Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke

Chris T. Volinsky; David Madigan; Adrian E. Raftery; Richard A. Kronmal

Applied Statistics, Vol. 46, No. 4 (1997), 433-448.

Stable URL:
http://links.jstor.org/sici?sici=0035-9254%281997%2946%3A4%3C433%3ABMAIPH%3E2.0.CO%3B2-K

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Applied Statistics is published by Royal Statistical Society. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Applied Statistics
©1997 Royal Statistical Society

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

http://www.jstor.org/
Tue Oct 14 20:21:26 2003

Appl. Statist. (1997)
46, *No.* 4, *pp.* 433–448

Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke

By CHRIS T. VOLINSKY†, DAVID MADIGAN, ADRIAN E. RAFTERY and RICHARD A. KRONMAL

University of Washington, Seattle, USA

[Received January 1996. Final revision November 1996]

SUMMARY

In the context of the Cardiovascular Health Study, a comprehensive investigation into the risk factors for strokes, we apply Bayesian model averaging to the selection of variables in Cox proportional hazard models. We use an extension of the leaps-and-bounds algorithm for locating the models that are to be averaged over and make available S-PLUS software to implement the methods. Bayesian model averaging provides a posterior probability that each variable belongs in the model, a more directly interpretable measure of variable importance than a *P*-value. *P*-values from models preferred by stepwise methods tend to overstate the evidence for the predictive value of a variable and do not account for model uncertainty. We introduce the *partial predictive score* to evaluate predictive performance. For the Cardiovascular Health Study, Bayesian model averaging predictively outperforms standard model selection and does a better job of assessing who is at high risk for a stroke.

Keywords: Bayesian model averaging; Model uncertainty; Stroke risk factors; Variable selection

1. Introduction

Strokes are the third leading cause of death among adults in the USA. Much recent research has attempted to identify stroke-related risk factors. Some of these factors, such as smoking, are life style choices that can be changed. Others, such as hypertension, are conditions that can be treated non-invasively. Both types of factor 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 a high priority should be placed on finding the risk factors for strokes and thus on identifying the people whose future strokes can be prevented.

Traditional analysis of the time until a stroke identifies risk factors or other independent variables as 'significant' by invoking a procedure for the selection of variables. Subsequent prediction utilizes the single model selected by the procedure.

†*Address for correspondence*: Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195, USA.
E-mail: volinsky@stat.washington.edu

© 1997 Royal Statistical Society 0035–9254/97/46433

Any such procedure ignores model uncertainty. 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 (Draper, 1995; Chatfield, 1995; Madigan and Raftery, 1994). Bayesian model averaging (BMA) selects a subset of all possible models and averages over the subset for all inference and prediction. This paper aims to show that BMA leads to a better evaluation of the risk factors for strokes, as well as improved risk assessment for potential stroke victims.

Section 2 introduces the Cardiovascular Health Study (CHS), a highly censored data set which investigates the risk factors for strokes in elderly people. Section 3 reviews the standard approach to modelling survival data of this type. Section 4 describes BMA as it applies to the selection of variables in the Cox proportional hazard model. We define the subset of model space to be averaged over and we identify the subset efficiently by using a modified leaps-and-bounds algorithm. Section 5 interprets the results of the BMA analysis of the CHS. With two different methods used to assess predictive performance, BMA is shown to be better at prediction and risk assessment than the standard measures. Section 6 discusses some of the interesting results of the study and how they relate to the prevention of strokes, and describes how to access S-PLUS functions to implement these methods. We conclude that taking account of model uncertainty by using BMA can enhance predictive performance in survival analysis.

Although we know of no other work applying BMA to the selection of variables in survival analysis, there has been much 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 could be useful for uncertainty about the model for the baseline hazard in parametric survival analysis. Clyde *et al.* (1995) looked 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^3 algorithm of Madigan and York (1995). This was developed for linear regression but could perhaps be extended to survival analysis.

2. Cardiovascular Health Study

Risk factors for strokes have been extensively studied in the medical literature. In the USA, several population-based studies have identified what are believed to be the main risk factors: smoking, hypertension and irregular heart contractions known as atrial fibrillation (Kannel *et al.*, 1976; Matsumoto *et al.*, 1973). However, the effects of these factors may weaken with increasing age, or even have a protective effect in the elderly (Kannel *et al.*, 1976). The CHS is a longitudinal observational study, funded by the National Institutes of Health, and started in June 1989 to study cardiovascular disease in people aged 65 years and over. The 5201 participants are from four US counties: Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; Allegheny County, Pennsylvania. It is the most comprehensive source of data on stroke incidence in elderly people.

