Bayesian variable selection using a cost-penalised approach, with application to cost-effective measurement of quality of health care

D. Fouskakis*, I. Ntzoufras† and D. Draper‡

1 September 2006

Summary: In the field of quality of health care measurement, patient sickness at admission is traditionally assessed by using logistic regression of mortality within 30 days of admission on a fairly large number of sickness indicators (perhaps on the order of 100) to construct a sickness scale, employing classical variable selection methods to find an "optimal" subset of 10–20 indicators. Such "benefit-only" methods ignore the considerable differences among the sickness indicators in cost of data collection, an issue that is crucial when admission sickness is used to drive programmes (now implemented or under consideration in several countries, including the U.K. and U.S.) that attempt to identify substandard hospitals by comparing observed and expected mortality rates (given admission sickness). When both data-collection cost and accuracy of prediction of 30-day mortality are considered, a large variable-selection problem arises in which costly variables that do not predict well enough should be omitted from the final scale. In this paper we use posterior model odds for the evaluation of models and variables. We propose a prior setup which accounts for the cost of each variable and results in a set of posterior model probabilities which correspond to a generalised cost-modified version of BIC. We use reversible-jump Markov chain Monte Carlo (MCMC) methods to search the model space and check the stability of our findings with two variants of the MCMC model composition (MC^3) algorithm. Initially we reduce our model space by dropping variables with low marginal posterior probabilities and we then estimate posterior model probabilities in the reduced space. Our cost-benefit approach results in a set of models with a noticeable reduction in cost and dimensionality, and only a minor decrease in predictive performance, when compared with models arising from the standard benefit-only analysis. Our results are phrased in the language of health policy but apply with equal force to other quality assessment settings with dichotomous outcomes, such as the examination of drop-out rates in education, the study of retention rates in the workplace and the creation of cost-effective credit scores in business.

Keywords: Input-output analysis; Quality of health care; Sickness at hospital admission; Cost-benefit analysis; Laplace approximation; Reversible-jump Markov chain Monte Carlo (MCMC) methods; MCMC model composition (MC^3); Bayesian Information Criterion (BIC); Cost-modified BIC.

1 Introduction

An important topic in health policy is the assessment of the quality of health care offered to hospitalised patients. Quality of care is usually thought to depend mainly on three ingredients

*D. Fouskakis is with the Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens 15780 Greece; email `fouskakis@math.ntua.gr`

†I. Ntzoufras is with the Department of Statistics, Athens University of Economics and Business, 76 Patision Street, Athens 10434 Greece; email `ntzoufras@aueb.gr`

‡D. Draper is with the Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz CA 95064 USA; email `draper@ams.ucsc.edu`

(e.g., Donabedian and Bashshur, 2002): (i) *process*, which is what health care providers do on behalf of patients, (ii) *outcomes*, which are what happens to patients as a result of the care they receive, and (iii) patient *sickness at admission*, because the appropriateness of outcomes cannot be judged without taking account of the burden of illness brought to the hospital by its patients.

A direct audit of the processes of care is usually regarded as the single most informative component in an evaluation of quality, but process is much more expensive to measure than outcomes or admission sickness (e.g., Kahn, Rogers, *et al.*, 1990). Interest has therefore focused in recent years, in countries such as the United Kingdom and the United States, on an indirect method of assessment—which might be termed the *input-output* approach[1] (e.g., Draper, 1995; Goldstein and Spiegelhalter, 1996)—in which hospital outcomes (for instance, death within 30 days of admission) are compared after adjusting for differences in inputs (sickness at admission). The idea is to treat what goes on inside the hospital—process—as a black box, with the contents of the box inferred by examining its outputs after taking account of its inputs.

1.1 Indirect measurement of quality of health care

In practice, to indirectly measure quality of care at any given moment in time, this strategy proceeds by (a) taking a sample of hospitals and a sample of patients in the chosen hospitals, (b) obtaining mortality outcomes for the sampled patients (for example, from central government data bases), (c) extracting information on admission sickness from the medical records of these patients, (d) forming an expected mortality rate for each hospital based on (c), and (e) comparing observed and expected mortality rates to identify unusual hospitals (on both the "good" and "bad" ends of the spectrum). Since this would involve abstracting data from the charts of many thousands of patients if it were attempted on a large scale, the *cost-effective* measurement of admission sickness is crucial to this approach. Progress is being made in the U.S. (see, e.g., CMS, 2004, for details on Medicare's plans to compile a Uniform Clinical Data Set) and elsewhere on routine (automated) data collection of clinically richer sets of process and sickness variables for hospital patients than those previously available from administrative data bases, but it is likely to remain true for at least the next decade that cost-effective collection of primary data will be relevant to the design of quality of care studies in health policy (see, e.g., NDNQI, 2004, and CalNOC, 2004, for current examples, in the field of nursing quality assessment, where extensive non-automated primary data collection is both ongoing and planned).

Quality of care assessment is a highly disease-specific activity: for instance, the right admission sickness variables to examine for pneumonia would be quite different from those for heart attack. With any given disease there will be on the order of 100 separate variables potentially available in the medical record that are directly or indirectly related to admission sickness. In the case of pneumonia, for example, on which we focus in this paper, a list of the important variables from a clinical perspective would include such things as systolic blood pressure on day 1 of admission, the presence or absence of shortness of breath, and the blood urea nitrogen level (a measure of kidney functioning).

1.2 Standard benefit-only variable-selection approach

The standard method for creating an expected mortality rate from these admission sickness inputs is logistic regression, with 30-day death as the outcome, and using a nationally-representative sample of patients to normalise the expectation to average care across the nation. Typically a frequentist variable-selection method—such as all-subsets regression—is employed to find a parsimonious and clinically reasonable subset of the available sickness variables. In a major

[1] In the U.K. this approach is also referred to as *league-table quality assessment*, by analogy with the process of ranking football teams; in the U.S. and elsewhere it is also called *provider profiling* (e.g., Normand, *et al.*, 1997).

Table 1: *The Rand admission sickness scale for pneumonia ($p = 14$ variables), with the marginal data collection costs per patient for each variable (in minutes of abstraction time).*

Variable	Cost (Minutes)
Total APACHE II Score (36–point scale)	10.0
Age	0.5
Systolic Blood Pressure Score (2–point scale)	0.5
Chest X-Ray Congestive Heart Failure Score (3–point scale)	2.5
Blood Urea Nitrogen	1.5
APACHE II Coma Score (3–point scale)	2.5
Serum Albumin (3–point scale)	1.5
Shortness of Breath (yes, no)	1.0
Respiratory Distress (yes, no)	1.0
Septic Complications (yes, no)	3.0
Prior Respiratory Failure (yes, no)	2.0
Recently Hospitalised (yes, no)	2.0
Ambulatory Score (3–point scale)	2.5
Temperature	0.5

U.S. study conducted by the Rand Corporation, of quality of hospital care for $n = 2,532$ elderly patients in the late 1980s (Kahn, Rubenstein, *et al.*, 1990), this approach was used to reduce the initial list of $p = 83$ available sickness indicators for pneumonia down to a core of 14 predictors (Keeler *et al.*, 1990).

