

# Bayesian Model Averaging in Proportional Hazard Models: Assessing Stroke Risk

Chris T. Volinsky, David Madigan, Adrian E. Raftery, and Richard A. Kronmal  
University of Washington, U.S.A.<sup>1</sup>

Technical Report no. 302  
Department of Statistics  
University of Washington

January 16, 1996

<sup>1</sup>Chris Volinsky is a Research Assistant, David Madigan is a Professor of Statistics and Adrian E. Raftery is a Professor of Statistics and Sociology, Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195. Richard A. Kronmal is a Professor of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195. Email correspondence: *volinsky@stat.washington.edu*

## **Abstract**

Evaluating the risk of stroke is important in reducing the incidence of this devastating disease. Here, we apply Bayesian model averaging to variable selection in Cox proportional hazard models in the context of the Cardiovascular Health Study, a comprehensive investigation into the risk factors for stroke. We introduce a technique based on the leaps and bounds algorithm which efficiently locates and fits the best models in the very large model space and thereby extends all subsets regression to Cox models. For each independent variable considered, the method provides the posterior probability that it belongs in the model. This is more directly interpretable than the corresponding P-values, and also more valid in that it takes account of model uncertainty. P-values from models preferred by stepwise methods tend to overstate the evidence for the predictive value of a variable. In our data Bayesian model averaging predictively outperforms standard model selection methods for assessing stroke risk.

**KEY WORDS:** Bayesian Model Averaging, Model Selection, Cardiovascular Health Study

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The Cardiovascular Health Study</b>	<b>1</b>
<b>3</b>	<b>Standard Methods</b>	<b>2</b>
<b>4</b>	<b>Methods</b>	<b>3</b>
4.1	Cox Proportional Hazards Model . . . . .	3
4.2	Bayesian Model Averaging . . . . .	4
4.3	Predictive Distribution and the MLE Approximation . . . . .	5
4.4	Integrated Likelihood and the BIC Approximation . . . . .	5
4.5	Identifying the Models in Occam’s Window . . . . .	6
4.6	Assessment of Predictive Performance . . . . .	8
<b>5</b>	<b>Application to the Cardiovascular Health Study</b>	<b>9</b>
5.1	Results . . . . .	9
5.2	Predictive Performance . . . . .	11
<b>6</b>	<b>Software Implementation</b>	<b>12</b>
<b>7</b>	<b>Discussion</b>	<b>13</b>

## List of Figures

1	P-values vs. posterior probabilities of a non-zero coefficient . . . . .	12
---	--	----

## List of Tables

1	Models with the highest posterior model probability . . . . .	10
2	Posterior parameter estimates, standard errors, and parameter probabilities for the CHS . . . . .	10
3	Partial predictive scores for different methods . . . . .	11
4	Test data cross-classification of assigned risk group vs. stroke occurrence . . . . .	13
5	Predictive discrimination of Bayesian model averaging, top PMP model and stepwise . . . . .	13

# 1 Introduction

Stroke is the third leading cause of death among adults in the United States. Much recent research has attempted to identify the risk factors associated with this deadly condition. Some of these factors, such as smoking, are lifestyle choices which can be changed. Others, such as hypertension, are conditions which can be treated in a non-invasive manner. Both types of risk factors are therefore controllable (unlike an uncontrollable factor such as heredity), and the strokes caused by these conditions could be prevented. In fact, many doctors believe that most of the variables which determine one’s risk for stroke are controllable. Gorelick (1995) estimated that as many as 80% of strokes could be prevented. This suggests that society should place a high priority on finding the risk factors for stroke and in doing so identifying the people whose future strokes we can prevent.

Traditional analysis identifies risk factors or other independent variables as “significant” by invoking a model selection procedure. A single model is used for prediction which includes these variables and ignores the variables not selected by the procedure. Any such model selection procedure ignores uncertainty due to this procedure. Such procedures can underestimate uncertainty about the parameters, overestimate confidence in a particular model being “correct”, and lead to riskier decisions and poorer predictive ability. Bayesian model averaging (BMA) selects a subset of all possible models, and uses the posterior probabilities of the models to perform all inference and prediction. The goal of this paper is to show that BMA leads to better evaluation of the risk factors for stroke, as well as improved risk assessment for potential stroke victims.

Section 2 introduces the Cardiovascular Health Study, a highly censored dataset which investigates the risk factors for stroke in elderly patients. Section 3 reviews the standard approach to modelling this type of survival data. In Section 4 we describe Bayesian model averaging as it applies to variable selection in the Cox proportional hazard model. Using Occam’s Window, we define the subset of model space to average over and using a modification of the leaps and bounds algorithm we identify those models efficiently.

Section 5 returns to the Cardiovascular Health Study, and interprets the results of the BMA analysis. By using two different methods to assess predictive performance, BMA is shown to be better at prediction and risk assessment than the standard measures. Section 6 describes `bic.surv`, a function available free from Statlib to implement BMA for survival analysis.

In Section 7 we discuss some of the interesting results of the study and how they relate to stroke prevention. We also discuss some of the open questions and relate our results here to other work in the literature. Our conclusion is that taking account of model uncertainty using Bayesian model averaging can enhance predictive performance in survival analysis.

