Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2009)
58, *Part* 3, *pp.* 383–403

Population-based reversible jump Markov chain Monte Carlo methods for Bayesian variable selection and evaluation under cost limit restrictions

D. Fouskakis,

National Technical University of Athens, Greece

I. Ntzoufras

Athens University of Economics and Business, Greece

and D. Draper

University of California, Santa Cruz, USA

[Received July 2007. Revised October 2008]

Summary. The measurement and improvement of the quality of health care are important areas of current research and development. A judgement of appropriateness of medical outcomes in hospital quality-of-care studies must depend on an assessment of patient sickness at admission to hospital. Indicators of patient sickness often must be abstracted from medical records, and some variables are more expensive to measure than others. Quality-of-care studies are frequently undertaken in an environment of cost restriction; thus any scale measuring patient sickness must simultaneously respect two optimality criteria: high predictive accuracy and low cost. Here we examine a variable selection strategy for construction of a scale of sickness in which predictive accuracy is optimized subject to a bound on cost. Conventional model search algorithms (such as those based on standard reversible jump Markov chain Monte Carlo (RJMCMC) sampling) in our setting will often fail, because of the existence of multiple modes of the criterion function with movement paths that are forbidden because of the cost restriction. We develop a population-based trans-dimensional RJMCMC (population RJMCMC) algorithm, in which ideas from the population-based MCMC and simulated tempering algorithms are combined. Comparing our method with standard RJMCMC sampling, we find that the population-based RJMCMC algorithm moves successfully and more efficiently between distant neighbourhoods of ‘good’ models, achieves convergence faster and has smaller Monte Carlo standard errors for a given amount of central processor unit time. In a case-study of $n=2532$ pneumonia patients on whom $p=83$ sickness indicators were measured, with marginal costs varying from smallest to largest across the predictor variables by a factor of 20, the final model chosen by population RJMCMC sampling, on the basis of both highest posterior probability and specifying the median probability model, was clinically sensible for pneumonia patients and achieved good predictive ability while capping data collection costs.

Keywords: Bayesian model comparison; Cost restriction–benefit analysis; Health care evaluation; Population-based Markov chain Monte Carlo algorithms; Reversible jump Markov chain Monte Carlo methods; Simulated tempering

Address for correspondence: D. Fouskakis, Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens 15780, Greece.
E-mail: fouskakis@math.ntua.gr

0035–9254/09/58383

1. Health care evaluation under cost restrictions

Evaluation of health services for hospitalized patients is an important area of current research and development (e.g. Ohlssen *et al.* (2007)). One leading indirect method for quality assessment (e.g. Goldstein and Spiegelhalter (1996) and Zhang *et al.* (2006)) involves the comparison of health outcomes, such as death within 30 days of admission, after adjusting for differences in sickness at admission. 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 patients sampled (e.g. from central government databases),
(c) extracting information on admission sickness from the medical records of these patients,
(d) forming an expected mortality rate for each hospital on the basis of (c) and
(e) comparing observed and expected rates of mortality to identify unusual hospitals.

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 at present in some countries, including the UK and USA (e.g. the National Institute for Health and Clinical Excellence (`www.nice.org.uk`) and the Centers for Medicare and Medicaid Services (`www.cms.hhs.gov`)) on realtime electronic data collection of clinically richer sets of sickness variables for hospital patients than those previously available from administrative databases, but it is likely to remain true for at least the next decade that the cost-effective collection of data from non-electronic medical records will be relevant to the design of quality-of-care studies in health policy. This is particularly true in countries with an interest in quality-of-care measurement but insufficient resources to be at the cutting edge in medical informatics.

The assessment of quality of care in this way depends strongly on the disease outcome; for example, the appropriate admission sickness variables for pneumonia would be quite different from those for heart attack. For any disease under evaluation, of the order of 100 potential sickness indicators may be available from hospital records. For pneumonia, on which we focus in this paper, in the data set with which we work there are 83 sickness variables, such as systolic blood pressure on the first day after admission, presence or absence of shortness of breath and blood urea nitrogen level (a measure of kidney functioning). Logistic regression of an adverse outcome, such as dead or alive within 30 days of admission, on the available sickness indicators is a common method for creating a sickness scale from which expected mortality rates can be estimated; standard variable selection methods, such as backward selection from the model with all predictors, are typically used to find a parsimonious and clinically reasonable model composed of variables that predict mortality well.

In this paper we use data from a major US study, which was conducted by the RAND Corporation in the late 1980s (Kahn *et al.*, 1990), of quality of hospital care for $n=2532$ elderly patients suffering from pneumonia. Backward selection, as described above, was used to reduce the initial list of $p=83$ available pneumonia predictors to a subset of 14 variables (Keeler *et al.*, 1990). Table 1 lists the full set of 83 sickness variables, 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 by using the prevailing wage rate for qualified data abstraction personnel, but there is nothing to be gained from such a transformation). The RAND scale is identified in the fourth column; the fifth column specifies another variable subset which was chosen by the methods of Section 3 and further described in Section 4. The 14-variable scale resulting from RAND's backward selection approach has merit with respect to simplicity and ease of clinical communication, but—when the goal is the creation of a scale of sickness that may be used prospectively to measure quality of care on a new set of patients not yet examined—the RAND scale may not be optimal, because it pays no attention to differences in the cost of data collection among the available predictors (which varied for pneumonia from 30 s to 10 min of abstraction time per variable); in fact, there was a general feeling among health policy experts whom we consulted that, at 31 min of abstraction time per patient, the 14-variable RAND scale was too expensive to be useful for large-scale quality-of-care assessment.

When cost and predictive accuracy must both be considered in seeking an optimal subset of predictors, there are two ways forward: either

(a) both criteria can be placed on a common scale, trading one against the other, and optimization can occur on that scale, or
(b) one criterion can be optimized, subject to a bound on the other.

Elsewhere (Fouskakis and Draper, 2008; Fouskakis *et al.*, 2009) we explore strategy (a); here we develop a method for implementing strategy (b), through a cost restriction–benefit analysis. The practical relevance of the selected variable subsets by using the method of this paper is ensured by enforcing an overall limit on the total data collection cost of each subset: the search is conducted only among models whose cost does not exceed this budgetary restriction. See Lindley (1968) and Brown *et al.* (1998, 2002) for other approaches to incorporating data collection costs in regression settings.

The structure of the paper is as follows. In the next section, the problem of cost-restricted variable selection in health evaluation is formulated in the Bayesian paradigm. In Section 3, we describe the proposed population-based Markov chain Monte Carlo (MCMC) algorithm, whereas implementation details and experimental results on the pneumonia data set are presented in Section 4. Section 5 concludes the paper with a brief discussion.

2. Bayesian model comparison for health care evaluation

As an abbreviation we denote a model in this context by $\gamma=(\gamma_1,\ldots,\gamma_p)$, where γ_j is a binary indicator taking the value 1 if variable j is included in the model and 0 otherwise, and p is the total number of predictors. We further denote the likelihood of this model by $f(\mathbf{y}|\boldsymbol{\beta}_\gamma,\gamma)$, the prior distribution of model parameters by $f(\boldsymbol{\beta}_\gamma|\gamma)$ and the corresponding prior model weight (probability) by $f(\gamma)$, where $\mathbf{y}$ is the vector of outcomes and $\boldsymbol{\beta}_\gamma$ is a parameter vector under model γ, i.e. $\boldsymbol{\beta}_\gamma=(\beta_i:\gamma_i=1, i=0,1,\ldots,p)$. The posterior model probabilities $f(\gamma|\mathbf{y})$ are the main tool in Bayesian inference for comparing models (in this case, variable subsets). These posterior model probabilities are rarely analytically tractable; Markov chain Monte Carlo methods are usually adopted, such as the reversible jump MCMC (RJMCMC) (Green, 1995) algorithm. In this approach $f(\gamma|\mathbf{y})$ is estimated by sampling from the joint posterior distribution

$$f(\boldsymbol{\beta}_\gamma,\gamma|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta}_\gamma,\gamma)\,f(\boldsymbol{\beta}_\gamma|\gamma)\,f(\gamma). \tag{1}$$

In the problem that was described in Section 1, we use a logistic regression model with response y_i equal to 1 if patient i suffering from pneumonia dies within 30 days of admission and 0 otherwise. We further denote by X_{ij} the sickness predictor variable j for patient i. The model formulation can be summarized as

Table 1. Full set of 83 variables, together with their data collection costs per patient and their status according to the RAND and population RJMCMC approaches†