As good as the resulting scale may be on grounds of simplicity and ease of clinical communication, we take the view in this paper that—when the goal is the creation of a sickness scale that may be used prospectively to measure quality of care on a new set of patients not yet examined—the original Rand approach is sub-optimal, because it takes no account of differences in the *cost of data collection* among the available predictors (which varied for pneumonia from 30 seconds to 15 minutes of abstraction time per variable). The Rand approach represents a kind of benefit-only analysis; we propose a cost-benefit analysis, in which variables are chosen for the final scale only when they predict mortality well enough given how much they cost to collect. The relevance of this cost-benefit perspective is seen by noting that in practice the amount of money devoted to quality assessment will almost invariably be constrained, so that money wasted on excess data collection costs could be better spent on obtaining (for example) a larger sample size at the patient and/or hospital levels.

Table 1 lists the 14 variables chosen by the benefit-only Rand approach, together with their marginal data collection costs per patient (expressed in minutes of data abstraction time; this could be linearly transformed to a monetary scale using the prevailing wage rate for qualified data abstraction personnel, but there is nothing to be gained from such a transformation). The full list of all 83 sickness indicators for pneumonia, together with their costs, can be found in Fouskakis (2001).

1.3 Cost-benefit analysis

Weighing data-collection costs against the accuracy of prediction creates a large variable-selection problem. With $p = 83$ it is necessary to compare $2^{83} \doteq 9.7 \cdot 10^{24}$ subsets of sickness variables in order to find the optimal subset. Solving this problem by brute-force examination of all 10^{25} models is deeply infeasible given contemporary computing resources.

Following Fouskakis (2001), suppose (a) the 30-day mortality outcome Y_i and data on p sickness indicators $(X_{i1}, \ldots, X_{ip})$ have been collected on n individuals sampled randomly from a population $\mathcal{P}$ of patients with a given disease, and (b) the goal is to predict the death outcome for n^* new patients who will in the future be sampled randomly from $\mathcal{P}$, (c) on the basis of some or all of the predictors X_j, when (d) the marginal costs of data collection per patient $c_1, \ldots, c_p$ for the X_j vary considerably. What is the best subset of the X_j to choose based on both the quality and the cost of obtaining the predictions?

Draper and Fouskakis (2000) and Fouskakis and Draper (2002, 2006) proposed a solution to this problem based on Bayesian decision theory. They used stochastic optimisation methods—including simulated annealing, genetic algorithms, and tabu search—to find (near-) optimal subsets of predictor variables that maximise an expected utility function which trades off data collection cost against predictive accuracy. They concluded that optimal subsets of variables that achieve a cost-benefit compromise have the potential to generate large cost savings in quality assessment programmes. (Brown *et al.*, 1998, presented an application of decision theory to variable selection in multivariate regression which is motivated by somewhat similar cost-benefit considerations in a quite different setting; Lindley, 1968, used squared-error loss to measure predictive accuracy while recommending a cost-benefit tradeoff in variable selection in a less problem-specific framework than the one presented here.)

In this paper we investigate an alternative approach, based on posterior model odds for the evaluation of models and variables. In order to incorporate preferences based on costs of the variables we use a Laplace approximation to obtain cost-based penalties for each variable. After setting up the prior model and variable probabilities we use reversible-jump Markov chain Monte Carlo to search the model space. The data on which we demonstrate our method in this paper consist of the representative sample of 2,532 elderly American patients hospitalised in the period 1980–86 with pneumonia taken from the Rand study described above.

The plan of the paper is as follows. In Section 2 we describe the approach we investigate in this paper, and Section 3 provides details concerning the computation. Section 4 illustrates the experimental results on the pneumonia data set using the proposed cost-benefit analysis, and in Section 5 we conclude the paper with a brief discussion of some statistical and quality assessment implications of our work.

2 A Bayesian approach to cost-effective variable selection

Bayesian model comparison and variable selection are based on specifying a model m, its likelihood $f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)$, the prior distribution of model parameters $f(\boldsymbol{\theta}_m|m)$ and the corresponding prior model weight (or probability) $f(m)$, where $\boldsymbol{\theta}_m$ is a parameter vector under model m and $\boldsymbol{y}$ is the data vector. Parametric inference is based on the posterior distribution $f(\boldsymbol{\theta}_m|\boldsymbol{y}, m)$, and quantifying model uncertainty by estimating the posterior model probability $f(m|\boldsymbol{y})$ is also an important issue. Hence, when we consider a set of competing models $\mathcal{M} = \{m_1, m_2, \cdots, m_{|\mathcal{M}|}\}$, we focus on the posterior probability of model $m \in \mathcal{M}$, defined as

$$f(m|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|m)f(m)}{\sum_{m_l \in \mathcal{M}} f(\boldsymbol{y}|m_l)f(m_l)} = \left(\sum_{m_l \in \mathcal{M}} PO_{m_l,m} \right)^{-1} = \left(\sum_{m_l \in \mathcal{M}} B_{m_l,m} \frac{f(m_l)}{f(m)} \right)^{-1}, \quad (1)$$

where $PO_{m_i,m_j} = f(m_i|\boldsymbol{y})/f(m_j|\boldsymbol{y})$ is the posterior model odds and B_{m_i,m_j} is the Bayes factor for comparing models m_i and m_j. When we limit ourselves in the comparison of only two models we typically focus on PO_{m_i,m_j} and B_{m_i,m_j}, which have the desirable property of insensitivity to the selection of the model space $\mathcal{M}$. By definition the Bayes factor is the ratio of the posterior model odds over the prior model odds; thus large values of B_{m_i,m_j} (usually greater than 12, say) indicate

strong posterior support of model m_i against model m_j (for details see, e.g., Raftery, 1996). The posterior model probabilities and integrated likelihoods $f(\boldsymbol{y}|m_i)$ in (1) are rarely analytically tractable; we use a combination of Laplace approximations (e.g., Bernardo and Smith, 1994) and Markov Chain Monte Carlo (MCMC) methodology (e.g., Green, 1995; Han and Carlin, 2001; Chipman *et al.*, 2001; Dellaportas *et al.*, 2002; Lopes, 2002) to approximate posterior odds and Bayes factors.