## 2 The Cardiovascular Health Study

The Cardiovascular Health Study (CHS) is an NIH-funded longitudinal study, started in June 1989, whose aim is to study cardiovascular disease in people aged 65 and over. Risk factors for stroke have been extensively studied in the medical literature. In the United States, several population-based studies have identified what are believed to be the main

risk factors for stroke: smoking, hypertension and atrial fibrillation (Kannel, McGee, and Gordon 1976; Matsumoto, Whisnant, et al. 1973). However, there is some evidence that the effects of these factors weaken with increasing age. Indeed, some studies have hinted that the factors for increased stroke in younger populations may have no effect or even a protective effect in the elderly. The CHS is a longitudinal, observational study of 5201 patients in four U.S. counties: Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Allegheny County, Pennsylvania. It is the most comprehensive source of data on stroke incidence in elderly people.

The CHS collected 23 measures of sub-clinical disease which may be related to heart attacks and/or strokes. For the purposes of this technical report, these variables will be called V1-V23. Fried et al. (1991) describe the complete sample design and study methods as well as specific protocols for the classification of some of the independent variables. The data used for this analysis consists of 4501 patients who were free of stroke at baseline, and who had complete data. The followup was for between 3.5 and 4.5 years with an average of 4.1 years. Those who had not experienced a stroke at the end of the study were censored at that time. Only 172 of those patients experienced a stroke in the study time period.

The CHS contains a high proportion of censored data (96%) and a large number of potential risk factors. We therefore expect a fair amount of model uncertainty. In this paper we account for this model uncertainty using BMA.

### 3 Standard Methods

Methods for analyzing survival data such as the CHS data often focus on modelling the hazard rate, also called the instantaneous failure rate. The most popular way of doing this is to use the Cox proportional hazards model (Cox 1972), which allows different hazard rates for patients with different covariate vectors, and leaves the underlying common baseline hazard rate unspecified. The parameters are estimated and standard errors found using the partial likelihood (Section 4.1).

Often when many variables are collected, the Cox regression model is used in conjunction with a variable selection method. Such a method aims to choose the subset of the full variable list which best fits the data collected. The most popular such procedures are *stepwise methods*, which iteratively propose models that differ from the current model by one variable and accept or reject those models based on a significance test. In *forward selection* the stepwise procedure starts at the empty model and adds variables, while in *backward selection* the initial model is the full model and variables are deleted. Efron (1960) combined forward and backward selection to create *stepwise selection*, where each step can be an addition or a deletion of a variable. This seems to be the most popular method for Cox proportional hazard models (e.g. Fleming and Harrington 1991).

There is a large literature on variable selection procedures in the context of multiple linear regression. Many of the popular regression textbooks (e.g. Neter et al. 1990; Weisberg 1985) support the use of stepwise methods in conjunction with other preliminary data analyses to arrive at a single acceptable model. However, there is now much literature exposing the problems with this general approach. Derksen and Keselman (1992) provide a literature review of studies which critique the use of stepwise methods. These criticisms also apply to

any selection procedure which settles on a single model. The criticisms can be summarized as follows:

1. *Credibility of model exaggerated.* Use of a selection method leads to a model which appears to have more explanatory power than it really does. Selection methods which minimize a particular criterion tend to overestimate the true population value. For linear regression, stepwise methods often give highly significant values for  $R^2$  when the explanatory variables are pure noise (Freedman 1983).
2. *Level of testing procedure unknown.* Stepwise procedures are in essence a sequential application of significance tests at a specified level (a level pre-designated by the analyst). Therefore, the overall probability of a Type I error in the family of tests far exceeds the specified level for an individual test. It is difficult to calculate the true level of the entire stepwise procedure.
3. *Criterion level.* There is no agreement on the best criterion for the addition and deletion of variables in a stepwise procedure. Ideally, such a threshold will exclude noise variables and include relevant ones. Procedures suggested for survival analysis include setting a threshold for the Wald statistic or a variation thereof, using its asymptotic distribution (Peduzzi, Hardy, and Holford 1980) so that its P-value is no greater than a pre-determined level. There is no consensus on what that level should be.
4. *Selection of noise variables.* The greater the number of models investigated, the more likely it is that one of the models is going to model the chance variation well. Several studies (Freedman 1983; Derksen and Keselman 1992) show that stepwise procedures tend to select models which label pure noise as significant, especially when the dependent variables are correlated.
5. *Significance of selected variables.* Often users attach importance to the first variable included or relative importance based on the order of entry or deletion from the model. This can be misleading. In fact, Hocking (1976) claimed that the first variable added in a forward selection can also be the first one deleted in a backward elimination!

Several simulation studies have exposed the problems associated with stepwise methods in linear regression (Freedman 1983; Derksen and Keselman 1992; Raftery, Madigan, and Hoeting 1993). No such study exists for survival analysis, although there is little reason to think that the problems would vanish in this class of models.

## 4 Methods

### 4.1 Cox Proportional Hazards Model

We use the Cox (1972) proportional hazards model, which specifies the hazard rate for subject  $i$  with covariate vector  $\mathbf{x}_i$  to be

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i^T \theta)$$

where  $\lambda_0(t)$  is the baseline hazard function at time  $t$ , left unspecified in Cox’s formulation, and  $\theta$  is a vector of unknown parameters.

The estimation of  $\theta$  is commonly based on the partial likelihood, namely

$$PL(\theta) = \prod_{i=1}^n \left( \frac{\exp(\mathbf{x}_i^T \theta)}{\sum_{\ell \in R_i} \exp(\mathbf{x}_\ell^T \theta)} \right)^{w_i}, \quad (1)$$

where  $R_i$  is the risk set at time  $t_i$  (i.e. the set of subjects who have not yet experienced an event) and  $w_i$  is an indicator for whether or not patient  $i$  is censored. Equation (1) assumes that there are no ties between the times at which deaths occur; when there are ties modifications are necessary, but for simplicity we do not consider these modifications here. There are several ways of checking the proportional hazard assumptions, mostly based on the martingale residuals (Fleming and Harrington 1991).

