

Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance *

Adrian E. Raftery, David Madigan and Chris T. Volinsky
Department of Statistics, GN-22
University of Washington
Seattle, WA 98195, USA.

June 1, 1994; revised August 1, 1994

Abstract

Survival analysis is concerned with finding models to predict the survival of patients or to assess the efficacy of a clinical treatment. A key part of the model-building process is the selection of the predictor variables. It is standard to use a stepwise procedure guided by a series of significance tests to select a single model, and then to make inference conditionally on the selected model. However, this ignores model uncertainty, which can be substantial.

We review the standard Bayesian model averaging solution to this problem and extend it to survival analysis, introducing partial Bayes factors to do so for the Cox proportional hazards model. In two examples, taking account of model uncertainty enhances predictive performance, to an extent that could be clinically useful.

1 Introduction

From 1974 to 1984 the Mayo Clinic conducted a double-blinded randomized clinical trial involving 312 patients to compare the drug DPCA with a placebo in the treatment of primary biliary cirrhosis of the liver (PBC) (Dickson *et al.*, 1989; Grambsch *et al.*, 1989; Markus *et al.*, 1989; Fleming and Harrington, 1991). DPCA turned out not to have a substantial effect (Dickson *et al.*, 1985), but the data on other variables are still useful, because they allow one to develop a natural history model of the disease. Such a model is useful for prediction (counseling patients and predicting the course of PBC in untreated patients) and inference (historical control information to assess new therapies). Fleming and Harrington (1991)

*This work was supported by ONR Contract no. N00014-91-J-1074. The authors are grateful to Jennifer Hoeting for helpful discussions.

developed such a model. Starting with 14 candidate independent variables, they selected a Cox regression model with five of the variables.

The analysis of Fleming and Harrington (1991) certainly represents the state of the art in survival analysis. However, we argue here that model selection procedures such as theirs ignore model uncertainty, and can thus underestimate uncertainty about quantities of interest, leading to decisions that are riskier than one thinks they are. In problems such as this of variable selection in regression, the number of models initially considered can be huge. Fleming and Harrington (1991, p. 160) implicitly considered $2^{18} \approx 260,000$ models (in addition to the 14 independent variables, transformations of four of them were considered), and selected a single one, conditioning all subsequent inferences on the selected model. We will show that the ignored model uncertainty is large.

In Section 2 we review the standard Bayesian model averaging solution to this problem and in Section 3 we show how practical difficulties with it can be overcome using the Occam’s Window algorithm. We claim that accounting for model uncertainty in this way improves out-of-sample predictive performance, and discuss how to assess this claim.

Survival analysis is hard mainly because observations are often censored. One model that can accommodate this is log-linear exponential regression (Prentice, 1973), and in Section 4 we show how to account for model uncertainty in this model class. This is illustrated by a lung cancer clinical trial data set.

In Section 5 we consider the Cox (1972) proportional hazards model. Previous methods do not apply directly because this is typically a semi-parametric model, and partial likelihood is used rather than the full likelihood on which Bayesian methods rely. We introduce the concept of “partial Bayes factors” to implement Bayesian model averaging in this context, and develop the “partial predictive score” to assess performance. In Section 5.3 we return to the PBC data set and apply these ideas to it.

Our conclusion is that taking account of model uncertainty using Bayesian model averaging can indeed enhance predictive performance in survival analysis. However, the methods described here leave room for improvement, and in Section 6 we discuss some of the open questions and relate our results here to other work in the literature.

2 Bayesian Model Averaging

2.1 General Framework

A typical approach to data analysis is to carry out a model selection exercise leading to a single “best” model and to then make inference as if the selected model were the true model. However, this 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 Miller (1984, 1990), Regal and Hook (1991), Madigan and York (1993), Raftery (1993a), and Kass and Raftery (1994).

There is a standard Bayesian solution to this problem. If $\mathcal{M} = \{M_1, \dots, M_k\}$ denotes the set of all models considered and if Δ is the quantity of interest such as a future observation

or the utility of a course of action, then the posterior distribution of Δ given the data D is

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

(Leamer, 1978). This is an average of the posterior distributions under each model weighted by the corresponding posterior model probabilities. In (1), the predictive distribution of Δ given a particular model M_k is

$$\text{pr}(\Delta | M_k, D) = \int \text{pr}(\Delta | \theta_k, M_k, D) \text{pr}(\theta_k | M_k, D) d\theta_k, \quad (2)$$

and the posterior probability of model M_k is given by

$$\text{pr}(M_k | D) = \frac{\text{pr}(D | M_k) \text{pr}(M_k)}{\sum_{l=1}^K \text{pr}(D | M_l) \text{pr}(M_l)}, \quad (3)$$

where

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

is the integrated likelihood of model M_k , θ_k is the vector of parameters of model M_k , $\text{pr}(\theta_k | M_k)$ is the prior density of θ_k under model M_k , $\text{pr}(D | \theta_k, M_k)$ is the likelihood, and $\text{pr}(M_k)$ is the prior probability that M_k is the true model. All probabilities are implicitly conditional on \mathcal{M} , the set of all models being considered.

This strategy has several elements that must be specified or computed, each of which can present a difficulty. We discuss these elements in the following sections. They are as follows.

- (i) The prior distribution of the parameters in each model, $\text{pr}(\theta_k | M_k)$, used in (4) (Section 2.2).
- (ii) The posterior model probabilities in (3) (Section 2.2): The required integrals are often of high dimension and can be hard to compute. We use a simple approximation that corresponds to a reasonable choice of prior parameter distributions.
- (iii) The prior model probabilities, $\text{pr}(M_k)$ (Section 2.3): We make the simple choice of giving all models equal prior probability and argue that this is reasonable. We also discuss the elicitation of more informative priors.
- (iv) The predictive distribution of the quantity of interest for each model, $\text{pr}(\Delta | M_k, D)$ in (1) (Section 2.4): This is defined by the integral (2), which can be formidable to evaluate. We use a simple approximation.
- (v) The sum in (1): The number of models, K , can be too big for direct evaluation of the sum to be feasible, and two practical solutions are discussed in Section 3.

- (vi) The most used model in survival analysis is the Cox proportional hazards model, for which the distribution of the observations, and hence the likelihood, are not fully specified; partial likelihood is used instead. This poses a problem both for posterior model probabilities and for assessing predictive performance. In Section 5 we develop simple solutions in the spirit of partial likelihood.

Together, these provide a workable strategy for taking account of model uncertainty in survival analysis. Elements (i)–(iv) can be found in Taplin (1990) and element (v) in Madigan and Raftery (1991, 1994) and Madigan and York (1993) in the context of other applications.

2.2 Posterior Model Probabilities

In regression models for survival analysis, analytic evaluation of the integrals (4) is not possible except in some very simple special cases. Thus 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 best way to approximate the integral (4) is usually via the Laplace method. The Laplace method for integrals is based on a Taylor series expansion of the real-valued function $f(u)$ of the d -dimensional vector u , and yields the approximation

$$\int e^{f(u)} du \approx (2\pi)^{d/2} |A|^{-1/2} \exp\{f(u^*)\}, \quad (5)$$

where u^* is the value of u at which f attains its maximum, and A is minus the inverse Hessian of f evaluated at u^* . When applied to equation (4) it yields

$$p(D|M_k) = (2\pi)^{d_k/2} |\Psi_k|^{-1/2} \text{pr}(D|\tilde{\theta}_k, M_k) \text{pr}(\tilde{\theta}_k|M_k) O(n^{-1}), \quad (6)$$

where d_k is the dimension of θ_k , $\tilde{\theta}_k$ is the posterior mode of θ_k , and Ψ_k is minus the inverse Hessian of $h(\theta_k) = \log\{\text{pr}(D|\theta_k, M_k)\text{pr}(\theta_k|M_k)\}$, evaluated at $\theta_k = \tilde{\theta}_k$.

This can be simplified by noting that when n is large, $\tilde{\theta}_k \approx \hat{\theta}_k$, where $\hat{\theta}_k$ denotes the MLE, and $\log |\Psi_k| = -d_k \log n + O(1)$, where d_k is the dimension of model M_k , i.e. the number of independently estimable parameters. Thus

$$\log \text{pr}(D|M_k) = \log \text{pr}(D|\hat{\theta}_k, M_k) - d_k \log n + O(1), \quad (7)$$

a result derived by Schwarz (1978) in another way. In (7), n 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). This is because Ψ_k is asymptotically equal to the inverse of the observed information matrix, which in turn is asymptotically equal to the number of deaths times a constant matrix, for both exponential regression and the Cox proportional hazards model (Kalbfleisch and Prentice, 1980, p.49 and equation (4.10)).