In the problem described in Section 1, we use a simple logistic regression model with response $Y_i = 1$ if patient i dies and 0 otherwise. We further denote by X_{ij} the sickness predictor variable j for patient i and by γ_j an indicator, often used in Bayesian variable selection problems (e.g., George and McCulloch, 1993; Kuo and Mallick, 1998; Brown *et al.*, 1998; Dellaportas *et al.*, 2002), taking the value 1 if variable j is included in the model and 0 otherwise. Thus in this case $\mathcal{M} = \{0,1\}^p$, where p is the total number of variables. In order to map the set of binary model indicators $\boldsymbol{\gamma}$ onto a model m we can use a representation of the form $m(\boldsymbol{\gamma}) = \sum_{i=1}^{p} 2^{i-1}\gamma_i$. Hence the model formulation can be summarised as

$$\begin{aligned} (Y_i \mid \boldsymbol{\gamma}) &\overset{\text{indep}}{\sim} \text{Bernoulli}[p_i(\boldsymbol{\gamma})], \\ \eta_i(\boldsymbol{\gamma}) = \log\left[\frac{p_i(\boldsymbol{\gamma})}{1-p_i(\boldsymbol{\gamma})}\right] &= \sum_{j=0}^{p} \beta_j \gamma_j X_{ij}, \\ \boldsymbol{\eta}(\boldsymbol{\gamma}) &= \boldsymbol{X}\,\text{diag}(\boldsymbol{\gamma})\,\boldsymbol{\beta} = \boldsymbol{X}_{\boldsymbol{\gamma}}\,\boldsymbol{\beta}_{\boldsymbol{\gamma}} \end{aligned} \tag{2}$$

defining $X_{i0} = 1$ for all $i = 1, \ldots, n$ and $\gamma_0 = 1$ with prior probability one since here the intercept is always included in all models. Here $p_i(\boldsymbol{\gamma})$ is the death probability (which may be thought of as the sickness score) for patient i under model $\boldsymbol{\gamma}$, $\boldsymbol{\eta}(\boldsymbol{\gamma}) = [\eta_1(\boldsymbol{\gamma}), \ldots, \eta_n(\boldsymbol{\gamma})]^T$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, and $\boldsymbol{X} = (X_{ij}, i = 1, \ldots, n; j = 0, 1, \ldots, p)$; the vector $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ stands for the subvector of $\boldsymbol{\beta}$ which is included in the model specified by $\boldsymbol{\gamma}$, i.e., $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\beta_i : \gamma_i = 1, i = 0, 1, \ldots, p)$, and is equivalent to the $\boldsymbol{\theta}_m$ vector defined at the beginning of this section. Similarly $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is the submatrix of $\boldsymbol{X}$ with columns corresponding to variables included in the model specified by $\boldsymbol{\gamma}$.

In the remainder of this section we illustrate how to build a prior distribution to accommodate in the posterior distribution a penalty function for the increased cost of expensive predictor variables. To this end we first build a minimally informative prior for the model parameters based on the ideas of Ntzoufras *et al.* (2003). Then we employ a Laplace approximation (e.g., Tierney and Kadane, 1986) to examine the penalty (indirectly) imposed upon the model likelihood using the Bayesian approach. Finally, we specify prior model weights (probabilities) in such a way that the posterior model probabilities in effect result from a likelihood penalised according to the cost of each variable in the model.

2.1 Prior on model parameters

One important problem in Bayesian model evaluation using posterior model probabilities is their sensitivity to the prior variance of the model parameters: large variance of the $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ (used to represent prior ignorance) will increase the posterior probabilities of the simpler models considered in the model space $\mathcal{M}$ (Bartlett, 1957; Lindley, 1957; Shafer, 1982; Robert, 1993; Kass and Raftery, 1995; Sinharay and Stern, 2002). Therefore, specifying the prior distribution is pivotal for the *a posteriori* support of the models examined. We address this issue by using ideas proposed by Ntzoufras *et al.* (2003): we use a prior distribution of the form

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = N(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) \tag{3}$$

with prior covariance matrix given by $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = n\left[\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\right]^{-1}$, where n is the total sample size and $\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ is the information matrix

$$\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \boldsymbol{X}_{\boldsymbol{\gamma}}^{\mathcal{T}} \boldsymbol{W}_{\boldsymbol{\gamma}} \boldsymbol{X}_{\boldsymbol{\gamma}};$$

here $\boldsymbol{W}_{\boldsymbol{\gamma}}$ is a diagonal matrix which in the Bernoulli case (e.g., McCullagh and Nelder, 1983) takes the form

$$\boldsymbol{W}_{\boldsymbol{\gamma}} = \text{diag}\left\{p_i(\boldsymbol{\gamma})[1 - p_i(\boldsymbol{\gamma})]\right\}.$$

This is the *unit information prior* introduced by Kass and Wasserman (1996), which corresponds to adding one data point to the data. Here we use this prior as a base, but we specify $p_i(\boldsymbol{\gamma})$ in the information matrix according to our prior information. In this manner we avoid (even minimal) reuse of the data in the prior.

When little prior information is available, a reasonable prior mean for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{0}$. This corresponds to a prior mean on the log-odds scale of zero, from which a sensible prior estimate for all model probabilities is $p_i(\boldsymbol{\gamma}) = 1/2$; with this choice (3) becomes

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = N\left[\mathbf{0}, 4n\left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\right)^{-1}\right]. \tag{4}$$

This prior distribution can also be motivated by combining the idea of imaginary data with the power prior approach of Chen *et al.* (2000). After observing the design matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ for any model $\boldsymbol{\gamma}$, we consider a set of imaginary data $\boldsymbol{y}_i^* = (y_{i1}^* = 1,\ y_{i2}^* = 0), i = 1, \ldots, n$ that assigns probabilities $1/2$ for all i and therefore supports the simplest (constant) model. We consider a prior that is generated using the likelihood of these imaginary data,

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y}^*) \propto \left\{\prod_{i=1}^{n} p_i(\boldsymbol{\gamma})[1 - p_i(\boldsymbol{\gamma})]\right\}^{(2n)^{-1}}, \tag{5}$$

where $\boldsymbol{y}^* = (\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*)$. Using the above prior the posterior becomes

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y}) \propto \prod_{i=1}^{n} p_i(\boldsymbol{\gamma})^{y_i + \frac{1}{2n}}[1 - p_i(\boldsymbol{\gamma})]^{(1+\frac{1}{n}) - (y_i + \frac{1}{2n})};$$

therefore this is equivalent to obtaining information from $\sum_{i=1}^{n}(1 + \frac{1}{n}) = n + 1$ data points, instead of n data points when using a flat prior. Thus the proposed prior (5) introduces additional information to the posterior equivalent to adding one data point to the likelihood and therefore we support *a priori* the simplest model with a weight of one data point.

Using a Laplace approximation to (5) (see, e.g., Bernardo and Smith, 1994, p. 286), we obtain

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y}^*) \dot{\sim} N\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, 2n\,\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})^{-1}\right],$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the maximum likelihood estimate if the imaginary data $\boldsymbol{y}_i^*$ were observed and $\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})$ is the observed information matrix given by

$$\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = \boldsymbol{X}_{\boldsymbol{\gamma}}^T\,\text{diag}\left\{2\,\hat{p}_i^*(\boldsymbol{\gamma})[1 - \hat{p}_i^*(\boldsymbol{\gamma})]\right\}\boldsymbol{X}_{\boldsymbol{\gamma}},$$

in which $\hat{p}_i^*(\boldsymbol{\gamma}) = \left[1 + \exp(-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\right]^{-1}$ is the fitted success probability for all i under model $\boldsymbol{\gamma}$ when observing data $\boldsymbol{y}^*$. Under the above imaginary data, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \mathbf{0}$ and $\hat{p}_i(\boldsymbol{\gamma}) = 1/2$ for all i, yielding $\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = \frac{1}{2}\left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\right)$ and therefore leading to the prior given by (4). This approach is also sensible in terms of the parsimony principle. Posterior model odds (and Bayes factors) penalise the model likelihood for deviations of the actual data from the prior distribution (see Raftery, 1996, equation 12). Since the above prior can be generated using a set of minimally-weighted imaginary data that fully support the constant model, it will provide sensible *a priori* support for more parsimonious models.

