Variable Selection and Bayesian Model Averaging in Case-Control Studies

Valerie Viallefont	Adrian E. Raftery	Sylvia Richardson
INSERM, France	University of Washington	INSERM, France

Technical Report no. 343

Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195-4322, USA.

November 1998

Abstract

Covariate and confounder selection in case-control studies is most commonly carried out using either a two-step method or a stepwise variable selection 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 underestimates 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 1,000 subjects, of variables declared to be "significant" with p-values between .01 and .05, only 49% 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 previously published case-control study of cervical cancer.

Contents

List of Tables

1 Introduction

Case-control studies ([1], [2]) represent a high proportion of epidemiological practice. For example, at least 49 such studies were published in the *American Journal of Epidemiology* alone in 1996 (see below).

The aim of case-control studies is to test the existence of possible risk factors of interest, and to estimate their association with the presence or absence of a disease, after adjusting for possible confounders. A sample of n_1 cases and n_2 controls is taken, where often n_2 is roughly an integer multiple of n_1 ($n_1 \approx n_2$ is common). Although the sample is drawn based on the disease outcome, the sampling plan is much more efficient than random sampling. Remarkably, consistent and near-optimal estimates of adjusted relative risks can be obtained if the model used is logistic regression, namely

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

where Y is 1 if the disease is present and 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. An attraction of this model is that the adjusted relative risk, $\exp(\beta_1)$, is the same for all values of the confounders, and does not involve any coefficients other than β_1. This greatly facilitates reporting and interpretation of the results. This adjusted relative risk can be approximated by the odds ratio, namely

$$\frac{\text{Odds}(Y=1|X_1=1, X_2, \ldots, X_q)}{\text{Odds}(Y=1|X_1=0, X_2, \ldots, X_q)}, \tag{2}$$

where Odds = Probability/(1−Probability).

The choice of the confounders, $X_2, \ldots, X_q$, to include is a major issue. In most studies, the potential confounders are numerous, including demographic, socioeconomic, familial disease history, smoking and other lifestyle variables, as well as medical measurements; often something like 20 to 50 such variables are initially considered. It is vital not to omit important confounders, which would point towards including all confounders considered, but doing so tends to lead to inefficient estimation, both in theory (e.g. [3], chap. 9) and in practice [4]. Thus, investigators have tended to use statistical methods to choose among the many confounders indicated by substantive considerations. The two most commonly used methods are a two-stage method [4] and backwards stepwise regression; both of these are described, for example, in the influential textbook [5].

Investigators typically carry out tests and compute confidence intervals [6] conditionally on the selected logistic regression model, 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 (e.g. [7], [8]).

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 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. In addition, Bayesian model averaging point estimates of adjusted relative risks are more accurate on average than those from the variable selection methods. We illustrate the method with an application to a case-control study investigating risk factors for cervical cancer [9].

2 MATERIALS AND METHODS

2.1 Bayesian Model Averaging

2.1.1 General Principles

Bayesian model averaging (BMA) is the Bayesian solution to the problem of inference in the presence of multiple competing models ([10], [11], [12], [13], [14], [15], [16], [17]). For general introductions to Bayesian inference, see [18], [19] and [20].

BMA starts by acknowledging that in the situation of equation (1) there are up to $K = 2^q$ possible models (assuming that no interaction exists 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, i.e. 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, which here is an adjusted relative risk, but could also be a future observation, or the utility of a course of action. Its posterior distribution, taking account of model uncertainty, is

$$p(Q|D) = \sum_{k=1}^{K} p(Q|D, M_k)\ p(M_k|D), \tag{3}$$

where D denotes the data at hand (here, observations on disease, 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

$$p(M_k|D) \propto p(D|M_k)p(M_k), \tag{4}$$

where the constant of proportionality is chosen so that the posterior model probabilities add up to one. In equation (4), $p(M_k)$ is the prior model probability of model M_k; these are often chosen to be equal so as not to favor one model over another *a priori*. The quantity $p(D|M_k)$ is the *integrated likelihood* of model M_k, which follows again from the law of total probability as the multiple integral

$$p(D|M_k) = \int p(D|\beta^k, M_k)p(\beta^k|M_k)d\beta^k, \tag{5}$$

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 (5) 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 using the Laplace method [15]. 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 www.research.att.com/˜volinsky/software/glib, or at lib.statlib.cmu.edu/S/glib.