Previous studies have focused on the conventional risk factors mentioned above, and the CHS augments them with non-invasive measurements of subclinical disease

which may be related to heart attacks and strokes. Table 1 shows the explanatory and dependent variables. Fried *et al.* (1991) described the complete sample design and study methods as well as specific protocols for the classification of some of the explanatory variables (e.g. congestive heart failure and abnormal ejection fraction). The data used for this analysis are for 4504 people who were free of stroke at the base-line, and who had complete data. The follow-up was for between 3.5 and 4.5 years, with an average of 4.1 years. Those who had not had a stroke by the end of the study were censored at that time. Only 172 people experienced a stroke during the period of the study.

The CHS contains a high proportion of censored data (96%) and many potential risk factors. We therefore expect considerable model uncertainty. This paper uses BMA to account for this model uncertainty.

3. Standard Methods

Methods for analysing survival data such as the CHS data often focus on modelling the hazard rate, also called the instantaneous failure rate. The most

TABLE 1
Variables for the CHS†

Variable	*Range*	*Mean*
Age (years)	65–100	72
Diuretic use	0 ≡ no, 1 ≡ yes	0.24
Regular aspirin use	0 ≡ no, 1 ≡ yes	0.24
Self-reported atrial fibrillation	0 ≡ no, 1 ≡ yes	0.04
Diastolic blood pressure (mmHg)	0–116	70
Systolic blood pressure (mmHg)	79–230	135
Congestive heart failure	0 ≡ no, 1 ≡ yes	0.02
Creatinine (mg dl^{-1})	0.4–7.3	1.05
Diabetes	0 ≡ normal, 1 ≡ impaired glucose tolerance, 2 ≡ diabetes	0.70
Electrocardiogram atrial fibrillation	0 ≡ no, 1 ≡ yes	0.03
Left ventricular hypertrophy	0 ≡ no, 1 ≡ yes	0.04
Fibrinogen (mg dl^{-1})	109–872	320
Sex	0 ≡ male, 1 ≡ female	0.42
Fasting glucose (mg dl^{-1})	53–657	109
High density lipoprotein (mg dl^{-1})	20–249	54
Low density lipoprotein (mg dl^{-1})	28–340	133
Antihypertension medication	0 ≡ no, 1 ≡ yes	0.43
Insulin (microunits dl^{-1})	3–400	16
Abnormal ejection fraction	0 ≡ no, 1 ≡ yes	0.09
Stenosis of common carotid artery	0 ≡ low, . . ., 3 ≡ high	0.74
History of cardiovascular disease	0 ≡ no, 1 ≡ yes	0.29
Pack-years smoked	0–204	18
Timed walk	0 ≡ fast, 1 ≡ medium, 2 ≡ slow	0.29
Time observed (days)	7–1480	1209
Status	0 ≡ censored, 1 ≡ stroke	0.04

†$n = 4504$ patients; time observed until stroke is the dependent variable; status is the censoring variable.

popular way of doing this is to use the Cox proportional hazards model (Cox, 1972), which allows different hazard rates for people with different covariate vectors and leaves the underlying common base-line hazard rate unspecified. The parameters are estimated and standard errors found by using the partial likelihood (Section 4.1).

Often, when many variables are collected, the Cox regression model is used in conjunction with a selection of variables method to choose the subset of the full variables list which gives the best fit in modelling the time to stroke. The most popular such procedures are *stepwise methods* (Efroymson, 1960)

There are now many references exposing the problems with this general approach. Freedman (1983) and Flack and Chang (1987) demonstrated the tendency of stepwise methods to select pure noise variables as significant, especially when there is multicollinearity among the explanatory variables or when the ratio of the number of variables to the number of records is high. Derksen and Keselman (1992) provided a literature review of studies which critique stepwise methods for linear regression. In short, goodness-of-fit estimates are overestimated and the predictive power of the models is poor.

Altman and Andersen (1989) described a bootstrap experiment of stepwise methods for Cox models. When applied to their full data set (primary biliary cirrhosis), stepwise selection labelled 6 of the 17 explanatory variables as significant, yet different bootstrap samples provided models which included many different variables. Four of the six 'significant' variables were contained in fewer than 75% of the bootstrap models, and each of the 17 variables appeared in some models. Clearly there is model uncertainty here which one application of stepwise methods does not account for. Although Altman and Andersen (1989) compared confidence intervals of various methods, they did not have a direct comparison of predictive performance as we have here.