Index	*Variable*		*Method*	
	Name‡	*Cost* c_j *(min)*	*RAND*	*Population RJMCMC*
1	Systolic blood pressure score	0.5	§	§
2	Age	0.5	§	§
3	Blood urea nitrogen	1.5	§	§
4	APACHE II coma score	2.5	§	§
5	Shortness of breath day 1?	1.0	§	§
6	Serum albumin score	1.5	§	
7	Respiratory distress?	1.0	§	
8	Septic complications?	3.0	§	
9	Prior respiratory failure?	2.0	§	
10	Recently hospitalized?	2.0	§	
11	Racbilateral process score	1.5		
12	Initial temperature	0.5	§	§
13	Heart rate day 1	0.5		
14	Chest pain day 1?	0.5		
15	Cardiomegaly score	1.5		
16	Plural effusion score	1.5		
17	Chest X-ray congestive heart failure score	2.5	§	
18	Ambulatory score	2.5	§	
19	Endocarditis at admission?	1.5		
20	Creatine phosphokinase score	2.0		
21	Prior antibiotics?	0.5		
22	Prior interstitial lung disease?	0.5		
23	Home oxygen use?	1.0		
24	Prior pneumonectomy?	0.5		
25	Prior tracheostomy?	0.5		
26	Prior aminophylline score	0.5		
27	Haematologic history score	1.5		
28	Cancer score	1.5		
29	APACHE heart rate score	1.5		
30	Corodaker score	1.0		
31	Disease of thorax?	1.0		
32	Multiple myeloma?	0.5		
33	Immunocompromised?	0.5		
34	Residence score	1.0		
35	Hepatobiliary history?	0.5		
36	Renal history score	1.0		
37	APACHE respiratory rate score	1.0		§
38	New lung score	1.0		
39	Comorbid aspiration score	0.5		
40	APACHE sodium score	2.0		
41	APACHE haematocrit score	1.5		
42	APACHE white blood cell score	1.5		
43	APACHE oxygenation score	1.5		
44	Cardiovascular accident score	1.0		
45	APACHE potassium score	1.0		
46	Admission systolic blood pressure	0.5		
47	Congestive heart failure chest X-ray score	2.5		
48	Total APACHE II score	10.0	§	
49	Respiratory rate day 1	0.5		

(*continued*)

Table 1 (*continued*)

Index	*Variable*		*Method*	
	Name‡	*Cost* c_j *(min)*	*RAND*	*Population* *RJMCMC*
50	Diastolic blood pressure day 1	0.5		
51	Confusion day 1?	0.5		
52	Pulmonary vascular congestion score	0.5		
53	APACHE venous bicarbonate score	1.5		
54	Pulmonary of oedema score	0.5		
55	Sum of congestive heart failure components	5.5		
56	Influenza score	0.5		
57	Arrest in emergency room score	0.5		
58	Biliribin score	1.5		
59	Positive blood culture?	0.5		
60	Positive urine culture?	0.5		
61	Wheezing at admission?	0.5		
62	Body system count	2.5		§
63	Morbid prior chronic obstructive pulmonary disease score	0.5		
64	Morbid pulmonary hospitalization score	0.5		
65	Comorbid cirrhosis score	0.5		
66	Comorbid congestive heart failure score	0.5		
67	Comorbid arrhythmias score	0.5		
68	Comorbid smoking score	0.5		
69	Comorbid alcoholism score	0.5		
70	APACHE acidity score	1.0		
71	Comorbid nasogastric tubes score	0.5		
72	Comorbid steroids score	0.5		
73	Morbid + comorbid score	7.5		
74	Cardiac history score	0.5		
75	Neurologic history score	0.5		
76	Oncologic history score	0.5		
77	Immunologic history score	0.5		
78	Musculoskeletal score	0.5		
79	APACHE temperature score	1.0		
80	APACHE mean blood pressure score	1.0		
81	APACHE creatinine score	1.0		
82	Diagnoses score	1.0		
83	Sex of patient	0.5		

†The fourth and fifth columns are explained in the text. Variables with a question mark in their names were dichotomous answers to yes–no questions, scored 1, yes, and 0, no; all other variables (except variable 1, which was also dichotomous) were measured on quantitative scales with three or more possible values.

‡APACHE: acute physiology and chronic health evaluation.

§Variable chosen.

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

defining $X_{i0} = 1$ for all $i = 1, \ldots, n$ and $\gamma_0 = 1$ with prior probability 1 since the intercept is always included in all models. Here

(a) $p_i(\gamma)$ is the probability of death (which may be thought of as the sickness score) for patient i under model $\gamma \in \mathcal{M} = \{0,1\}^p$,
(b) $\boldsymbol{\eta}(\gamma) = (\eta_1(\gamma), \ldots, \eta_n(\gamma))^{\mathrm{T}}$,
(c) $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$,
(d) $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$,
(e) $\mathbf{X} = (X_{ij}, i = 1, \ldots, n; j = 0, 1, \ldots, p)$ and
(f) $\mathbf{X}_\gamma$ is the submatrix of $\mathbf{X}$ with columns corresponding to variables included in the model that are specified by γ.

In Fouskakis *et al.* (2009) we specified this model structure by using a strategy (a) approach (trading off cost and predictive accuracy on a common scale), by incorporating the cost of each variable in the modelling procedure via a cost-penalized prior model probability. The problem was handled in two stages: first we identified variables with high predictive ability relative to their cost and then we identified cost-effective models by restricting the model search to only the 'good' variables from stage 1, by using the MC^3 (Madigan and York, 1995) and RJMCMC algorithms as our model search tools. In this paper we ensure the practical relevance of the final variable subsets that we discover in a different way, by enforcing an overall limit on the amount of money that it would cost to collect the data with each subset: the search is conducted only among models whose cost does not exceed this budgetary restriction. Trying to implement model search algorithms such as MC^3 and RJMCMC with this approach will frequently fail if the best model (with no overall monetary limit) exceeds the cost restriction and we also have collinear predictors with high predictive ability; the reason for this failure is the multiple modes that do not communicate since their movement paths are forbidden because of the cost restrictions. Therefore in this paper we develop population-based trans-dimensional algorithms that are based on the approach of Jasra *et al.* (2007a,b).

To complete the Bayesian model formulation, we use the prior on model parameters

$$f(\boldsymbol{\beta}_\gamma|\gamma) = N\{\mathbf{0}, 4n(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\} \tag{3}$$

that was motivated by Fouskakis *et al.* (2009) on the basis of unit information prior considerations (Kass and Wasserman, 1996), and a uniform prior on cost-restricted model space, i.e.

$$f(\gamma) \propto I\left\{\gamma \in \mathcal{M} : c(\gamma) = \sum_{j=1}^{p} c_j \gamma_j \leqslant C\right\}, \tag{4}$$

where c_j is the marginal cost per observation for variable X_j and C is the overall budgetary restriction.

3. Population-based trans-dimensional Markov chain Monte Carlo schemes

By the nature of our approach, models γ with total cost larger than C should be *a priori* excluded, resulting in the significantly reduced model space $\mathcal{M}_C = \{\gamma \in \{0,1\}^p : \Sigma_{j=1}^p c_j \gamma_j \leqslant C\}$. In variable selection model search the usual, and simplest, proposed moves are *1-bit flips* involving the addition or removal of a single variable, but under a global cost constraint a more elaborate strategy is needed to avoid becoming trapped in local optima. To see why this is so, imagine beginning at the null model; the search algorithm adds some low cost variables until the budgetary restriction C is reached. Then, using 1-bit flips, the algorithm can never include a good variable of high expense, since this will cause the total cost to exceed C. One way to overcome this is to facilitate proposed moves to models with identical properties to those of the current model (i.e. the same cost or quality of fit) from remote neighbourhoods. This can be achieved with the assistance of population-based MCMC algorithms (Jasra *et al.*, 2007a, b), in which the difficulty that was identified above can be overcome by running multiple chains and performing swaps between them.

The main idea of population-based MCMC algorithms is to generate $k=1,\ldots,N$ parallel auxiliary chains, each of them raised to a different power $t_k>0$, which is called the *temperature*, to explore the model space more efficiently. Low values (below 1) of the temperature will result in flatter target distributions, and therefore the algorithm will explore the model space extensively by moving to regions that have not been visited much; large values (above 1) of the temperature will result in steeper target distributions and will yield chains that focus the search around local modes.

