STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:3215–3230 (DOI: 10.1002/sim.976)

Variable selection and Bayesian model averaging in case-control studies

Valerie Viallefont[1,*], Adrian E. Raftery[2] and Sylvia Richardson[1,‡]

[1]*INSERM-U. 170, 16 av P. Vaillant-Couturier, 94 807 Villejuif Cedex, France*
[2]*Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, U.S.A.*

SUMMARY

Covariate and confounder selection in case-control studies is often carried out using a statistical variable selection method, such as a two-step method or a stepwise method in logistic regression. Inference is then carried out conditionally on the selected model, but this ignores the model uncertainty implicit in the variable selection process, and so may underestimate uncertainty about relative risks. We report on a simulation study designed to be similar to actual case-control studies. This shows that p-values computed after variable selection can greatly overstate the strength of conclusions. For example, for our simulated case-control studies with 1000 subjects, of variables declared to be 'significant' with p-values between 0.01 and 0.05, only 49 per cent actually were risk factors when stepwise variable selection was used. We propose Bayesian model averaging as a formal way of taking account of model uncertainty in case-control studies. This yields an easily interpreted summary, the posterior probability that a variable is a risk factor, and our simulation study indicates this to be reasonably well calibrated in the situations simulated. The methods are applied and compared in the context of a case-control study of cervical cancer. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

In this paper we consider variable selection in epidemiological case-control studies. Epidemiologists conduct case-control studies in order to test the existence of possible risk factors of interest [1, 2], and to estimate their association with the presence or absence of a disease, after adjusting for possible confounders. A model often used is logistic regression, namely

$$\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q \tag{1}$$

*Correspondence to: Valerie Viallefont, Department of Mathematics and Statistics, Faylds College, Lancaster University, Lancaster LA1 HYF, U.K.
‡ Presently at: Department of Epidemiology and Public Health, Imperial College, School of Medicine, St Mary's Campus, Norfolk Place, London W21PG, U.K.

Contract/grant sponsor: Office of Naval Research; contract/grant number: N00014-96-1-0192

Received May 1999
Accepted March 2001

where $Y = 1$ if the disease is present and $Y = 0$ if it is absent, X_1 is a dichotomous risk factor of interest, $X_2, \ldots, X_q$ are confounders, and $\beta_0, \beta_1, \ldots, \beta_q$ are regression parameters. The choice of the confounders to include is a major issue. In many studies, the potential confounders are numerous, including demographic, socio-economic, familial disease history, smoking and other lifestyle variables, as well as medical measurements. Thus, investigators have tended to use statistical methods to choose among the many confounders indicated by substantive considerations. Two commonly used methods are a two-stage method [3] and backwards stepwise regression; both of these are described, for example, in the influential textbook by Hosmer and Lemeshow [4].

Investigators typically carry out tests and compute confidence intervals conditionally on the selected logistic regression model [5], without taking account of the fact that variable selection has been done. It has been shown, particularly in the linear regression context, that doing this can yield misleading results, often tending to reject null hypotheses more often than the nominal levels would suggest, and to produce confidence intervals that are too narrow (for example, references [6, 7]).

Here we present a method, *Bayesian model averaging*, that provides a formal way of taking account of this uncertainty in both tests and confidence intervals. We illustrate the method with an application to a case-control study investigating risk factors for cervical cancer [8]. Then we carry out a simulation study designed to be representative of actual case-control studies, and show that p-values computed after two-stage or stepwise variable selection can be quite misleading, while the posterior probabilities from Bayesian model averaging achieve roughly their nominal error rates.

2. BAYESIAN MODEL AVERAGING

2.1. General principles

Bayesian model averaging (BMA) is a Bayesian solution to the problem of inference in the presence of multiple competing models [9–17]. For general introductions to Bayesian inference, see references [18–20].

BMA starts by acknowledging that in the situation of equation (1), there are up to $K = 2^q$ possible models (assuming that there are no interactions between the risk factors) defined by allowing each of $X_1, \ldots, X_q$ to be either in or out of the model. We denote these models by $M_1, \ldots, M_K$. We do not know in advance which of these is the best model, and so there is model uncertainty. BMA simply propagates this uncertainty through to inference about any quantity of interest in the same way as the Bayesian approach propagates any other form of uncertainty. This is done using the law of total probability, that is, by summing or integrating over the quantities that are not of primary interest and about which there is uncertainty.

Suppose that Q is a quantity of interest that has the same interpretation in each of the models considered. Here this is an adjusted odds ratio (OR), but it could also be a future observation, or the utility of a course of action. Its posterior distribution, taking account of model uncertainty, is

$$p(Q|D) = \sum_{k=1}^{K} p(Q|D, M_k)\, p(M_k|D) \tag{2}$$

where D denotes the data at hand (here, observations of disease status, risk factors and confounders), $p(Q|D,M_k)$ is the posterior distribution of Q under model M_k, and $p(M_k|D)$ is the posterior probability of model M_k given the data. Thus the overall posterior distribution of Q is a weighted average, or mixture, of its model-specific posterior distributions, where the weights are the posterior model probabilities.