Alternatives to stepwise methods include ridge regression (Hoerl and Kennard, 1970) and shrinkage models, which use all available explanatory variables but shrink the estimates towards 0. The new biased estimates are often substantial improvements over least squares. Miller (1990) used the ridge trace to reduce the number of explanatory variables and to do subset selection. Another alternative is to investigate all possible subsets of the available explanatory variables, known as *all subsets regression*. When calculating all subsets is not feasible, a branch-and-bound procedure, first used by Beale *et al.* (1967), is often used. Also, there is the widely used leaps-and-bounds algorithm of Furnival and Wilson (1974). Lawless and Singhal (1978) extended leaps and bounds to non-linear models and Kuk (1984) applied this extension to Cox regression. All these methods focus on optimal ways of finding the single best model and neglect model uncertainty.

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

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

where $\lambda_0(t)$ is the base-line hazard function at time t, left unspecified in Cox's formulation, and θ is a vector of unknown parameters.

The estimation of θ is commonly based on the partial likelihood, namely

$$\mathrm{PL}(\theta) = \prod_{i=1}^{n} \left\{ \exp(\mathbf{x}_i^{\mathrm{T}}\theta) \Big/ \sum_{l \in R_i} \exp(\mathbf{x}_l^{\mathrm{T}}\theta) \right\}^{w_i}, \tag{1}$$

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 subject i is censored. Equation (1) assumes that there are no ties between times of death; for simplicity, we do not consider modifications needed when there are ties.

4.2. *Bayesian Model Averaging*

Cox proportional hazards modelling entails selecting explanatory 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 uncertainty about the model itself, and so 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).

The standard Bayesian solution to this problem is to use a set of models, instead of just one, for making all prediction and inference. Let Δ be any quantity of interest such as a future observation or the utility of a course of action. If $\mathcal{M} = \{M_1, \ldots, M_K\}$ denotes the set of all models considered, then the posterior distribution of Δ given the data D is

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

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

For a problem with p potential explanatory variables, the number of models, K, in the finite sum (2) can be enormous ($K = 2^p$ in the absence of other constraints). However, most of these models have very little support from the data. Following Madigan and Raftery (1994), we average only over the best models as an approximation to averaging over all 2^p models, where best is determined by the posterior model probability. Thus, only models belonging to the set

$$\mathcal{A} = \left\{ M_k : \frac{\max_l\{\mathrm{pr}(M_l|D)\}}{\mathrm{pr}(M_k|D)} \leqslant C \right\} \tag{3}$$

should be included in the sum in equation (2). We use $C = 20$ and hence average only over models with posterior model probability at least 1/20 of that of the best model. The value of C will depend on the context of the problem at hand.

Equation (2) has three components, each posing its own computational difficulties, which are discussed in the following sections. The predictive distribution $\mathrm{pr}(\Delta|M_k, D)$ requires integrating out the model parameter θ_k (Section 4.3). The posterior model probabilities $\mathrm{pr}(M_k|D)$ similarly involve the calculation of an integrated likelihood (Section 4.4). Finally, the models that fall into $\mathcal{A}$ must be located and evaluated efficiently (Section 4.5).

4.3. *Predictive Distribution and Maximum Likelihood Estimation Approximation*

In equation (2) the predictive distribution of Δ given a particular model M_k is found by integrating out the model parameter θ_k:

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

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

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

In the context of model uncertainty, this approximation was used 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 Bayesian Information Criterion Approximation*

The posterior probability of model M_k is given by

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

where

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

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

In regular statistical models (roughly speaking, those in which the MLE is consistent and asymptotically normal), the integral in equation (7) can often be approximated via the Laplace method (Raftery, 1996). This method yields

$$\log \text{pr}(D|M_k) = \log \text{pr}(D|\hat{\theta}_k, M_k) - (d_k/2)\log n + O(1), \tag{8}$$

where n is the number of records in the data and d_k is the number of parameters in model M_k. This is the Bayesian information criterion (BIC) approximation derived by Schwarz (1978). In fact, equation (8) is much more accurate for many practical purposes than its $O(1)$ error term suggests. Kass and Wasserman (1995) showed 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 equation (8) is $O(n^{-1/2})$, under certain assumptions, rather than $O(1)$. Raftery (1996) gave further empirical evidence for the accuracy of this approximation.

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. 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 (Raftery *et al.*, 1997; Madigan and Raftery,

1994; Madigan *et al.*, 1996). When prior information about the importance of a variable is available, a prior probability on model M_i can be specified as

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

where $\pi_j \in [0, 1]$ is the prior probability that $\theta_j \neq 0$ (for the CHS, $j = 1, \ldots, 23$) and δ_{ij} is an indicator of whether or not variable j is included in model M_i. Assigning $\pi_j = 0.5$ for all j corresponds to a uniform prior across the model space, whereas $\pi_j < 0.5$ 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, the elicitation of prior probabilities for models is straightforward and avoids the need to elicit priors for a large number of models. For an alternative approach when expert information is available, see Madigan *et al.* (1995).