To link the different chains we use the following augmented posterior distribution:

$$f(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\beta}_{(1)},\boldsymbol{\gamma}_{(1)},\ldots,\boldsymbol{\beta}_{(N)},\boldsymbol{\gamma}_{(N)}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})\, f(\boldsymbol{\beta}|\boldsymbol{\gamma})\, f(\boldsymbol{\gamma}) \times \prod_{k=1}^{N} \{f(\mathbf{y}|\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\, f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}_{(k)})\, f(\boldsymbol{\gamma}_{(k)})\}^{t_k}, \qquad (5)$$

where $\boldsymbol{\gamma}_{(k)}$ and $\boldsymbol{\beta}_{(k)}$ are the model indicator and its parameter vector respectively in chain k. In the above posterior, the marginal target distribution $f(\boldsymbol{\gamma}|\mathbf{y})$ remains the same but we can now exchange information between different chains. An important implementation issue is the specification of the number of chains N and the different temperatures t_k in each chain. A large number of parallel chains is usually considered (see, for example, Jasra *et al.* (2007a,b)), to use a sufficient range of temperatures, leading to a computationally expensive algorithm. An extensive number of chains can be avoided if we combine ideas from the population-based MCMC and simulated tempering algorithms; for details of the latter see Geyer and Thompson (1995). We propose only $N=2$ additional auxiliary chains (as in population-based MCMC sampling) but with temperatures that will vary stochastically (as in simulated tempering). This will enable the two chains to use a variety of temperatures, allowing them to move in different model space regions. To achieve an effective exploration of the space, we use large values of the temperature for the first chain ($t_1>1$)—therefore, as mentioned above, tending to search in neighbourhoods that are closer to the highest probability models—and low temperature values for the second chain ($0<t_2<1$), to visit low probability regions.

The incorporation of stochastic temperatures can be achieved by using pseudopriors $g_k(t_k)$. With this approach the posterior distribution becomes

$$f(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\beta}_{(1)},\boldsymbol{\gamma}_{(1)},\boldsymbol{\beta}_{(2)},\boldsymbol{\gamma}_{(2)},t_1,t_2|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})\, f(\boldsymbol{\beta}|\boldsymbol{\gamma})\, f(\boldsymbol{\gamma}) \times \prod_{k=1}^{2} \{f(\mathbf{y}|\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\, f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}_{(k)})\, f(\boldsymbol{\gamma}_{(k)})\}^{t_k} g_k(t_k). \qquad (6)$$

In this manner we can use standard trans-dimensional MCMC algorithms for variable selection (see, for example, Dellaportas *et al.* (2002)) to generate values of $(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\beta}_{(1)},\boldsymbol{\gamma}_{(1)},\boldsymbol{\beta}_{(2)},\boldsymbol{\gamma}_{(2)}, t_1,t_2)$ from the above joint posterior distribution, in a three-step process. Specifically,

(a) the model indicators $(\boldsymbol{\gamma},\boldsymbol{\gamma}_{(1)},\boldsymbol{\gamma}_{(2)})$ and their corresponding model parameters $(\boldsymbol{\beta},\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)})$ are updated by using RJMCMC steps.

(b) After specifying the model structure in each chain, to ensure the mixing of the algorithm, the model parameters are updated from the corresponding conditional posterior distributions.

(c) Finally, the temperature t_k is generated, in Gibbs sampling, from the conditional posterior distribution

$$f(t_k|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\beta}_{(1)},\boldsymbol{\gamma}_{(1)},\boldsymbol{\beta}_{(2)},\boldsymbol{\gamma}_{(2)},t_{\backslash k},\mathbf{y}) \propto \{f(\mathbf{y}|\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\gamma}_{(k)})\}^{t_k}\,g_k(t_k). \qquad (7)$$

The pseudopriors $g_k(t_k)$ in this proposed scheme must be carefully specified so that the temperatures generated will help the mixing of the MCMC algorithm that is described below. When a flat prior for the temperatures t_k is used, the conditional distribution (7) is simply an increasing function of those temperatures. Therefore, we propose to use directly the marginal posterior distribution of the temperatures t_k, $f(t_k|\mathbf{y})$, in the sampling scheme. We use ideas of Geyer and Thompson (1995) to achieve the desired posterior marginal distribution for the temperatures t_k, given by

$$\begin{aligned} f(t_k|\mathbf{y}) &= \sum_{\boldsymbol{\gamma}_{(k)}\in\mathcal{M}_C} \int f(\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)},t_k|\mathbf{y})\,\mathrm{d}\boldsymbol{\beta}_{(k)} \\ &\propto \sum_{\boldsymbol{\gamma}_{(k)}\in\mathcal{M}_C} \int \{f(\mathbf{y}|t_k,\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\gamma}_{(k)})\}^{t_k}\,g_k(t_k)\,\mathrm{d}\boldsymbol{\beta}_{(k)} \qquad (8) \\ &\propto Z_k(\mathbf{y},t_k)\,g_k(t_k), \end{aligned}$$

where $Z_k(\mathbf{y},t_k)$ is the marginal likelihood over all possible models for chain k given by

$$Z_k(\mathbf{y},t_k) = \sum_{\boldsymbol{\gamma}_{(k)}\in\mathcal{M}_C} \int \{f(\mathbf{y}|t_k,\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\gamma}_{(k)})\}^{t_k}\,\mathrm{d}\boldsymbol{\beta}_{(k)}. \qquad (9)$$

Since the $g_k(t_k)$ are pseudopriors, we can set

$$g_k(t_k) \propto h_k(t_k)/Z_k(\mathbf{y},t_k), \qquad (10)$$

where $h_k(t_k)$ can be chosen to be convenient density functions that are easy to simulate from, yielding

$$f(t_k|\mathbf{y}) = h_k(t_k). \qquad (11)$$

In the Gibbs sampler that is included in the algorithm described below, we generate the temperatures directly from the distributions $h_k(t_k)$. With the pseudoprior distributions proposed we do not need to calculate the normalizing constants $Z_k(\mathbf{y},t_k)$; these constants just provide a justification for simulating the stochastic temperatures from easily specified distributions.

For the selection of $h_k(t_k)$ we use

$$\begin{aligned} h_1(t_1) &= \mathrm{gamma}(t_1-1;a_1,b_1), \\ h_2(t_2) &= \mathrm{beta}(t_2;a_2,b_2), \end{aligned} \qquad (12)$$

where beta$(t;a,b)$ and gamma$(t;a,b)$ are the density functions of the beta and gamma distributions respectively, with parameters a and b, evaluated at point t. The specification of the pseudoparameters (a_1,b_1,a_2,b_2) can be based on pilot tuning until we achieve appropriate acceptance rates (e.g. 20%).

Our algorithm can be summarized as follows.

Step 1: select initial values for $(\boldsymbol{\beta},\boldsymbol{\beta}_{(1)},\boldsymbol{\beta}_{(2)})$ and $(\boldsymbol{\gamma},\boldsymbol{\gamma}_{(1)},\boldsymbol{\gamma}_{(2)})$.
Step 2: for $l=1,\ldots,L$ (where L is the number of iterations), repeat the following cycle.

(a) Generate t_1 and t_2 from $f(t_1|\mathbf{y})=h_1(t_1)$ and $f(t_2|\mathbf{y})=h_2(t_2)$ respectively.
(b) For $k=0,1,2$,
  (i) sample $\boldsymbol{\beta}_{(k)}$ by using Gibbs steps and
  (ii) sample $\boldsymbol{\gamma}_{(k)}$ by using RJMCMC steps by proposing to change each component sequentially; thus, for every $j\in\{1,\ldots,p\}$ (in a random scan),

(A) with probability 1 propose $\gamma'_{(k)}:\gamma'_{j,(k)}=1-\gamma_{j,(k)}$ and $\gamma'_{l,(k)}=\gamma_{l,(k)}$ for all $l\neq j$,

(B) if $\gamma_{j,(k)}=0$ then propose $\beta'_{j,(k)}$ from $q_{j,k}(\beta'_{j,(k)})$ and set $\beta'_{l,(k)}=\beta_{l,(k)}$ for $l\neq j$,

(C) accept the proposed move with probability $\alpha=\min\{1,A\}$, where

$$A=\left\{\frac{f(\mathbf{y}|\boldsymbol{\beta}'_{(k)},\boldsymbol{\gamma}'_{(k)})\,f(\boldsymbol{\beta}'_{(k)}|\boldsymbol{\gamma}'_{(k)})\,f(\boldsymbol{\gamma}'_{(k)})}{f(\mathbf{y}|\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\gamma}_{(k)})}\right\}^{t_k}\frac{q_{j,k}(\beta_{j,(k)})^{\gamma_{j,(k)}}}{q_{j,k}(\beta'_{j,(k)})^{1-\gamma_{j,(k)}}}. \tag{13}$$

In these steps, $\boldsymbol{\beta}_{(0)}$ and $\boldsymbol{\gamma}_{(0)}$ correspond to the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of the original chain, and $t_0=1$ is the temperature of the original chain.