The posterior model probability, $p(M_k|D)$, of model M_k given the data, is given by

$$p(M_k|D) \propto p(D|M_k)\,p(M_k) \tag{3}$$

where the constant of proportionality is chosen so that the posterior model probabilities add up to one. In equation (3), $p(M_k)$ is the prior model probability of model M_k; these are often chosen to be equal so as not to favour one model over another *a priori*. The quantity $p(D|M_k)$ is the *integrated likelihood* of model M_k, namely

$$p(D|M_k) = \int p(D|\beta^k, M_k)\,p(\beta^k|M_k)\,\mathrm{d}\beta^k \tag{4}$$

where β^k is the vector of regression parameters for model M_k, $p(D|\beta^k, M_k)$ is its (ordinary) likelihood and $p(\beta^k|M_k)$ is its prior distribution, both under model M_k.

Equation (4) poses two problems. The first is the evaluation of the integral, which does not usually have an analytic form and can be of high dimension. Fortunately, for logistic regression, and indeed for generalized linear models more broadly, an accurate and quite tractable approximation is available via the Laplace method [14]. This can be implemented, and BMA carried out for logistic regression, using the `glib` software, which runs under S-Plus and is available on the web at the BMA Home Page (www.research.att.com/~volinsky/bma.html) or at Statlib (http://lib.statlib.cmu.edu/S/glib).

The second problem is the choice of the prior distribution of the parameters, β^k. We will use a prior distribution in which each of the β_j^k is independent and normal, with mean zero for all the parameters except the intercept, and with a prior standard deviation equal to a common scale parameter, ϕ, divided by the standard deviation of X_j if X_j is continuous. This seems capable of representing a reasonable range of prior distributions while requiring the specification of only a single prior parameter, ϕ; see reference [14]. We will discuss the choice of ϕ below.

BMA has several appealing theoretical long-run properties. First, for the prediction of new observations, it gives better predictive performance, on average, than any single model that could reasonably have been selected [10]; this theoretical result has been widely verified in practice [21]. Second, inferences are well calibrated in the sense that, for example, confidence intervals have the right coverage on average [22].

This general approach has been used in several previous analyses of medical and epidemiological data [10, 21, 23–26].

2.2. BMA inference for an adjusted odds ratio

BMA provides hypothesis tests, point estimates and confidence intervals for an adjusted OR, all of which take account of model uncertainty.

For hypothesis testing, from a Bayesian point of view, the question becomes ‘what is the posterior probability that X_1 is a risk factor, that is, that β_1, the adjusted log OR, is not equal

to zero?'. This is simply the sum of the posterior probabilities of the models that include X_1, namely

$$\Pr[\beta_1 \neq 0|D] = \sum_{M_k : X_1 \in M_k} p(M_k|D) \tag{5}$$

We choose the notation $\Pr[\beta_1 \neq 0|D]$ to denote this probability, the expression $(\beta_1 \neq 0)$ indicating solely that β_1 is included and estimated in at least one model and not necessarily to be taken literally. In particular, this expression can be viewed as an approximation to the statement that β_1 is small, where 'small' typically means less than about one-half of a standard error [27].

Conventional rules of thumb for interpreting this quantity verbally are that if $\Pr[\beta_1 \neq 0|D]$ is less than 50 per cent, there is no evidence for X_1 being a risk factor, if it is between 50 per cent and 75 per cent there is weak evidence for X_1 being a risk factor, if it is between 75 per cent and 95 per cent there is positive evidence, between 95 per cent and 99 per cent the evidence is strong, and beyond 99 per cent the evidence is very strong (reference [28], Appendix B; [12]).

A Bayesian point estimate of β_1 is its posterior mean, given that X_1 is in the model, namely

$$E[\beta_1|D] = \sum \hat{\beta}_1^k \, p(M_k|D) \tag{6}$$

where $\hat{\beta}_1^k$ is the posterior mean of β_1 under model M_k, which is zero if X_1 is not in M_k. This is a weighted average of the model-specific point estimates, where once again the weights are the posterior model probabilities. A Bayesian standard error, the posterior standard deviation of β_1, is equal to the square root of

$$\mathrm{var}(\beta_1|D) = \sum \{(\mathrm{var}(\beta_1|D, M_k) + (\beta_1^k)^2)\, p(M_k|D) - E[\beta_1|D]^2\} \tag{7}$$

Inference about β_1, including testing, point estimation and approximate interval estimation, is summarized by equations (5), (6) and (7). A difficulty is that each of these equations involves summation over all $K = 2^q$ possible models, and K will often be impracticably large. To get around this, we approximate the full sum by excluding models that are far less probable *a posteriori* than the best model, an approach known as *Occam's window* [10]. Here we adopt the convention of excluding models whose posterior probability is less than one-twentieth of that of the best model. Implementation details are given in the Appendix.

The BMA approach requires the specification of the following prior quantities:

(i) the prior probabilities $P(M_k)$ of the models, which we will take to be equal, so as not to favour any model *a priori*;
(ii) the prior distribution of the parameters β_i. As suggested in reference [14] for the case of weak prior information, we assume that the distribution of β is a centred Gaussian with a standard deviation ϕ to be specified.

To choose a value of ϕ, we note that e^{β_i} is the odds ratio (OR) corresponding to exposure to a dichotomous X_i. Hence a choice of ϕ can be translated into probability statements about the distribution of typical ORs. We chose ϕ so that the prior probability of finding an OR greater than 7 will be less than 5 per cent. This was motivated by the shared expectation, based on the consideration of the range of ORs found in a review of case-control studies published in 1996 in the *American Journal of Epidemiology* (see Section 3.2.1) and on discussions with

epidemiologists at INSERM, that most ORs investigated in case-control studies will be less than 7, or greater than 1/7 in the case of a protective factor. One reason for this is that, for risk factors with greater relative risks than this, whether they are risk factors or not tends not to be an issue, and research interest focuses on quantifying the association rather than establishing it. This is more likely to be done by a dose-response study than a case-control one. Note that this prior choice can be easily modified by simply changing the value of ϕ.

