripheral cholestasis	16
MITO antibodies	14
Lymphoid follicles	14
Jaundice	11
Enlarged liver	11
Enlarged spleen	10
Incapacitation index > 25	8
Xanthoma	6

Table IV. Distribution of the number of variables in M_{step}^i and the number of variables included from model M_0

Number of variables in M_{step}^i	Number of variables in M_0 included in M_{step}^i					Total
	2	3	4	5	6	
4	1	4	2			7
5		2	8	6		16
6		3	8	16	1	28
7		1	7	13	8	29
8			1	8	5	14
9				2	2	4
10					2	2
Total	1	10	26	45	18	100

about two-thirds of the models, and these six variables were selected much more frequently than any of the other 11 variables. Of these, only pigmentation appeared in more than a third of the models. It is interesting that all 17 variables appeared in some models. The order of entry of the 6 variables into M_0 , as shown in Table I, is almost exactly reflected in the frequencies shown in Table III.

Of the 100 models only one comprised the same 6 variables as M_0 , although a further 17 included those 6 variables. Table IV shows the distribution of the number of variables in M_{step}^i and the number of variables from M_0 included in these models.

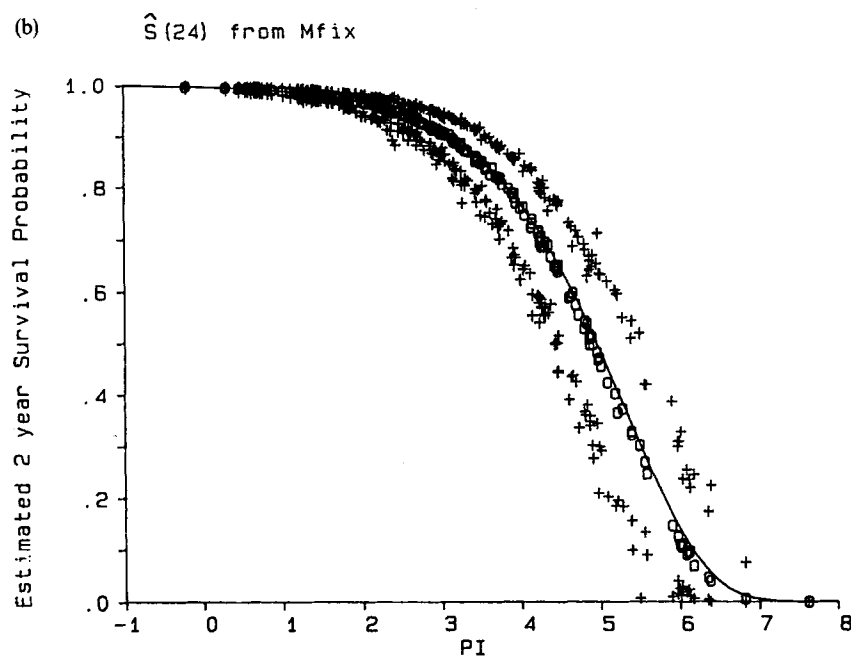
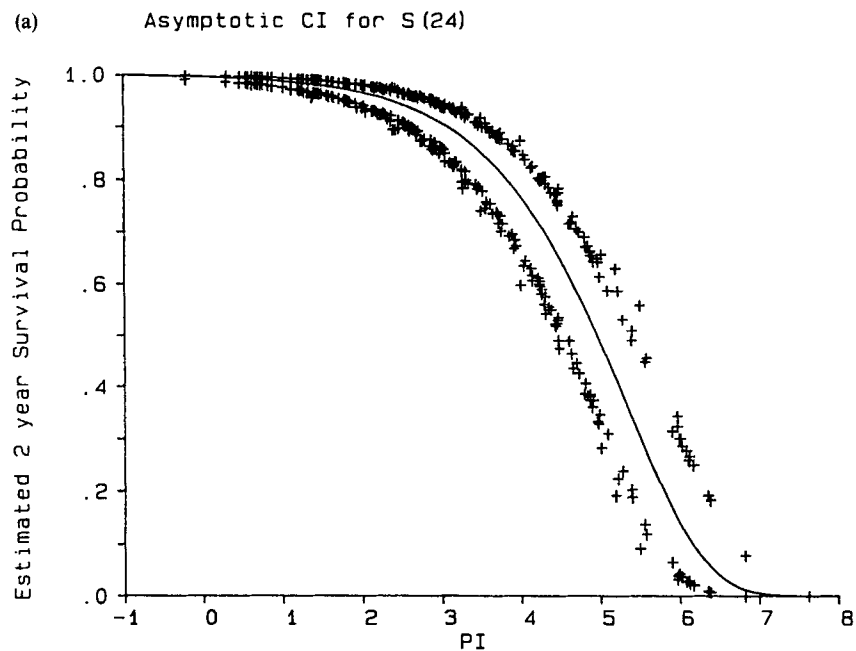