## 4.2 Bayesian Model Averaging

Once the Cox model has been chosen, modelling consists of choosing the independent variables. The typical approach is to find the “best” model (from now on, the term ‘model’ refers to the specific collection of variables selected for use in the Cox model). However, this approach ignores a major component of uncertainty, namely uncertainty about the model itself. As a consequence, uncertainty about quantities of interest can be underestimated. For striking examples of this see Regal and Hook (1991), Madigan and York (1995), Kass and Raftery (1995), and Raftery (1996).

There is a standard Bayesian solution to this problem. All prediction and inference is made using a set of models instead of just one. Let  $\Delta$  be any quantity of interest such as a future observation or the utility of a course of action. If  $\mathcal{M} = \{M_1, \dots, M_k\}$  denotes the set of all models considered, then the posterior distribution of  $\Delta$  given the data  $D$  is

$$\text{pr}(\Delta | D) = \sum_{k=1}^K \text{pr}(\Delta | M_k, D) \text{pr}(M_k | D). \quad (2)$$

This is an average of the posterior distributions under each model weighted by the corresponding posterior model probabilities.

For a problem with  $p$  potential covariates, the number of models,  $K$ , in the finite sum (2) can be enormous ( $K=2^p$  in the absence of other constraints). Usually only a small number of these models will have much support from the data. To take advantage of this, we use the Occam’s Window approach of Madigan and Raftery (1994). Occam’s Window includes only the models with the highest posterior model probabilities in the sum above. The Occam’s Window approach argues that if a model is far less likely given the data than the most likely model, then it has effectively been discredited and should no longer be considered. Thus, only models belonging to the set

$$\mathcal{A} = \left\{ M_k : \frac{\max_l \{\text{pr}(M_l | D)\}}{\text{pr}(M_k | D)} \leq C \right\}, \quad (3)$$

should be included in the sum in equation (2), where  $C$  is chosen by the data analyst. We have found  $C = 20$  to be a reasonable choice.

Equation (2) has three components, each posing its own computational difficulties, which will be discussed in the following sections. The predictive distribution  $\text{pr}(\Delta|M_k, D)$  requires integrating out the model parameter  $\theta_k$  (Section 4.3). The posterior model probabilities  $\text{pr}(M_k|D)$  involve the calculation of an integrated likelihood (Section 4.4). Finally, there is a need to efficiently locate and evaluate the models in the finite sum which fall into Occam's Window (Section 4.5).

### 4.3 Predictive Distribution and the MLE Approximation

In (2) the predictive distribution of  $\Delta$  given a particular model  $M_k$  is found by integrating out the model parameter  $\theta_k$ :

$$\text{pr}(\Delta | M_k, D) = \int \text{pr}(\Delta|\theta_k, M_k, D) \text{pr}(\theta_k|M_k, D) d\theta_k, \quad (4)$$

This integral does not have a closed form solution for Cox models. Here we use the MLE approximation:

$$\text{pr}(\Delta|M_k, D) \approx \text{pr}(\Delta|M_k, \hat{\theta}_k, D). \quad (5)$$

This was used in the model uncertainty context by Taplin (1993) who found it to give an excellent approximation in his time series regression problem; it was subsequently used by Taplin and Raftery (1994) and Draper (1995).

### 4.4 Integrated Likelihood and the BIC Approximation

The posterior probability of model  $M_k$  is given by

$$\text{pr}(M_k | D) \propto \text{pr}(D | M_k) \text{pr}(M_k), \quad (6)$$

where

$$\text{pr}(D | M_k) = \int \text{pr}(D | \theta_k, M_k) \text{pr}(\theta_k | M_k) d\theta_k \quad (7)$$

is the integrated likelihood of model  $M_k$ , and  $\text{pr}(\theta_k | M_k)$  is the prior density of  $\theta_k$  under model  $M_k$ . In regression models for survival analysis, analytic evaluation of the integral (7) is not possible in general, and so some kind of analytic or computational approximation is necessary.

In regular statistical models (roughly speaking, those in which the MLE is consistent and asymptotically normal), the integral in (7) can often be approximated via the Laplace method. This method provides an approximation which when applied to equation (7) yields

$$\log \text{pr}(D|M_k) = \log \text{pr}(D|\hat{\theta}_k, M_k) - d_k \log n + O(1), \quad (8)$$

which is the Bayesian information criterion (BIC) approximation derived by Schwarz (1978) in another way. In fact, (8) is much more accurate for many practical purposes than its  $O(1)$  error term suggests. Kass and Wasserman (1995) have shown, under certain assumptions, that when  $M_j$  and  $M_k$  are nested and the amount of information in the prior distribution is equal to that in one observation, then the error in approximating the Bayes factor between

the two models with (8) is  $O(n^{-\frac{1}{2}})$  rather than  $O(1)$ . Raftery (1996) gave some further empirical evidence for the accuracy of this approximation.

In (8),  $n$  is usually taken to be the total number of cases. However, for survival analysis we take  $n$  to be the total number of *uncensored* cases (i.e. deaths or events); we have found this choice empirically to be the most effective. Proving that this provides the best approximation remains an open problem.