3. APPLICATION

To illustrate the application of Bayesian model averaging to an epidemiological observational study, we re-analyse a case-control study on cervical cancer. It was conducted and analysed by Peters *et al.*, and we will refer to their results published in reference [8]. It included 400 subjects: 200 women with invasive squamous cell carcinoma of the uterine cervix and 200 controls matched on age, neighbourhood of residence and preferred language (Spanish or English).

3.1. Classical analyses

In reference [8], this data set was analysed as a set of matched pairs, considering 35 risk factors classified into categories (sexual history, method of contraception, genital infection, risk inducing behaviour and demographic characteristics). The authors did not include any interaction terms, and did not focus on any specific factor of the 35. As a selection strategy, they first performed a univariate analysis for each of the 35 variables, of which 25 had two-sided p-values smaller than 0.10. Before performing a multivariate analysis on the selected variables, they constructed meaningful combinations of some of the 35 variables, for example those related to smoking, leaving a total of 28 variables. Of these 28 variables, 18 had been selected by the previous univariate analysis step. Starting with these 18 variables, they selected a multiple logistic regression model using a forward stepwise method. This led to a final model with eight variables. Seven of these eight variables were statistically significant at the 5 per cent level, and the eighth, 'Years from menarche to first intercourse', was significant only at the 10 per cent level. Because excluding it from the final subset consistently changed the estimates of most of the seven other logistic coefficients, the authors decided to keep it in their 'final model', reproduced here in Table I. The successive analyses and considerations

Table I. Cervical cancer study: adjusted effects as published in reference [8].

	$\hat{\beta}$	SE	p-value
Sexual partners before age 20	0.62	0.21	<0.01
Years using barrier contraceptives	-0.142	0.043	<0.01
Episodes of genital warts	1.32	0.45	<0.01
Years of education	-0.252	0.073	<0.001
Years since last PAP smear (log)	0.86	0.18	<0.001
Cumulative smoking exposure	0.00229	0.00063	<0.001
Cumulative douche use	0.0081	0.0022	<0.001
Years from menarche to first intercourse	-0.078	0.049	<0.10

that eventually yielded the 'final model' illustrate the difficulty of selecting among a large number of potential risk factors.

To complete this classical analysis, we also ran a stepwise forward analysis starting from all the 28 available variables. Unfortunately, two of them ('estimated number of other partners of current partner', and 'douched with water-vinegar, (yes/no)') were no longer available, but these two missing variables had not been statistically significant in the univariate analysis [8]. This left a total of 26 available variables. We thus carried out a stepwise procedure starting with all the 26 variables; this led to ten of them being included in the final model, including 'Years from menarche to first intercourse'. Table II presents the logistic regression coefficients and standard errors estimated in this second 'final' model. It is interesting to note the differences between these two 'final' models, both selected by reasonable and well-recognized methods.

3.2. Bayesian model averaging

For comparability, we used the same variable definitions as in reference [8], we did not consider interactions, and we analysed the data as matched pairs. The BMA approach starts by considering all possible combinations of the 26 available variables, yielding an initial set of 2^{26} models, reduced by the use of Occam's window (see Appendix for details).

The prior distribution of β_i is a centred Gaussian distribution whose standard deviation, ϕ, was chosen so that the OR between unexposed and exposed subjects would fall in the interval $[1/7, 7]$ with prior probability 95 per cent. For a dichotomous variable, this OR is e^{β_i}, and the interval $[1/7, 7]$ corresponds to choosing $\beta_i \sim \mathrm{N}(0, 1)$ *a priori*. For a continuous exposure variable X_i, we chose the prior standard deviation of β_i so that the OR between subjects at the 25th and 75th percentiles would also fall in the interval $[1/7, 7]$ with probability 95 per cent. The resulting prior standard deviations thus depend on the variability, specifically the interquartile range, of the corresponding continuous variables.

There were 76 models in Occam's window, and these were used to calculate the BMA estimates of the regression coefficients (see the right hand side of Table II). The posterior probability of being a risk factor, given by $\Pr(\beta_i \neq 0|D)$, is also shown. Note that a reported posterior probability of 100 per cent results from the use of the Occam's window approximation and indicates that all the models that were plausible *a posteriori* contained the corresponding variable. If full BMA were carried out, averaging over all possible models, the posterior probability would likely be close to, but not exactly equal to 100 per cent. Table II summarizes the estimation of the logistic coefficients as well as the associated probabilities: the p-values for the 'classical' analyses and the $\Pr(\beta_i \neq 0|D)$ for the Bayesian analysis. The variables that do not figure in Table II had both a non-significant p-value and a posterior probability lower than 5 per cent.

We now compare the results of the classical analysis with those of BMA. Six variables had p-values below 0.01, and these all had posterior probabilities of 100 per cent, so that for these variables the two approaches were in agreement.

Two variables had p-values between 0.01 and 0.05. For one of these, the number of genital warts episodes, the posterior probability was 97 per cent, so the two analyses agreed. For the other one, 'other cervicitis', the posterior probability was 82 per cent, so that the p-value seems too decisive once model uncertainty is taken into account. An interesting point about the 'other cervicitis' variable is that it did not figure in the published results (Table I), because

Table II. Classical and Bayesian analyses of the cervical cancer study.