The second problem is the choice of the prior distribution of the parameters, β^k. 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, seems capable of representing a reasonable range of prior distributions while requiring the specification of only a single prior parameter, ϕ [15]. We will discuss the choice of ϕ below.

BMA has three appealing theoretical long-run properties. First, for the prediction of new observations, it gives better predictive performance on average (using a log score to measure performance), than any single model that could reasonably have been selected [11]. Second, inferences are well calibrated in the sense that, for example, confidence intervals have the right coverage on average [21]. The third property relates to Bayesian hypothesis testing using *Bayes factors*, when the null model is rejected against an alternative if it has lower posterior model probability. This test has lower total error rate (i.e. sum of Type I and Type II error rates) than any other test on average over the models and the prior distributions ([22], p. 396), and hence can be viewed as an automatic way of choosing a significance level so as to optimally balance power and significance.

This general approach has been used in several previous analyses of medical and epidemiological data. Racine et al [23] showed how this method may be used to make inference about a treatment effect in the presence of uncertainty about the existence of a carryover effect. An epidemiological study of fat and alcohol consumption as risk factors for breast cancer [24] was reanalyzed using Bayesian model averaging [25].

Similar analyses of coronary heart dise ase risk factors and the diagnosis of scrotal swellings have been reported [11]. In their examples, the authors found that out-of-sample predictive performance was better if one took account of model uncertainty than if one conditioned on any single model that might reasonably have been selected.

2.1.2 BMA Inference for an Adjusted Relative Risk

BMA provides hypothesis tests, point estimates and confidence intervals for an adjusted relative risk, 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, i.e. that β_1, the adjusted log relative risk, is not equal to zero?” This is simply

$$\Pr[\beta_1 \neq 0|D] = \sum_{M_k : X_1 \in M_k} p(M_k|D). \tag{6}$$

Conventional rules of thumb for interpreting this quantity verbally are that if $\Pr[\beta_1 \neq 0|D]$ is less than 50%, there is no evidence for X_1 being a risk factor, if it is between 50% and 75% there is weak evidence for X_1 being a risk factor, if it is between 75% and 95% there is positive evidence, between 95% and 99% the evidence is strong, and beyond 99% the evidence is very strong ([22], Appendix B; [13]).

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

$$E[\beta_1|D] = \sum_{M_k : X_1 \in M_k} \hat{\beta}_1^k p(M_k|D), \tag{7}$$

where $\hat{\beta}_1^k$ is the posterior mean of β_1 under model 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, or posterior standard deviation of β_1, is the square root of

$$\mathrm{Var}(\beta_1|D) = \sum_{M_k : X_1 \in M_k} \left\{ \left(\mathrm{Var}(\beta_1|D, M_k) + \left(\beta_1^k\right)^2 \right) \, p(M_k|D) - E[\beta_1|D]^2 \right\}. \tag{8}$$

If the sample size is moderate or large, $\hat{\beta}_1^k$ will often be well approximated by the maximum likelihood estimator, and $\mathrm{Var}(\beta_1|D, M_k)$ will be well approximated by the square of the usual standard error under model M_k.

Inference about β_1, including testing, point estimation and confidence intervals, is summarized by equations (6), (7) and (8). A difficulty is that each of these equations involves summation over all $K = 2^q$ possible models, and K will often be impracticably large. For example, in the application we present later, $q = 35$, and so K could be as large as 2^{35}, which is about 34 billion. To get around this, we approximate the full sum by excluding models that are far less probable *a posteriori* than the best model; here we adopt the convention of excluding models whose posterior probability is less than one-twentieth of that of the best model. This device, known as *Occam's window* [11], reduces the number of models enormously but still seems to provide a good approximation to the full sum [26]. There are also scientific arguments supporting its use in its own right, and not just as an approximation

[11]. The models in Occam's window can be found rapidly using a generalization of the leaps and bounds algorithm ([27], [12], [28]). Details are given in Appendix.