Equation (6) requires the specification of model priors. When there is little prior information about the relative plausibility of the models considered, taking them all to be equally likely *a priori* is a reasonable “neutral” choice: we adopt this choice for the CHS analysis. In our experience with very large model spaces (up to  $10^{12}$  models) involving several kinds of model and about 20 data sets, we have found no perverse effects from putting a uniform prior over the models even in such a large model space (Raftery, Madigan, and Hoeting 1993; Madigan and Raftery 1994; Madigan, Perlman, and Volinsky 1996). When prior information about the importance of a variable is available, a prior probability on model  $M_i$  can be specified as:

$$\text{pr}(M_i) = \prod_{j=1}^p \pi_j^{\delta_{ij}} (1 - \pi_j)^{1 - \delta_{ij}}. \quad (9)$$

where  $\pi_j \in [0, 1]$  is the prior probability that  $\theta_j \neq 0$  (for the CHS,  $j = 1 \dots 23$ ),  $\delta_{ij}$  is an indicator of whether or not variable  $j$  is included in model  $M_i$ . Assigning  $\pi_j = .5$  for all  $j$  corresponds to a uniform prior across model space, while  $\pi_j < .5$  for all  $j$  imposes a penalty for large models. Using  $\pi_j = 1$  ensures that variable  $j$  is included in all models. Using this framework, elicitation of prior probabilities for models is straightforward and avoids the need to elicit priors for a large number of models. For an alternate approach when expert information is available, see Madigan, Gavrin, and Raftery (1995).

## 4.5 Identifying the Models in Occam’s Window

The Occam’s Window approach requires that we identify the best model and average over only those models which fall within a factor,  $C$ , of the posterior probability of that best model. We define the best model as the model with the largest BIC. But how can we identify the best model without visiting all models? If there were a way to quickly screen the models without fitting them all, targeting those that are close in posterior probability to the best one, we could then frugally run BMA on this reduced set of models.

Such a procedure does exist, but only for linear regression. Regression by leaps and bounds (Furnival and Wilson 1974) is an efficient algorithm which allows the user to move through model space proceeding quickly towards the best model, deleting large subsets of model space as it goes. The leaps and bounds algorithm provides the top  $q$  models of each model size, where  $q$  is designated by the user, plus the MLE  $\hat{\theta}_k$ ,  $\text{var}(\hat{\theta}_k)$ , and  $R^2$  for each model  $k$  returned. The method draws on this simple fact: for two models  $A$  and  $B$ , where  $A$  and  $B$  are each subsets of the full parameter set, if  $A \subset B$  then  $\text{RSS}(A) > \text{RSS}(B)$ . Using this fact, the method is able to eliminate large portions of model space by performing sweep operations on the matrix:

$$\begin{pmatrix} X'X & X'y \\ y'X & y'y \end{pmatrix}. \quad (10)$$

So how can we adapt the leaps and bounds algorithm to the survival analysis setting? Using a method of Lawless and Singhal (1978) developed for nonlinear regression models, a modification of the leaps and bounds algorithm provides an approximate likelihood ratio test statistic (and therefore an approximate BIC value). From this, model space is substantially reduced by excluding models that are unlikely to be in  $\mathcal{A}$ .

The method proceeds as follows: Let  $\theta$  be the parameter vector of the full model and let  $\theta_k$  be the vector for a given submodel  $k$ . Rewrite  $\theta_k$  as  $(\theta_1, \theta_2)$  so that Model  $M_k$  corresponds to the submodel  $\theta_2 = 0$ . Also, let

$$V = \mathcal{I}^{-1} = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{pmatrix}$$

denote the inverse observed information matrix. If  $L(\hat{\theta})$  is the maximized likelihood under the full (unrestricted) model, and  $L(\tilde{\theta})$  is the maximized likelihood under  $\theta_2 = 0$ , then

$$\Lambda = -2[\log L(\tilde{\theta}) - \log L(\hat{\theta})]$$

is the usual likelihood ratio statistic for the test of the submodel versus the full model while

$$\Lambda' = \hat{\theta}_2' V_{22}^{-1} \hat{\theta}_2$$

is an approximation to  $\Lambda$  based on the Wald statistic. Finally, replace the matrix in (10) with

$$\begin{pmatrix} \mathcal{I} & \mathcal{I}\hat{\theta} \\ \hat{\theta}'\mathcal{I} & \hat{\theta}'\mathcal{I}\hat{\theta} \end{pmatrix}$$

and perform the same matrix sweep operators from the leaps and bounds algorithm on this matrix. As a result the function provides

- an estimate of the best  $q$  proportional hazards models for each model size,
- the LRT approximation  $\Lambda'$  for each model,
- an approximation to  $\tilde{\theta}$ , the MLE for the parameters of the submodel, and
- the asymptotic covariance matrix  $V_{11}^{-1}$ .

As long as  $q$  is large enough, this procedure returns the models in  $\mathcal{A}$  plus many models not in  $\mathcal{A}$ . We can use the approximate LRT to reduce the remaining subset of models to those most likely to be in  $\mathcal{A}$ . This reduction step keeps only the models whose posterior probabilities fall within a factor  $C'$  of the model with the best PMP, where  $C'$  is greater than  $C$ . We use  $C' = C^2$ , which we have found to be large enough to be almost sure that no models in  $\mathcal{A}$  will be lost. Values of  $C^{1.5}$  or less will usually suffice, but perhaps with the loss of some low posterior probability models. We fit the remaining models iteratively using standard survival analysis software, as a result, the value of  $C'$  may have a big effect on the computer time needed.

We fit the remaining models by any standard survival analysis program, calculate the exact BIC value for each one, and eliminate those models not in  $\mathcal{A}$ . For the models in  $\mathcal{A}$ ,

we calculate a posterior model probability by normalizing over the model set. We calculate parameter estimates and standard errors of those estimates by taking weighted averages of the estimates and errors from the individual models, using the posterior model probabilities as weights.