	Classical variable selection			Bayesian model averaging		
	$\hat{\beta}$	SE	p-value	$E(\beta\|D)$	$\mathrm{SD}(\beta\|D)$	$\Pr(\beta_i \neq 0\|D)$
Sexual partners before age 20	0.70	0.25	0.004	0.61	0.23	100%
Years using barrier contraceptives	−0.15	0.046	0.0007	−0.12	0.044	100%
Episodes of genital warts	1.48	0.48	0.002	0.93	0.44	97%
Years of education	−0.26	0.079	0.001	−0.24	0.076	100%
Years since last PAP smear (log)	0.99	0.20	<0.0001	0.71	0.18	100%
Cumulative smoking exposure	0.0029	0.00073	<0.0001	0.0024	0.0007	100%
Cumulative douche use	0.0089	0.0023	0.0001	0.0078	0.0023	100%
Years from menarche to first intercourse	−0.078	0.051	0.13	−0.074	0.05	39%
Genital herpes (yes/no)				0.31	0.69	13%
Age at first intercourse				−0.047	0.065	12%
Income				−0.085	0.12	9%
Partner with genital warts				0.11	0.51	7%
Age at first regular intercourse				−0.03	0.04	6%
Gonorrhoea or syphilis (yes/no)				−0.88	0.61	66%
Other cervicitis (yes/no)	1.23	0.54	0.022	0.73	0.46	82%
Intra-uterine devices	−0.14	0.074	0.052	−0.11	0.073	13%

the authors initially excluded variables whose p-values was greater than 0.10 in a univariate analysis, which was the case for 'other cervicitis'.* Thus it seems that the classical analysis either indicated significant evidence for 'other cervicitis' being a risk factor, or excluded it completely, depending on the precise methodology used. BMA, on the other hand, consistently indicated positive but not strong evidence for this variable.

The variables that were 'not significant' in the classical analysis generally had posterior probabilities below 50 per cent, in most cases well below. There was one notable exception, however, namely 'gonorrhoea or syphilis', whose posterior probability was 66 per cent. Thus the classical analysis missed the (weak) evidence in the data for an effect of this variable. This fairly high posterior probability indicates that this variable may be a good marker of past sexual risk factors.

The only other non-significant variable for which the posterior probability approached 50 per cent was 'years from menarche to first intercourse' (39 per cent). Again, BMA gives a more nuanced result than the classical approach, not indicating evidence for this variable but not ruling it out either. If this were an important variable, this result would point towards the need for more research on its possible effect.

*The results published in reference [8] were based on stepwise variable selection starting from 18 rather than 26 variables.

The logistic regression parameter estimates from the BMA analysis are shrunk towards zero in comparison with the classical analyses; this is common in Bayesian estimation, with or without model averaging.

4. SIMULATION STUDY

We have illustrated the use of BMA for the analysis of a case-control study. We have seen that the results differ in some respects from those given by a classical analysis. Now we compare the performance of the BMA and non-Bayesian approaches via a simulation study that is designed to be realistic, that is, to resemble case-control studies carried out in practice. First we describe the three strategies for variable selection that we compare, namely two-step, stepwise and BMA. Then we describe the simulation design, and finally we present the results.

4.1. Methods

The difficulties of variable selection techniques have been discussed by many authors [1,4, 29–31]. In order to characterize variable selection approaches commonly used for logistic regression in epidemiology, we reviewed the case-control studies published in the *American Journal of Epidemiology* in 1996, as well as the recent methodological literature. The results of case-control studies published in the *American Journal of Epidemiology* in 1996 were usually presented in terms of just one single model: the choice of this *single final model* corresponds most of the time to a complex mixture of statistical and epidemiological arguments. We cannot model the full range of epidemiological considerations which influence the choice of the final model. We will oversimplify this procedure by considering the choice of models to be based only on significance level criteria, in two 'classical' strategies.

4.1.1. The two-step procedure. This procedure, described in reference [4], has been considered and examined in reference [3] as a possible strategy for identifying confounders. A first step of selection is carried out by univariate logistic regression. In a second step, each variable with a p-value in the first stage less than a threshold is retained for inclusion in a subsequent multiple logistic regression. The threshold is frequently taken as $\alpha = 0.20$. The associations interpreted are essentially those corresponding to variables having p-values less than $\alpha = 0.05$ in the second stage.

4.1.2. Stepwise backwards selection. In this strategy, all the variables are included in the first model, then some are eliminated in a stepwise manner. One iteration of the procedure consists of the evaluation of the p-value of each variable in a multiple logistic regression, and the elimination of the variable which has the highest p-value. The process stops when the p-values corresponding to all the remaining variables are less than 0.05, or some other threshold. The coefficients of the remaining variables are then re-estimated in this model.

4.1.3. Bayesian model averaging. Its principles have been presented in Section 1, and applied to real data in Section 2. Here, the variables are dichotomous, so the prior for each logistic regression coefficient is a Gaussian with mean 0 and variance 1. All the models are equiprobable *a priori*. The models included in the model averaging are those in Occam's window (see

the details in the Appendix). The quantities (means and variance of the regression coefficient for each variable) are estimated by BMA on the selected models.

4.2. Design of the simulation study

4.2.1. Basis for the simulation study design: The case-control studies in the *American Journal of Epidemiology* in 1996. Determining a few cases of 'typical' data sets to be simulated and analysed is a delicate task and we have based our choice on a review of case-control studies published in the *American Journal of Epidemiology* in 1996. We have selected 50 case-control studies, which are used essentially as providing useful guidelines for the design of our simulation study.