In fact, (7) is much more accurate for many practical purposes than its $O(1)$ error term suggests. For the evaluation of (1), particularly when it is approximated using Occam's

Window (Section 3.1), what matters most are comparisons between models with similar posterior probabilities. These models are often fairly similar to each other (in the sense of sharing many of the same independent variables), and are often nested, or nearly so. Model comparisons are based on the *Bayes factor*, $B_{jk} = \text{pr}(D|M_j)/\text{pr}(D|M_k)$, and (7) will produce an $O(1)$ error in the approximation to $\log B_{jk}$, in general.

However, Kass and Wasserman (1992) have shown, roughly speaking, 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 $\log B_{jk}$ with (7) is $O(n^{-\frac{1}{2}})$ rather than $O(1)$. This holds under the “local alternative” assumption, namely that the posterior under M_k is concentrated close to the value assumed under M_j . This will be the case for most comparisons of interest to us; when it does not hold, the Bayes factor will quickly become decisive and the less likely model will typically contribute little (or, in the Occam’s Window case, nothing) to the sum in (1). Empirical evidence for the accuracy of (7) in generalized linear models is given by Raftery (1993a).

2.3 Prior Model Probabilities

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. When the number of models is small or moderate, this is intuitively appealing and understandable. When the number of models is very large, however, one could worry about whether this choice might have unintended perverse consequences, as has happened before with other “uniform” priors. 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 such perverse effects (Raftery, Madigan and Hoeting, 1993; Madigan and Raftery, 1994; Madigan *et al.*, 1994). In addition, we have found inference using Occam’s Window (Section 3) to be quite robust to moderately large changes in prior model probabilities (e.g. halving or doubling the prior odds).

Spiegelhalter *et al.* (1993) and Lauritzen *at al.* (1994) provide a detailed analysis of the benefits of incorporating informative prior distributions in Bayesian knowledge-based systems, and demonstrate improved predictive performance with informative priors. If it is available, prior information can easily be taken into account by adjusting the prior model probabilities.

In variable selection problems, prior information often takes the form of prior evidence for the inclusion of a variable, rather than for an individual model. Suppose that this is the only kind of prior information available and that model M_k is specified by a vector $(\delta_{k1}, \dots, \delta_{kd})$, where $\delta_{ki} = 1$ if the i -th variable is included and 0 if not. Then, if π_i is the prior probability that the i -th variable has an effect, and if the prior information about different variables is approximately independent, it is reasonable to specify

$$\text{pr}(M_k) \propto \prod_{i=1}^d [\pi_i^{\delta_{ki}} (1 - \pi_i)^{(1-\delta_{ki})}] \quad (8)$$

(Madigan and Raftery, 1991).

Prior information often takes the form of strong evidence from previous studies for the inclusion of a particular variable. If the previous studies combined are much more informative about the variable than the present one, then it is a reasonable approximation to set the corresponding $\pi_i = 1$ in equation (8), and hence to consider only models that include that risk factor. This approximation is reasonable in the sense that the true prior probability π_i is so close to 1 that the posterior probability would also be close to 1, almost regardless of the data at hand. Thus setting $\pi_i = 1$ yields a posterior distribution of quantities of interest close to what would result from using the true π_i (which is very close to, but not quite equal to, 1). This provides a formal rationale for the common practice of “controlling” for particular independent variables even when the data at hand provide little evidence for their inclusion in the model.

In our experience with moderate to large data sets, we have found it sufficient to restrict attention to prior model probabilities of the form (8) with $\pi_i = 0, \frac{1}{2}$ or 1. This might not be enough, however, for small data sets such as small clinical trials, where more careful assessment and elicitation of prior model probabilities might be needed.

Madigan *et al.* (1994) proposed an elicitation method for Bayesian model averaging which allows domain experts to express knowledge in terms of observable quantities. The method starts with a uniform prior distribution on model space, updates it using imaginary data provided by the domain expert, and then uses the updated prior distribution as the prior distribution for the Bayesian analysis. Ibrahim and Laud (1994) and Laud *et al.* (1992) adopt a somewhat similar approach, but in the context of linear models. Gavasakar (1988) uses imaginary data to elicit a prior distribution for a binomial parameter. Madigan *et al.* (1994) demonstrate that incorporation of prior opinion in this manner provides a modest improvement in out-of-sample predictive performance for a challenging graphical models example.

Our experience is that observable quantities are easier for domain experts to think about than abstract entities such as parameters and models. See Kadane *et al.* (1980) for a persuasive presentation of this viewpoint. Researchers have developed methods for eliciting informative distributions via observable quantities for several different modeling situations, for example, the Bernoulli process (Winkler, 1967; Chaloner and Duncan, 1983), the normal linear model (Kadane *et al.*, 1980; Garthwaite and Dickey, 1990 and 1991), the analysis of variance (Laskey and Black, 1989), generalized linear models (Laud *et al.*, 1992), and survival analysis (Chaloner *et al.*, 1993).

2.4 The MLE Approximation and Components of Uncertainty

In (1) we use the relation

$$\text{pr}(\Delta|M_k, D) \approx \text{pr}(\Delta|M_k, \hat{\theta}_k, D), \quad (9)$$

to approximate the integral in (2). This was used in the model uncertainty context by Taplin (1990) who found it to give an excellent approximation in his time series regression problem;

it was subsequently used by Taplin (1993), Taplin and Raftery (1991, 1994) and Draper (1994).

Further research is needed to understand when and why it works well, but the following heuristics, based on a decomposition of the posterior variance, $\text{Var}[\Delta|D]$, may provide an initial framework for studying it. Let M denote a model chosen at random from \mathcal{M} according to the posterior distribution $\text{pr}(M|D)$, let θ_M denote its parameter, and let E , Var , E_M , Var_M , E_{θ_M} and Var_{θ_M} denote expectations and variances over sample space, model space and parameter space, respectively. All that follows in this subsection is conditional on the data D , and so for simplicity we drop D from the notation.

We have

$$\text{Var}[\Delta] = E_M[\text{Var}(\Delta|M)] + \text{Var}_M[E(\Delta|M)], \text{ and} \quad (10)$$

$$\text{Var}[\Delta|M] = E_{\theta_M}[\text{Var}(\Delta|\theta_M, M)] + \text{Var}_{\theta_M}[E(\Delta|M)]. \quad (11)$$

Substituting (11) into (10) gives

$$\begin{aligned} \text{Var}[\Delta] &= E_M[E_{\theta_M}\{\text{Var}(\Delta|\theta_M, M)\}] + E_M[\text{Var}_{\theta_M}\{E(\Delta|\theta_M, M)\}] + \text{Var}_M[E_{\theta_M}\{E(\Delta|\theta_M, M)\}] \\ &= U_S + U_P + U_M \\ &= \left(\begin{array}{c} \text{sampling} \\ \text{uncertainty} \end{array} \right) + \left(\begin{array}{c} \text{parameter} \\ \text{uncertainty} \end{array} \right) + \left(\begin{array}{c} \text{model} \\ \text{uncertainty} \end{array} \right), \end{aligned} \quad (12)$$

an expression similar to Draper (1994, equation (13)).

Heuristic calculations suggest that in the regression variable selection case, $U_S = O(1)$, $U_P = O(n^{-1})$, and $U_M = O(dn^{-1})$, where d is the number of candidate independent variables. The MLE approximation (9) involves setting $U_P = 0$, i.e. ignoring parameter uncertainty while taking account of model uncertainty. This is most reasonable when there is much more model uncertainty than parameter uncertainty, and our heuristics suggest that this is most likely to be the case when d is large.

The approximation also involves replacing U_S by $\hat{U}_S = E_M[\text{Var}(\Delta|\hat{\theta}_M, M)]$ and U_M by $\hat{U}_M = \text{Var}_M[E(\Delta|\hat{\theta}_M, M)]$, in each case introducing further $O(n^{-1})$ errors. It follows that the approximation is likely to be most accurate if $f(\theta_M) = E[\Delta|\theta_M, M]$ and $g(\theta_M) = \text{Var}(\Delta|\theta_M, M)$ are relatively insensitive to the precise value of θ_M , or, when n is large, if $E_M[f'(\hat{\theta}_M)^2]$ and $E_M[g'(\hat{\theta}_M)]$ are small.

3 Implementation and Evaluation

3.1 Implementation: Occam's Window and MC³

When the number of candidate independent variables, d , is large, the number of models, K , in (1) is enormous ($K = 2^d$ in the absence of other constraints), in which case (1) cannot be evaluated directly. Here we review two ways of getting around this difficulty.

The first way involves applying the Occam's Window algorithm of Madigan and Raftery (1994). Two basic principles underly this approach. First, we argue 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 models not belonging to

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l \{\text{pr}(M_l | D)\}}{\text{pr}(M_k | D)} \leq C_1 \right\}, \quad (13)$$

should be excluded from equation (1) where C_1 is chosen by the data analyst. In the examples we used $C_1 = 20$.