Finally, we compute the posterior probability that the regression coefficient for a variable is non-zero, by adding the posterior probabilities of the models which contain that variable. Standard rules of thumb for interpreting this posterior probability are as follows:  $< 50\%$ : evidence against the effect;  $50 - 75\%$ : weak evidence for the effect;  $75 - 95\%$ : positive evidence;  $95 - 99\%$ : strong evidence; and  $> 99\%$ : very strong evidence (Kass and Raftery 1995).

## 4.6 Assessment of Predictive Performance

We assess the value of our method (BMA) as opposed to that of competing methods (e.g. stepwise variable selection) on the basis of their predictive performance. To measure predictive performance we randomly split the data into two halves. We build the model by applying the procedures outlined above to the first half of the data, and then we predict the survival of each member of the second half of the data. Ideally, we would then assess predictive performance by a log score based on a predictive ordinate (Good 1952). If we denote the first half by  $D^B$  (build data), and the second half by  $D^T$  (test data), a predictive log score for any given model  $M_k$  is

$$\sum_{d \in D^T} \log \text{pr}(d \mid M_k, D^B). \quad (11)$$

Similarly, the predictive log score for model averaging is

$$\sum_{d \in D^T} \log \left\{ \sum_{M \in \mathcal{M}} \text{pr}(d \mid M, D^B) \text{pr}(M \mid D^B) \right\}, \quad (12)$$

where  $\mathcal{M}$  is the set of selected models.

However, under the partial likelihood framework predictive distribution functions usually take the form of a step function (Breslow 1975), so the predictive ordinate  $\text{pr}(d \mid M, D^B)$  is difficult to compute. In the spirit of the log score, and of Cox's partial likelihood (1), we developed the *partial predictive log score*. Informally, partial likelihood asks for each failure of subject  $i$  at time  $t$ : "Given that someone in the risk set dies at time  $t$ , what is the probability of that person being subject  $i$ ?" Partial predictive score asks the same question of the test data by substituting

$$\text{pr}(d \mid M_k, D^B) = \left( \frac{\exp(\mathbf{x}_i^T \hat{\theta}_k)}{\sum_{\ell \in R_i} \exp(\mathbf{x}_\ell^T \hat{\theta}_k)} \right)^{w_i}$$

into (11) and (12) above. Using this method we can compare BMA to any single model that could have been chosen. The partial predictive score is greater for the method which gives higher probability to the events that occur in the test set.

We also compare methods based on their *predictive discrimination*, namely how well they sort the subjects in the test set into discrete risk categories (high, medium, low risk). We assess predictive discrimination of a single model as follows:

1. Fit the model to the build data to get estimated coefficients  $\hat{\theta}$ .
2. Calculate risk scores ( $\mathbf{x}_i^T \hat{\theta}$ ) for each patient in the build data.
3. Define low, medium and high risk groups for the model by the empirical 33rd and 66th percentiles of the risk scores.
4. Calculate risk scores for the test data and assign each patient to a risk group.
5. Extract the patients who are assessed as being in a higher risk group by one method than by another, and tabulate what happened to those patients over the study period.

To assess predictive discrimination for BMA, we need to take into account the multiple models that we average over. We replace the first steps above with

- 1'. Fit each model  $M_1 \dots M_K$  in  $\mathcal{A}$  to get estimated coefficients  $\hat{\theta}_k$ .
- 2'. Calculate risk scores ( $\mathbf{x}_i^T \hat{\theta}_k$ ) under each model in  $\mathcal{A}$  for each patient in the build data. A patient's risk score under BMA is the weighted average of these:  $\sum_{k=1}^K (\mathbf{x}_i^T \hat{\theta}_k) \text{pr}(M_k | D^B)$ .

A method is better if it consistently assigns higher risks to the patients who actually had strokes.

## 5 Application to the Cardiovascular Health Study

### 5.1 Results

The CHS, with over 95% censoring provides an opportunity to compare BMA with model selection methods in the presence of heavy censoring. The model chosen by a stepwise (backward elimination) procedure included 10 variables. The model with the highest approximate posterior probability, was the same as the stepwise model except that V11 was not included.

In the list of models provided by BMA, the stepwise model places fourth. Table 1 contains the variables included in any of the top five models. None of the 12 other variables in the original variable list (V1-V23) appears in these models. Inference about independent variables is expressed in terms of the posterior probability that the parameter does not equal zero. Table 2 contains the posterior means, standard deviations and posterior probabilities of the variables.

Figure 1 shows the posterior probability that each regression coefficient is non-zero, plotted against the corresponding  $P$  value from stepwise variable selection. Overall, the posterior probabilities imply weaker evidence for effects than do the  $P$  values. This is partly due to the fact that  $P$  values overstate confidence because they ignore model uncertainty. However, even when there is no model uncertainty, it has been argued that  $P$  values overstate the

Table 1: *The top five models (in terms of BIC) chosen by Bayesian Model Averaging, with posterior model probabilities. ( $\star$  indicates the model chosen by stepwise variable selection).*

Model	V1	V2	V3	V6	V8	V9	V10	V19	V20	V23	V11	PMP(in %)
1	✓	✓	✓	✓	✓	✓	✓		✓	✓		1.7
2	✓		✓	✓	✓	✓	✓		✓	✓	✓	1.6
3	✓		✓	✓	✓	✓	✓		✓	✓		1.4
4 $\star$	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	1.4
5	✓		✓	✓	✓	✓	✓	✓	✓	✓		1.1

Table 2: *Posterior parameter estimates (means), standard deviations and parameter probabilities for the variables in the CHS dataset. Means and standard deviations are averaged over all models included in the BMA analysis. The parameter probability is the posterior probability that the parameter is non-zero.*