4.5. *Identifying Models in* $\mathcal{A}$

Our approach requires that we identify the best model and average over only those models with posterior probability of at least $1/C$ of the posterior probability of the best model (equation (3)). We define the best model as the model with the largest BIC, corresponding to the model with the highest posterior model probability. A quick way of screening the models, without fitting them all, and of targeting those that are close in posterior probability to the best model is needed. We can then average over this reduced set of models.

Such a procedure exists for linear regression. Regression by leaps and bounds (Furnival and Wilson, 1974) is an efficient algorithm which 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 M_k returned. 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 eliminates large portions of model space by sweep operations on the matrix

$$\begin{pmatrix} X'X & X'y \\ y'X & y'y \end{pmatrix}. \tag{10}$$

Lawless and Singhal (1978) developed a modification of the leaps-and-bounds algorithm for non-linear regression models, which provides an approximate likelihood ratio test (LRT) statistic (and therefore an approximate BIC value). The method proceeds as follows: let θ be the parameter vector of the full model and let θ_k be the vector for a given submodel k. Rewrite θ_k as (θ_1, θ_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 whereas

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

is an approximation to Λ based on the Wald statistic. Finally, replace the matrix (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

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

As long as q is sufficiently large, this procedure returns the models in $\mathcal{A}$ plus many models that are 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 are at least $1/C'$ of the posterior model probability PMP of the best model, where C' is greater than C. We use $C' = C^2$, which we have found to be sufficiently large virtually to guarantee that no models in $\mathcal{A}$ will be lost.

Kuk (1984) first applied this algorithm to Cox models to find the single best model. We use it to help to locate the models in $\mathcal{A}$ that are to be averaged over. We fit the returned models by a standard survival analysis program, calculate the exact BIC value for each one (which corresponds to a posterior model probability by equation (8)) and eliminate those models that are not in $\mathcal{A}$. The posterior model probabilities are then normalized over the model set. We calculate BMA parameter estimates and standard errors 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: less than 50%, evidence against the effect; 50–75%, weak evidence for the effect; 75–95%, positive evidence; 95–99%, strong evidence; greater than 99%, very strong evidence (Kass and Raftery, 1995).

4.6. *Assessment of Predictive Performance*

We assess the relative values of our method (BMA) and those of competing methods on the basis of their predictive performances. To assess performance, we randomly split the data into two halves and use an analogue of the logarithmic scoring rule of Good (1952). First, we apply each model selection method to the first half of the data, called the *build data* (D^{B}). The corresponding coefficient estimates define a predictive density for each person in the second half of the data (*test data*, or

D^{T}). Then, a log-score for any given model M_k is based on the observed ordinate of the predictive density for the subjects in D^{T}:

$$\sum_{d \in D^{\mathrm{T}}} \log \mathrm{pr}(d|M_k, D^{\mathrm{B}}). \tag{11}$$

Similarly, the predictive log-score for BMA is

$$\sum_{d \in D^{\mathrm{T}}} \log \left\{ \sum_{M \in \mathcal{M}} \mathrm{pr}(d|M, D^{\mathrm{B}}) \, \mathrm{pr}(M|D^{\mathrm{B}}) \right\}, \tag{12}$$

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

However, the Cox model does not directly provide a predictive density. Rather it provides an estimated predictive cumulative density function which is a step function (Breslow, 1975) and therefore does not lead to differentiation into a density. In the spirit of Cox's partial likelihood (1), we have designed an alternative to the predictive density:

$$\mathrm{pr}(d|M_k, D^{\mathrm{B}}) = \left\{ \exp(\mathbf{x}_i^{\mathrm{T}} \hat{\theta}_k) \Big/ \sum_{l \in R_i} \exp(\mathbf{x}_l^{\mathrm{T}} \hat{\theta}_k) \right\}^{w_i}.$$

By substituting this into expressions (11) and (12), we now have an analogue to a log-score called the *partial predictive score* (PPS). Using the PPS, we can compare BMA with any single model selected. The PPS 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 and low risk). We assess predictive discrimination of a single model as follows.

(a) Fit the model to the build data to obtain estimated coefficients $\hat{\theta}$.
(b) Calculate risk scores ($\mathbf{x}_i^{\mathrm{T}} \hat{\theta}$) for each subject in the build data.
(c) Define low, medium and high risk groups for the model by the empirical 33rd and 66th percentiles of the risk scores.
(d) Calculate risk scores for the test data and assign each subject to a risk group.
(e) Extract the subjects who are assessed as being in a higher risk group by one method rather than by another, and tabulate what happened to those subjects over the study period.