Second, appealing to Occam's razor, we exclude models which receive less (or much less) support from the data than any of their simpler submodels. Formally, we also exclude from (1) models belonging to:

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}, M_l \subset M_k, \frac{\text{pr}(M_l | D)}{\text{pr}(M_k | D)} > C_2 \right\}. \quad (14)$$

Typically $1 \leq C_2 \leq C_1$. We have found two versions of Occam's Window to be useful: (a) a parsimonious version in which $C_2 = 1$, and (b) a symmetric version in which $C_2 = C_1$.

Equation (1) is then replaced by

$$\text{pr}(\Delta | D) = \frac{\sum_{M_k \in \mathcal{A}} \text{pr}(\Delta | M_k, D) \text{pr}(D | M_k) \text{pr}(M_k)}{\sum_{M_k \in \mathcal{A}} \text{pr}(D | M_k) \text{pr}(M_k)} \quad (15)$$

where

$$\mathcal{A} = \mathcal{A}' \setminus \mathcal{B}. \quad (16)$$

This greatly reduces the number of models in the sum in equation (1). Typically, in our experience, the parsimonious version of Occam's Window reduces the number of terms in (1) to fewer than 25, and often to as few as two. The symmetric version of Occam's Window usually involves a larger but still manageable set of models. Madigan and Raftery (1994) describe an efficient search algorithm that can be used to find the models in Occam's Window when the number of models initially considered is too large to calculate all their posterior model probabilities.

The second approach is to approximate (1) using the Markov chain Monte Carlo model composition (MC³) approach of Madigan and York (1993). MC³ generates a stochastic process which moves through model space. We do not use this in the examples below, but it could easily be adapted to the survival analysis context.

3.2 Assessing Predictive Performance

We use the predictive ability of the selected models for future observations to measure the effectiveness of a model selection strategy. Our specific objective is to compare the quality of the predictions based on model averaging with the quality of predictions based on any single model that an analyst might reasonably have selected.

To measure performance we randomly split the complete data set into two subsets. We evaluate (1) directly, and also (15) via Occam's Window, using one subset of the data, called

the build set, and denoted by D^B . We evaluate performance using the test data set made up of the remaining half of the data, $D^T = D \setminus D^B$.

Within this framework, we assess predictive performance using the logarithmic scoring rule of Good (1952) which is based on the conditional predictive ordinate (Geisser, 1980). Specifically, we measure the predictive ability of an individual model, M , with its predictive score, namely

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

Similarly, we measure the predictive performance for model averaging with

$$\sum_{d \in D^T} \log \left\{ \sum_{M \in \mathcal{A}} \text{pr}(d \mid M, D^B) \text{pr}(M \mid D^B) \right\}, \quad (17)$$

where \mathcal{A} is the full set of models, or, for Occam's Window, the set of selected models. In (17) we use the approximation (9).

For the Cox proportional hazards model we use a modification of this method, described below in Section 5.2.

4 Model Uncertainty in Exponential Regression with Censoring

4.1 Exponential Regression with Censoring

Let $f(t)$ be the probability density function of survival times, $F(t)$ the cumulative distribution function, $S(t) = 1 - F(t)$ the survivor function, and $h(t) = f(t)/S(t)$ the hazard rate. In the exponential regression model, the hazard rate is constant for each individual, but depends on covariates. We will adopt the log-linear formulation $h_i(t) = \exp(\mathbf{x}_i^T \beta)$, where \mathbf{x}_i is the vector of covariates for the i th individual.

We now consider estimation. Suppose we have data $(t_1, w_1), \dots, (t_n, w_n)$, where t_i is the length of time for which the i th individual was observed, $w_i = 1$ if he or she died, and $w_i = 0$ if the observation was censored. In the absence of censoring, this is a generalized linear model with dependent variable t_i , independent variables \mathbf{x}_i , logarithmic link and exponential (or gamma) error.

In the presence of censoring, this is no longer the case, but it can still be written as a generalized linear model, albeit one with a completely different form. Now the dependent variable is w_i , the independent variables are \mathbf{x}_i , the link function is logarithmic, the error is Poisson, and there is an offset, $\log t_i$ (Aitkin *et al.*, 1989, p. 270). MLEs can then be found using any generalized linear model estimation program, and the approximate approach to accounting for model uncertainty of Sections 2 and 3 can be implemented.

4.2 Example: Lung Cancer Clinical Trial

4.2.1 Background

We reanalyzed the Veteran’s Administration Lung Cancer Trial data reported by Prentice (1973). This randomized clinical trial was designed to assess a test chemotherapy treatment. It dealt with advanced inoperable cases, and most patients were followed until death; only 9 of the 137 survival times are censored.

Five independent variables were considered for inclusion in the model. These included one treatment variable (test or standard chemotherapy), and four variables intended to control for heterogeneity between patients, namely age, whether or not prior therapy had been received, performance rating as measured by the Karnofsky score, and type of tumor (a categorical variable with four categories: squamous, small cell, large cell, adeno). We use the following notation: c = cell type, t = treatment, k = Karnofsky score, a = age, and p = prior therapy.

Prentice (1973) and Kalbfleisch and Prentice (1980) concluded that these data are well fit by an exponential regression model with log-linear dependence of the hazard rate on the covariates. We therefore adopt this model class and restrict attention here to the selection of independent variables.

4.2.2 Results for the full data set

The results for the full data set are shown in Table 1. The variables highlighted as significant by Kalbfleisch and Prentice (1980, p.61), namely c and k , are also those that appear in the model with the highest posterior probability, ck , which has 63% of the total posterior probability. There is model uncertainty, however: the symmetric Occam’s Window contains five of the 32 models, which together account for 95% of the total posterior probability. There is not enough data to decisively rule out the possibility that t , a or p have an effect, but the evidence is against it. The parsimonious Occam’s Window contains only two models, ck and k .

The posterior probability that a coefficient is non-zero is equal to the sum of the posterior probabilities of the models that include the coefficient. These are shown in the bottom panel of Table 1, both summing over all models and only over those in Occam’s Window; the results from the two methods are similar. The c and k coefficients are likely to be non-zero, while the corresponding posterior probabilities for t , a and p are small. In particular, the posterior probability that the treatment is effective is just 13% (or 11% from Occam’s Window).

4.2.3 Predictive performance

To assess predictive performance, we proceed as described in Section 3.2. We randomly split the data in half and use the build set to develop prediction rules corresponding to individual models and to different forms of model averaging. We then evaluate and compare these different prediction rules by calculating the predictive score of each one for the test data.

Table 1: Lung cancer trial: Results from full data. PMP denotes the posterior model probability. Only the five models in the symmetric Occam’s Window are shown.

Model	c	t	k	a	p	PMP	Deviance	df
ck	•		•			.63	139.1	132
k			•			.10	157.4	135
ctk	•	•	•			.10	138.0	131
cka	•		•	•		.06	138.9	131
ckp	•		•		•	.06	138.9	131
$\Pr_{\text{MA}}[\beta_i \neq 0]$.86	.13	1.00	.09	.09			
$\Pr_{\text{Occam}}[\beta_i \neq 0]$.89	.11	1.00	.06	.06			

Table 2: Lung cancer trial: Predictive results for split data. PMP denotes the posterior model probability from the build data set. Pred Score denotes the predictive score from the test data set, plus 350. Only the eight models in the symmetric Occam’s Window are shown.

Model	c	t	k	a	p	PMP	Deviance	df	Pred Score
ck	•		•			.46	55.3	62	−11.4
k			•			.17	69.6	65	−11.5
ckp	•		•		•	.08	54.7	61	−12.7
ctk	•	•	•			.07	55.0	61	−11.1
cka	•		•	•		.06	55.2	61	−11.3
kp			•		•	.05	68.1	64	−12.9
ka			•	•		.03	68.8	64	−13.2
tk		•	•			.03	69.3	64	−12.3

Based on the build data alone, eight of the $2^5 = 32$ models considered have posterior model probabilities within a factor of 20 of the most likely single model and so fall within the symmetric Occam’s Window with $C_1 = C_2 = 20$. Together they account for 95% of the total posterior probability; see Table 2. There are more models in the symmetric Occam’s Window for the build data set than for the full data set; this is because less data is used and so there is more model uncertainty.

As for the full data set, the most likely single model is *ck*, which would also be selected by any reasonable forwards, backwards or stepwise procedure at either the 5% or the 1% level, or by the AIC. Thus all model selection procedures in common use would agree on the best *single* model to be selected. Nevertheless, the best model accounts for less than half of the total posterior probability, and substantial model uncertainty remains.