In the set-up of our simulation study, all the variables are uniformly considered as potential risk factors, with no distinction between them and no specific focus on one of them.

The 50 selected studies had numbers of subjects ranging from 118 to 9913, with a median of 894. We chose two different cases, each representative of substantial numbers of published studies: 'Simulation 1' includes 300 subjects and 'Simulation 2' includes 1000, each with equal numbers of cases and controls.

4.2.2. The variables. The total number of variables initially considered is rarely fully reported in the articles. A detailed examination of the 50 studies, with special attention to the ones where variables were all considered as potential risk factors, led us to choose $q = 32$, which is also close to the number of variables in the cervical cancer study that we re-analysed.

The number of variables actually associated with the health outcome, ideally corresponding to the typical dimension of a 'final model' found in case-control studies, was chosen as 10, again based on our review of the published studies. We split this group into a subgroup of five variables correlated with each other and five others independent of each other. Among the 22 remaining variables which are not linked to the health outcome, some are also correlated with each other. Our aim is to encompass typical classes of 'explanatory' variables recorded in epidemiological studies.

Before being entered in a logistic regression, the variables are almost always categorized or dichotomized. This is also the choice in our two simulations: the independent variables were Bernoulli, while the variables correlated with each other were simulated as centred multivariate normal with all correlations equal to 0.4, then dichotomized. The frequency of exposure to the risk factors was set at 20 per cent.

Table III shows the distribution of the recorded ORs (or of 1/OR if OR < 1), separately for the smaller and larger studies reported in the *American Journal of Epidemiology* in 1996, with 200–400 and 700–1300 subjects, respectively. The observed ORs were higher in the smaller studies. This is not unexpected as smaller studies have the power to detect only stronger

Table III. Distribution of the values of ORs of primary interest found in 50 case-control studies reported in the *American Journal of Epidemiology* in 1996, for smaller and larger studies separately.

	Minimum	1st quartile	Median	3rd quartile	90 per cent percentile	Maximum
$n \in [200–400]$ (19 ORs)	1.4	2.2	2.5	2.9	8.3	15.2
$n \in [700–1300]$ (41 ORs)	1.0	1.1	1.5	2.2	4.4	10.7

Table IV. Design of the simulations: correlation and strength of the simulated associations.

	$n = 300$			$n = 1000$	
Variables independent of Y					
X_{1-10}			Independent of each other		
X_{11-15}			Correlated with each other		
X_{16-20}			Correlated with each other		
X_{21} and X_{22}			Correlated with covariates X_{23}–X_{27}		
	β_i	OR		β_i	OR
Variables associated with Y					
			Correlated with each other		
X_{23}	0.5	1.6		0.3	1.3
X_{24}	0.8	2.2		0.5	1.6
X_{25}	1.1	3.0		0.7	2.0
X_{26}	1.4	4.1		0.9	2.5
X_{27}	1.7	5.5		1.1	3.0
			Independent of each other		
X_{28}	0.5	1.6		0.3	1.3
X_{29}	0.8	2.2		0.5	1.6
X_{30}	1.1	3.0		0.7	2.0
X_{31}	1.4	4.1		0.9	2.5
X_{32}	1.7	5.5		1.1	3.0

associations, and hence are often conducted when the association is expected to be strong. Correspondingly, we chose different coefficients β_i in the two simulated cases, with odds ratios in the interval [1.6–5.5] for Simulation 1 and in the interval [1.3–3.0] for Simulation 2.

The outcome variables were simulated using the model in equation (1), where β_0 was adjusted so as to yield a number of controls almost equal to the number of cases ($\beta_0 = -2$ in Simulation 1, and $\beta_0 = -1.5$ in Simulation 2). This can be considered as a short cut to the sampling of a case-control study within a cohort. For each simulation set-up, 200 data-sets were generated. Table IV summarizes the design choices.

4.3. Results: posterior probabilities and p-values

We analysed our simulated data sets using the two 'classical' methods and Bayesian model averaging. For the classical analyses, we recorded the p-values in the 'final single model' when X_i was listed among the selected variables. When X_i did not appear in the final model, we proceeded differently for the two methods. If X_i had been excluded in the first step of the two-step procedure, we kept the p-value from this previous univariate analysis. If X_i had been excluded during the stepwise process, we kept its p-value in the last model in which X_i figured among the variables. In the Bayesian approach, we simply recorded the posterior probability that X_i is associated with Y, $\Pr(\beta_i \neq 0|D)$, from equation (5).

We compare the observed and nominal performances of the methods. We thus consider the p-value associated with a coefficient for the classical methods, and the $\Pr(\beta_i \neq 0|D)$ for Bayesian model averaging. For each simulation design, the 200 repetitions give $32 \times 200 = 6400$

Table V. Two-step method: nominal significance levels and proportions of variables actually associated with the outcome.

Significance of an association with Y	Recorded p-values	Simulation 1 ($n = 300$)		Simulation 2 ($n = 1000$)	
		Observed proportion of variables with $\beta \neq 0$ by design	Number of p-values	Observed proportion of variables with $\beta \neq 0$ by design	Number of p-values
Barely significant	0.05–0.10	0.44	281	0.30	253
Significant	0.01–0.05	0.68	414	0.57	377
Highly significant	0.001–0.01	0.92	421	0.87	298
Very highly significant	<0.001	0.996	747	0.997	1161

Table VI. Stepwise method: nominal significance levels and proportions of variables actually associated with the outcome.