(c) For $k=1,2$,

(i) propose with probability 1 to swap $(\boldsymbol{\beta},\boldsymbol{\gamma})\leftrightarrow(\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})$ and

(ii) accept the proposed move with probability $\alpha=\min\{1,A\}$, where

$$A=\left\{\frac{f(\mathbf{y}|\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\,f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}'_{(k)})\,f(\boldsymbol{\gamma}'_{(k)})}{f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})\,f(\boldsymbol{\beta}|\boldsymbol{\gamma})\,f(\boldsymbol{\gamma})}\right\}^{1-t_k}. \tag{14}$$

This sampling scheme can be enriched with additional moves used in population MCMC sampling (such as mutation and crossover), but in our problem the exchange moves that were described above were sufficient to achieve good mixing.

In this algorithm, it remains to specify the proposal distributions $q_{j,k}(\beta_{j,(k)})$. We use Gaussian proposals of the form

$$q_{j,k}(\beta_{j,(k)})\sim N(\bar{\beta}_{j,(k)},\bar{\sigma}^2_{j,(k)}). \tag{15}$$

The proposal parameters for the original chain $\bar{\beta}_{j,(0)}$ and $\bar{\sigma}^2_{j,(0)}$ can be specified by a pilot study of the full model, by the maximum likelihood estimates of the full model (this is MCMC efficient only when the prior on the coefficients is diffuse), by a conditional maximization approach (see, for example, Dellaportas *et al.* (2002), for details on all these methods), or by more sophisticated techniques (Brooks *et al.*, 2003). Whatever approach is used for the specification of the original chain's proposal parameters, for the remaining parameters we can set

$$\bar{\beta}_{j,(k)}=\bar{\beta}_{j,(0)}\quad\text{and}\quad\bar{\sigma}^2_{j,(k)}=\bar{\sigma}^2_{j,(0)}/t_k\qquad\text{for } k=1,2. \tag{16}$$

The proposal specifications in expression (16) can be derived by considering the following approximation to the posterior distribution:

$$f(\boldsymbol{\beta}_{(k)}|\mathbf{y},\boldsymbol{\gamma}_{(k)})\cong N[\tilde{\boldsymbol{\beta}}_{(k)},\{-H_k(\tilde{\boldsymbol{\beta}}_{(k)})\}^{-1}], \tag{17}$$

where $\tilde{\boldsymbol{\beta}}_{(k)}$ is the value maximizing the heated or cooled log-posterior-density with temperature t_k, which is given by

$$\begin{aligned}\Lambda_k(\boldsymbol{\beta}_{(k)})&=t_k\log\{f(\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)}|\mathbf{y})\} \qquad (18)\\ &=c+t_k\log\{f(\mathbf{y}|\boldsymbol{\beta}_{(k)},\boldsymbol{\gamma}_{(k)})\}+t_k\log\{f(\boldsymbol{\beta}_{(k)}|\boldsymbol{\gamma}_{(k)})\}+t_k\log\{f(\boldsymbol{\gamma}_{(k)})\};\end{aligned}$$

here c is a constant and $H_k(\tilde{\boldsymbol{\beta}}_{(k)})$ is the Hessian of $\Lambda_k(\boldsymbol{\beta}_{(k)})$ evaluated at its maximum. The heated or cooled posterior mode $\tilde{\boldsymbol{\beta}}_{(k)}$ is equal to the mode $\tilde{\boldsymbol{\beta}}$ of the actual posterior density (with temperature 1), yielding the first part of expression (16). Moreover, $H_k(\tilde{\boldsymbol{\beta}}_{(k)})$ equals the second derivative of the posterior with temperature 1 multiplied by the temperature t_k, i.e.

Table 2. Preliminary results: variables with marginal posterior probabilities $f(\gamma_j=1|\mathbf{y})$ above 0.30 in at least one 100000 population RJMCMC run†

Index	*Variable name*	*Cost*	*Marginal posterior probabilities*	
			First run	*Second run*
1	Systolic blood pressure score	0.50	0.98	0.99
2	Age	0.50	0.97	0.95
3	Blood urea nitrogen	1.50	0.99	0.91
4	APACHE II coma score	2.50	0.55	1.00
5	Shortness of breath day 1	1.00	0.92	0.80
6	Serum albumin score	1.50	0.40	0.55
12	Initial temperature	0.50	0.91	0.93
37	APACHE respiratory rate score	1.00	0.72	0.79
46	Admission systolic blood pressure	0.50	0.45	0.25
49	Respiratory rate day 1	0.50	0.35	0.25
51	Confusion day 1	0.50	0.44	0.01
62	Body system count	2.50	0.55	0.33
70	APACHE acidity score	1.00	0.81	0.73

†Costs are expressed in minutes of abstraction time.

$$H_k(\tilde{\beta}_{(k)})=t_k\ H(\tilde{\beta}); \tag{19}$$

thus the proposal variance of the chain with temperature t_k can be defined as the variance of the chain with temperature 1 divided by t_k, as in the second part of expression (16).

4. Implementation and results

We used a total cost limit of 10 min of abstraction time, for two reasons:

(a) medical and health policy experts told us that this would lead to feasible implementation costs for widespread hospital screening based on comparisons of observed and expected mortality (with the latter value dependent on a sickness-at-admission scale chosen, for example, with the methods of this paper) and

(b) we had previously found (Fouskakis *et al.*, 2009) using a different method based on a cost–benefit trade-off (rather than the cost restriction–benefit analysis that is pursued here) that costs of 7–8 min of abstraction time were the right order of magnitude for optimizing cost and predictive accuracy in this problem with this data set.

Initially, the algorithm proposed was used to remove variables with posterior inclusion probabilities $f(\gamma_j|\mathbf{y})$ that were below a threshold value; Barbieri and Berger (2004) have shown that this approach may lead to the identification of models with better predictive abilities than approaches that are based on maximizing posterior model probabilities. A threshold value of 0.3 was used for $f(\gamma_j=1|\mathbf{y})$ to identify and eliminate variables that were not contributing to models with high posterior probabilities. The pseudoparameters of equation (12) were tuned, to achieve appropriate acceptance rates (around 20%) for swapping values between chains of different temperatures, resulting in $(a_1,b_1,a_2,b_2)=(2,4,7,3)$ for the RAND pneumonia data.

Table 3. Marginal posterior probabilities $f(\gamma_j = 1|\mathbf{y})$ in the reduced model space

Variable index	*Results for the population RJMCMC algorithm*						*Results for the simple RJMCMC algorithm*			
	100000 iterations		*200000 iterations*		*500000 iterations*		*500000 iterations*		*1.5 million iterations*	
	First run	*Second run*	*First run*	*Second run*	*First run*	*Second run*	*First run*	*Second run*	*First run*	*Second run*
1	0.97	0.98	0.97	0.97	0.97	0.96	0.95	0.96	0.95	0.97
2	0.95	0.96	0.95	0.95	0.95	0.95	0.96	0.95	0.96	0.96
3	0.87	0.87	0.87	0.85	0.85	0.87	0.91	0.87	0.89	0.88
4	1.00	0.85	0.90	1.00	0.96	0.99	1.00	1.00	1.00	1.00
5	0.89	0.89	0.92	0.89	0.89	0.88	0.89	0.87	0.87	0.83
6	0.07	0.06	0.06	0.05	0.05	0.08	0.00	0.00	0.01	0.00
12	0.94	0.95	0.94	0.94	0.93	0.93	0.92	0.94	0.93	0.95
37	0.83	0.89	0.85	0.84	0.84	0.83	0.82	0.90	0.86	0.89
46	0.22	0.31	0.28	0.22	0.26	0.24	0.19	0.19	0.19	0.18
49	0.18	0.16	0.18	0.16	0.17	0.18	0.18	0.11	0.14	0.11
51	0.02	0.17	0.10	0.02	0.06	0.03	0.01	0.01	0.00	0.01
62	0.93	0.94	0.94	0.95	0.95	0.92	1.00	1.00	0.99	1.00
70	0.38	0.48	0.42	0.38	0.41	0.42	0.30	0.33	0.33	0.35

We ran our algorithm for 100000 iterations twice, starting each time from a randomly selected different initial stage which satisfied the cost constraint; variables with posterior inclusion probabilities below 0.3 in both runs were eliminated. In this manner, the number of explanatory variables was reduced from 83 to 13. Table 2 presents the reduced set of variables together with their costs and marginal posterior probabilities in both runs. It is evident that, with only 100000 monitoring runs, for some variables there are differences between the marginal inclusion probabilities that were evaluated in the two runs. However, the purpose of this step in the algorithm (Section 2) is to eliminate variables with very low marginal inclusion probabilities, to reduce the size of the model space substantially without undue computation, and for this the algorithm's first step was successful.