Fig. 1 (a, b)

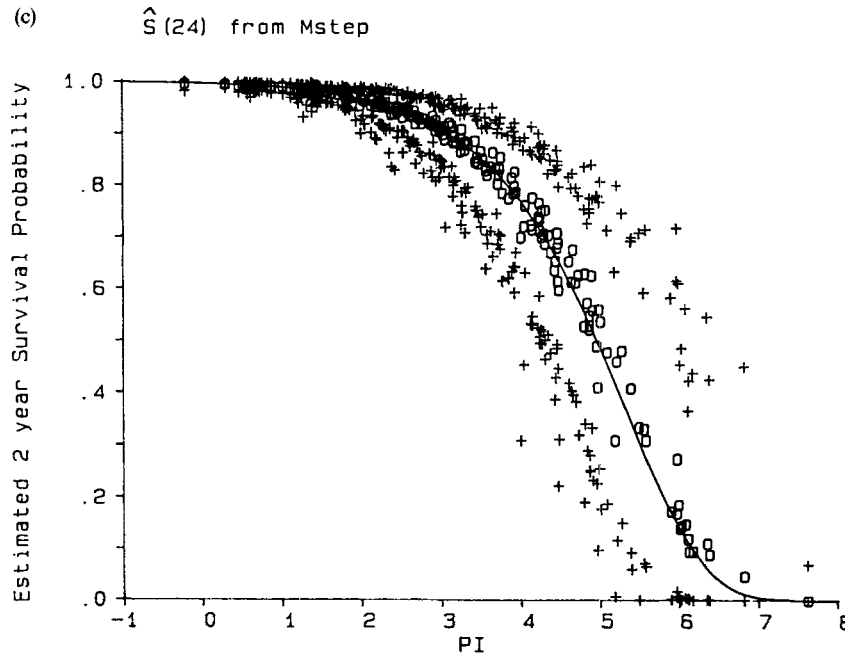


Figure 1. (a) Ninety per cent confidence intervals for the estimated probability of surviving two years for each of the 216 patients based on M_0 . The asymptotic variance was used for the confidence intervals (CI_0). The solid line shows $\hat{S}(24)$ from M_0 . The abscissa PI is the prognostic index based on M_0 (b) Ninety per cent percentile bootstrap confidence intervals for the probability of surviving two years for each of the 216 patients based on M_{fix}^i (CI_{fix}). The solid line shows $\hat{S}(24)$ from M_0 and the circles show the bootstrap estimates of $\hat{S}(24)$. The abscissa PI is the prognostic index based on M_0 (c) Ninety per cent percentile bootstrap confidence intervals for the probability of surviving two years for each of the 216 patients based on M_{step}^i (CI_{step}). The solid line shows $\hat{S}(24)$ from M_0 and the circles show the bootstrap estimates of $\hat{S}(24)$. The abscissa PI is the prognostic index based on M_0

Comparison of predictions by different models

For each of the 216 covariate patterns in the original data set, confidence intervals for the survival probability were calculated in three ways. Figure 1(a) shows 90 per cent confidence intervals for the estimated probability of surviving two years ($\hat{S}(24)$) using the asymptotic formula for the variance, and Figures 1(b) and 1(c) show corresponding 90 per cent bootstrap percentile confidence intervals based on M_{fix}^i and M_{step}^i respectively. The individual point estimates of $\hat{S}(24)$ based on M_{fix}^i and M_{step}^i are both very similar to the asymptotic results for M_0 . However, the confidence intervals are very different, as illustrated further in Figure 2, which shows the widths of the confidence intervals CI_0 and CI_{step} in relation to CI_{fix} for $\log[\Lambda(24)]$. Whereas CI_0 and CI_{fix} are very similar, CI_{step} is consistently larger (by about 60 per cent) than CI_{fix} .