Var	Mean	SD	$P(\theta \neq 0)$	Var	Mean	SD	$P(\theta \neq 0)$	Var	Mean	SD	$P(\theta \neq 0)$
V1	0.04	0.01	89	V9	0.42	0.10	100	V17	0.00	0.00	0
V2	0.40	0.17	59	V10	0.83	0.30	66	V18	0.00	0.00	24
V3	0.52	0.16	99	V11	0.60	0.25	50	V19	0.46	0.22	26
V4	0.62	0.31	18	V12	0.00	0.00	0	V20	0.34	0.12	90
V5	0.01	0.01	0	V13	0.37	0.18	21	V21	0.24	0.17	5
V6	0.02	0.00	100	V14	0.00	0.00	0	V22	0.00	0.00	0
V7	0.79	0.37	24	V15	-0.01	0.01	7	V23	0.47	0.10	100
V8	0.43	0.14	64	V16	0.05	0.23	0				

evidence for an effect (Edwards, Lindman, and Savage 1963; Berger and Delampady 1987; Berger and Sellke 1987).

For the four variables, V3, V6, V9, V23, the posterior probabilities and the  $P$  values agree that there is very strong evidence for an effect ( $P < .001$  and  $P(\theta \neq 0) > 99\%$ ). For the following eight variables, however, the two approaches lead to qualitatively different conclusions:

Var	$P$ value	$P(\theta \neq 0)$ (%)
V8	.002**	64
V20	.004**	90
V10	.005**	66
V1	.007**	89
V11	.022*	50
V2	.026*	59
V13	.053	21
V16	.058	24

In each case the  $P$  value overstates the evidence for an effect. In the first four cases, the  $P$  value would lead to the effect being called “highly significant” ( $P < .01$ ), while the posterior probability indicates the evidence to be positive but not strong. For the next two variables (V11 and V2), the  $P$  value is “significant” ( $P < .05$ ), but the posterior probabilities indicate the evidence for an effect to be weak. For the last two variables (V13 and V16), the  $P$  values are “marginally significant” ( $P < .06$ ), but the posterior probabilities actually indicate (weak) evidence *against* an effect.

For the remaining 11 variables,  $P$  values and posterior probabilities agree in saying that there is little or no evidence for an effect. However, posterior probabilities enable one to make one distinction that  $P$  values cannot. One may fail to reject the null hypothesis of “no effect” because either (a) there are not enough data to detect an effect, or (b) the data provide evidence *for* the null hypothesis.  $P$  values cannot distinguish between these two situations, but posterior probabilities can. Thus, for example, for V19,  $P(\theta \neq 0) = 26\%$ , so that the data are indecisive, while for V5,  $P(\theta \neq 0) = 0\%$ , indicating strong evidence *for* the null hypothesis of no effect. Note that the posterior probability of “no effect” can be viewed as an approximation to the posterior probability of the effect being “small”, namely  $P(|\theta| < \varepsilon)$ , provided that  $\varepsilon$  is at most about one-half of a standard error (Berger and Delampady 1987).

For V10 the posterior probability is 0.66, indicating weak evidence for an effect. This weak result is probably due to the fact that only about 3% of the patients had this condition. However, this variable is known from previous studies to have an effect (Wolf, Abbott, and Kannel 1987), and so the prior probability of an effect should be set much higher, perhaps equal to or near 1, as discussed in Section 4.4. Had this been done the posterior probability would have been equal to or near 1 as well.

## 5.2 Predictive Performance

For assessing predictive performance, the data were randomly split into two parts, in such a way that an equal number of events (86 strokes) occurred in each part. We compare the results for BMA as compared to stepwise model selection and to the single model with the highest posterior probability. This contrasts BMA’s performance with both Bayesian and frequentist single model methods. Table 3 shows the predictive scores (PPS) for the competing methods. A higher score (less negative) indicates better predictive performance. Note that the top model and stepwise model may be different than those in the previous section since they are built using only half the data.

Table 3: *Partial predictive scores for the model with the highest posterior model probability, the stepwise chosen model, and Bayesian Model Averaging.*

Method	PPS
Top Model	-641.6
Stepwise	-632.8
BMA	-629.9

The difference in PPS of 11.7 can be viewed as an increase in predictive performance *per event* by a factor of  $\exp(11.7/86) = 1.15$  or by about 15%. We would interpret this as saying that BMA predicts who is at risk for stroke 15% more effectively than a method which picks the model with the best posterior model probability (or 3.5% more effectively than a stepwise method).

Predictive discrimination shows the benefit of using Bayesian model averaging in another way. Table 4 shows the classification of the 2252 patients in the test data, and whether or not those patients had a stroke in the study period. Table 5 summarizes the outcome for