In the reduced model space, various runs of our population RJMCMC algorithm were performed using two replications each of 100000, 200000 and 500000 monitoring iterations. For comparison, we also ran the simple RJMCMC algorithm twice for 500000 and 1.5 million iterations; in central processor unit (CPU) clock time the population RJMCMC approach is almost exactly a third of the speed of the simple RJMCMC algorithm, so a fair comparison in clock time can be achieved by comparing the 500000 population and 1.5 million simple RJMCMC results (Appendix A gives further computing details). To ensure convergence, we started each chain from randomly selected models, all satisfying the cost constraint, and we removed the first 10000 iterations as a burn-in. The effect of the cost constraint on the model space was evident by monitoring the total cost of the models visited. For instance, in both 500000 population RJMCMC runs, about 90% of the models visited had costs equal to the cost limit of 10 min, and fewer than 1.5% models visited had cost less than 9.5 min.

Table 3 presents the marginal posterior probabilities for all runs (Rao–Blackwellization made almost no difference in this case; we report non-Rao–Blackwellized estimates for simplicity). Differences in the marginal posterior inclusion probabilities between the two runs at each number of monitoring iterations indicate lack of convergence for population RJMCMC sampling

Table 4. Reduced model space: MCSEs (in percentage points) of marginal inclusion posterior probabilities†

Type	*Run*	*Iterations*	X_1	X_2	X_3	X_4	X_5	X_6	X_{12}	X_{37}	X_{46}	X_{49}	X_{51}	X_{62}	X_{70}
P	1	500000	0.5	0.2	0.6	1.8	1.0	1.1	0.6	1.1	1.2	1.1	1.7	1.1	1.2
P	2	500000	0.4	0.1	0.5	0.7	1.1	1.2	0.5	0.9	0.9	1.0	0.8	1.2	1.5
P	1	200000	0.5	0.2	0.9	2.0	1.3	1.7	0.6	1.3	1.6	1.3	2.0	1.7	1.9
P	2	200000	0.5	0.3	1.1	0.0	1.3	1.2	0.7	1.4	1.1	1.4	0.2	1.2	1.5
P	1	100000	0.9	0.3	1.3	0.0	1.9	2.1	1.0	2.0	1.5	1.9	0.3	2.1	2.5
P	2	100000	0.5	0.2	1.0	2.3	1.8	1.5	0.6	1.3	2.0	1.5	2.3	1.5	2.8
S	1	500000	2.3	0.9	1.9	0.0	3.7	0.0	2.7	3.5	2.8	3.4	0.3	0.0	4.1
S	2	500000	2.0	0.4	1.5	0.0	3.7	0.0	2.7	4.0	2.2	4.0	0.3	0.0	4.0
S	1	1.5 million	1.2	0.2	1.2	0.0	2.6	1.4	1.6	2.4	1.4	2.4	0.1	1.4	2.8
S	2	1.5 million	1.3	0.4	1.0	0.0	3.3	0.0	1.7	2.3	1.5	2.3	0.2	0.0	3.1

†*S* refers to the simple and *P* to the population RJMCMC runs.

Table 5. Reduced model space: relative comparisons (ratios of the 1.5 million simple (*S*) RJMCMC MCSEs over the MCSEs of each population (*P*) RJMCMC run)

	P iterations	X_1	X_2	X_3	X_4	X_5	X_6	X_{12}	X_{37}	X_{46}	X_{49}	X_{51}	X_{62}	X_{70}
First 1.5 million	500000	2.4	1.0	2.0	0.0	2.6	1.3	2.7	2.2	1.2	2.2	0.1	1.3	2.3
S run *versus P*	200000	2.4	1.0	1.3	0.0	2.0	0.8	2.7	1.9	0.9	1.9	0.1	0.8	1.5
	100000	1.3	0.7	0.9	—	1.4	0.7	1.6	1.2	0.9	1.3	0.3	0.7	1.1
Second 1.5 million	500000	3.2	4.0	2.0	0.0	3.0	0.0	3.4	2.6	1.7	2.3	0.3	0.0	2.1
S run *versus P*	200000	2.6	1.3	0.9	—	2.5	0.0	2.4	1.6	1.4	1.6	1.0	0.0	2.1
	100000	2.6	2.0	1.0	0.0	1.8	0.0	2.8	1.8	0.8	1.5	0.1	0.0	1.1

based on the 100000 and 200000 runs, but minimal differences are observed after running the algorithm for 500000 iterations. By contrast, for the 500000 simple RJMCMC runs, variables X_{37} and X_{49} have marginal posterior inclusion probability differences above 6 percentage points, whereas variable X_6 was never included in any model visited and variable X_4 was always included in all models visited. With 1.5 million iterations the simple RJMCMC algorithm is more stable according to this measure of convergence.

We estimated the Monte Carlo standard errors (MCSEs) for the inclusion probabilities, on the basis of the batch mean method (Geyer, 1992) with 50 batches, to examine the variability due to simulation. Table 4 presents these MCSEs for all the runs; the ratios of the simple (*S*) 1.5 million RJMCMC MCSEs over the corresponding values produced in each population (*P*) RJMCMC run are also given in Table 5. (Zero values in the relative comparisons indicate no variability between the batches in the simple RJMCMC runs, whereas the dashes (—) indicate no variability between the batches in the population RJMCMC runs. Absence of variability between batches may be considered as a sign of a possibly poor exploration of the model space.) The median MCSE ratio S/P with 1.5 million and 500000 iterations for simple and population RJMCMC sampling respectively—a comparison that (as noted above) holds CPU clock time constant for the two algorithms—was 2.03; in other words, it would require simple RJMCMC sampling approximately $2.03^2 = 4.1$ times more clock time than population RJMCMC sampling to achieve the same Monte Carlo accuracy in estimating posterior marginal inclusion probabilities for the variables. The corresponding median ratio values for 1.5 million (*S*) *versus* 200000 and 100000 (*P*) were 1.36 and 1.11 respectively, so even the 100000 population RJMCMC runs

Table 6. Reduced model space: posterior model probabilities above 3% and posterior odds (PO_{1k}) of the best model within each analysis *versus* the current model k†

k	m	*Common variables*	*Additional variables*	*Model probability*	PO_{1k}
Population RJMCMC: first run (500000 iterations)					
1	m_1	$X_1+X_{12}+X_{37}$	$+X_3+X_5$ $+X_{62}$	0.4872	1.00
2	m_2		$+X_5$ $+X_{46}+X_{62}+X_{70}$	0.1202	4.05
3	m_3		$+X_3$ $+X_{70}$	0.0894	5.45
4	m_4		$+X_3+X_5+X_6$ $+X_{70}$	0.0344	14.16
Population RJMCMC: second run (500000 iterations)					
1	m_1	$X_1+X_{12}+X_{37}$	$+X_3+X_5$ $+X_{62}$	0.4879	1.00
2	m_2		$+X_5$ $+X_{46}+X_{62}+X_{70}$	0.1052	4.63
3	m_3		$+X_3$ $+X_{62}+X_{70}$	0.0982	4.97
4	m_4		$+X_3+X_5+X_6$ $+X_{70}$	0.0498	9.80
Simple RJMCMC: first run (1.5 million iterations)					
1	m_1	X_{62}	$+X_1+X_3+X_5+X_{12}+X_{37}$	0.6159	1.00
2	m_3		$+X_1+X_3$ $+X_{12}+X_{37}$ $+X_{70}$	0.1061	5.80
3	m_2		$+X_1$ $+X_5+X_{12}+X_{37}+X_{46}$ $+X_{70}$	0.0926	6.65
4	m_5		$+X_3+X_5$ $+X_{46}+X_{49}+X_{70}$	0.0403	15.28
Simple RJMCMC: second run (1.5 million iterations)					
1	m_1	$X_1+X_{12}+X_{37}+X_{62}$	$+X_3+X_5$	0.5912	1.00
2	m_3		$+X_3$ $+X_{70}$	0.1525	3.88
3	m_2		$+X_5$ $+X_{46}$ $+X_{70}$	0.1041	5.68

†The second column refers to the model indicator of the five different models appearing in the table. Variables X_2 and X_4 were common to all models. All models appearing in the table had a total cost of 10 min (the cost limit).

had higher median Monte Carlo accuracy than the 1.5 million simple RJMCMC runs (and the clock time for the former was a fifth that of the latter).