All subsets regression

All subsets regression showed that the best model was that containing the first seven variables in Table III, that is M_0 plus pigmentation. However, this model had only a marginally lower value of C_p than M_0 (1.44 compared with 1.54).

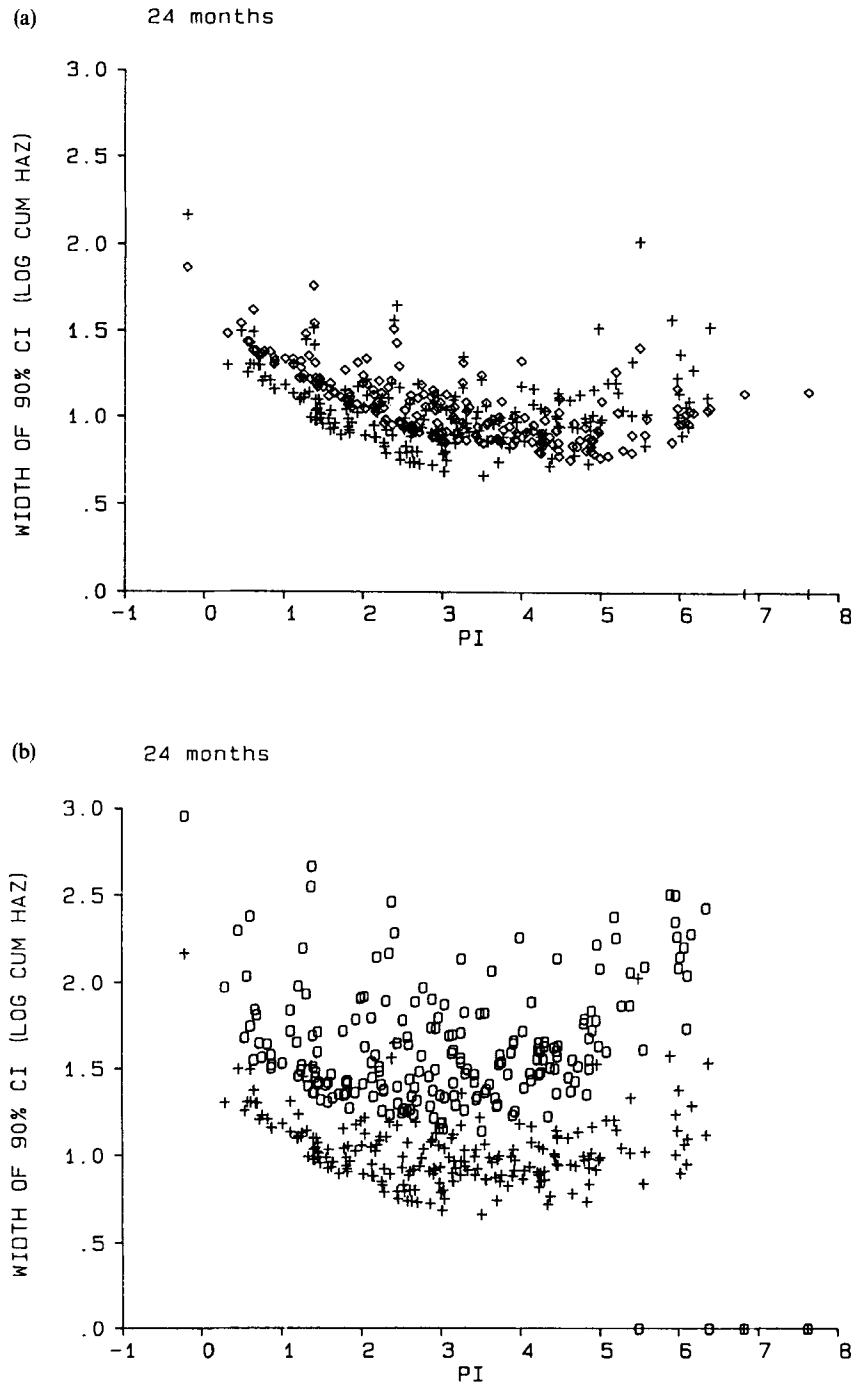


Figure 2. (a) Comparison of the widths of CI_0 (\diamond) and CI_{fix} ($+$) for $\log[\Lambda(24)]$ for each of the 216 patients. (The width is shown as zero when both estimated survival probabilities were less than 0.0005.) The abscissa PI is the prognostic index based on M_0 . (b) Comparison of the widths of CI_{fix} ($+$) and CI_{step} (\circ) for $\log[\Lambda(24)]$ for each of the 216 patients. (The width is shown as zero when both estimated survival probabilities were less than 0.0005.) The abscissa PI is the prognostic index based on M_0 .