Predictive scores for the various prediction methods considered are shown in Table 3. Model averaging performs better than *any* of the 32 models individually, including the “best”

Table 3: Lung cancer trial: Predictive performance comparison. 350 has been added to the predictive scores.

Method	Predictive Score
Best single model (<i>ck</i>)	-11.4
Model averaging (all models)	-9.8
Occam’s Window	-9.8

model, *ck*. The symmetric Occam’s Window, which includes eight of the 32 models, performs as well as averaging over all models.

The improvement in predictive score for Occam’s Window over the best single model is $\delta = 1.6$ points. The test data has $n_{\text{test}} = 70$ cases, and so this means that, on average, the predictive probability of what was actually observed was bigger on average for model averaging than for the best single model by a factor of $\exp(\delta/n_{\text{test}}) = 1.023$, or by about 2.3%.

This improvement can be obtained at little cost to the user, and may be clinically useful, as the following “biased coin” analogy shows. Suppose that we have to estimate the probability of success in a Bernoulli trial when the true probability is π . In the absence of any other information, we would guess $\hat{\pi} = 1/2$, for which our expected predictive score per observation would be $\pi \log \hat{\pi} + (1 - \pi) \log(1 - \hat{\pi}) = -\log 2$. Getting π exactly right (i.e. guessing $\hat{\pi} = \pi$), yields an improvement in predictive score per observation of $\pi \log \pi + (1 - \pi) \log(1 - \pi) + \log 2$, and the value of π for which this is equal to δ/n_{test} gives an intuitive interpretation of the gain in predictive ability due to model averaging. For the lung data this is $\pi = .61$, an improvement of 0.11 on the biased coin prediction scale. This is fairly large, which is all the more striking since model selection seems clearcut in these data and one would not necessarily expect model averaging to improve things by much.

In order to check that these results were not due to the particular random split used, we redid the analysis for a further 100 random splits. In each case we calculated the difference between the predictive scores for model averaging (both averaging over all 32 models and using the symmetric Occam’s Window) and for the model with the highest posterior probability. The results are shown in Table 4, and show that the split which we have reported in detail gave results well within the range of the 100 splits. Model averaging had better predictive ability than the single model with highest posterior probability 81 times out of 100 when all models were averaged over, and 69 times out of 100 when Occam’s Window was used.

Table 4: Lung cancer trial: Predictive performance comparison for 100 random splits.

Predictive Score Difference	Min	1st Quartile	Median	3rd Quartile	Max	% Positive
Model averaging – Best model	−2.0	0.1	0.9	2.2	7.6	81
Occam’s Window – Best model	−2.0	−0.1	0.7	2.0	7.9	69

5 Model Uncertainty in the Cox Proportional Hazards Model

5.1 The Cox Proportional Hazards Model and Partial Bayes Factors

In the Cox (1972) proportional hazards model, the hazard rate for the i th individual is $h_i(t) = \lambda_0(t) \exp(\mathbf{x}_i^T \beta)$, in the notation of Section 4.1, where $\lambda_0(t)$ is an unknown baseline hazard rate, common to all individuals. Estimation of β is commonly based on the partial likelihood, namely

$$PL(\beta) = \prod_{i=1}^n \left(\frac{\exp(\mathbf{x}_i^T \beta)}{\sum_{\ell \in R_i} \exp(\mathbf{x}_\ell^T \beta)} \right)^{w_i}, \quad (18)$$

where R_i is the set of individuals at risk at time t_i (often called the risk set) (Cox, 1972, 1975). Equation (18) 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 this here.

The parameter for the model is $\theta = (\beta, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = \{\lambda_0(t) : t \in \mathbf{R}_+\}$. Here we use the partial likelihood (18) as the likelihood for β with $\boldsymbol{\lambda}$ integrated out, and replace (4) by

$$\text{pr}(D|M_k) = \int PL(\beta_k) \text{pr}(\beta_k|M_k) d\beta_k. \quad (19)$$

We refer to ratios of the form $B_{jk} = \text{pr}(D|M_j)/\text{pr}(D|M_k)$ as *partial Bayes factors* when (19) is used, because they are based on the partial likelihood rather than the full likelihood.

There are two heuristic justifications for the approximation (19). The first is due to Kalbfleisch (1978), who showed that if the prior for $\boldsymbol{\lambda}$ is a diffuse gamma process, then, to a first order approximation,

$$\text{pr}(\beta|D) = \int \text{pr}(\beta, \boldsymbol{\lambda}|D) d\boldsymbol{\lambda} \approx K PL(\beta), \quad (20)$$

(Kalbfleisch and Prentice, 1980, equation (8.33)). Thus (19) amounts approximately to integrating over $\boldsymbol{\lambda}$ as well as β_k to form the integrated likelihood, modulo the constant K in (20). This constant cancels when posterior model probabilities and Bayes factors are computed, because these involve ratios of integrated likelihoods.

The second justification is that the partial likelihood (18) actually becomes a full likelihood if a part of the data is discarded, namely the times at which deaths occur. It is a full likelihood for the part of the data consisting of the *order* in which individuals die and the risk sets, R_i , corresponding to each death. When derived this way it is often referred to as the marginal likelihood (Savage, 1957; Kalbfleisch and Prentice, 1973). (This is not to be confused with the integrated likelihood of (4), which is sometimes also called the marginal likelihood). Then replacing (4) by (19) amounts to using the standard model uncertainty formulation of Section 2, the only difference being that a reduced data set is used. The part of the data discarded is typically rather uninformative about β , which is what the competing models differ most about (Efron, 1977; Oakes, 1977).

Although they did not give an explicit justification, Ritter and Tanner (1992) and Carlin *et al.* (1993) also used the partial likelihood in place of the full likelihood for Bayesian estimation of the Cox proportional hazards model.

5.2 Assessing Performance: The Partial Predictive Score

The predictive score of Section 3.2 cannot be computed easily for Cox models because it requires the predictive *density* of survival times in the test set, which is not explicitly estimated as part of the Cox model fitting process. While the survivor function, $S(t)$, can be estimated fairly directly (e.g. Breslow, 1975), estimating the density (or equivalently the hazard rate itself) is much trickier.

Instead, in the spirit of partial likelihood, we avoid the problem altogether by introducing the partial predictive score. The basic idea is that we measure how well a prediction rule predicts which of the individuals in the test data set who are in the risk set at a failure time dies, given that one of them does. This is in contrast to the predictive score of Section 3.2, which measures how well each survival time in the test data set is predicted. Thus we measure the predictive ability of a particular Cox proportional hazards model M by its partial predictive score, namely

$$\sum_{t_j \in D^T} \log \Pr[\text{individual dies at } t_j | R_j^{(T)}, \text{ death in test set at } t_j, M, D^B],$$

and the predictive performance of model averaging by

$$\sum_{t_j \in D^T} \log \left\{ \sum_{M \in \mathcal{A}} \Pr[\text{individual dies at } t_j | R_j^{(T)}, \text{ death in test set at } t_j, M, D^B] \Pr(M | D^B) \right\}.$$

5.3 Example: The PBC Data

5.3.1 Data

We now return to the PBC data discussed earlier in Section 1. Of the 312 patients, we omitted two because of missing data. Of the remaining 310 patients, 124 were followed until death and the other 186 observations were censored. Of the 14 variables initially considered

by Fleming and Harrington (1991) — hereafter FH — six clearly had no effect, and for simplicity we restricted attention to the following eight:

Age Age in years

Edema Presence of edema (0=no edema, 0.5=edema resolved with therapy, 1=edema not resolved with therapy)

bili Serum bilirubin (mg/dl)

albu Albumin in gm/dl

copp Urine copper in $\mu\text{g}/\text{day}$

SGOT SGOT in U/dat (aspartate aminotransferase)

thromb Prothrombin time, in seconds

hist Histologic stage of disease (graded 1, 2, 3 or 4)

Also, following FH, we used logarithmic transformations of bili, albu and thromb. We thus initially considered $2^8 = 256$ models.

FH used a backwards elimination variable selection method, and concluded that the best model was the one with the five independent variables age, edema, bili, albu and thromb.

5.3.2 Results for the full data set

The model with the highest posterior probability is the same as that selected by FH using a different method; see Table 5. In spite of this agreement, there is substantial model uncertainty: the best model has only 13% of the total posterior probability and is not decisively better than its nearest rivals. The symmetric and parsimonious Occam's Windows contain 21 and 5 models, respectively.

From the bottom panel of Table 5 we see that the data provide strong evidence for the predictive usefulness of bili, age and albu, and positive evidence for that of edema and thromb; these are the five variables in the best single model. The data appear indecisive about the usefulness of copp and hist and, on balance, to provide evidence against the inclusion of SGOT.