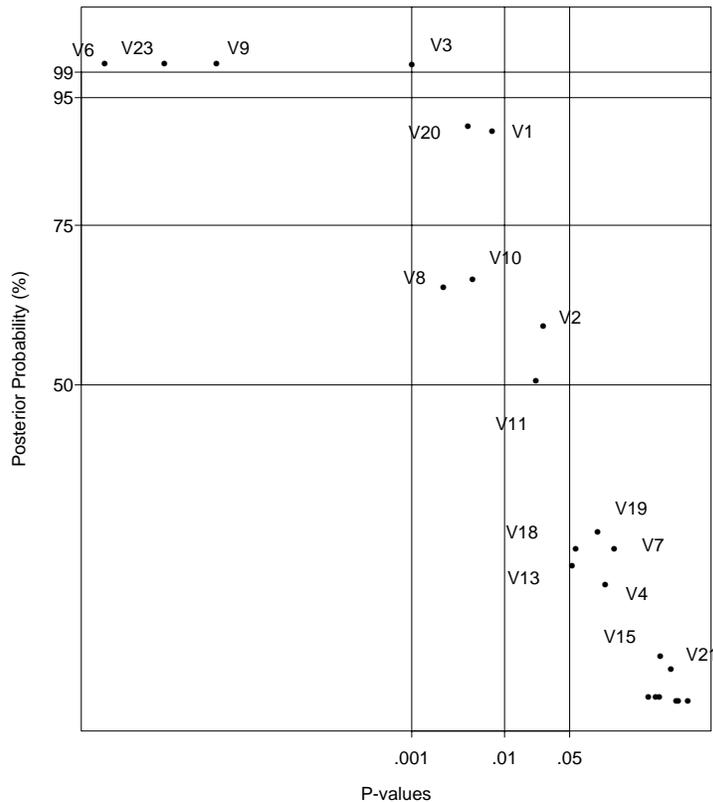


Figure 1: *P-values versus posterior probabilities that the coefficient is non-zero for the variables in the CHS. The lines at P-values of .05, .01, and .001 indicate the standard limits for significant, highly significant and very highly significant. Lines at 50%, 75%, 95% and 99% posterior probability indicate weak evidence, positive evidence, strong evidence and very strong evidence.*

the patients who were given higher risk by one method over another. The patients to whom which BMA assigned a higher risk had more strokes and were stroke-free for a shorter period of time. Both tables show that Bayesian model averaging does a better job of indicating which patients are at risk of strokes.

## 6 Software Implementation

We have written a series of S-PLUS functions to implement Bayesian model averaging for proportional hazard models using the approximations outlined here. The function `bic.surv` performs the Bayesian model averaging and outputs all information about the models selected, while `summary.bs` condenses this information to output only posterior model probabilities, posterior variable probabilities, and the variables included in the selected models. A modification of the internal S-PLUS leaps and bounds function, called `leaps.bs`, is needed

Table 4: *Test data cross-classification of assigned risk group vs. stroke occurrence. Stroke risk for each subject in the test data is assessed from the 33rd and 66th percentiles of the risk scores from the build data. Risk groups are determined separately for the three methods: model averaging, stepwise and the top posterior model probability.*

		Model Averaging		Stepwise		Top PMP	
		Stroke-free	Stroke	Stroke-free	Stroke	Stroke-free	Stroke
Assigned risk group	Low	751	7	750	8	724	10
	Med	770	24	799	27	801	28
	High	645	55	617	51	641	48

Table 5: *Predictive discrimination of Bayesian model averaging compared with the top posterior probability model and the stepwise model. This is a breakdown of the patients which fell into different risk groups (low, medium, or high) for two competing methods. Mean surv time is the mean stroke-free time for the patients in that group.*

Estimated Risk	# strokes	# patients	Mean surv time
BMA > top PMP	11	147	1169
BMA < top PMP	1	160	1243
BMA > stepwise	10	165	1191
BMA < stepwise	5	133	1243

to use these functions.

`bic.surv` requires a few user-specified parameters. The parameters “nbest”, “OR”, and “OR.fix” correspond to the variables  $q$ ,  $C$ , and  $C'$  in Section 4.5. The defaults for these parameters are set to minimize the loss of models with relatively high posterior probability. The values of these variables can be changed if computer time is an important factor. Prior information about variables can be specified through the function parameter “prior.param” as a vector of length equal to the number of independent variables. Each element of the vector must be between 0 and 1, corresponding to the prior probability that the coefficient is non-zero. By default the prior on all variables is .5, leading to a uniform prior over model space.

All the software can be obtained free of charge by sending the message “send bic.surv from S” to [statlib@stat.cmu.edu](mailto:statlib@stat.cmu.edu) or on the World Wide Web at the URL: <http://lib.stat.cmu.edu/S/bic.surv>.

## 7 Discussion

We have extended all subsets regression to Cox models, and developed a Bayesian way of accounting for model uncertainty. This is an alternative to standard stepwise and other variable selection methods. For each variable, our Bayesian model averaging method yields the posterior probability that the variable has an effect on survival. This is more directly

interpretable than the corresponding P-value, and it is also more valid because it takes account of model uncertainty, which the P-value does not. It is by now well established that P-values can be very misleading when used with stepwise and other standard variable selection methods. The P-value tends to overstate the evidence for the predictive value of a variable.

While stepwise methods lead to hard inclusion or exclusion of each variable, the posterior probability can be more informative. There may be some evidence for the predictive value of a variable, even if it is not in the “best” selected model. For example, the variable Ejection is not in either the stepwise selected model or the model with the highest posterior probability, but the posterior probability that its parameter is non-zero is an indecisive 0.26, indicating that it should not necessarily be discarded in future work, and in fact it is included in any prediction with weight proportional to this posterior probability. This indecisive result may be due to a lack of power in the data rather than to its not having appreciable predictive value.

Our method also has somewhat better predictive performance than single model methods, both frequentist and Bayesian. Thus, it may ultimately allow practitioners to better target patients susceptible to stroke so that preventive resources can be spent more efficiently.

Our approach is approximate in several respects. The method could be improved by using a better approximation to posterior model probabilities than that based on (8). The models considered here can be written as generalized linear models (Aitkin et al. 1989), and so the more accurate approximations of Raftery (1996) should be applicable. Also, the MLE approximation (5) to the predictive distribution could be improved, perhaps by using a Laplace approximation to the integral (4) or by a Monte-Carlo method.