To assess predictive discrimination for BMA, we must take account of the multiple models that we average over. We replace the first steps above with the following.

(a′) Fit each model $M_1, \ldots, M_K$ in $\mathcal{A}$ to obtain estimated coefficients $\hat{\theta}_k$.
(b′) Calculate risk scores ($\mathbf{x}_i^{\mathrm{T}} \hat{\theta}_k$) under each model in $\mathcal{A}$ for each person in the build data. A person's risk score under BMA is the weighted average of these:

$$\sum_{k=1}^{K} (\mathbf{x}_i^{\mathrm{T}} \hat{\theta}_k) \, \mathrm{pr}(M_k|D^{\mathrm{B}}).$$

A method is better if it consistently assigns higher risks to the people in the test set who actually had strokes.

5. Application to 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, starting with the variables in Table 1, included the following 10 variables: age, diuretic, aspirin, systolic blood pressure, creatinine, diabetes, atrial fibrillation (by electrocardiogram), stenosis, timed walk and left ventricular (LV) hypertrophy. The model with the highest approximate posterior probability was the same as the stepwise model except that LV hypertrophy was not included.

In the list of models provided by BMA, the stepwise model appears fourth. Table 2 contains the variables included in any of the top five models. Inference about explanatory variables is expressed in terms of the posterior probability that the parameter does not equal 0. Table 3 contains the posterior means, standard deviations and posterior probabilities of the variables.

Fig. 1 shows the posterior probability that each regression coefficient is non-zero, plotted against the corresponding P-value from stepwise selection of variables. 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, P-values arguably overstate the evidence for an effect (Edwards *et al.*, 1963; Berger and Delampady, 1987; Berger and Sellke, 1987).

For the four variables systolic blood pressure, timed walk, diabetes and daily aspirin, the posterior probabilities and the P-values agree that there is very strong evidence for an effect ($P < 0.001$ and $\text{pr}(\theta \neq 0) > 99\%$). For the eight variables in Table 4, however, the two approaches lead to qualitatively different conclusions. Each P-value overstates the evidence for an effect. For the first four of the variables, the P-value would lead to the effect being called 'highly significant' ($P < 0.01$), whereas the posterior probability indicates the evidence to be positive but not strong. For the next two variables (LV hypertrophy and diuretic use), the P-value is 'significant' ($P < 0.05$), but the posterior probabilities indicate the evidence for an effect to be weak. For the last two variables (sex and low density lipoprotein), the P-values are 'marginally significant' ($P < 0.06$), but the posterior probabilities actually indicate (weak) evidence *against* an effect.

TABLE 2
Top five models (in terms of BIC) chosen by BMA, with PMPs†

Model	*Age*	*Diuretic*	*Aspirin*	*Systolic blood pressure*	*Creat-inine*	*Diabetes*	*Atrial fibrilla-tion*	*Ejection*	*Stenosis*	*Timed walk*	*LV hyper-trophy*	*PMP (%)*
1	✓	✓	✓	✓	✓	✓	✓		✓	✓		1.7
2	✓		✓	✓	✓	✓	✓		✓	✓	✓	1.6
3	✓		✓	✓	✓	✓	✓		✓	✓		1.4
4†	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	1.4
5	✓		✓	✓	✓	✓	✓	✓	✓	✓		1.1

†Model chosen by stepwise selection of variables.

TABLE 3
Posterior parameter estimates (means), standard deviations and probabilities that the coefficients are non-zero for the variables in the CHS data set†

Variable	*Mean*	*Standard deviation*	$pr(\theta \neq 0)$
Age	0.04	0.01	89
Diuretic use	0.40	0.17	59
Aspirin	0.52	0.16	99
Self-reported atrial fibrillation	0.62	0.31	18
Diastolic blood pressure	0.01	0.01	0
Systolic blood pressure	0.02	0.00	100
Congestive heart failure	0.79	0.37	24
Creatinine	0.43	0.14	64
Diabetes	0.42	0.10	100
Electrocardiogram atrial fibrillation	0.83	0.30	66
LV hypertrophy	0.60	0.25	50
Fibrinogen	0.00	0.00	0
Sex	0.37	0.18	21
Glucose	0.00	0.00	0
High density lipoprotein	−0.01	0.01	7
Antihypertensives	0.05	0.23	0
Insulin	0.00	0.00	0
Low density lipoprotein	0.00	0.00	24
Ejection	0.46	0.22	26
Stenosis	0.34	0.12	90
Disease	0.24	0.17	5
Smoking	0.00	0.00	0
Timed walk	0.47	0.10	100

†Means and standard deviations are averaged over all models included in the BMA analysis.