5.3.3 Predictive performance

As before, we split the data randomly into two equal parts, the build set and the test set. Partial predictive scores for the top ten models (as determined from the build set) are shown in Table 6. The model with the highest posterior probability is the same for the build set as for the entire data set, but the other results are not the same. We also applied the backwards elimination procedure of FH to the build data, and the selected model is the third one in Table 6, which contains all the variables in the model with the highest posterior probability,

Table 5: PBC trial: Results for full data set. PMP denotes the posterior model probability. Only the 10 models with the highest PMP values are shown.

	Age	Edema	bili	albu	copp	SGOT	thromb	hist	PMP	Log Lik	df
	•	•	•	•			•		.13	-534.1	305
	•	•	•	•	•		•		.11	-531.8	304
	•	•	•	•			•	•	.07	-532.3	304
	•	•	•	•				•	.07	-534.8	305
	•	•	•	•	•		•	•	.06	-530.1	303
	•	•	•	•	•			•	.06	-532.6	304
	•	•	•	•		•	•		.05	-530.3	303
	•		•	•			•		.05	-537.5	306
	•		•	•	•		•		.04	-535.2	305
	•	•	•	•		•	•		.04	-532.8	304
$\text{Pr}_{\text{MA}}[\beta_i \neq 0]$.98	.81	1.00	.96	.43	.25	.72	.48			
$\text{Pr}_{\text{Occam}}[\beta_i \neq 0]$	1.00	.82	1.00	1.00	.42	.23	.74	.46			

with SGOT in addition. The symmetric and parsimonious Occam’s Windows contained 17 and 5 models respectively, based on the build data set.

The predictive scores for the different prediction rules are shown in Table 7. Once again, model averaging performs better than the single-model methods. Model averaging beats the model with the highest posterior probability by 3 points of partial predictive score, for $n_{\text{test}} = 155$. Thus the predicted probability of what was observed was 2% higher for model averaging than for the most likely single model, which corresponds to an improvement of about 0.10 on the “biased coin” scale of Section 4.2.3.

To check that our results were not unduly influenced by the particular random split used, we replicated the experiment ten times with different random splits. The results for all ten experiments were similar to those for the one we have reported. We also checked that none of the six omitted variables would, if included, have improved any of the models with the highest posterior probabilities.

6 Discussion

We have described a general strategy for taking account of model uncertainty in survival analysis and given two examples where it gives better predictive performance than single models selected by standard (Bayesian or frequentist) methods. Our two examples were previously analyzed by eminent survival analysts and there is general agreement in each case about what the best single model is. Nevertheless there is model uncertainty, and taking account of it improves predictive performance to an extent that could be clinically useful.

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 (7). The models considered here can be written as generalized linear models (Aitkin *et al.*, 1989),

Table 6: PBC trial: Predictive results from the split data. PMP denotes the posterior model probability from the build data; only the models with the top ten PMP values are shown. Pred Score denotes the partial predictive score from the test data, plus 190.

Variable								PMP	LogLik	df	Pred Score
Age	Edema	bili	albu	copp	SGOT	thromb	hist				
•	•	•	•			•		.22	-228.7	150	-6.6
•	•	•				•		.08	-231.8	151	-8.7
•	•	•	•		•	•		.07	-227.8	149	-6.0
•	•	•	•					.07	-232.0	151	-5.0
•	•	•	•			•	•	.06	-227.8	149	-5.4
•	•	•	•				•	.06	-230.0	150	-3.1
•	•	•			•	•		.05	-230.2	150	-8.7
•	•	•					•	.05	-232.3	151	-1.3
•	•	•				•	•	.04	-230.3	150	-5.3
•	•	•			•	•	•	.04	-228.3	149	-4.5

Table 7: PBC trial: Predictive performance comparison from split data. 190 has been added to the partial predictive scores.

Method	Partial Predictive Score
Highest posterior probability model	-6.6
Backwards elimination model (as FH)	-6.0
Model averaging	-3.6
Occam's Window	-3.9

Table 8: Summary of improvements in predictive performance from model averaging (via Occam’s Window), relative to the model with highest posterior probability.

Data	Model	δ	n_{test}	% increase in pred. prob.	% improvement (biased coin)
1. Coronary risk factors	Discrete graphical	29.8	1381	2.2	10
2. Women and mathematics	Discrete graphical	5.0	892	0.6	6
3. Scrotal swellings	Discrete graphical	11.1	224	5.1	16
4. Crime and punishment	Linear regression	11.0	23	61.3	45
5. Lung cancer trial	Exponential regression	1.6	70	2.3	11
6. PBC trial	Cox regression	2.7	155	1.8	10

NOTE: δ is the improvement in (partial) predictive score;

n_{test} is the number of individuals in the test data set;

% increase in pred. prob. = $100(\exp(\delta/n_{\text{test}}) - 1)$; and

% improvement (biased coin) is its equivalent on the biased coin scale of Section 4.2.3.

SOURCES: Data sets 1, 2, 3: Madigan and Raftery (1994); Data set 4: Raftery, Madigan and Hoeting (1993); Data sets 5 and 6: this article.

and so the more accurate approximations of Raftery (1993a) should be applicable. Also, the MLE approximation (9) to the predictive distribution could be improved, perhaps by using a Laplace approximation to the integral (2) or by a Monte-Carlo method. Finally, automating the Occam’s Window and MC³ algorithms for survival analysis would allow bigger problems to be tackled, such as the full 14-variable version of the PBC problem.

We have considered only one component of model uncertainty: which independent variables to include in the model. There are other components also: uncertainty about functional forms (e.g. which transformations of the independent variables to use), and, in the case of fully parametric survival analysis, uncertainty about the model for the baseline hazard rate function. Our approach could be extended to take account of those.

The benefits of taking account of model uncertainty have now been assessed for several different model classes. The results for six data sets analyzed to date are summarized in Table 8 using the absolute increase in predictive score as well as its interpretation in terms of the increase in average predicted probability, and the “biased coin” scale discussed in Section 4.2.3. In each case model averaging improved predictive performance, by amounts that range from modest to substantial.

We know of no other work on taking account of model uncertainty in survival analysis. However, there is much work on model uncertainty in general that might be applicable to survival analysis; see Kass and Raftery (1994) and Draper (1994) for reviews. Draper’s (1994) 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. Several standard failure time models (exponential, Weibull, gamma) are special cases of Stacy’s (1962) generalized gamma distribution, which could be used for continuous model expansion; the generalized F distribution of Prentice (1975)

would be an even more general choice. George and McCulloch (1993) have developed the Stochastic Search Variable Selection (SSVS) method, which is similar in spirit to the MC³ algorithm of Madigan and York (1993). This was developed for linear regression but could probably be extended to survival analysis.

References

- Aitkin, M., Anderson, D.A., Francis, B. and Hinde, J. (1989). *Statistical Modelling in GLIM*. Oxford: Clarendon Press.
- Breslow, N. (1975). Analysis of survival data under the proportional hazards model. *Int. Statist. Rev.* **43**, 45–58.
- Carlin, B.P., Chaloner, K.M., Louis, T.A., Rhame, F.S. (1993) Elicitation, monitoring, and analysis for an AIDS clinical trial. Tech Rep 93-004 Uni. of Minnesota, Division of Biostatistics. To appear in *Case Studies in Bayesian Statistics, vol. 2* (C. Gatsonis *et al.*, eds.), New York: Springer-Verlag.
- Chaloner, K.M. and Duncan, G.T. (1983) Assessment of a beta prior distribution. *The Statistician*, **32**,174–180.
- Chaloner, K.M., Church, T., Louis, T.A., and Matts, J.P. (1993) Graphical elicitation of a prior distribution for a clinical trial. *The Statistician*, **42**,341–353.
- Cox, D.R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc. B* **34**, 187–220.
- Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.
- Dickson, E.R. *et al.* (1985). Trial of penicillamine in advanced primary biliary cirrhosis. *New England J. Med.* **312**, 1011–1015.
- Dickson, E.R. *et al.* (1989). Prognosis in primary biliary cirrhosis: Model for decision-making. *Hepatology* **10**, 1–7.
- Draper, D. (1994). Assessment and propagation of model uncertainty (with Discussion). *Journal of the Royal Statistical Society B*, **56**, to appear.
- Efron, B. (1977). Efficiency of Cox’s likelihood function for censored data. *J. Amer. Statist. Ass.* **72**, 557–565.
- Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Garthwaite, P.H. and Dickey, J.M. (1990) Quantifying expert opinion in linear regression models. *Journal of the Royal Statistical Society (Series B)*, **50**,462–474.
- Garthwaite, P.H. and Dickey, J.M. (1991) An elicitation method for multiple linear regression models. *Journal of Behavioral Decision Making*, **4**,17–31.
- Gavasakar, U. (1988) A comparison of two elicitation methods for a prior distribution for a binomial parameter. *Management Science*, **34**,784–790.
- George, E.I. and McCulloch, R.E. (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 Statistical Society (Series B)*, **14**, 107–114.