We have considered only one component of model uncertainty: which independent variables to include in the model. There are other components also including uncertainty about functional forms of the independent variables and uncertainty about the Cox model itself. Our approach could be extended to take account of those.

The benefits of using BMA to account for model uncertainty have now been assessed for several different model classes, including linear regression, exponential survival models, logistic regression and discrete graphical models. The results of these studies are summarized in Raftery, Madigan, and Volinsky (1995). In each case model averaging improved predictive performance, by amounts that ranged from modest to substantial.

Although we are not aware of other work applying Bayesian model averaging to variable selection in survival analysis, there has been a good deal of recent work on model uncertainty; see Kass and Raftery (1995), Chatfield (1995) and Draper (1995) for reviews. Draper’s (1995) idea of model expansion is not applicable directly to uncertainty about regression variable selection, but it could well be useful for uncertainty about the model for the baseline hazard in parametric survival analysis. Clyde et al. (1995) look at model mixing through orthogonalization of the independent variables. George and McCulloch (1993) have developed the Stochastic Search Variable Selection (SSVS) method, which is similar in spirit to the MC<sup>3</sup> algorithm of Madigan and York (1995). This was developed for linear regression but could probably be extended to survival analysis.

## Acknowledgements

We thank the U.S. Office of Naval Research (grant N00014-96-1-0192), which funded the contribution of Volinsky and Raftery. The National Science Foundation funded Madigan's work. The CHS study is funded by contracts NO1-HC-85079 to NO1-HC-85086 from the National Heart, Lung and Blood Institute.

## References

- Aitkin, M., D. Anderson, B. Francis, and J. Hinde (1989). *Statistical Modelling in GLIM*. Oxford: Clarendon Press.
- Berger, J. O. and M. Delampady (1987). Testing precise hypotheses. *Statistical Science* 2, 317–52.
- Berger, J. O. and T. Sellke (1987). Testing a point null hypothesis (with discussion). *Journal of the American Statistical Association* 82, 112–22.
- Breslow, N. (1975). Analysis of survival data under the proportional hazards model. *Int. Statist. Rev.* 43, 45–8.
- Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference (with discussion). *Journal of the Royal Statistical Society A* 158, 419–66.
- Clyde, M., H. DeSimone, and G. Parmigiani (1995). Prediction via orthogonalized model mixing. Technical report, Institution of Statistics and Decision Sciences–Duke University.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 34, 187–220.
- Derksen, S. and H. Keselman (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 45, 262–282.
- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Ser. B* 57, 45–97.
- Edwards, W., H. Lindman, and L. J. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review* 70, 193–242.
- Efroymson, M. (1960). Multiple regression analysis. In A. Ralston and H. Wilf (Eds.), *Mathematical Methods for Digital Computers*. New York: Wiley.
- Fleming, T. R. and D. H. Harrington (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician* 37, 152–155.
- Fried, L. P., N. O. Borhani, et al. (1991). The Cardiovascular Health Study: Design and rationale. *Annals of Epidemiology* 1, 263–276.

- Furnival, G. M. and R. W. Wilson (1974). Regression by leaps and bounds. *Technometrics* 16, 499–511.
- George, E. and R. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88, 881–889.
- Good, I. J. (1952). Rational decisions. *Journal of the Royal Stat. Soc. B* 14, 107–14.
- Gorelick, P. B. (1995). Stroke prevention. *Archives of Neurology* 52, 347–355.
- Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* 32, 1–49.
- Kannel, W. B., D. McGee, and T. Gordon (1976). A general cardiovascular risk profile: The Framingham study. *The American Journal of Cardiology* 38, 46–51.
- Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90, 773–795.
- Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses with large samples. *Journal of the American Statistical Society* 90, 928–934.
- Lawless, J. and K. Singhal (1978). Efficient screening of nonnormal regression models. *Biometrics* 34, 318–27.
- Madigan, D., J. Gavrinn, and A. E. Raftery (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics - Theory and Methods* 24, 2271–92.
- Madigan, D., M. Perlman, and C. T. Volinsky (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics*. to appear.
- Madigan, D. and A. E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam’s Window. *Journal of the American Statistical Association* 89, 1535–46.
- Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.* 63, 215–32.
- Matsumoto, N., J. P. Whisnant, et al. (1973). Natural history of stroke in Rochester, Minnesota, 1955 through 1969: An extension of a previous study, 1945 through 1954. *Stroke* 4, 20–25.
- Neter, J., W. Wasserman, and M. H. Kutner (1990). *Applied Linear Statistical Models* (3d ed.). Irwin.
- Peduzzi, P. N., R. J. Hardy, and T. R. Holford (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics* 36, 511–6.
- Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*. to appear.
- Raftery, A. E., D. Madigan, and J. Hoeting (1993). Model selection and accounting for model uncertainty in linear regression models. Technical Report 262, University of Washington.

- Raftery, A. E., D. Madigan, and C. T. Volinsky (1995). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 5*. Oxford University Press. to appear.
- Regal, R. and E. B. Hook (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistical Medicine* 10, 717–21.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.
- Taplin, R. H. (1993). Robust likelihood calculation for time series. *Journal of the Royal Statistical Society B* 55, 829–36.
- Taplin, R. H. and A. E. Raftery (1994). Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Biometrics* 50, 764–781.
- Weisberg, S. (1985). *Applied Linear Regression* (2d ed.). New York: Wiley.
- Wolf, P., R. Abbott, and W. Kannel (1987). Atrial fibrillation: a major contributor to stroke in the elderly. The Framingham study. *Archives of Internal Medicine* 147, 1561–4.