TABLE 4
P-values and posterior probabilities that the coefficient is non-zero for eight of the explanatory variables in the CHS

Variable	*P-value*	$pr(\theta \neq 0)$ *(%)*
Creatinine	0.002†	64
Stenosis	0.004†	90
Electrocardiogram atrial fibrillation	0.005†	66
Age	0.007†	89
LV hypertrophy	0.022‡	50
Diuretic use	0.026‡	59
Sex	0.053	21
Low density lipoprotein	0.058	24

†$P < 0.01$.
‡$P < 0.05$.

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 us to make one distinction that *P*-values cannot. We 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.

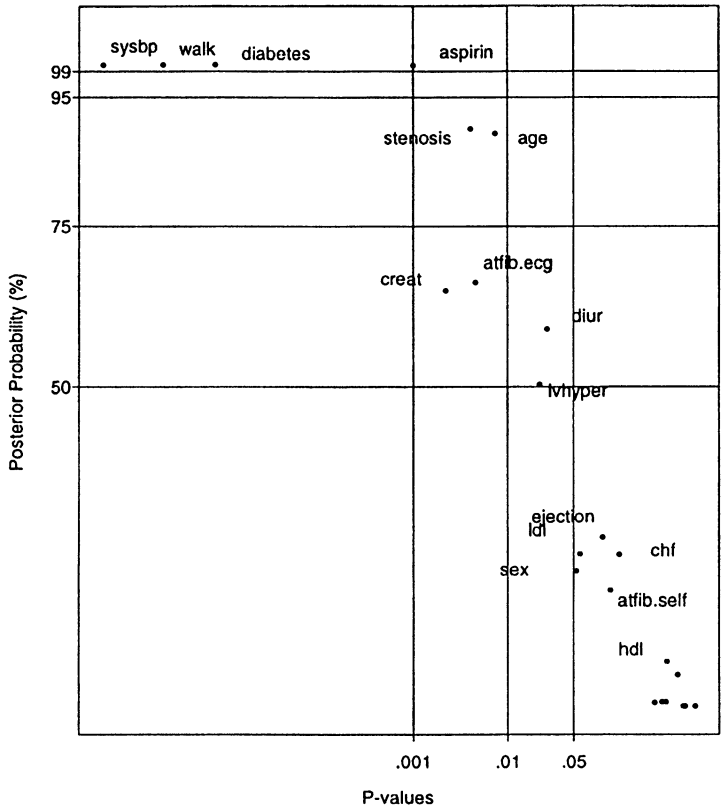

Fig. 1. *P*-values *versus* posterior probabilities that the coefficient is non-zero for the variables in the CHS: the lines at *P*-values of 0.05, 0.01 and 0.001 indicate the standard limits for significant, highly significant and very highly significant; lines at 50%, 75%, 95% and 99% posterior probability indicate weak evidence, positive evidence, strong evidence and very strong evidence

P-values cannot distinguish between these two situations, but posterior probabilities can. Thus, for example, for the ejection fraction, $\mathrm{pr}(\theta \neq 0) = 26\%$, so the data are indecisive, whereas, for diastolic blood pressure, $\mathrm{pr}(\theta \neq 0) = 0\%$, indicating strong evidence *for* the null hypothesis of no effect. The posterior probability of no effect can be viewed as an approximation to the posterior probability of the effect being 'small', namely $\mathrm{pr}(|\theta| < \epsilon)$, provided that ϵ is at most about half of a standard error (Berger and Delampady, 1987).

Six of the variables are known to be risk factors for the general population: age, systolic blood pressure, diabetes, atrial fibrillation, antihypertensive medication and smoking (Kannel *et al.*, 1976). We performed an analysis increasing the prior probability for these variables to 75%. The results showed no change in the conclusion for diabetes and blood pressure (both stayed at 100%) or for smoking and antihypertensive medication (both remained under 10%). However, for atrial fibrillation as measured by electrocardiograms, the posterior probability is 92% (in the original analysis it was 66%). This sensitivity is probably because only about 3% of the subjects had this condition. The posterior probability for age jumps from 90% in the original analysis to 99%.

5.2. *Predictive Performance*

For assessing predictive performance, we randomly split the data into two parts such that an equal number of events (86 strokes) occurred in each part. We compare the results for BMA with those for stepwise model selection and for the single model with the highest posterior probability. Table 5 shows the PPSs for the competing methods; a higher score (less negative) indicates better predictive performance. The top model and stepwise model may be different from those in the previous section since they are built using only half the data.