- Grambsch, P.M. *et al.* (1989). Extramural cross-validation of the Mayo primary biliary cirrhosis survival model establishes its generalizability. *Hepatology* **10**, 846–850.
- Ibrahim, J.G. and Laud, P.W. (1994) A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, to appear.
- Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. (1980) Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, **75**, 845–854.
- Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *J. R. Statist. Soc. B* **40**, 214–221.
- Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox’s regression and life model. *Biometrika* **60**, 267–278.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kass, R.E. and Raftery, A.E. (1994). Bayes factors. *Journal of the American Statistical Association*, to appear.
- Kass, R.E. and Wasserman, L. (1992b). A reference Bayesian test for nested hypotheses with large samples. Technical Report no. 567, Department of Statistics, Carnegie Mellon University.
- Laskey, K.B. and Black, P.K. (1989) Models for elicitation in Bayesian analysis of variance. In *Computer Science and Statistics: Proceedings of the Eight Conference on the Interface*, 242–247.
- Laskey, K.B. (1993) Sensitivity analysis for probability assessments for Bayesian networks. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, D. Heckerman and A. Mamdani, (Eds.), Morgan Kaufman, San Mateo, 136–137.
- Laud, P.W., Ibrahim, J.G., Gopalan, R., and Ramgopal, P. (1992) Predictive variable selection in generalized linear models. *Technical Report*, Division of Statistics, Northern Illinois University.
- Lauritzen, S.L., Thiesson, B., and Spiegelhalter, D.J. (1994) Diagnostic systems created by model selection methods - A case study. In *Selecting Models from Data: AI and Statistics IV* (P. Cheeseman and R.W. Oldford, eds.), New York: Springer-Verlag, pp. 143–152.
- Leamer, E.E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.
- Madigan, D., Gavrin, J., and Raftery, A.E. (1994) Enhancing the predictive performance of Bayesian graphical models. Technical Report 270, Department of Statistics, University of Washington.
- Madigan, D. and Raftery, A.E. (1991) Model selection and accounting for model uncertainty in graphical models using Occam’s window. Technical Report no. 213, Department of Statistics, University of Washington.
- Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association*, to appear.
- Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M., and Almond, R.G. (1994). Strategies

- for graphical model selection. In *Selecting Models from Data: AI and Statistics IV* (P. Cheeseman and R.W. Oldford, eds.), New York: Springer-Verlag, pp. 91-100.
- Madigan, D. and York, J. (1993) Bayesian graphical models for discrete data. Technical Report no. 259, Department of Statistics, University of Washington.
- Miller, A.J. (1984) Selection of subsets of regression variables (with Discussion). *J. R. Statist. Soc. (ser. A)*, 147, 389–425.
- Miller, A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.
- Oakes, D. (1977). The asymptotic information in censored survival data. *Biometrika* **64**, 441–448.
- Prentice, R.L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika* **60**, 279–288.
- Prentice, R.L. (1975). Discrimination among some parametric models. *Biometrika* **62**, 607–614.
- Raftery, A.E. (1993a). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Technical Report no. 255, Department of Statistics, University of Washington.
- Raftery, A.E., Madigan, D.M. and Hoeting (1993). Model selection and accounting for model uncertainty in linear regression models. Technical Report no. 262, Department of Statistics, University of Washington.
- 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.
- Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *J. Amer. Statist. Ass.* **87**, 861–868.
- Savage, I.R. (1957). Contributions to the theory of rank order statistics. The “trend” case. *Ann. Math. Statist.* **28**, 968–977.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
- Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., and Cowell, R.G. (1993) Bayesian analysis in expert systems. *Statistical Science*, **8**, 219–283.
- Stacy, E.W. (1962). A generalization of the gamma distribution. *Ann. Math. Statist.* **33**, 1187–1192.
- Taplin, R.H. (1990). Modeling agricultural field trials in the presence of outliers and fertility jumps. *Ph.D. Dissertation*, Department of Statistics, University of Washington.
- 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. (1991). Analysis of agricultural field trials in the presence of outliers and fertility jumps. Technical Report 218, Department of Statistics, University of Washington.
- Taplin, R.H. and Raftery, A.E. (1994). Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Biometrics*, to appear.
- Winkler, R.L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, **62**, 1105–1120.

Discussion

E.I. GEORGE (*University of Texas, USA*)

TeX version not received (hard copy attached).

M. A. CLYDE, H. DeSIMONE, G. PARMIGIANI (*Duke University, USA*)

We feel that the problem addressed by the authors is extremely important and we agree very much with the general philosophy of their work. Posterior model probabilities can be a very effective tool in communicating model uncertainty to clients. In our experience, we find it very difficult to estimate such posterior probabilities accurately, both based on Occam's window and on MCMC (we use a hybrid of MC³ and George and McCulloch's 1994 integrated SSVS in Clyde and Parmigiani, 1994). In particular, we suggest that posterior probabilities obtained by renormalizing only over the models in Occam's window, or even over models in relatively large MC samples, may be substantially higher than the actual values, especially for data sets with a moderate or large number of predictors.

To illustrate this, we considered the crime data of Raftery, Madigan and Hoeting (1993) (RMH), where it is not hard to evaluate all models by enumeration. We compared the estimates obtained by normalizing the model probabilities within Occam's window with the exact ones, and obtained values for the most likely model of .067 and 0.012 respectively. We used the Occam cutoffs of RMH, and applied them to the full set of models. This, and a slightly different choice of prior distribution, results in a larger number of models in the window compared to RMH. In complex problems it is very common to have a large mass of probability outside the window, or outside the MC sample: in the crime data, the probability of being in Occam's window is .181, based on our prior distribution. In general, however, the predictive performance of model mixing does not seem to be severely affected by this; in MC algorithms, in particular, the predictive distributions appear to converge much faster than the probabilities over model space.

D. DRAPER (*University of Bath, UK*)

I have two comments on this excellent paper.

- First, it is important in judging the practical effects of model uncertainty to measure these effects on substantively meaningful scales. I have not found the log scoring method employed here and previously by these authors to be very useful in this regard, and the biased coin analogy introduced in this paper does not seem to help much, since it transfers the discourse from survival times to a less relevant domain. One idea that might work in the context of survival analysis would be to pretend at random that some of the uncensored survival times were censored, construct predictive distributions for them using the single "best" model and model averaging, and calculate predictive z -scores $((\text{actual} - \text{predictive mean})/(\text{predictive SD}))$ on (say) the log survival times and actual versus nominal coverage of predictive intervals. Showing that the single "best"

model produces predictive z -scores with SD 1.4 (say) and nominal 90% intervals that only predict correctly 75% of the time, with corresponding values 1.1 and 85% (say) for model averaging, would be a more direct demonstration of the calibrative improvement arising from propagation of model uncertainty.

- Second, it seems more natural in variable selection to regard the problem as one of continuous uncertainty about the entire β vector, rather than a series of discrete decisions about whether components of β are zero or not. Readers of this paper may imagine what the marginal posterior distributions for components of β would look like with the authors' method (mixtures of point mass at 0 and continuous distributions on the real line) and ask themselves if this is typically an accurate scientific summary of what is known about β in light of the data. The continuous posterior on the full β vector obtained by hierarchically incorporating clinical prior information on the direction and monotonic ordering of the standardized coefficients (e.g., Draper, 1994) would seem to provide a more satisfying solution to the joint problems of inference and prediction, and utility considerations (e.g., Lindley, 1968) weighing factors such as data collection costs for the predictors and parsimony preferences against predictive performance will then cause variables to drop out if they don't predict well enough given how much they cost. An inferential and predictive comparison of this approach with that of the authors would be interesting.

S.E. FIENBERG (*Carnegie Mellon University, USA*)

The authors' approach is interesting and, at least heuristically, has appealing properties. Nonetheless, the examples raise a number of as yet unaddressed issues. In the lung cancer trial example, there is a key independent variable – chemotherapy treatment – and the question of interest is whether it has a “causal” effect on survival. Further, the meaning of the coefficient in the exponential regression model or the Cox regression model changes as we add and delete variables.

Suppose you have the results of a new randomized clinical trial on the basis of which the manufacturer is going to seek approval to market a new drug from the U. S. Food and Drug Administration (FDA). Your colleagues have measured 50 covariates as well as data on treatment and outcome and ask you to look at the effects of treatment on survival via a Cox model and to prepare a report for the FDA. Given your model averaging approach, how would you address the issue of the causal effects of the new drug treatment in a manner that the statisticians at the FDA would find convincing?