Table 6 presents the models with posterior model probabilities above 3% (in descending order) for all 500000 and 1.5 million population and simple RJMCMC runs. Posterior odds of the highest posterior probability model compared with the other models are also provided. For the two 500000 population RJMCMC runs, the same highest probability models were obtained, with exactly the same order and with minor differences between their posterior probabilities. In contrast, there is rather less agreement between the 1.5 million simple RJMCMC runs. Specifically, three highest probability models are common in both runs, but with rather different estimated probabilities, whereas one additional model was indicated by the first run. Other differences between population and simple RJMCMC runs are also evident: for example, the fourth highest probability model of population RJMCMC sampling was never visited by simple RJMCMC sampling, and there are large differences between the posterior probabilities of the common best models obtained by the two algorithms, resulting in some cases in different rank ordering.

Results from Table 6 can be used for the implementation of Bayesian model averaging (see, for example, Draper (1995)), or for the selection of a single model, based on the highest posterior probability. Using the 500000 population RJMCMC runs, the latter approach specifies the inclusion of variables $\{1,2,3,4,5,12,37,62\}$ (identified in the fifth column in Table 1). Alternatively, the median probability model (Barbieri and Berger, 2004), which incorporates all variables with marginal posterior inclusion probabilities that are greater than 0.5, can be selected. According to Table 3, for the 500000 population RJMCMC runs, this method leads

Table 7. Reduced model space: MCSEs of the posterior model probabilities for models ($m_1,\ldots,m_4$) from Table 6 in all runs†

Type	*Run*	*Iterations*	*MCSEs (percentage points)*			
			m_1	m_2	m_3	m_4
P	1	500000	1.2	0.5	0.9	0.7
P	2	500000	1.5	0.4	1.0	0.7
P	1	200000	1.9	0.8	1.1	1.2
P	2	200000	1.6	1.0	1.1	0.9
P	1	100000	2.5	1.2	1.7	1.5
P	2	100000	2.7	0.9	1.6	1.2
S	1	500000	4.2	1.3	3.2	0.0
S	2	500000	4.2	1.7	3.6	0.0
S	1	1.5 million	2.9	1.1	2.1	1.0
S	2	1.5 million	3.1	0.9	3.1	0.0

†*S* refers to the simple and *P* to the population RJMCMC runs.

Table 8. Reduced model space: relative comparisons (ratios of the 1.5 million simple (*S*) RJMCMC MCSEs over the MCSEs of each population (*P*) RJMCMC run)

	P iterations	*Relative comparisons*			
		m_1	m_2	m_3	m_4
First 1.5 million	500000	2.4	2.2	2.3	1.4
S run *versus P*	200000	1.5	1.4	1.9	0.8
	100000	1.2	0.9	1.2	0.7
Second 1.5 million	500000	2.1	2.3	3.1	0.0
S run *versus P*	200000	1.9	0.9	2.8	0.0
	100000	1.2	1.0	1.9	0.0

to the same single model as before. Physician experts have told us that this model is clinically sensible for pneumonia: it examines the cardiovascular system of the patient (through a systolic blood pressure score), the kidney function (through a blood urea nitrogen measurement), the patient's responsiveness and neurological function (through a coma score), the severity of the patient's pneumonia (through a measurement of shortness of breath, a respiratory rate score and a temperature reading: low body temperature is a bad sign for pneumonia patients that the infection is not being fought) and two overall measures of function (the patient's age and a count of how many body systems are compromised by the primary illness and comorbidities).

Tables 7 and 8 present the Monte Carlo standard errors (again by using the batch mean method) of the posterior model probabilities for the best models that were obtained by the 500000 population RJMCMC method in all runs, together with the ratios of the simple (*S*) 1.5 million RJMCMC MCSEs over the corresponding values produced in each population (*P*) RJMCMC run. The median MCSE ratio S/P with 1.5 million and 500000 iterations for simple and population RJMCMC sampling was 2.23, so (similarly to the previous conclusion about marginal inclusion probabilities) it would require simple RJMCMC sampling approximately $2.23^2 = 4.9$ times more clock time than population RJMCMC sampling to achieve the same Monte Carlo accuracy in estimating posterior model probabilities. The corresponding median ratio values for 1.5 million (S) *versus* 200000 and 100000 (P) were 1.45 and 1.07 respectively, so (once again) even the 100000 population RJMCMC runs had higher median Monte Carlo accuracy than the 1.5 million simple RJMCMC runs (and required less clock time by a factor of 5).

Table 9 explores the cost–model dimension–accuracy trade-offs between the five models that were identified in Table 6 and the RAND 14-variable scale, by summarizing the deviance (calculated on the entire data set), total cost (in minutes of abstraction time) and number of variables. Two observations are noteworthy:

(a) model m_5, with posterior probability in excess of 4% according to one of the 1.5 million simple RJMCMC runs, has the same cost and dimension as m_1 and m_3 (two of the four best models that were identified in a stable manner by both of the 500000 population RJMCMC runs) but a substantially worse deviance value, and

(b) the RAND scale achieves a deviance that is about 1% lower than the corresponding values for models m_1–m_4 (the high posterior probability models from population RJMCMC sampling), but with a data collection cost that is 210% higher and a model dimension that is 56–75% higher.

It may be thought, from an examination of the population RJMCMC algorithm that was summarized in Section 3—in particular, step 2(b)(ii) of that algorithm, which is based on 1-bit flips—that the algorithm might be improved by enriching the neighbourhood structure with more complicated moves (such as *2-bit swaps*, in which a variable is added and another is simultaneously removed). However, because at every iteration our algorithm does not simply propose a single 1-bit flip but instead proposes p such moves (one for each variable, as in step 2(b)(ii)(A), which is embedded in a loop across all the variables), in fact our neighbourhood structure is already much richer than that induced by single 1-bit flips; if the dimension of the model at stage t is k the dimension of the model at stage $t+1$ can easily be much bigger or smaller. Fig. 1 gives density and time series plots of the model dimension for one of the 1.5 million simple RJMCMC and one of the 500000 population RJMCMC runs (with the former thinned by a factor of 3 to produce visual comparability; the conclusions are the same with the entire 1.5 million series). The mean and standard deviation of dimension for the simple and population RJMCMC runs were (8.12, 0.38) and (8.27, 0.54) respectively; the population RJMCMC standard deviation is

Table 9. Comparison of the models identified in Table 6 and the RAND 14-variable model, on deviance, total cost and model dimension

Model	*Deviance*	*Total cost*	*Dimension*
m_1	1610.0	10	8
m_2	1606.7	10	9
m_3	1612.8	10	8
m_4	1608.6	10	9
m_5	1616.5	10	8
RAND	1587.3	31	14