2.2 A cost-penalised prior on model space

The aim of this section is to specify a set of prior model probabilities (or odds) that accounts for the prior preference based on the cost of each variable. To make this more explicit we use two subsections. In the first we describe some preliminary results concerning the posterior model probabilities $f(\boldsymbol{\gamma}|\boldsymbol{y})$ and the corresponding model odds using the prior distribution (4) specified in the previous section, when no assumption is made for the prior model probability $f(\boldsymbol{\gamma})$. In the second subsection we specify a prior on the model space which takes into account prior preferences based on the cost of the variables. In order to achieve this we use a penalty-based interpretation of the prior $f(\boldsymbol{\gamma})$ imposed on the log-likelihood which directly results from the first subsection. We then use this cost-penalised model prior to calculate the posterior model probabilities and odds.

2.2.1 Preliminary results: posterior probabilities and model odds in the general setup

Let us denote by $PO_{k\ell}$ the posterior odds of model $\boldsymbol{\gamma}^{(k)}$ versus model $\boldsymbol{\gamma}^{(\ell)}$. Then we have

$$-2\log PO_{k\ell} = -2\left[\log f(\boldsymbol{\gamma}^{(k)}|\boldsymbol{y}) - \log f(\boldsymbol{\gamma}^{(\ell)}|\boldsymbol{y})\right]. \tag{6}$$

Following the approach of Raftery (1996), we can approximate the posterior distribution of a model $\boldsymbol{\gamma}$ using the following Laplace approximation:

$$\begin{aligned} -2\log f(\boldsymbol{\gamma}|\boldsymbol{y}) &= -2\log f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}) - 2\log f(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) - d_{\boldsymbol{\gamma}}\log(2\pi) \\ &\quad - \log|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}| - 2\log f(\boldsymbol{\gamma}) + O(n^{-1}), \end{aligned} \tag{7}$$

where $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the posterior mode of $f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{y},\boldsymbol{\gamma})$, $d_{\boldsymbol{\gamma}} = \sum_{j=0}^{p}\gamma_j$ is the dimension of the model $\boldsymbol{\gamma}$, and $\boldsymbol{\Psi}_{\boldsymbol{\gamma}}$ is minus the inverse of the Hessian matrix of $h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \log f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}) + \log f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})$ evaluated at the posterior mode $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$. Under the model formulation given by equation (2) and the prior distribution (4) we have that

$$\begin{aligned} \boldsymbol{\Psi}_{\boldsymbol{\gamma}} &= \left[-\left.\frac{\partial^2\log f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}_{\boldsymbol{\gamma}}^2}\right|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}=\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}} - \left.\frac{\partial^2\log f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}_{\boldsymbol{\gamma}}^2}\right|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}=\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}}\right]^{-1} \\ &= \left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\,\mathrm{diag}\left\{\frac{\exp\left(\boldsymbol{X}_{\boldsymbol{\gamma},i}\,\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)}{\left[1+\exp\left(\boldsymbol{X}_{\boldsymbol{\gamma},i}\,\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)\right]^2} + \frac{1}{4n}\right\}\boldsymbol{X}_{\boldsymbol{\gamma}}\right)^{-1}, \end{aligned} \tag{8}$$

where $\boldsymbol{X}_{\boldsymbol{\gamma},i}$ is row i of the matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ for $i = 1,\ldots,n$.

By substituting the prior (4) in expression (7) we get

$$-2\log f(\boldsymbol{\gamma}|\boldsymbol{y}) = -2\log f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}) + \phi(\boldsymbol{\gamma}) - 2\log f(\boldsymbol{\gamma}) + O(n^{-1}), \tag{9}$$

where

$$\phi(\boldsymbol{\gamma}) = \frac{1}{4n}\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} + d_{\boldsymbol{\gamma}}\log(4n) + \log\frac{|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^{-1}|}{|\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}|}. \tag{10}$$

From the above expression it is clear that the logarithm of a posterior model probability can be regarded as a penalised log-likelihood evaluated at the posterior mode of the model, in which the term $\phi(\boldsymbol{\gamma}) - 2\log f(\boldsymbol{\gamma})$ can be interpreted as the penalty imposed upon the log-likelihood. In

pairwise model comparisons, we can directly use the posterior model odds (6), which can now be written as

$$-2\log PO_{k\ell} = -2\log\left\{\frac{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}},\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}},\boldsymbol{\gamma}^{(\ell)})}\right\} + \phi\left(\boldsymbol{\gamma}^{(k)}\right) - \phi\left(\boldsymbol{\gamma}^{(\ell)}\right) - 2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(n^{-1}). \quad (11)$$

Therefore, the comparison of the two models is based on a penalised log-likelihood ratio, where the penalty is now given by

$$\psi(\boldsymbol{\gamma}^{(k)},\boldsymbol{\gamma}^{(\ell)}) = \phi(\boldsymbol{\gamma}^{(k)}) - \phi(\boldsymbol{\gamma}^{(\ell)}) - 2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})};$$

for more details see Ntzoufras (1999, chapter 6).

Each penalty term is divided into two parts: $\phi(\boldsymbol{\gamma})$ and $-2\log f(\boldsymbol{\gamma})$. The first term, $\phi(\boldsymbol{\gamma})$, has its source in the marginal likelihood $f(\boldsymbol{y}|\boldsymbol{\gamma})$ of model $\boldsymbol{\gamma}$ and can be thought of as a measure of discrepancy between the data and the prior information for the model parameters. The second part comes from the prior model probabilities $f(\boldsymbol{\gamma})$. Indifference on the space of all models, usually expressed by the uniform distribution (i.e., $f(\boldsymbol{\gamma}) \propto 1$), eliminates the second term from the model comparison procedure, since the penalty term in (11) will then be based only on the difference of the first penalty terms $\phi(\boldsymbol{\gamma}^{(k)}) - \phi(\boldsymbol{\gamma}^{(\ell)})$. For this reason the penalty term $\phi(\boldsymbol{\gamma})$ is the imposed penalty which appears in the penalised log-likelihood expression of the Bayes factor $BF_{k\ell}$ with a uniform prior on model space.