R.E. KASS (*Carnegie Mellon University, USA*)

In this paper and others the authors have contributed substantially to what is in principle an appealingly simple approach to an important problem. They have, however, omitted from their discussion an issue that is central to much epidemiological and biostatistical practice.

The typical survival problem involves a treatment variable together with a set of covariates. The coefficient of the treatment variable is the main quantity of interest, being interpreted as the effect of the treatment after adjustment for the effects of covariates included in the specific model under consideration. When models are averaged we lose this simple interpretation: within each model the coefficient has a different meaning (because its meaning depends on the covariates appearing in that model), so when we average the posteriors of these coefficients across models we get a marginal distribution of the treatment coefficient that has a rather subtle interpretation whose relevance is less clear.

General remarks about the wisdom of reporting conditional as well as marginal results (e.g., Box and Tiao, 1973, pp. 71-72), and about difficulties in treatment effect interpretation (e.g., Mosteller and Tukey, 1977, Chapter 13) apply here, and one must be careful in practice not to neglect them simply because they may become less obvious with model averaging.

In conversation after his presentation, Adrian Raftery reminded me that the marginal posterior distribution of the treatment coefficient takes account of the complete collection of covariates in the study (via assignment of prior probability one-half on inclusion of each covariate). If this is to be emphasized in interpreting results then the usual phrase describing the (conditional) treatment coefficient for a specific model as “the effect of the treatment after adjustment for covariate effects” might be modified, in describing the (marginal) distribution after averaging, to something like “the effect of the treatment after adjustment for *the possibility of* covariate effects.” The authors may wish to elaborate on this point in their reply.

J.W. PRATT (*Harvard University, USA*)

LaTeX version not received (hard copy attached).

R.L. WINKLER (*Duke University, USA*)

LaTeX version not received (hard copy attached).

Reply to the Discussion

We are grateful to all nine discussants for their stimulating and supportive comments, and also to those who made equally useful spoken comments but did not submit written versions (Nick Patterson, Joe Sedransk and Arnold Zellner).

Concern was voiced by some (especially in the spoken discussion) about the cumulative effect of our approximations. We would remind them that the measures of predictive performance provide an assessment of the entire procedure, approximations and all. The point was also made that our procedure is conditional on a set of models, but that real data are unlikely to be generated exactly by one of these models. However, we assessed our procedure using real, not simulated, data.

Treatment Effects

Fienberg and Kass both raised the important issue of inference about treatment effects. It is common practice in epidemiological studies to measure a large number of covariates, *to select a subset of these*, and then to base inference on the single model with the treatment effect and the selected covariates. Fleming and Harrington (1991, p. 162) do this for the PBC data in their Table 4.4.4, headed “Adjusted estimation of treatment effect”. Although this is a state of the art analysis, we would contend that it is not a fully adjusted estimate, because it fails to take account of model uncertainty. It is *not* desirable or commonly recommended to automatically include all measured covariates in the model when these are very numerous, which they often are (see Fienberg’s comment).

Inference about treatment effects is straightforward in our approach. How likely is it that the treatment has an effect¹? The answer is $\Pr[\beta_{\text{treatment}} \neq 0 | D]$. For the lung cancer example this is 13% from Table 1 (or 11% using Occam’s Window). For the PBC data, we calculated it by adding the treatment (DPCA) to the list of variables considered, and obtained 18% (8% from Occam’s Window). The conclusion that there is little evidence for a treatment effect is confirmed.

It might be objected that the question “how likely is it that there is a treatment effect?” does not precisely express what we want to know. It might be argued that we “know” that there always is a treatment effect (however small); the question is whether it is large enough to be useful. However, Berger and Delampady (1987) have shown that the latter question is also addressed by the posterior probability of the point null hypothesis in the sense that $\Pr[\beta_{\text{treatment}} \neq 0 | D] \approx \Pr[|\beta_{\text{treatment}}| < \varepsilon | D]$. This holds as long as ε is less than about one-half of a standard error, which is often big enough to be realistic.

How big is the treatment effect, given that there is one? The answer is given by the posterior effect of $\beta_{\text{treatment}}$, namely

$$\text{pr}(\beta_{\text{treatment}} | D) = \sum_{\mathcal{T}} \text{pr}(\beta_{\text{treatment}} | M_k, D) \text{pr}(M_k | D), \quad (21)$$

where $\mathcal{T} = \{M_k : \beta_{\text{treatment}} \neq 0\}$. This follows from equation (1). It can be summarized by the posterior mean and standard deviation of $\beta_{\text{treatment}}$, which can be approximated by

$$\begin{aligned} E[\beta_{\text{treatment}} | D] &\approx \sum_{\mathcal{T}} \hat{\beta}_{\text{treatment},k} \text{pr}(M_k | D), \\ \text{SD}^2(\beta_{\text{treatment}} | D) &\approx \sum_{\mathcal{T}} \left(\text{se}_k^2 + \hat{\beta}_{\text{treatment},k}^2 \right) \text{pr}(M_k | D) - E[\beta_{\text{treatment}} | D]^2, \end{aligned}$$

where $\hat{\beta}_{\text{treatment},k}$ and se_k are the MLE of $\beta_{\text{treatment}}$ and its standard error under model M_k (e.g. Raftery, 1993b).

In Table 9 we show the results for the PBC data; these can be compared with Table 4.4.4 of Fleming and Harrington (1991). The most likely single set of covariates in our analysis is

¹As Pratt pointed out, we use the word “effect” somewhat loosely, and it would be more accurate to ask whether the treatment is associated with survival after controlling for measured covariates. However, for brevity we stick with “effect”.

Table 9: PBC Trial: Posterior results for $\hat{\beta}_{treatment}$ given its inclusion in the model.

Top 5 PMP Models	$\hat{\beta}_{treatment}$	SD($\hat{\beta}_{treatment}$)	PMP
(F-H Model)	.1594	.1865	.149
	.1324	.1871	.113
	.1458	.1866	.075
	.1538	.1870	.072
	.1157	.1874	.051
Model Averaging	.1298	.1894	
Occam's Window	.1423	.1882	

NOTE: PMP denotes posterior model probability (given that treatment is included in the model). The top five models are the same as the top five models in Table 5 with the treatment effect also included. The top model is the same as the one in Fleming and Harrington (1991, Table 4.4.4) but the results are slightly different because our results are based on two fewer cases than theirs because of missing data.

the same as the set for which Fleming and Harrington adjusted. In this case the estimated treatment effect is robust to model choice, and taking account of model uncertainty increases the standard error of the treatment effect only slightly (model uncertainty accounts for only about 3% of the total posterior variance). Even in this very robust case, however, there are sizeable differences between the estimated treatment effects under different plausible models.

The difference between our approach and the standard one of selecting a single set of covariates is that we have adjusted for the *possibility* of the effects of *all* the covariates. Here the numerical difference is small, but that is not always the case; for an instance in which the difference is considerable see Raftery (1993a).

Kass mentioned one possible objection to this solution, namely that $\beta_{treatment}$ has different meanings in different models (depending on what covariates are included), and so it does not make sense to combine inferences from different models. We do not find this to be a fatal flaw, and indeed Kass summarized our response to it.

Equation (21) can be viewed in two ways, the first of which is the one we have emphasized, as a mixture across different models. This is the one that Kass sees as hard to interpret. It can also, however, be viewed as the posterior distribution from the single model with all covariates, but with a prior distribution that assigns probability $\frac{1}{2}$ to each coefficient being equal to zero. When viewed this way, we see no problem with the interpretation of $\text{pr}(\beta_{treatment}|D)$ in equation (21), as the posterior distribution distribution of $\beta_{treatment}$ controlling for *all* the covariates, but allowing for the possibility that they have no effect. Kass's suggested terminology for it, "the effect of the treatment after adjustment for the *possibility* of (all) the covariate effects" seems useful and appropriate.

“True Models”

In his thought-provoking comments, Winkler raises the issue of how $\text{pr}(M_k|D)$ should be interpreted. The standard interpretation is that it is “the probability that M_k is the true model given the data”. In our application, however, we do not necessarily believe that any of the models considered is literally the “true model” (although we do think there is a good chance that at least one of them is a good approximation), and so the standard interpretation is questionable.

We are conditioning on a set of models, $\cup_{k=1}^K M_k$, and the modified interpretation “the probability that M_k is the true model given the data *and* given that one of M_1, \dots, M_k is the true model” seems more palatable. Most scientific inference is conditional on models, and the difference between single-model methods and our model averaging approach is that we condition on a set of models rather than just one.