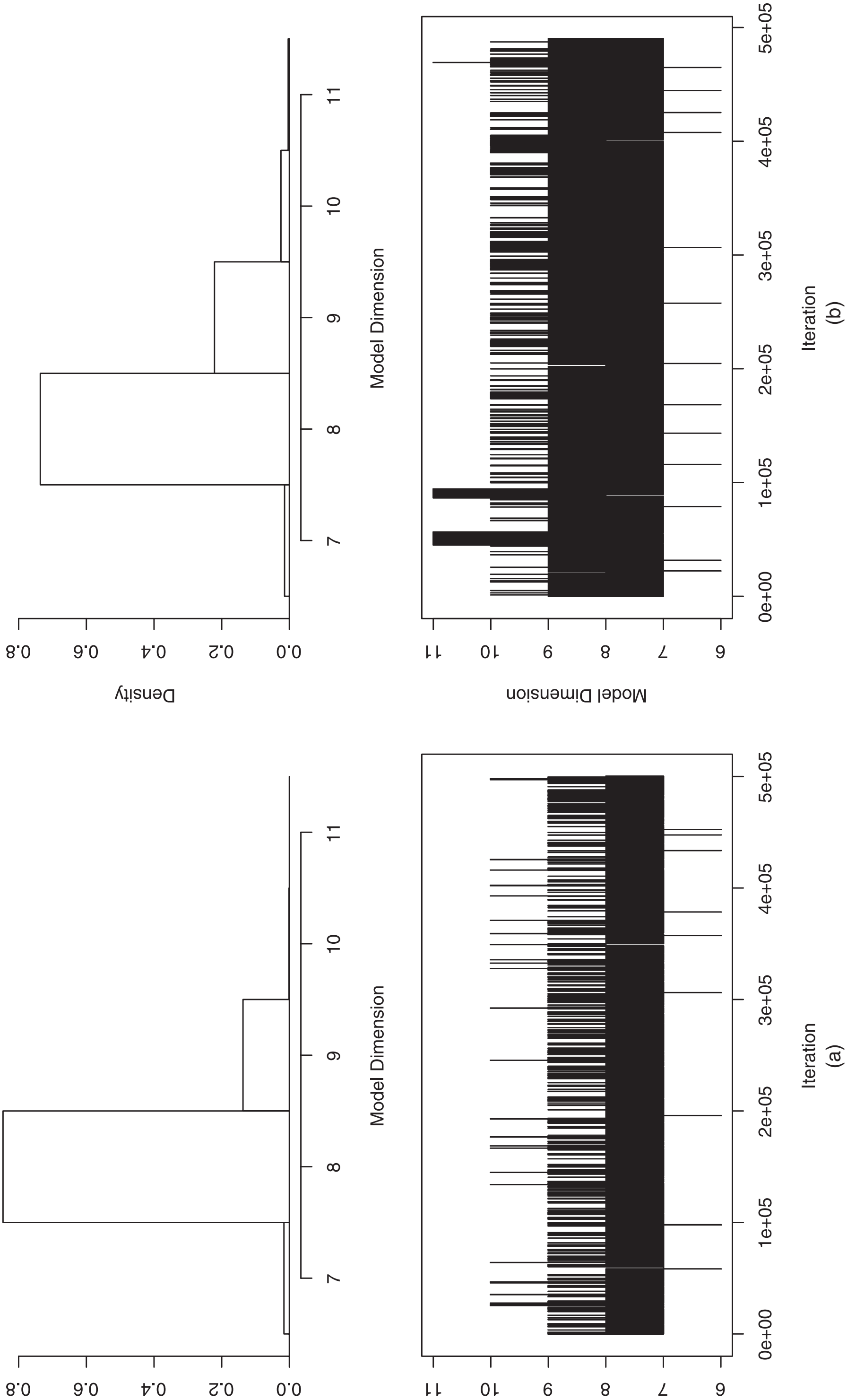
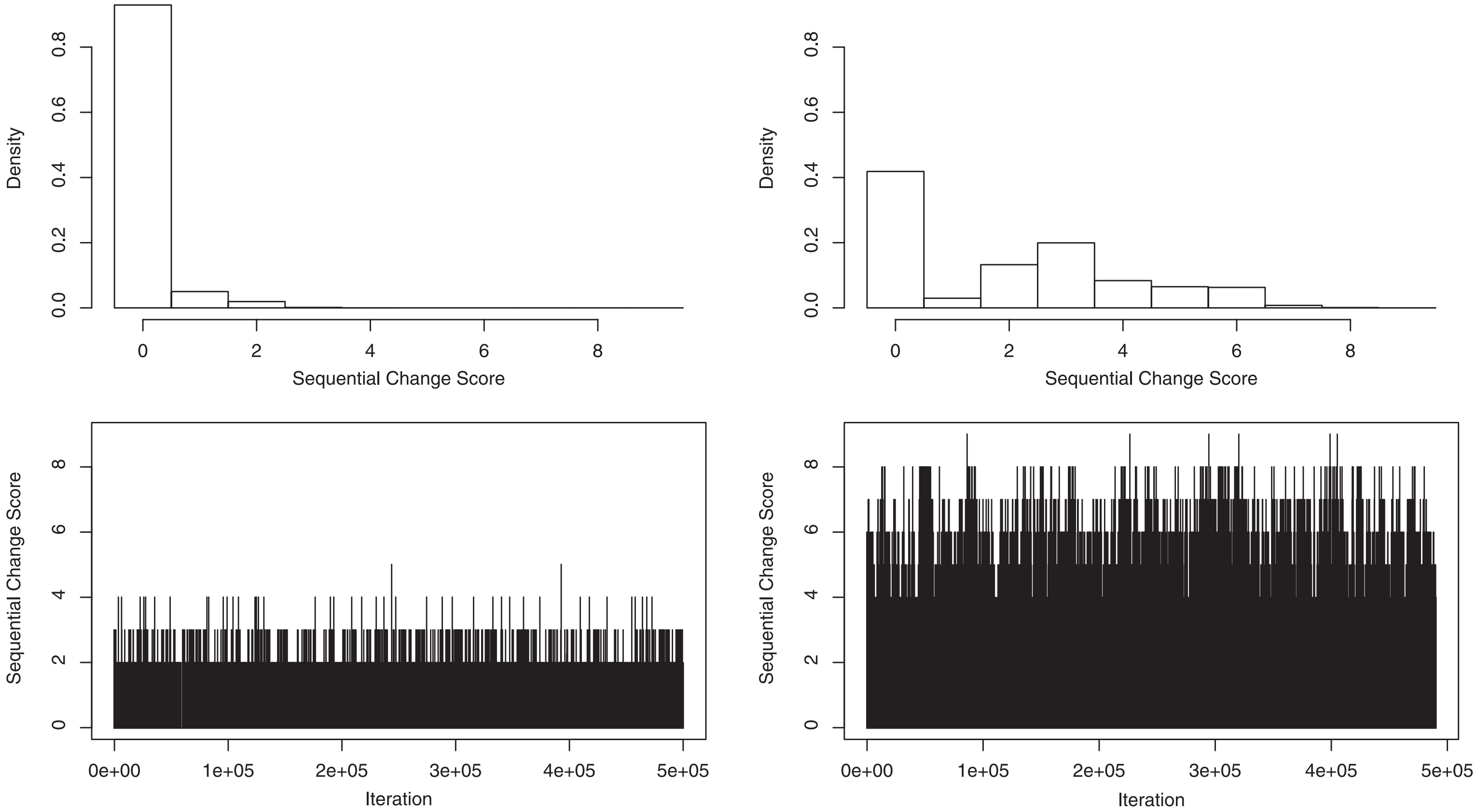


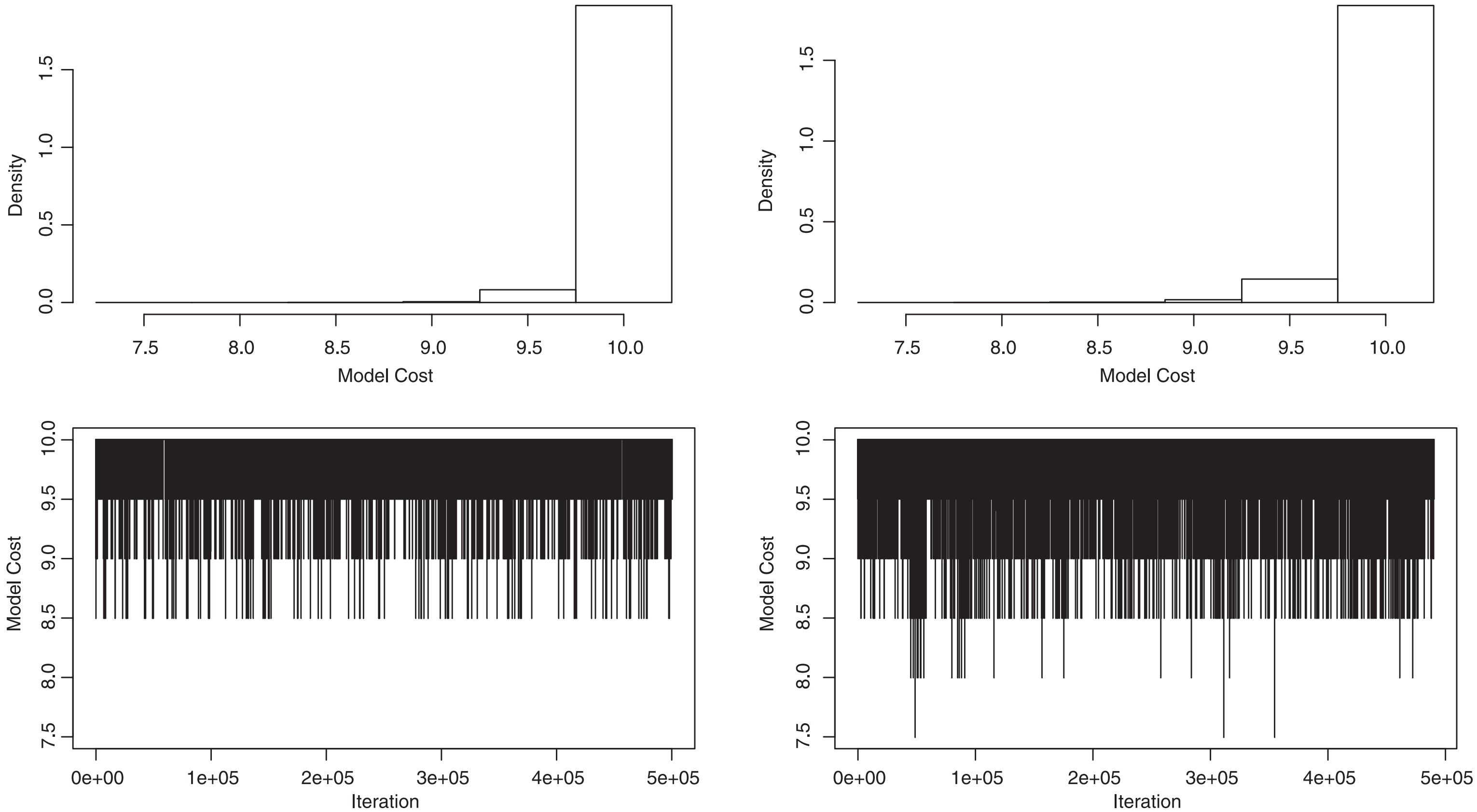
Fig. 1. Density and time series plots of model dimension for (a) the simple and (b) the population RJMCMC algorithms: the simple RJMCMC data are from one of the 1.5 million runs, thinned by a factor of 3, and the population RJMCMC data are from one of the 500000 runs