The BMA approach requires the specification of some prior quantities:

- the prior probabilities $P(M_k)$ of the models, which will be taken to be equal, so as not to favor any model *a priori*, and

- the prior distribution of the parameters β_i. As suggested in [15] for the case of weak prior information, we assume that the distribution of β is a centered Gaussian with a standard deviation ϕ to be specified.

To choose a value of ϕ, we noted 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 relating to the distribution of typical OR's. We chose ϕ so that the prior probability of finding an OR greater than 7 will be less than 5 %. This was motivated by the shared expectation, based on our own experience and on discussions with epidemiologists at INSERM, that most OR investigated in case-control studies will be less than 7. This is in part because stronger associations will often already have been identified, and with strong associations interest is more likely to focus on the strength of the association rather than its existence, with dose-response or other specifically designed studies. Our choice was reinforced by consideration of the range of OR's found in our review of case-control studies published in 1996 (see the next sections and Table 1). Note that this prior choice can be easily modified by simply changing the value of ϕ.

2.2 Standard Methods

The difficulties of variable selection techniques have been discussed by many authors ([29], [30], [31], [1], [5]). 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.

2.2.1 The Two-Step Procedure

This procedure, described in [5], has been considered and examined in [4] as a possible strategy to identify 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 $\alpha = .20$. The associations interpreted are essentially those corresponding to variables having p-values less than $\alpha = .05$ in the second stage.

2.2.2 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 .05, or some other threshold. The coefficients of the remaining variables are then re-estimated in this model.

2.3 Design of the Simulation Study

2.3.1 Basis for the Simulation Study Design: The Case-Control Studies in AJE, 1996

Determining a few cases of "typical" datasets to be simulated and analyzed 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 49 case-control studies, which are used essentially as providing useful guidelines for the design of our simulation study.

Most of the studies were conducted to establish the link between a specific set of variables and a health outcome, or to evaluate the strength of the association for a variable already established as a risk factor. In this case, many other variables are entered in the multiple logistic regression as potential confounders. Other studies were more exploratory in nature and included a substantial number of variables, all considered to be potential risk factors.

In the setup of our simulation study, we have not focused on a particular variable, but have considered them all symmetrically.

The 49 selected studies had numbers of subjects ranging from 118 to 9,913, 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," 1,000. In most studies, the number of controls was close to the number of cases and we chose equal numbers of cases and controls in each of the two simulation designs.

2.3.2 The Variables

The total number of variables initially considered is rarely fully reported in the articles. Indeed, the focus on one particular variable, or simply the need to summarize the results, often leads to the presentation of a small number of tables, including only some of the variables mentioned in the "Data Collection" paragraph. In the 49 studies, the total number of variables examined in the tables was between 3 and 39, with a mean of 16. This is likely to be an underestimate, due partly to the fact that the tables presented in a published study are a summary of the analysis itself and therefore frequently do not include all the variables examined, especially in the case where a particular risk factor is of prime interest. A detailed examination of the 49 studies, with special attention to the ones where variables were considered symmetrically 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 reanalyze later.

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. This is aimed 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 : independent variables were Bernoulli, while the variables correlated with each other were simulated as centered multivariate normal with all correlations equal to 0.4, then dichotomized. The frequency of exposure to the risk factors has been set at 20%, which is, for example, the approximate frequency of exposure to solvents in the general population, and seems reasonable to mimic the frequency of exposure in many case-control studies.