DISCUSSION

The interpretation of the frequency of selection of different variables in stepwise regression analyses of 100 bootstrap samples is not obvious. In this study the six variables included in the original model M_0 were selected most frequently (Table III), but two variables not in M_0 were entered in over a quarter of the models. Our results are broadly similar to those of Chen and George,¹⁰ who studied survival of 224 children with acute lymphocytic leukaemia. In their study stepwise regression yielded a model containing six of the nine original variables, one of which was only marginally significant and highly correlated with another entered variable. Stepwise regression of 100 bootstrap samples showed that five of their originally included variables were entered in over half of the models (64 to 82 per cent) while the suspect sixth variable appeared in only 33 per cent of the models. One of the originally excluded variables was included in 45 per cent of the models. The bootstrap analysis was performed to help selection of a final model by (arbitrarily) entering those variables selected by more than 50 per cent of the analyses of bootstrap data sets, that is the original model minus the suspect variable. By contrast, our bootstrap analysis was performed to confirm (or refute) a model established by a previous detailed analysis. Application of the bootstrap approach of Chen and George to our data would have arrived at model M_0 .

A contrasting example is the stepwise logistic regression analysis of survival of 155 patients with acute chronic hepatitis.²⁵ Nineteen variables were included in the analysis of 500 bootstrap data sets. The four variables chosen by the analysis of the original data were selected in 59, 48, 37 and 35 per cent of the bootstrap analyses. Diaconis and Efron⁴ commented that 'the variables identified in the original analysis cannot be taken very seriously' but they did not discuss the ability of the models to predict the outcome variables. One reason for the uncertainty associated with their best model is that there were only 33 deaths among the 155 patients.

How predictable is the ordering of frequency of variables in the 100 bootstrap models? Table V shows that the order of entry of variables into the full model using forward stepwise regression is very similar to the rank order of the frequency of inclusion in M_{step}^i . We might expect that with a larger number of bootstrap samples we would have seen closer agreement. The agreement with the significance of each variable in the full model was less good, but as expected there is no relation with the marginal significance of each variable. In particular, therapy was the least important variable when considered on its own, while peripheral cholestasis, not in M_0 , was second only to serum bilirubin. In the study of Chen and George¹⁰ the bootstrap frequency of inclusion was similar to both the order of entry into the full model and the order of marginal significance (Table VI). For their variable E+, the bootstrap frequency reflected its order of entry into the full model.

Good agreement was observed between the estimated regression coefficients and the corresponding 90 per cent confidence intervals based on M_0 and on the M_{fix}^i ($i = 1, \dots, 100$) (Tables I and II).

The predictions based on M_{fix}^i in the form of estimated survival probabilities and 90 per cent confidence intervals agreed very well with M_0 (Figures 1(a), 1(b), and 2(a)). Due to the firm theoretical basis for inference from bootstrap samples when B is large (see for example, Hjort¹⁶), we take these findings as an argument for the validity of the large-sample approximations^{20, 26} in our sample of 216 patients and 103 deaths. This interpretation agrees with the conclusions from the simulation study by Lee *et al.*²⁷ concerning size and power of test statistics in the Cox regression model. Also the estimated survival probabilities for the individual subjects based on M_{step}^i agreed well with M_0 (Figure 1(c)). The fact that the predictions from M_{fix}^i and M_{step}^i were so similar shows that for the purpose of giving a prognosis the choice of variables in a multiple

Table V. Rank order of 17 variables according to (a) frequency of inclusion in M_{step}^i (b) order of entering the full model using forward stepwise regression (c) $\hat{\beta}/\text{SE}(\hat{\beta})$ in full model (d) marginal significance (univariate analyses)
(columns (c) and (d) are both based on analysis of 211 cases with full data)