A simpler but less accurate approximation of $\log PO_{k\ell}$ can be obtained following the arguments of Schwartz (1978):

$$\begin{aligned} -2\log PO_{k\ell} &= -2\log\left[\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}},\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}},\boldsymbol{\gamma}^{(\ell)})}\right] + \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)\log n - 2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(1) \\ &= BIC_{k\ell} - 2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(1), \end{aligned} \quad (12)$$

where $BIC_{k\ell}$ is the Bayesian Information Criterion (e.g., Kass and Wasserman, 1996; Raftery, 1996) for choosing between models $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\gamma}^{(\ell)}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the vector of maximum likelihood estimates of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Since $BIC_{k\ell}$ is an $O(1)$ approximation, it might diverge from the exact value of the logarithm of the Bayes factor even for large samples. Even so, it has often been shown to provide a reasonable measure of evidence (for finite n) and its straightforward calculation has encouraged its widespread use in practice (see Kass and Raftery, 1995, for details).

2.2.2 Accounting for the cost of variables via prior model weights

Following the previous section and equations (9) and (11) it is clear that an additional penalty can be directly imposed on the posterior model probabilities and odds via the prior model probabilities $f(\boldsymbol{\gamma})$. Therefore we may use prior model probabilities to induce prior preferences for specific variables depending on their costs. For this reason we propose to use prior model probabilities of the form

$$f(\gamma_j) \propto \exp\left[\frac{\gamma_j}{2}\left(\frac{c_0 - c_j}{c_0}\right)\log n\right] \text{ for } j = 1,\ldots,p, \quad (13)$$

where c_j is the differential cost per observation for variable X_j and c_0 is a baseline cost per variable for each collected observation. We further assume that the constant term is included in all models by specifying $f(\gamma_0 = 1) = 1$, resulting in

$$-2\log f(\boldsymbol{\gamma}) = \sum_{j=1}^{p}\gamma_j\frac{c_j}{c_0}\log n - d_{\boldsymbol{\gamma}}\log n + 2\sum_{j=1}^{p}\log\left[1 + n^{-\frac{1}{2}\left(1-\frac{c_j}{c_0}\right)}\right]. \quad (14)$$

If all variables have the same cost or we are indifferent concerning the cost then we can set $c_j = c_0$ for $j = 1, \ldots, p$, which reduces to the uniform prior on model space ($f(\boldsymbol{\gamma}) \propto 1$) and posterior odds equal to the Bayes factor. When we set unequal costs, a natural choice for the baseline cost is $c_0 = \min\{c_j, j = 1, \ldots, p\}$.

When comparing two models $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\gamma}^{(\ell)}$, the additional penalty imposed on the log-likelihood ratio due to the cost-adjusted prior model probabilities is given by

$$\begin{aligned} -2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} &= \sum_{j=1}^{p}\left(\gamma_j^{(k)} - \gamma_j^{(\ell)}\right)\frac{c_j}{c_0}\log n - \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)\log n \\ &= \left[\frac{C_{\boldsymbol{\gamma}^{(k)}} - C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0} - \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)\right]\log n, \end{aligned} \tag{15}$$

where $C_{\boldsymbol{\gamma}} = \sum_{j=1}^{p}\gamma_j c_j$ is the total cost of model $\boldsymbol{\gamma}$; thus two models of the same dimension and cost will have the same prior weight. In the simpler case where we compare two nested models that differ only on the status of variable j, the prior model ratio simplifies to

$$-2\log\frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} = \left(\frac{c_j}{c_0} - 1\right)\log n, \tag{16}$$

where $\boldsymbol{\gamma}_{\setminus j}$ is the vector of $\boldsymbol{\gamma}$ excluding element γ_j. The above expression can be viewed as a prior penalty for including the variable j in the model, while the term $\left(\frac{c_j}{c_0} - 1\right)$ can be interpreted as the proportional additional penalty imposed upon $(-2\log BF)$ if the variable X_j is included in the model due to its increased cost.

Using the prior model odds (15) in the approximate posterior model odds (11) we obtain

$$-2\log PO_{k\ell} = -2\log\left[\frac{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})}\right] + \psi(\boldsymbol{\gamma}^{(k)}, \boldsymbol{\gamma}^{(\ell)}) + O(n^{-1}), \tag{17}$$

where the penalty term is given by

$$\begin{aligned} \psi(\boldsymbol{\gamma}^{(k)}, \boldsymbol{\gamma}^{(\ell)}) &= \frac{1}{4n}\left(\tilde{\beta}^T_{\boldsymbol{\gamma}^{(k)}}\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(k)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(k)}}\tilde{\beta}_{\boldsymbol{\gamma}^{(k)}} - \tilde{\beta}^T_{\boldsymbol{\gamma}^{(\ell)}}\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(\ell)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(\ell)}}\tilde{\beta}_{\boldsymbol{\gamma}^{(\ell)}}\right) + \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)\log(4) \\ &\quad + \log\frac{|\boldsymbol{\Psi}^{-1}_{\boldsymbol{\gamma}^{(k)}}|}{|\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(k)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(k)}}|} - \log\frac{|\boldsymbol{\Psi}^{-1}_{\boldsymbol{\gamma}^{(\ell)}}|}{|\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(\ell)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(\ell)}}|} + \frac{C_{\boldsymbol{\gamma}^{(k)}} - C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0}\log n. \end{aligned} \tag{18}$$

Finally we consider the BIC-based approximation (12) to the logarithm of the posterior model odds with the prior model odds (15), yielding

$$-2\log PO_{k\ell} = -2\log\left[\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})}\right] + \frac{C_{\boldsymbol{\gamma}^{(k)}} - C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0}\log n + O(1). \tag{19}$$

The penalty term $d_{\boldsymbol{\gamma}}\log n$ of model $\boldsymbol{\gamma}$ used in (12) has been replaced in the above expression by the cost-dependent penalty $c_0^{-1}C_{\boldsymbol{\gamma}}\log n$; when no costs are considered, $c_j = c_0$ for all j, yielding $c_0^{-1}C_{\boldsymbol{\gamma}} = d_{\boldsymbol{\gamma}}$, the original BIC expression. Therefore, we may interpret the quantity $\log n$ as the imposed (baseline) penalty for each variable included in the model $\boldsymbol{\gamma}$ when having no costs (or when having equal costs). Moreover, this baseline penalty term is inflated proportionally to the cost ratio $\frac{c_j}{c_0}$ for each variable X_j; for example, if the cost of a variable X_j is twice the baseline cost ($c_j = 2c_0$) then the imposed penalty is equivalent to adding two variables with the baseline cost. For this reason, (19) can be considered as a cost-modified generalization of BIC when cost-adjusted prior model probabilities of type (13) are adopted.

3 MCMC implementation

As noted earlier, in our quality of care study with $p = 83$ predictors there are on the order of 10^{25} possible models. In such situations, sampling algorithms will not be able to estimate posterior model probabilities with high accuracy in a reasonable amount of CPU time due to the large model space. For this reason, we implemented the following two-step method:

(1) First we use a model search tool to identify variables with high marginal posterior inclusion probabilities $f(\gamma_j|\boldsymbol{y})$, and we create a reduced model space consisting only of those variables whose marginal probabilities are above a threshold value. According to Barbieri and Berger (2004) this method of selecting variables based on their marginal probabilities may lead to the identification of models with better predictive abilities than approaches based on maximising posterior model probabilities. Although Barbieri and Berger proposed 0.5 as a threshold value for $f(\gamma_j = 1|\boldsymbol{y})$, we used the lower value of 0.3, since our aim was only to identify and eliminate variables not contributing to models with high posterior probabilities.