In order to choose the values of the logistic coefficients β_i, we have listed the OR of the variables of interest in the 49 studies selected: we considered the value(s) quoted in the summary when there were any, and if not, we took the estimated OR's corresponding to the risk factor investigated from the most complete model. Some OR's came from a univariate logistic regression, but most of them came from a multivariate model. Because of an obvious reversibility, we decided not to consider the case of protective factors and simply inverted their coefficients. Table 1 shows the distribution of the estimated OR, separately for the smaller and larger studies, 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 associations. 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.

Table 1: Distribution of the values of OR's of primary interest found in 49 case-control studies, for smaller and larger studies separately

	Minimum	1st Quartile	Median	3rd Quartile	90% 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

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). For each simulation setup, 200 datasets were generated. Table 2 summarizes the design choices.

3 RESULTS

We analyzed 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

Table 2: Design of the simulations

	Simulation 1	Simulation 2
number of subjects, n	300	1000
total number of variables	32	32
TYPE OF VARIABLES		
variables associated with Y and independent of each other	5	5
variables associated with Y and with each other	5	5
variables independent of Y but correlated with the 5 last ones	2	2
variables independent of Y but correlated with each other	2 groups of 5	2 groups of 5
variables independent of Y and of each other	10	10
Simulated odds ratios	1.6–5.5	1.3–3.0

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 (6).

3.1 Posterior Probabilities and p-Values

We first 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 $32 \times 200 = 6400$ p-values to consider for each of the two classical methods, and 6400 values of $Pr(\beta \neq 0|D)$ for the Bayesian approach.

3.1.1 Standard Methods

For the two-step and stepwise methods, we selected the p-values that were less than .10, and classified them between the bounds 0.10, 0.05, 0.01, 0.001 and 0, yielding four intervals. This type of classification corresponds to usual epidemiological practice. It is conventional to view p-values between .05 and .10 as “barely significant”, “nearly significant” or “just missing significance”, those between .01 and .05 as “significant”, those between .001 and .01 as “highly significant”, and those that are less than .001 as “very highly significant”.

We recorded the number of p-values falling in each of these intervals (columns 4 and 6 of

Tables 3 and 4). This number corresponds to variables that are *declared* by the method to be associated with Y at the given significance level. Then we calculated the proportion of these variables which are *actually* associated by design. This ratio figures in columns 3 and 5 of Tables 3 and 4 and gives the observed proportion of explanatory variables *declared associated* with this level of confidence which were *actually associated* with Y in our simulation.

Table 3: 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	# of p-values	Observed proportion of variables with $\beta \neq 0$ by design	# 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

As an example, consider Simulation 2, the two-step analysis, and the [0.01-0.05] interval. If the p-value for a coefficient was less than .05 but greater than .01 (i.e. "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 figured in a table summarizing the single final model marked with "p-value < 0.05". 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%. Thus commonly used interpretations such as "the probability of such a significant result occurring by chance is less than 5%" can be misleading when two-step (or stepwise) variable selection is carried out first; here the empirical probability of such a significant result occurring by chance was 43%. Hence the observed proportion of being truly a risk factor for X_i when the quoted p-value belongs to [0.01-0.05] is only 57%. One might well have expected a much higher value.

For the two classical methods, the observed proportion was well below 1, and also far below one minus the nominal significance level, for p-values in the range .10 right down to .001. Only when the p-value was very small, below .001, was the chance of a false association small and close to its nominal value. The two classical methods gave similar results, although the two-step method had somewhat better performance than the stepwise one. Hence, for

Table 4: 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	# of p-values	Observed proportion of variables with $\beta \neq 0$ by design	# 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

most of the range of p-values usually viewed as "significant", the p-value resulting from the classical variable selection methods is very different from the actual proportion of false positive significant results. Thus we should be wary of interpreting p-values in terms of the probability that a coefficient corresponds to an actual association.

3.1.2 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, $\Pr(\beta_i \neq 0|D)$, the ones which led to the declaration of a link between X_i and Y, i.e. the $\Pr(\beta \neq 0|D) > 50\%$. To categorize these posterior probabilities, we used the bounds cited in [13] : 50, 75, 95 and 99%, corresponding to weak, positive, strong and very strong evidence for an association with Y. The results are presented in Table 5.