Variable	(a)	(b)	(c)	(d)
$\log_{10}(\text{serum bilirubin})$	1	1	1	1
$\exp(\text{age})$	2	2	2	10
Cirrhosis	3	3	6	5
Therapy	4.5	4	3	17
Albumin	4.5	5	5	6
Central cholestasis	6	6	4	3
Pigmentation	7	7	7	7
Gastrointestinal haemorrhage	8	10	10	13
Fibrosis	9	8	8	9
Peripheral cholestasis	10	12	11	2
MITO antibodies	11.5	11	12	16
Lymphoid follicles	11.5	14	14	12
Jaundice	13.5	9	9	4
Enlarged liver	13.5	13	13	15
Enlarged spleen	15	17	17	8
Incapacitation index > 25	16	15	15	11
Xanthoma	17	16	16	14

Table VI. Rank order of nine variables in the study of Chen and George¹⁰ according to (a) frequency of inclusion in bootstrap models (b) order of entry into full model (c) marginal significance (univariate analysis)

Variable	(a)	(b)	(c)
DI	1	1	1
MM	2	4	2
AGE	3	2	3
WBC	4	3	4
CG	5	5	5
RACE	6	*	6
SEX	7	*	8
CNS	8	*	9
E+	9	6	7

* Not given

regression model appears of limited importance. As we would expect, however, the confidence intervals based on the stepwise regression models were considerably wider than those based on the models with the variables fixed in advance. The asymptotic variance formula for the estimated survival probabilities seems to express very closely the uncertainty of the estimate when *the same regression model* is refitted to a large number of independent samples. However, it seems to give

smaller values than a stepwise (forward) selection procedure on a large number of independent samples. Thus, our results indicate a risk of underestimation of the uncertainty of predictions from a single multiple regression model.

Selection of a 'best' regression model from many candidate independent variables poses many difficulties, especially in the absence of an independent set of data for validation. There are several strategies available for choosing the best subset of variables, notably forward or backward stepwise selection and investigation of all subsets. Stepwise approaches are generally based on significance tests, where the arbitrary 5 per cent level usually taken is not a true significance level because multiple dependent tests are performed. All subsets regression is often interpreted using a criterion such as Mallows's C_p , which is based on unexplained variation but incorporates a penalty for adding each extra variable to the model. The magnitude of the penalty is also arbitrary. Kuk²¹ showed that forward stepwise and all subsets regression can give dramatically different results. However, in our study all subsets analysis showed only a slight difference from the forward stepwise solution. The results using stepwise selection to fit regression models to the 100 bootstrap data sets appeared to agree well with the model fitted to the original data (Table III), but there is little experience in interpreting such results.

Any form of data-dependent variable selection is likely to lead to overoptimistic goodness of fit; we expect a worse fit to a new set of data. It is possible to shrink the estimated regression estimates towards zero¹ to compensate for the bias, and one interesting possibility is to use the bootstrap frequencies (Table III) as shrinkage factors; we have not investigated this. However, it is unlikely that shrinkage of regression estimates will be used much in practice. It may be felt that little harm would be done by including extra variables in a regression model, but Flack and Chang²⁸ have shown that noise variables are often included in stepwise regression models when the sample is small and entry criteria lax, especially when the number of candidate variables is large relative to the number of observations.

Bootstrapping is an appealing method for evaluating a regression model, as it allows investigation of the consistency of the inclusion of each variable in the regression model. It is, however, expensive in computing time. Where possible the chosen regression model should be fitted to a new set of data, but again it is not clear how the results of the validation analysis should be interpreted. Fitting multiple regression models seems likely to remain problematic, in particular when all the candidate variables are treated completely symmetrically, as when using automatic model selection procedures. Usually the variables will *not* be equally important from a scientific point of view, a fact which should be reflected in the statistical analysis.

CONCLUSIONS

There is no reason to doubt the validity of the original model (M_0). Bootstrap analysis with fixed variable selection (M_{fix}) confirmed the large-sample approximations for Cox regression in relation to estimates of regression coefficients and survival probabilities. However, bootstrapping with stepwise variable selection (M_{step}) gave similar individual predictions but larger confidence intervals for estimated survival probabilities.

ACKNOWLEDGEMENTS

We thank Erik Christensen, Niels Keiding, Patrick Royston, Willi Sauerbrei and Martin Schumacher for helpful comments on earlier versions of this paper. We thank Olive Waldron for typing the manuscript.