Significance of an association with Y	Recorded p-values	Simulation 1 ($n = 300$)		Simulation 2 ($n = 1000$)	
		Observed proportion of variables with $\beta \neq 0$ by design	Number of p-values	Observed proportion of variables with $\beta \neq 0$ by design	Number of p-values
Barely significant	0.05–0.10	0.34	350	0.28	302
Significant	0.01–0.05	0.55	448	0.49	414
Highly significant	0.001–0.01	0.89	432	0.86	283
Very highly significant	<0.001	0.99	858	0.995	1204

p-values to consider for each of the two classical methods, and 6400 values of $\Pr(\beta \neq 0|D)$ for the Bayesian approach.

4.3.1. Standard methods. For the two-step and stepwise methods, we selected the p-values that were less than 0.10, and classified them between the bounds 0.10, 0.05, 0.01, 0.001 and 0, yielding four intervals, often referred to as 'Barely significant', 'Significant', 'Highly significant' and 'Very highly significant', or similar terms. This type of classification corresponds to usual epidemiological practice.

We recorded the number of p-values falling in each of these intervals (columns 4 and 6 of Tables V and VI). This number corresponds to variables that were *declared* by the method to be associated with Y at the given significance level. Then we calculated the proportion of these variables that are *actually* associated by design. This ratio figures in columns 3 and 5 of Tables V and VI and gives the observed proportion of explanatory variables *declared associated* with this level of confidence that were *actually associated* with Y in our simulation.

As an example, consider Simulation 2 with $n = 1000$, the two-step analysis, and the [0.01–0.05] interval. If the p-value for a coefficient was less than 0.05 but greater than 0.01 (that is, 'significant'), how likely was it that the corresponding variable was actually associated with the outcome in the situation we simulated? We found 377 p-values in the [0.01–0.05] interval. In an article reporting a case-control study, the corresponding X_i would typically have

Table VII. Bayesian model averaging: posterior probabilities and proportions of variables actually associated with the outcome.

Evidence for an association with Y	$\Pr(\beta \neq 0 \mid D)$	Simulation 1 ($n = 300$)		Simulation 2 ($n = 1000$)	
		Observed proportion of variables with $\beta \neq 0$ by design	Number of posterior probabilities	Observed proportion of variables with $\beta \neq 0$ by design	Number of posterior probabilities
Weak	50–75%	0.65	156	0.57	161
Positive	75–95%	0.77	206	0.72	169
Strong	95–99%	0.92	109	0.90	93
Very strong	>99%	0.98	1037	0.99	1310

figured in a table summarizing the single final model marked with 'p-value < 0.05' (often denoted by one star). In reality, only 214 of these 377 p-values corresponded to variables that were actually associated with Y by design. Thus the observed proportion of 'significant' variables with p-values in the range [0.01–0.05] that were actually associated with Y was only 57 per cent ($= 214/377$), the value shown in Table V. One might well have expected a much higher value.

The two classical methods gave similar results: the observed proportion was well below 1, and also far below one minus the nominal significance level, for p-values in the range 0.10 right down to 0.001. Perhaps the most striking result is that for studies with 1000 subjects, of variables declared to be 'significant' with p-values between 0.01 and 0.05, only 49 per cent actually were risk factors when stepwise variable selection was used. In other words, far from being a likely risk factor, a variable whose coefficient would merit one star in an epidemiological article was about as likely to actually be one as a flipped coin to turn up heads. Only when the p-value was very small, below 0.001, was the chance of a false association small and close to its nominal value. The two-step method had somewhat better performance than the stepwise one.

4.3.2. Bayesian model averaging. For the Bayesian approach, we repeated our comparison by recording, among the 6400 posterior probabilities of a variable being a risk factor, $\Pr(\beta_i \neq 0 \mid D)$, the ones which led to the declaration of a link between X_i and Y, that is, those for which $\Pr(\beta_i \neq 0 \mid D) > 50$ per cent. To categorize these posterior probabilities, we used the bounds cited in reference [12]: 50, 75, 95 and 99 per cent, corresponding to weak, positive, strong and very strong evidence for an association with Y. The results are presented in Table VII, which can be loosely compared to Tables V and VI if one considers how respective results from classical and Bayesian analyses are declared and interpreted.

For example, among the 6400 values of $\Pr(\beta \neq 0 \mid D)$ obtained in Simulation 1, 206 were in the [0.75–0.95] interval, 159 of them corresponding to the last ten variables, which were actually associated with Y in the simulation. This gives 77 per cent ($= 159/206$) as the proportion of times that X_i was actually a risk factor for Y, among the times when its posterior probability was in the interval [0.75–0.95].

Overall in Table VII, we see reasonably good agreement between nominal and observed probabilities of an association. The interpretation of the uncertainty concerning the effect of

a potential explanatory variable, quantified by its posterior probability $\Pr(\beta \neq 0|D)$ is thus transparent and direct in the Bayesian results, in contrast to that of the p-value given by the classical analyses.

5. DISCUSSION

In observational cohort or case-control studies with many potential confounders, it is common to carry out statistical confounder selection using a two-step or stepwise procedure, and then to make inference using the selected model as if standard statistical methods were valid after variable selection. Our simulation study has shown that among variables with a given range of p-values, for example the range $[0.01, 0.05]$ commonly declared to be 'significant', the proportion of variables that are actually risk factors tends to be much lower than one minus the p-value (only 49 per cent actually were risk factors in our simulated case-control studies with 1000 subjects when stepwise variable selection was used). Thus, commonly made statements such as, 'The probability of obtaining such a result by chance is less than one in twenty', are somewhat misleading in this context.