A different interpretation views equation (1) as a predictive distribution in a way that avoids the concept of “true model”. Each model is viewed as a way of generating a predictive distribution, and equation (1) is seen as a combined predictive distribution in which the predictive distributions considered are averaged, with weights proportional to their predictive scores based on the available data.

We have $\text{pr}(M_k|D) \propto \text{pr}(D|M_k)\text{pr}(M_k)$. We now argue that this can be interpreted as a predictive score, because both $\text{pr}(D|M_k)$ and $\text{pr}(M_k)$ can be viewed in this way, and thus so can their product.

The integrated likelihood, $\text{pr}(D|M_k)$, can be interpreted as a predictive score, as follows (Kass and Raftery, 1994). Suppose that $D = \{y_1, \dots, y_n\}$ and that, for each i , we form a predictive distribution $\hat{\text{pr}}_i(\cdot)$ of y_i given the already available data $\{y_1, \dots, y_{i-1}\}$. As in our paper, we use Good’s (1952) logarithmic score, $\log \hat{\text{pr}}_i(y_i)$, to measure the quality of the predictive distribution $\hat{\text{pr}}_i(\cdot)$. Then the overall score of any rule that generates such predictive distributions is $LS = \sum_i \log \hat{\text{pr}}_i(y_i)$. In particular, if the prediction rule is derived from the model M_k (i.e. likelihood and prior), then $\log \text{pr}(D|M_k) = \sum_i \log \text{pr}(y_i|y_{i-1}, \dots, y_1, M_k) = LS_k$, so that the logarithmic predictive score is precisely the log-integrated likelihood. Thus the integrated likelihood can be viewed as measuring the success of M_k at predicting the data. This is related to *prequential analysis* (Dawid, 1984) and also to *stochastic complexity* (Rissanen, 1987); the connections are discussed by Dawid (1992) and Hartigan (1992).

The prior model probabilities, $\text{pr}(M_k)$, can be viewed in the same way. In the absence of any information, the principle of insufficient reason would lead us to use equal prior probabilities for all models, as in our paper. Different prior probabilities can be viewed as derived from previous data and representing the relative success of the models in predicting those previous data. This may apply even when prior model probabilities represent apparently subjective expert opinion. In other research, we elicited prior model probabilities from physicians by asking them to generate a fictitious data set representing patients typical of their experience, and our prior model probabilities were simply the posterior model probabilities (i.e. predictive scores) from the analysis of these fictitious data (Madigan, Gavrin and Raftery, 1994). This additional information yielded improved predictive performance.

Thus $\text{pr}(M_k|D)$ can be viewed as the predictive score for model M_k . Each model consid-

ered, M_k , generates a predictive distribution of Δ , $\text{pr}(\Delta|M_k, D)$, and these are then combined using a weighted average with weights proportional to their predictive scores. But there are many possible ways of combining predictive distributions; why should we choose the one given by equation (1)? One justification is given by Madigan and Raftery (1994), who show that this choice provides better average predictive performance than any individual model that could reasonably have been chosen, in a certain sense.

Model Uncertainty: Discrete or Continuous?

Draper contends that model uncertainty should be accounted for in a continuous manner, using all the covariates in a hierarchical model, and that variable selection should be based on utility considerations. Of course, any fully defined proposal along those lines would be an alternative way of generating predictive distributions, and could be compared with its competitors (including our approach) by assessing its predictive performance. We look forward to such comparisons.

As a practical matter, Draper’s proposal requires more information than ours, namely utilities (e.g. data collection costs), and in that sense demands more of the user. This information may not be easily available. Also, by basing a method on utilities, one is introducing a whole new layer of sensitivity concerns.

We suspect that whether model uncertainty should be treated discretely or continuously depends on the discipline and the question asked. In the early stages of research on a topic there is often much disagreement, and rival models represent genuinely conflicting theories. In this case, models do not easily blend into one another and the discrete approach is natural. Scientists who accept posterior distributions that take account of model uncertainty are “agreeing to disagree”. For a case where such a posterior distribution was agreed in spite of bitter disagreement about the model to use, see Raftery and Schweder (1993); there the discrete approach was natural and we doubt that any continuous approach would have been accepted.

When a research topic has matured there may be more agreement about the basic framework and the main variables to be included. Then most model uncertainty may be due to technical choices such as functional forms, in which case a continuous approach may be more natural.

At bottom, however, the use of even a large set of models represents a discrete choice, and there is no way of averaging continuously over all possible models. The definition of the set of “all covariates” boils down to an unavoidable, and discrete, choice.

Predictive Score and Calibration

Draper questions our use of the log score to measure predictive performance. We agree that thorough performance assessment requires a multifaceted approach. Draper suggests predictive z -scores and coverage assessment as possible tools and these are clearly worthwhile. We and others have also found calibration plots (Raftery, Madigan, and Hoeting, 1993),

ROC curves (Madigan and Raftery, 1994), and calibration scores (Hoeting, 1994; Wiper *et al.*, 1994) to be useful.

The log score, however, *is* reasonable: models which assign large predictive probability to the events that actually occur return a large log score. The biased coin device merely translates the log scores on to an alternative scale that some may find more intuitive.

Occam's Window

We agree with George, and with Clyde, DeSimone & Parigiani that the total posterior probability of the models in Occam's Window may be small, typically because there are many other models with tiny posterior probabilities individually, but which add up to a lot because there are so many of them. The key idea behind Occam's Window is to exclude such models, and our main argument for this is that it accords well with scientific practice and intuition (Madigan and Raftery, 1994). We do not claim that Occam's Window necessarily provides a good approximation to the posterior probabilities of individual models when all models are considered, but it does seem often to provide a good approximation to the predictive distributions of QUOIs. In the end it is predictive performance that matters, and there Occam's Window seems to do well.

George pointed out that when the goal is understanding rather than prediction, it may be enough to show a small set of plausible models. We agree, and indeed this is common practice in sociology, for instance. Some sociologists have reacted positively to Occam's Window, precisely because it provides a reasonable way of choosing the models to show.

Other Points

Pratt asked whether equation (1) predates Leamer (1978). We have not been able to find any earlier statement of it, although related ideas about combining forecasts from different time series models were around earlier, as pointed out by Zellner in his spoken discussion. We would certainly like to know about any earlier reference, but for the moment we will continue to cite Leamer.

George pointed out that our procedures could be evaluated using frequentist risk, and this would certainly be worthwhile. A first effort to do something along those lines is in Madigan and Raftery (1994), where an inequality favoring model averaging is established. The one problem with the idea of using risk is that it depends on postulating an underlying "true" distribution, and we have preferred to evaluate predictive performance using real data, where the "true model" is unknown.

Finally, the connection to multiple shrinkage pointed out by George is interesting and potentially fruitful, as well as reassuring. Much of the work on the assessment of model averaging so far has been empirical, and given the positive results, the time seems ripe for theoretical investigation. The ideas in George's discussion suggest things worth trying in this nascent research.

Additional References in the Discussion

- Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses (with Discussion). *Statist. Sci.* **2**, 317–352.
- Box and Tiao, (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Mass.
- Clyde, M.A. and Parmigiani, G. (1994) Protein Construct Storage: Bayesian Variable Selection and Prediction with Mixtures. ISDS DP94-14, Duke University, Durham, NC.
- Dawid, A.P. (1984). Present position and potential developments: some personal views. Statistical theory. The prequential approach (with Discussion). *J. R. Statist. Soc., series A*, **147**, 178–292.
- Dawid, A.P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In *Bayesian Statistics 4* (Bernardo, J.M. *et al.*), Oxford Science Publications, pp. 109-125.
- Draper, D. (1994). Hierarchical models and variable selection. Paper presented at *Fifth International Meeting on Bayesian Statistics*, Alicante, June 1994.
- George, E.I. and McCulloch, R.E. (1994) Integrated Bayes Variable selection. Unpublished manuscript.
- Hartigan, J.A. (1992). Locally uniform prior distributions. Technical Report, Department of Statistics, Yale University.
- Hoeting, J.A. (1994) Accounting for model uncertainty in linear regression. Unpublished PhD dissertation, Department of Statistics, University of Washington.
- Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, Series B* **30**, 31–66.
- Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, Mass.
- Raftery, A.E. (1993b). Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (eds. K.A. Bollen and J.S. Long), Beverly Hills: Sage, pp. 163–180.
- Raftery, A.E. and Schweder, T. (1993). Inference about the ratio of two parameters, with application to whale censusing. *The American Statistician*, **47**, 259-264.
- Rissanen, J. (1987) Stochastic complexity. *J. R. Statist. Soc., series B*, **49**, 223–239.
- Wiper, M.A., French, S., and Cooke, R. (1994). Hypothesis-based calibration scores. *The Statistician*, **43**, 231-236.