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%. This means that BMA predicts who is at risk of strokes 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 BMA in another way. Table 6 shows the classification of the 2252 people in the test data and whether or not those people had a stroke in the study period. The people assigned to the high risk group by BMA had more strokes than did those assigned high risk by other methods; similarly those assigned to the low risk group by BMA had fewer strokes. Table 7 summarizes the outcome for the people who were given higher risk by one method over another. The people assigned to a higher risk group by BMA had more strokes and were stroke free for a shorter period of time. Both Table 6 and Table 7 show that BMA is better at indicating which people are at risk of strokes.

TABLE 5
PPSs for the model with the highest PMP, the stepwise chosen model and BMA

Method	*PPS*
Top model	−641.6
Stepwise	−632.8
BMA	−629.9

TABLE 6
Test data cross-classification of assigned risk group versus stroke occurrence†

Assigned risk group	*No. of people under the following classifications:*					
	Model averaging		*Stepwise*		*Top PMP*	
	Stroke free	*Stroke*	*Stroke free*	*Stroke*	*Stroke free*	*Stroke*
Low	751	7	750	8	724	10
Medium	770	24	799	27	801	28
High	645	55	617	51	641	48

†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 BMA, stepwise and top PMP.

TABLE 7
Predictive discrimination of BMA compared with the top PMP model and the stepwise model†

Estimated risk	*No. of strokes*	*No. of subjects*	*Mean survival time*
BMA > top PMP	11	147	1169
BMA < top PMP	1	160	1243
BMA > stepwise	10	165	1191
BMA < stepwise	5	133	1243

†The table is a breakdown of the patients into different risk groups (low, medium or high) for two competing methods. The mean survival time is the mean stroke-free time for the patients in that group.

6. Discussion

Although there are many references regarding the dangers of model selection and of stepwise methods in general, a glance at popular survival analysis texts (Fleming and Harrington, 1991; Kalbfleisch and Prentice, 1980) indicates that the stepwise method is still the *status quo* for the selection of variables in Cox models. Similarly, *P*-values from the final model are the standard reported evidence for a variable's importance. By applying BMA to Cox models, we have accounted for the model uncertainty ignored by stepwise methods, and we have provided a function `bic.surv` to facilitate the application. BMA shifts the focus away from selecting a single model. Whereas stepwise methods lead to hard inclusion or exclusion of each variable, the posterior probability that the parameter is non-zero supplied by BMA can be more informative.

Also, 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 fraction' is not in either the stepwise selected model or the model with the highest posterior probability, but the posterior probability that the parameter is non-zero is an indecisive 0.26, indicating that it should not necessarily be discarded in future work. In fact it is included in any prediction with weight proportional to its posterior probability. This indecisive result may be because of a lack of power in the data rather than because the variable does not have appreciable predictive value.

The PPS allows a direct comparison of predictive performance of one Cox model with another, and for the CHS it shows increased predictive performance for BMA over model selection techniques. Altman and Andersen (1989) claimed that the choice of model used for prediction 'appears of limited importance' in their example. On the contrary, for the CHS the predictive discrimination analysis appears to show that a substantive difference can be made in evaluating who is at high risk for strokes. Thus, it may ultimately allow better targeting of people who are susceptible to strokes so that preventive resources can be allocated 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 equation (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 explanatory variables to include in the model. There are other components also including uncertainty about functional forms of the explanatory variables and uncertainty about the Cox model itself. Our approach could be extended to take account of those, as has been studied for linear regression by Raftery *et al.* (1997).

The benefits of using BMA to account for model uncertainty have now been assessed for several different classes of model, e.g. linear regression, exponential survival models, logistic regression and discrete graphical models. The results of these studies were summarized in Raftery *et al.* (1995). For each class model averaging improved predictive performance, by amounts that ranged from modest to substantial. We emphasize that the use of BMA for any model class should be backed up by standard diagnostic checks of goodness of fit, as averaging over a model class is only beneficial when that model class is appropriate.

We have written a series of S-PLUS functions using the approximations outlined here to implement BMA for proportional hazard models. The function `bic.surv` performs the BMA and outputs all information about the models selected, whereas `summary.bs` condenses this information to output only posterior model probabilities, posterior variable probabilities and the variables included in the selected models. The functions allow user specification of parameters to limit the number of models returned and of priors for independent variables. All the software can be obtained by sending the message 'send bic.surv from S' to `statlib@stat.cmu.edu` or at the universal resource locator: `http://lib.stat.cmu.edu/S/bic.surv` on the World Wide Web.

Acknowledgements

We thank the US 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 is funded by contracts NO1-HC-85079–NO1-HC-85086 from the National Heart, Lung and Blood Institute.

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM*. Oxford: Clarendon.