(2) Then we use a model search tool in the reduced model space to estimate posterior model probabilities (and the corresponding odds).

To ensure stability of our findings we explored the use of two model search tools in step (1):

- a reversible-jump MCMC algorithm (RJMCMC; Green, 1995), as implemented for variable selection in generalised linear models by Dellaportas *et al.* (2002) and Ntzoufras *et al.*(2003); and
- the MCMC model composition (MC^3) algorithm (Madigan and York, 1995).

More specifically, we implemented reversible-jump moves within Gibbs for the model indicators γ_i, by proposing the new model to differ from the current one in each step by a single term i with probability one (see Dellaportas *et al.*, 2002, for details). The algorithm can be summarized as follows:

1. For $j = 1, \ldots, p$, use RJMCMC to compare the current model $\boldsymbol{\gamma}$ with the proposed one $\boldsymbol{\gamma}'$ with components $\gamma'_j = 1 - \gamma_j$ and $\gamma'_k = \gamma_k$ for $k \neq j$ with probability one. The updating sequence of γ_j is randomly determined in each step.
2. For $j = 0, \ldots, p$, if $\gamma_j = 1$ then generate model parameters β_j from the corresponding posterior distribution $f(\beta_j|\boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}, \boldsymbol{y})$, otherwise set $\beta_j = 0$.

In our context the MC^3 algorithm may be summarised by the following steps:

1. For $j = 1, \ldots, p$, propose a move from the current model $\boldsymbol{\gamma}$ to a new one $\boldsymbol{\gamma}'$ with components $\gamma'_j = 1 - \gamma_j$ and $\gamma'_k = \gamma_k$ for $k \neq j$ with probability one. The updating sequence of γ_j is randomly determined in each step.
2. Accept the proposed model $\boldsymbol{\gamma}'$ with probability

$$\alpha = \min\left[1, \frac{f(\boldsymbol{\gamma}'|\boldsymbol{y})}{f(\boldsymbol{\gamma}|\boldsymbol{y})}\right] = \min\left(1, PO_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}\right).$$

Since the posterior model odds $PO_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$ used in MC^3 are not analytically available here, we also explored two methods for calculating them—approximating the acceptance probabilities with Laplace (equation 17) and with BIC (equation 19; cf. Raftery, 1995; Hoeting *et al.*, 1999)—and in

Table 2: *Preliminary results: variables with marginal posterior probabilities* $f(\gamma_j = 1|\boldsymbol{y})$ *above 0.30; costs are expressed in minutes of abstraction time.*

Variable			Marginal Posterior Probabilities	
			Benefit-Only	Cost-Benefit
Index	Name	Cost	Analysis	Analysis
1	Systolic Blood Pressure Score	0.50	0.99	0.99
2	Age	0.50	0.99	0.99
3	Blood Urea Nitrogen	1.50	1.00	0.99
4	Apache II Coma Score	2.50	1.00	
5	Shortness of Breath	1.00	0.97	0.79
8	Septic Complications	3.00	0.88	
12	Temperature	0.50	0.98	0.96
13	Heart Rate	0.50		0.34
14	Chest Pain	0.50		0.39
15	Cardiomegaly Score	1.50	0.71	
27	Hematologic History Score	1.50	0.45	
37	Apache Respiratory Rate Score	1.00	0.95	0.32
46	Admission SBP	0.50	0.68	0.90
49	Respiratory Rate	0.50		0.81
51	Confusion	0.50		0.95
70	Apache PH Score	1.00	0.98	0.98
73	Morbid + Comorbid	7.50	0.96	
78	Musculoskeletal Score	1.00		0.54

addition we further explored one additional form of sensitivity analysis: initializing the MCMC runs at the null model (with no predictors) and the full model (with all predictors). All of this was done both for the benefit-only analysis and the cost-benefit approach.

In moving from the full to the reduced model space to implement step (1) of our two-step method, for both the benefit-only and cost-benefit analyses we found a striking level of agreement—across (a) the two model search tools, (b) the two methods to approximate the acceptance probabilities in MC^3, and (c) the two choices for initializing the MCMC runs—in the subset of variables defining the reduced model space; this made it unnecessary to perform a similar sensitivity analysis in step (2). Results in the next section are therefore presented only for RJMCMC (starting from the full model). Convergence of the RJMCMC algorithm was checked using ergodic mean plots of the marginal inclusion probabilities for the full model space and the posterior model probabilities for the reduced space. Additional computing details are available in the Appendix.

4 Experimental results

Table 2 presents the marginal posterior probabilities of the variables that exceeded the threshold value of 0.30, in each of the benefit-only and cost-benefit analyses, together with their data collection costs. In both the benefit-only and cost-benefit situations our methods reduced the initial list of $p = 83$ available candidates down to 13 predictors. Note from Table 2 that expensive variables with high marginal posterior probabilities in the benefit-only analysis were absent from the set of promising variables in the cost-benefit analysis (e.g., the Apache II Coma Score, variable 4).

Table 3: *Reduced model space: posterior model probabilities above 0.03, posterior odds (PO_{1k}) of the best model within each analysis versus the current model k, and model costs.*

Benefit-Only Analysis

k	Common Variables Within Each Analysis	Additional Variables	Model Cost	Posterior Probabilities	PO_{1k}
1	$X_4 + X_{15} + X_{37} + X_{73}$	$+X_8 + X_{27} + X_{46}$	22.5	0.3066	1.00
2		$+X_8 + X_{27}$	22.0	0.1969	1.56
3		$+X_8$	20.5	0.1833	1.67
4		$+X_{27} + X_{46}$	19.5	0.0763	4.02
5			17.5	0.0383	8.00

Cost-Benefit Analysis

k	Common Variables Within Each Analysis	Additional Variables	Model Cost	Posterior Probabilities	PO_{1k}
1	$X_{46} + X_{51}$	$+X_{49} + X_{78}$	7.0	0.1460	1.00
2		$+X_{14} + X_{49} + X_{78}$	7.5	0.1168	1.27
3		$+X_{13} + X_{49} + X_{78}$	7.5	0.0866	1.69
4		$+X_{13} + X_{14} + X_{49} + X_{78}$	8.0	0.0665	2.20
5		$+X_{14} + X_{49}$	7.0	0.0461	3.17
6		$+X_{49}$	6.5	0.0409	3.57
7		$+X_{37} + X_{78}$	7.5	0.0382	3.82
8		$+X_{13} + X_{14} + X_{49}$	7.5	0.0369	3.96
9		$+X_{13}$	6.5	0.0344	4.25

Common variables in both analyses: $X_1 + X_2 + X_3 + X_5 + X_{12} + X_{70}$

Similarly, some inexpensive variables with low marginal posterior probabilities in the benefit-only analysis were included in most of the models visited in the cost-benefit analysis (e.g., Confusion, variable 51). Note that there is not a strong degree of overlap between the 14 variables chosen in the original Rand benefit-only analysis summarised in Table 1 and the 13 variables with high marginal posterior probabilities in the benefit-only part of Table 2; we return to this point below.