We have proposed Bayesian model averaging as a formal way of accounting for model uncertainty in observational studies analysed using logistic regression. In our simulations, it was well calibrated, unlike the classical p-value methods considered. BMA can be easily implemented, and S-Plus functions to do it automatically are available on the Web. The posterior probabilities, $\Pr(\beta_i \neq 0|D)$, have a clear interpretation, which our simulation suggests is a valid one.

It should be emphasized that BMA is not a substitute for careful incorporation of available scientific knowledge, or for careful data analysis. These together should lead to a set of candidate confounders, or potential risk factors in a more exploratory study. The role of BMA is merely to account for the uncertainty remaining at the end of the scientific and data analysis; model uncertainty should be minimized on the basis of scientific considerations to the extent possible. However, the model uncertainty that remains should be taken into account when final conclusions are drawn. Note that BMA could be used to assess and take account of the uncertainty about different codings of a potential risk factor or confounder, and possibly to choose one of them. Raftery and Richardson [25] did this with the variable 'Alcohol consumption' in their Bayesian analysis of a case-control study of breast cancer.

In any Bayesian analysis, the prior distribution is important. For BMA, there are two major components to the prior distribution. The first consists of the prior model probabilities, and here we have taken all 2^q models to be equally likely *a priori*. This corresponds to no model being favoured *a priori* over others, but if prior expert knowledge is available that some variables or combinations of variables are more likely to be important than others, this can and should be incorporated via the prior model probabilities. However, our experience has been that the results tend to be relatively insensitive to deviations from this specification. Of course, if a variable C is known from other studies to be undoubtedly related to the disease, then C should be included in the model; this can be thought of as assigning prior probability zero to all models that do not include C.

The results also depend on the prior distribution of β_i, and more precisely on the prior scale parameter ϕ of the centred Normal chosen as a prior distribution in the Bayesian analysis. The choice of ϕ determines a prior interval for the quantity of interest (here the OR, for

which we have suggested the prior 95 per cent interval $[1/7, 7]$). Note that the context can be useful for tuning ϕ more precisely to the kind of study at hand. For example, if we regularly observed ORs as high as 14 (instead of 7) for variables in case-control studies similar to the one being analysed, we might increase the value of ϕ from 1.0 to 1.7, yielding a 95 per cent prior interval for the OR of $[1/14, 14]$. In the cervical cancer application, we tested the influence of the choice of prior variances by changing these to the ones described above (that is, OR in the interval $[1/14, 14]$). The posterior probabilities and estimates of the coefficients were essentially unchanged from those of Table II (results not shown).

As a final comment on the influence of the prior variance for β_i, we note that the shrinkage effect on β_i induced by the prior distribution of this coefficient is more marked when the maximum likelihood estimate of β_i is less precise. For example, in our application, this happens for dichotomous variables with low frequency of exposure, such as 'Genital warts', 'Gonorrhoea or syphilis' or 'Other cervicitis'.

One objection that might be raised against BMA is the following. The view might be taken that the interpretation of an OR depends on the confounders for which it has been adjusted, and hence that BMA, by combining results from models with different sets of confounders, is really mixing apples and oranges. We believe that this objection does not apply when the quantity of interest Q in equation (2) has the same interpretation for each model considered. This will be the case if, for example, Q can be interpreted as an observable quantity to be predicted. An adjusted log-odds ratio such as β_1 can be cast in this framework.

Another way of looking at this issue is as follows. BMA can be thought of as averaging over a collection of models. However, it can also be thought of as carrying out inference based on just one model, the full model with all variables $X_1, \ldots, X_q$, but with a rather special prior. This prior assigns non-zero probability to the events $\{\beta_i = 0\}$ for each i, that is, it allows for the possibility that the coefficients might be zero. Thus, since BMA can be thought of as a way of making Bayesian inference about a single model, the resulting inference about one of its coefficients can be interpreted as inference controlling for *all* the other variables.

One result from our simulation study is the difficulty of interpreting the p-value in classical stepwise and two-step procedures. A smaller than expected proportion of the variables declared to be associated with the disease outcome actually are. On the other hand, Bayesian model averaging provides a transparent statement of the probability that a variable is associated with a health outcome, through the posterior probability $\Pr(\beta_i \neq 0 | D)$. Such an approach could be helpful with the difficult task of choosing confounders, and was shown to have good performance in a realistic simulation study.

APPENDIX: REDUCING THE SET OF MODELS $\{M_k, k = 1, \ldots, K\}$

With q explanatory variables, and no interactions, the initial number of possible models will be equal to 2^q. This set will be very large in most epidemiological studies where q is frequently 20 or more, and it needs to be reduced. We will use the Occam's window approximation [10], which consists of:

(i) calculating the posterior probabilities of all models, using a workable fast approximation;
(ii) identifying the 'best' model M_b; (that is, the one with the highest posterior probability),

(iii) eliminating models that are unlikely *a posteriori*, that is, that are more than δ times less probable than the best one.