Altman, D. G. and Andersen, P. K. (1989) Bootstrap investigation of the stability of a Cox regression model. *Statist. Med.*, **8**, 771–783.

Beale, E. M. L., Kendall, M. G. and Mann, D. (1967) The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357–366.

Berger, J. O. and Delampady, M. (1987) Testing precise hypotheses. *Statist. Sci.*, **2**, 317–352.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis (with discussion). *J. Am. Statist. Ass.*, **82**, 112–122.

Breslow, N. (1975) Analysis of survival data under the proportional hazards model. *Int. Statist. Rev.*, **43**, 45–48.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference (with discussion). *J. R. Statist. Soc.* A, **158**, 419–466.

Clyde, M., DeSimone, H. and Parmigiani, G. (1995) Prediction via orthogonalized model mixing. *J. Am. Statist. Ass.*, **91**, 1197–1208.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Derksen, S. and Keselman, H. (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br. J. Math. Statist. Psychol.*, **45**, 262–282.

Draper, D. (1995) Assessment and propagation of model uncertainty. *J. R. Statist. Soc.* B, **57**, 45–97.
Edwards, W., Lindman, H. and Savage, L. J. (1963) Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193–242.
Efroymson, M. (1960) Multiple regression analysis. In *Mathematical Methods for Digital Computers* (eds A. Ralston and H. Wilf). New York: Wiley.
Flack, V. F. and Chang, P. C. (1987) Frequency of selecting noise variables in subset regression analysis: a simulation study. *Am. Statistn*, **41**, 84–88.
Fleming, T. R. and Harrington, D. H. (1991) *Counting Processes and Survival Analysis*. New York: Wiley.
Freedman, D. A. (1983) A note on screening regression equations. *Am. Statistn*, **37**, 152–155.
Fried, L. P., Borhani, N. O., Enright, P., Furberg, C. D., Gardin, J. M., Kronmal, R. A., Kuller, L. H., Manolio, T. A., Mittelmark, M. B., Newman, A., O'Leary, D. H., Psaty, B., Rautaharju, P., Tracy, R. P. and Weiler, P. G. (1991) The Cardiovascular Health Study: design and rationale. *Ann. Epidem.*, **1**, 263–276.
Furnival, G. M. and Wilson, R. W. (1974) Regression by leaps and bounds. *Technometrics*, **16**, 499–511.
George, E. and McCulloch, R. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.
Good, I. J. (1952) Rational decisions. *J. R. Statist. Soc.* B, **14**, 107–114.
Gorelick, P. D. (1995) Stroke prevention. *Arch. Neurol.*, **52**, 347–355.
Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
Kalbfleisch, J. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*. New York: Wiley.
Kannel, W. B., McGee, D. and Gordon, T. (1976) A general cardiovascular risk profile: the Framingham study. *Am. J. Cardiol.*, **38**, 46–51.
Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.
Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses with large samples. *J. Am. Statist. Ass.*, **90**, 928–934.
Kuk, A. Y. C. (1984) All subsets regression in a proportional hazards model. *Biometrika*, **71**, 587–592.
Lawless, J. and Singhal, K. (1978) Efficient screening of nonnormal regression models. *Biometrics*, **34**, 318–327.
Madigan, D., Andersson, S. A., Perlman, M. and Volinsky, C. T. (1996) Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communs Statist. Theory Meth.*, **25**, 2493–2520.
Madigan, D., Gavrin, J. and Raftery, A. E. (1995) Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communs Statist. Theory Meth.*, **24**, 2271–2292.
Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's Window. *J. Am. Statist. Ass.*, **89**, 1535–1546.
Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, 215–232.
Matsumoto, N., Whisnant, J. P., Kurland, L. T. and Okazaki, H. (1973) Natural history of stroke in Rochester, Minnesota, 1955 through 1969: an extension of a previous study, 1945 through 1954. *Stroke*, **4**, 20–25.
Miller, A. J. (1990) *Subset Selection in Regression*. London: Chapman and Hall.
Raftery, A. E. (1996) Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251–266.
Raftery, A. E., Madigan, D. and Hoeting, J. (1993) Bayesian model averaging in linear regression models. *J. Am. Statist. Ass.*, **92**, 179–191.
Raftery, A. E., Madigan, D. and Volinsky, C. T. (1995) Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 323–349. Oxford: Oxford University Press.
Regal, R. and Hook, E. B. (1991) The effects of model selection on confidence intervals for the size of a closed population. *Statist. Med.*, **10**, 717–721.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Taplin, R. H. (1993) Robust likelihood calculation for time series. *J. R. Statist. Soc.* B, **55**, 829–836.
Taplin, R. H. and Raftery, A. E. (1994) Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Biometrics*, **50**, 764–781.