Table 3 presents models with posterior model probabilities above 0.03 (in descending order), as well as posterior odds of the model with the highest posterior probability compared to the remaining ones. In both types of analysis, the variables Systolic Blood Pressure Score (X_1), Age (X_2), Blood Urea Nitrogen (X_3), Shortness of Breath (X_5), Temperature (X_{12}) and Apache PH Score (X_{70}) were included in all the highest probability models, with costs (in minutes) 0.5, 0.5, 1.5, 1.0, 0.5 and 1.0 respectively.

For the cost-benefit analysis, 9 models had posterior probabilities above 0.03. In all of these models Admission Systolic Blood Pressure (SBP; X_{46}) and Confusion (X_{51}) were present (both having the lowest cost of 0.5 minutes). Predictors Respiratory Rate (X_{49}) and Musculoskeletal Score (X_{78}) were frequently included in the top nine models (in 7 and 5 of the 9 cases, respectively). Both of these variables were present in the four highest probability models with quite close posterior probabilities and therefore with no substantial differences between them. This latter conclusion arises from the fact that models 2–4 (see the cost-benefit analysis of Table 3) have posterior odds compared to the highest probability model less than 3, indicating evidence "not worth more than a bare mention" in favor of model 1 (cf. Raftery, 1996). All variables included in the highest probability models had costs of at most one minute with the exception of Blood Urea Nitrogen

Table 4: *Comparison of measures of fit, cost and dimensionality between the visited models in the reduced model space of the benefit-only and cost-benefit analysis; percentage difference is in relation to benefit-only.*

	Analysis		Percentage
	Benefit-Only	Cost-Benefit	Difference
Min Deviance	1553.2	1616.1	+4.1
Median Deviance	1572.0	1643.8	+4.6
Median Cost	22.0	7.5	–65.9
Median Dimension	13	11	–15.4

(X_3), which had a cost of 1.5. These cost-benefit results are in rough, but not perfect, agreement with those of Fouskakis and Draper (2006) using the decision-theoretic approach described in Section 1.3; we intend to report elsewhere on a detailed analysis of the differences between the two approaches in both health policy conclusions and computational efficiency.

In the benefit-only analysis, 5 models had posterior probabilities above 0.03. In all of these models Apache II Coma Score (X_4), Cardiomegaly Score (X_{15}), Apache Respiratory Rate Score (X_{37}) and Morbid + Comorbid (X_{73}) were present, having costs of 2.5, 1.5, 1.0 and 7.5 minutes (respectively). Note that the costs of the best models in the benefit-only analysis are 2.2 to 3.5 times higher than the costs of the best models from the cost-benefit analysis.

Since in the cost-benefit analysis we increase the penalty of relatively expensive variables in the prior, we end up selecting more parsimonious models in terms of both dimensionality and cost. It is therefore interesting to examine the loss in terms of prediction and goodness of fit. We use the posterior distribution of the deviance statistic

$$D(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) = -2 \sum_{i=1}^{n} \log f(y_i | \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$$

(Dempster, 1974; Spiegelhalter *et al.*, 2002) as a measure of model fit. Usually attention focuses on the minimum value of this posterior distribution (which sometimes is poorly estimated by MCMC runs), but other posterior descriptive measures such as the median or mean provide adequate measures of fit (Spiegelhalter *et al.*, 1996).

In Table 4, we present the minimum and median values of the posterior distribution of the deviance statistic, together with the median cost and dimension of all visited models in both types of analysis. Two main points are worth noting.

- The deviance statistic for the benefit-only Rand model summarised in Table 1 turned out to be 1587.3 (achieved with 14 predictors), substantially worse than the median deviance (1572.0, achieved with a median of 13 predictors) of the models visited by the benefit-only approach examined in this paper, i.e., in this case study, all-subsets regression (the Rand approach) was substantially out-performed by Bayesian RJMCMC.

- The minimum and median values of the posterior distribution of the deviance statistic for the benefit analysis were lower by a relatively modest 4% and 4.5% compared to the corresponding values of the cost-benefit analysis, but the median cost of the visited models for the cost-benefit analysis was almost 66% lower than that for the benefit-only analysis. Similarly, the median dimensionality of the visited models for the cost-benefit analysis was about 15% lower than that for the benefit-only analysis. These values indicate that the loss of predictive accuracy that we face when choosing to perform the proposed cost-benefit

analysis is small compared to the substantial gain we achieve in terms of cost and reduced model complexity.

An alternative predictive measure of fit is the cross-validation log score LS_{CV}, following ideas of Geisser and Eddy (1979) (also see, e.g., Draper and Krnjajić 2006). It is based on leave-one-out predictive distributions $f(y_i|\boldsymbol{y}_{\setminus i})$ and is given by

$$LS_{CV}(\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\boldsymbol{y}_{\setminus i}, \boldsymbol{\gamma}),$$

where $\boldsymbol{y}_{\setminus i}$ is the vector of data $\boldsymbol{y}$ without observation i (larger values of LS_{CV} indicate greater predictive accuracy). This measure can be estimated directly from a single MCMC run using the formula

$$\widehat{LS}_{CV}(\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} \log \overline{f^{-1}(y_i|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})}$$

where $\overline{f^{-1}(y_i|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})}$ is the posterior mean of the inverse of the predictive density for observation i (for details see, e.g., Gelfand, 1996, pp. 154–155). We calculated $\widehat{LS}_{CV}$ for the models with the highest posterior probability for each analysis and obtained a value of -0.312 for the best model of the benefit-only analysis and -0.327 for that of the cost-benefit analysis; the latter is 4.8% smaller than the former, in line with the last column of Table 4, and as before this small loss in predictive accuracy is accompanied by the 66% drop in cost and 15% decrease in model complexity achieved by the cost-benefit approach.

5 Discussion

In this paper we have examined a relatively new perspective on Bayesian variable selection, when data collection costs need to be traded off against predictive accuracy in choosing an optimal subset of predictors. We propose a prior setup which accounts for the cost of each variable and we utilize traditional posterior model odds for the evaluation of models. This leads to a set of posterior model probabilities which approximately correspond to a generalised cost-modified version of BIC. Computation is performed using reversible-jump MCMC in two stages: firstly to reduce the model space by dropping variables with low marginal posterior probabilities and secondly to estimate posterior model probabilities in the reduced space. We have applied our methodology to the problem of cost-effective input-output quality measurement in a health policy setting, with a binary outcome and a large number ($p = 83$) of predictors which differ substantially in data-collection costs. The resulting models achieve dramatic gains in cost and noticeable improvement in model simplicity at the price of a small loss in predictive accuracy, when compared to the results of a more traditional benefit-only analysis.