REFERENCES

1. Copas, J. B. 'Regression, prediction and shrinkage (with discussion)', *Journal of the Royal Statistical Society B*, **45**, 311–354 (1983).
2. Miller, A. J. 'Selection of subsets of regression variables (with discussion)', *Journal of the Royal Statistical Society A*, **147**, 389–425 (1984).
3. Cox, D. R. 'A note on data-splitting for the evaluation of significance levels', *Biometrics*, **62**, 441–444 (1975).
4. Diaconis, P. and Efron, B. 'Computer-intensive methods in statistics', *Scientific American*, May, 96–108 (1983).
5. Efron, B. and Gong, G. 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *American Statistician*, **37**, 36–48 (1983).
6. Efron, B. and Tibshirani, R. 'Bootstrap methods for standard errors, confidence intervals, and other methods of statistical accuracy', *Statistical Science*, **1**, 54–77 (1986).
7. Efron, B. 'Nonparametric standard errors and confidence intervals', *Canadian Journal of Statistics*, **9**, 139–172 (1981).
8. Efron, B. 'Bootstrap methods: another look at the jackknife', *Annals of Statistics*, **7**, 1–26 (1979).
9. Efron, B. 'Censored data and the bootstrap', *Journal of the American Statistical Association*, **76**, 312–319 (1981).
10. Chen, C.-H. and George, S. L. 'The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model', *Statistics in Medicine*, **4**, 39–46 (1985).
11. Gong, G. 'Some ideas on using the bootstrap in assessing model variability', in K. W. Heiner, R. S. Sacher and J. W. Wilkinson (eds), *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, Springer, New York, 1982, 169–173.
12. Christensen, E., Neuberger, J., Crowe, J., Altman, D. G., Popper, H., Portmann, B., Doniach, D., Ranek, L., Tygstrup, N. and Williams, R. 'Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial', *Gastroenterology*, **89**, 1084–1091 (1985).
13. Altman, D. G. 'Comparability of randomised groups', *Statistician*, **34**, 125–136 (1985).
14. Cox, D. R. 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society B*, **34**, 187–220 (1972).
15. Cox, D. R. 'Partial likelihood', *Biometrika*, **62**, 269–276 (1975).
16. Hjort, N. L. *Bootstrapping Cox's Regression Model*. Report from Norwegian Computing Centre, Oslo, 1986.
17. Wichmann, B. A. and Hill, I. D. 'An efficient and portable pseudo-random number generator', *Applied Statistics*, **31**, 188–190 (1982).
18. Hopkins, A. 'Survival analysis with covariates – Cox model', in W. J. Dixon *et al.* (eds), *BMDP Statistical Software*, University of California Press, Berkeley, 1981.
19. Altman, D. G. and Andersen, P. K. 'A note on the uncertainty of a survival probability estimated from a Cox regression model', *Biometrika*, **73**, 722–724 (1986).
20. Tsiatis, A. A. 'A large sample study of Cox's regression model', *Annals of Statistics*, **9**, 93–108 (1981).
21. Kuk, A. Y. C. 'All subsets regression in a proportional hazards model', *Biometrika*, **71**, 587–592 (1984).
22. Frane, J. 'All possible subsets regression', in W. J. Dixon *et al.* (eds), *BMDP Statistical Software*, University of California Press, Berkeley, 1981.
23. Mallows, C. L. 'Some comments on C_p ', *Technometrics*, **15**, 661–675 (1973).
24. Royston, J. P. 'An extension of Shapiro and Wilk's W test for normality to large samples', *Applied Statistics*, **21**, 115–124 (1982).
25. Efron, B. and Gong, G. 'Statistical theory and the computer'. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer, New York, 1981, pp. 3–7.
26. Andersen, P. K. and Gill, R. D. 'Cox's regression model for counting processes: a large sample study', *Annals of Statistics*, **10**, 1100–1120 (1982).
27. Lee, K. L., Harrell, F. E., Tolley, H. D. and Rosati, R. A. 'A comparison of test statistics for assessing the effects of concomitant variables in survival analysis', *Biometrics*, **39**, 341–350 (1983).
28. Flack, V. F. and Chang, P. C. 'Frequency of selecting noise variables in subset regression analysis: a simulation study', *American Statistician*, **41**, 84–86 (1987).