(a)

(b)

Fig. 2. Density and time series plots of sequential change score for (a) the simple and (b) the population RJMCMC algorithms: this counts the number of variables in the model at iteration $t + 1$ that are different from those in the model at iteration t; the simple RJMCMC data are from one of the 1.5 million runs, thinned by a factor of 3, and the population RJMCMC data are from one of the 500000 runs



(a)

(b)

Fig. 3. Density and time series plots of model cost for (a) the simple and (b) the population RJMCMC algorithms: the simple RJMCMC data are from one of the 1.5 million runs, thinned by a factor of 3, and the population RJMCMC data are from one of the 500000 runs

44% larger, indicating substantially greater movement for this algorithm with respect to model dimension than for simple RJMCMC sampling (this is confirmed visually from the time series plots in Fig. 1). Fig. 2 gives density and time series plots for the *sequential change score* for both algorithms; this counts the number of variables in the model at iteration $t+1$ that are different from those in the model at iteration t. For simple RJMCMC sampling the mean of this score is 0.09, and 98% of the time there is no change in variables from one iteration to the next; for population RJMCMC sampling the mean is 1.99 and a sequential change score of 0 occurs only 42% of the time. Fig. 3 gives density and time series plots for the model cost for the two algorithms. It is evident that population RJMCMC sampling is considerably more adventurous than simple RJMCMC sampling in its willingness to move away from models with cost equal to the imposed limit (10), in pursuit of other models with cost 10 that have even better predictive performance: the standard deviation of cost for population RJMCMC sampling (0.15) is 44% larger than that for simple RJMCMC sampling (0.11), and simple RJMCMC sampling spends only 4% of its time with models of cost less than 10, whereas the corresponding value for population RJMCMC sampling is 8%.

To summarize the findings of this section, population RJMCMC sampling moved successfully between distant neighbourhoods of good models, achieving convergence in a reasonable number of iterations, whereas simple RJMCMC sampling explored the model space poorly, as indicated by the estimated posterior model probabilities, the Monte Carlo standard errors and the sequential change scores. The final model that was chosen by population RJMCMC sampling, both on the basis of highest posterior probability and specifying the median probability model, is clinically sensible for pneumonia patients and achieves good predictive ability while capping data collection costs.

5. Discussion

In this paper, we have addressed a Bayesian variable selection problem arising in a health evaluation study, accounting for the data collection cost of each predictor while imposing a budgetary constraint on the total cost. In such problems, the implementation of standard model search algorithms, such as simple RJMCMC sampling, will fail, since multiple modes may exist on the cost boundary restriction. Therefore, we developed a population-based trans-dimensional RJMCMC algorithm (population RJMCMC sampling), combining ideas from the population-based MCMC and simulated tempering algorithms. Computation is performed using population RJMCMC sampling 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. Comparing the proposed technique with simple RJMCMC sampling, we find that the population RJMCMC algorithm explores the model space efficiently and converges much faster, with lower Monte Carlo standard errors (for a given amount of CPU time) than simple RJMCMC sampling. The final model identified by population RJMCMC sampling achieves clinical plausibility and an effective cost restriction–benefit trade-off between data collection cost and predictive accuracy.

Future health policy work that is motivated by this study would include formulating the entire problem of quality-of-care monitoring on the basis of comparisons of observed and expected mortality in Bayesian decision theoretic terms: given a fixed budget for monitoring all the hospitals in an administrative region (e.g. a county, state or nation), what are the optimal numbers of hospitals and patients per hospital to sample, and what is the optimal subset of predictors of sickness to use when the problem is viewed in this broader context?

In terms of population-based MCMC algorithms, two useful extensions are as follows.

(a) It would be a further step forward to identify a Monte Carlo estimator that makes use of the generated data from all the parallel chains and not only from the original chain; see Coluzza and Frenkel (2005) and Gramacy *et al.* (2007) for suggestions on how this might be accomplished in contexts that are similar to ours. This would increase the efficiency of the MCMC sampler, reducing the Monte Carlo error and resulting in a computationally faster algorithm.

(b) Our algorithm can be extended in a straightforward manner to improve the mixing of the MC^3 algorithm (Madigan and York, 1995), which is used in graphical and normal linear models and models for qualitative data.

Acknowledgements

The authors are grateful to Dr Katherine Kahn for providing them with data from the RAND diagnosis-related group quality-of-care study, to Dr Leonardo Bottolo for his productive comments on an earlier version of the paper, and to the Joint Editor, Associate Editor and a referee for comments that substantially improved the content and presentation.

Appendix A

Using efficient and optimized C code running under LINUX on a Pentium Celeron machine with 3.66 GHz of CPU speed and 1 Gbyte of random-access memory, we estimate that the clock times for both 500000 monitoring iterations with the population RJMCMC method and 1.5 million iterations with simple RJMCMC sampling would be approximately 4725 min (about 3.3 days).

References

Barbieri, M. D. and Berger, J. O. (2004) Optimal predictive model selection. *Ann. Statist.*, **32**, 870–897.

Brooks, S. P., Giudici, P. and Roberts, G. O. (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. R. Statist. Soc.* B, **65**, 3–55.

Brown, P. J., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc.* B, **60**, 627–641.

Brown, P. J., Vannucci, M. and Fearn, T. (2002) Bayes model averaging with selection of regressors. *J. R. Statist. Soc.* B, **64**, 519–536.

Coluzza, I. and Frenkel D. (2005) Virtual-move parallel tempering. *ChemPhysChem*, **6**, 1779–1783.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002) On Bayesian model and variable selection using MCMC. *Statist. Comput.*, **12**, 27–36.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.

Fouskakis, D. and Draper, D. (2008) Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. *J. Am. Statist. Ass.*, to be published.

Fouskakis, D., Ntzoufras, I. and Draper, D. (2009) Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Ann. Appl. Statist.*, to be published.

Geyer, C. J. (1992) Practical Markov Chain Monte Carlo (with discussion). *Statist. Sci.*, **7**, 473–511.

Geyer, C. J. and Thompson, E. A. (1995) Annealing Markov Chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Ass.*, **90**, 909–920.

Goldstein H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.

Gramacy, R. B., Samworth, R. J. and King, R. (2007) Importance tempering. *Technical Report*. Statistical Laboratory, University of Cambridge, Cambridge.

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

Jasra, A., Stephens, D. A. and Holmes, C. C. (2007a) Population-based reversible jump MCMC. *Biometrika*, **94**, 787–807.

Jasra, A., Stephens, D. A. and Holmes, C. C. (2007b) On population-based simulation for static inference. *Statist. Comput.*, **17**, 263–279.

Kahn, K., Rubenstein, L., Draper, D., Kosecoff, J., Rogers, W., Keeler, E. and 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 (with editorial comments). *J. Am. Med. Ass.*, **264**, 1953–1997.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *J. Am. Statist. Ass.*, **91**, 1343–1370.

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

Lindley, D. V. (1968) The choice of variables in multiple regression (with discussion). *J. R. Statist. Soc.* B, **30**, 31–66.

Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, 215–232.

Ohlssen, D. I., Sharples, L. D. and Spiegelhalter, D. J. (2007) A hierarchical modelling framework for identifying unusual performance in health care providers. *J. R. Statist. Soc.* A, **170**, 865–890.

Zhang, M., Strawderman, R. L., Cowen, M. E. and Wells, M. T. (2006) Bayesian inference for a two-part hierarchical model: an application to profiling providers in managed health care. *J. Am. Statist. Ass.*, **101**, 934–945.