Our proposed methodology appears to hold significant promise for cost-effective input-output quality and performance assessment. It can be applied in any setting where the outcome is binary, such as in education (with outcomes such as drop-out during university study and employment following graduation) and business (with outcomes such as retention in the workplace and the default status of a loan), and can be implemented with minor modifications for any other generalised linear model. We believe that the scope of applications of regression methodology in which

(a) the purpose of the model-building is to create a predictive scale and

(b) future use of the scale created in (a) will take place in a cost-constrained environment

is sufficiently broad that methods like those examined here are worthy both of consideration now for practical adoption and of further study to promote, e.g., additional computational efficiency gains.

Appendix: computing details

With reference to the MCMC methods described in Section 3, it is worth noting that both the coding time and the running time of RJMCMC were substantially higher than with either variant of MC^3 to achieve comparable MCMC accuracy. All MC^3 runs in the full model space were based on 10,000 monitoring iterations after a burn-in (from either the null model or the full model) of 1,000 iterations; each of these runs took 2–3 days (on a Pentium 4 machine at 2.8 GHz with 512MB RAM) for the Laplace variant of MC^3 and 1–2 days for the BIC variant. To achieve reasonable running times for RJMCMC it was necessary to implement the algorithm in `C`. RJMCMC runs were based on 100,000 iterations, after discarding an initial 10,000 iterations as a burn-in; each of these runs took 2–3 days in in the full model space and 9 hours in the reduced space. The resulting `R` and `C` programs are available upon request from the first or second authors of this paper.

References

Barbieri MD, Berger JO (2004). Optimal predictive model selection. *Annals of Statistics*, **32**, 870–897.

Bartlett MS (1957). Comment on D.V. Lindley's statistical paradox. *Biometrika*, **44**, 533–534.

Bernardo JM, Smith AFM (1994). *Bayesian Theory*. New York: Wiley.

Brown PJ, Vannucci M, Fearn T (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, **60**, 627–641.

CalNOC (2004). The California Nursing Outcomes Coalition Database Project. Available at `www.calnoc.org`

Chen MH, Ibrahim JG, Shao QM (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, **84**, 121–137.

Chipman H, George EI, McCulloch RE (2001). The Practical implementation of Bayesian model selection (with discussion). *IMS Lecture Notes - Monograph Series*, **38**, 67–134.

CMS (2004). Centers for Medicare & Medicaid Services: Medicare Information Resource. Available at `www.cms.hhs.gov/medicare/`

Dellaportas P, Forster JJ, Ntzoufras I (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27–36.

Dempster AP (1974). The direct use of likelihood for significance testing. In *Proceedings of a Conference on Foundational Questions in Statistical Inference*, edited by Barndorff-Nielsen O, Blaesild P, Sihon G. University of Aarhus, 335–352. [Reprinted: 1997, *Statistics and Computing*, **7**, 247–252].

Donabedian A, Bashshur R (2002). *An Introduction to Quality Assurance in Health Care*. Oxford: University Press.

Draper D (1995). Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics*, **20**, 115–147, 233–239.

Draper D, Fouskakis D (2000). A case study of stochastic optimisation in health policy: problem formulation and preliminary results. *Journal of Global Optimization*, **18**, 399–416.

Draper D, Krnjajić M (2006). Bayesian model specification. Submitted.

Fouskakis D (2001). *Stochastic Optimisation Methods for Cost-Effective Quality Assessment in Health*. Ph.D. dissertation, Department of Mathematical Sciences, University of Bath, UK. Available at http://www.math.ntua.gr/∼fouskakis

Fouskakis D, Draper D. (2002). Stochastic optimisation: a review. *International Statistical Review*, **70**, 315–349.

Fouskakis D, Draper D (2006). Stochastic optimization methods for cost-effective quality assessment in health. Submitted.

Geisser S, Eddy WF (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

Gelfand AE (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice*, edited by Gilks WR, Richardson S, Spiegelhalter DJ. London: Chapman and Hall, 145–162.

George EI, McCulloch RE (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Goldstein H, Spiegelhalter DJ (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society, Series A*, **159**, 385–444.

Green P (1995). Reversible ump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711–732.

Han C, Carlin B (2001). MCMC methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, **96**, 1122–1132.

Hoeting JA, Madigan D, Raftery AE, Volinski CT (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382–417.

Kahn K, Rogers W, Rubenstein L, Sherwood M, Reinisch E, Keeler E, Draper D, Kosecoff J, Brook R. (1990). Measuring quality of care with explicit process criteria before and after implementation of the DRG-based Prospective Payment System. *Journal of the American Medical Association*, **264**, 1969–1973 (with editorial comment, 1995–1997).

Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990). The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association*, **264**, 1953–1955 (with editorial comment, 1995–1997).

Kass RE, Raftery AE (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kass RE, Wasserman L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Keeler E, Kahn K, Draper D, Sherwood M, Rubenstein L, Reinisch E, Kosecoff J, Brook R (1990). Changes in sickness at admission following the introduction of the Prospective Payment System. *Journal of the American Medical Association*, **264**, 1962–1968.

Kuo L, Mallick B (1998). Variable selection for regression models. *Sankhyā B*, **60**, 65–81.

Lindley DV (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

Lindley DV (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 31–66.

Lopes HF (2002). Bayesian model selection. *Technical Report*, Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Brazil.

Madigan D, York J (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.

McCullagh P, Nelder JA (1983). *Generalized Linear Models*. London: Chapman and Hall.

NDNQI (2004). National Database of Nursing Quality Indicators. Available at `www.nursingquality.org`

Normand S, Glickman M, Gatsonis C (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*, **92**, 803–814.

Ntzoufras I (1999). *Aspects of Bayesian Model and Variable Selection Using MCMC*. Ph.D. Thesis, Department of Statistics, Athens University of Economics and Business. Available at `www.stat-athens.aueb.gr/~jbn/publications.htm`

Ntzoufras I, Dellaportas P, Forster JJ (2003). Bayesian variable and link determination for generalized linear models. *Journal of Statistical Planning and Inference*, **111**, 165–180.

Raftery AE (1995). Bayesian model selection in social research. *Sociological Methodology 1995* (P. V. Marsden ed.). Oxford: Blackwell, **25**, 111–196.

Raftery AE (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, **83**, 251–266.

Robert CP (1993). A note on the Jeffreys-Lindley paradox. *Statistica Sinica*, **3**, 601–608.

Schwarz G (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Shafer J (1982). Lindley's paradox (with discussion). *Journal of the American Statistical Association*, **77**, 325–334.

Sinharay S, Stern HS (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, **56**, 196–201.

Spiegelhalter DJ, Best N, Carlin B, van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

Spiegelhalter DJ, Thomas A, Best N, Gilks W (1996). *BUGS 0.5: Bayesian Inference Using Gibbs Sampling*. Available at `www.mrc-bsu.cam.ac.uk/bugs`

Tierney L, Kadane JB (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.