Specifically, we retain the models M_k that satisfy

$$\frac{P(M_b|D)}{P(M_k|D)} < \delta \tag{A1}$$

In practice, it turns out that we can avoid calculating the posterior probabilities of all the models, instead using the leaps and bounds algorithm to identify the most likely models *a posteriori*. We thus run the algorithm in two stages:

(i) the first time to eliminate models whose posterior probabilities are much smaller than that of the best model, using the BIC approximation to the Bayes factor and a threshold window with $\delta = 100$. To do this, an adapted leaps and bounds algorithm is implemented within the `bic.logit` or `bic.glm` S-Plus functions [11].

(ii) Then, on the models selected, we use the GLIB approximation to the Bayes factor to re-evaluate more precisely the posterior probabilities of the models kept, and to select a thinner window with $\delta = 20$. This is implemented in the `glib` S-Plus function. Note that the treatment of matched data required some modification of the S-Plus functions used for the Bayesian analyses. The calculation of the posterior probabilities is usually based on a comparison of each model with the empty one, in which only the intercept figures. As matched analyses are performed on models with no intercept term, these comparisons need to be made with *the full model.*

This defines our Bayesian model selection procedure. The models in the last set are the ones on which our BMA method is based and with which inference about the regression coefficients is carried out.

ACKNOWLEDGEMENTS

The authors are grateful to R. Peters for providing us with the data from the cervical cancer case-control study re-analysed here. They would like to thank D. Hemon, Sir David Cox, R. Peters and D. Thomas for helpful comments on simulation work and epidemiological application. The views expressed here are, however, solely those of the authors. The research of Viallefont and Raftery was supported by the Office of Naval Research, grant N00014-96-1-0192.

REFERENCES

1. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol. 1*: *The Analysis of Case-control Studies.* IARC scientific publication no. 32: Lyon, 1980.
2. Breslow NE. Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association* 1996; **91**:14–28.
3. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology* 1989; **129**(1):125–137.
4. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* Wiley: New York, 1989.
5. Savitz DA, Tolo K-A, Poole C. Statistical significance testing in the American Journal of Epidemiology, 1970–1990. *American Journal of Epidemiology* 1994; **139**(10):1047–1052.
6. Freedman DA. A note on screening regression equations. *American Statistician* 1983; **37**:152–155.
7. Miller AJ. *Subset Selection in Regression.* Chapman and Hall: London, UK, 1990.
8. Peters RK, Thomas D, Hagan DG, Mack TM, Henderson BE. Risk factors for invasive cervical cancer among latinas and non-latinas in Los Angeles county. *Journal of the National Cancer Institute* 1986; **77**(5): 1063–1077.

9. Leamer EE. *Specification Searches*: *Ad Hoc Inference With Nonexperimental Data.* Wiley: New York, 1978.
10. Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 1994; **89**:1535–1546.
11. Raftery AE. Bayesian model selection in social research (with Discussion). In *Sociological Methodology 1995*, Marsden PV (ed.). Blackwell Publishers: Cambridge, Mass., 1995; 111–163.
12. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**:773–795.
13. Draper D. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:45–98.
14. Raftery AE. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* 1996; **83**:251–266.
15. Chatfield CC. Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society*, *Series A* 1995; **158**:419–466.
16. Raftery AE, Madigan D, Hoeting JA. Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association* 1997; **92**:179–191.
17. Hoeting JA, Raftery AE, Madigan D. A method for simultaneous variable selection and outlier identification in linear regression. *Journal of Computational Statistics* 1995; **22**:251–271.
18. Lee PM. *Bayesian Statistics*: *An Introduction.* Oxford University Press: Oxford, U.K., 1989.
19. Bernardo JM, Smith AFM. *Bayesian Theory.* Wiley: New York, 1994.
20. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Chapman and Hall: London, 1995.
21. Raftery AE, Madigan D, Volinsky CT. Accounting for model uncertainty in survival analysis improves predictive performance (with Discussion). In *Bayesian Statistics 5*, Bernardo JM, Berger JO, Dawid AP, Smith FM. (eds). Oxford University Press: Oxford, U.K., 1995; 323–349.
22. Rubin DB, Schenker N. Efficiently simulating the coverage properties of interval estimates. *Applied Statistics* 1986; **35**:159–167.
23. Racine A, Grieve AP, Fluhler H, Smith AFM. Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statistics* 1986; **35**:93–150.
24. Richardson S, Gerber M, Cénée S. The role of fat, animal protein and some vitamin consumption in breast cancer: a case-control study in southern France. *International Journal of Cancer* 1991; **48**:1–9.
25. Raftery AE, Richardson S. Model selection for generalized linear models via GLIB: application to nutrition and breast cancer. In *Bayesian Biostatistics.* Oxford University Press: Oxford, UK, 1996; Chapter 12, 321–353.
26. Volinsky CT, Madigan D, Raftery AE, Kronmal RA. Bayesian model averaging in proportional hazard models: predicting the risk of a stroke. *Applied Statistics* 1997; **46**:443–448.
27. Berger JO, Delampady M. Testing precise hypotheses (with discussion). *Statistical Science* 1987; **3**:317–352.
28. Jeffreys H. *Theory of Probability.* 3rd edn. Oxford University Press: Oxford, U.K., 1961.
29. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research*: *Principles and Quantitative Methods.* Lifetime Learning Publications: Belmont, CA, 1982.
30. Armitage P. *Statistical Methods in Medical Research.* Blackwell Scientific Publications: Oxford, UK, 1971.
31. Rothman KJ, Greenland S. *Modern Epidemiology.* 2nd edn. Lippincott, Williams & Wilkins: Philadelphia, 1998.