Table 5: Bayesian Model Averaging: Posterior probabilities and proportions of variables actually associated with the outcome

Evidence for an association with Y	$Pr(\beta \neq 0|D)$	Simulation 1 (n=300)		Simulation 2 (n=1000)	
		Observed proportion of variables with $\beta \neq 0$ by design	Number of post.prob.	Observed proportion of variables with $\beta \neq 0$ by design	Number of post.prob.
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

For example, among the 6400 values of $\Pr(\beta \neq 0|D)$ 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 0.77 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 5, 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.

3.2 Estimation of the Coefficients

Here we compare the estimation of the logistic regression coefficients, β_i, given by the different methods by summing the squared errors over the 200 simulations:

$$SSE = \sum_{j=1}^{200} \sum_{i} (\hat{\beta}_i - \beta_{Ti})^2$$

(where β_{Ti} is the true value of β_i in the simulation), calculated separately for the subgroups of variables associated and not associated with Y. The BMA estimate of $\hat{\beta}_i$ is given by equation (7). It was not meaningful to evaluate this sum for the stepwise selection method, as $\hat{\beta}_i$ is undefined if X_i is not retained in the model. Table 6 contains results for the two-step method and for BMA, for the two different simulation setups.

Table 6: Estimation of Logistic Regression Coefficients: Sums of Squared Errors

	Simulation 1 (n=300)		Simulation 2 (n=1000)	
Variables	Two-step	BMA	Two-step	BMA
X_1-X_{20}	438	288	108	121
X_{21}-X_{22}	65	41	15	14
X_{23}-X_{32}	387	307	82	71

For Simulation 1, BMA gives smaller sums of errors for all the types of variables. Hence in smaller studies, where the uncertainty due to model choice is greater, Bayesian estimation via BMA gives an estimate that is closer to the true value on average. In Simulation 2, the mean squared error is smaller, as expected when n increases, and the results of the two methods become closer. The only case where the two-step approach gives a smaller SSE than the Bayesian one is for the variables $X_1 - X_{20}$, simulated independently from Y, when $n = 1,000$. Note that, in this case, the values of $\Pr(\beta_i|D)$ for such variables are very low, and so these are estimates that one would not usually interpret anyhow.

Table 7: Cervical cancer study: Adjusted effects published in [9]

	$\hat{\beta}$	SE	p-value
Sexual partners before age 20	0.62	0.21	$< .01$
Years using barrier contraceptives	-0.142	0.043	$< .01$
Episodes of genital warts	1.32	0.45	$< .01$
Years of education	-0.252	0.073	$< .001$
Years since last Pap smear (log)	0.86	0.18	$< .001$
Cumulative smoking exposure	0.00229	0.00063	$< .001$
Cumulative douche use	0.0081	0.0022	$< .001$
Yrs. from menarche to first intercourse	-0.078	0.049	$< .10$

4 APPLICATION

We now present an application of Bayesian model averaging to a case-control study on cervical cancer conducted by R. Peters and collaborators [9]. It included 400 subjects: 200 women with invasive squamous cell carcinoma of the uterine cervix and 200 controls matched on age, neighborhood of residence and preferred language (Spanish or English). For comparison, we will refer to the results published in [9].

4.1 Classical analyses

In [9], this dataset was analyzed as matched pairs, considering 35 risk factors classified in categories (sexual history, method of contraception, genital infection, risk inducing behavior and demographic characteristics). The authors did not include any interaction terms. First, they performed a univariate analysis for each of the 35 variables, of which 25 had two-sided p-values smaller than .10. In the multivariate analysis, they then 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 two-sided p-values smaller than .10. Starting with these 18 variables, they then carried out a multiple logistic regression based on a stepwise forward analysis. This led to a final model with 8 variables. Seven of these 8 variables were statistically significant at the 5% level, and the eighth, “Years from menarche to first intercourse”, was significant only at the 10% 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 7. We will do the same in the following classical analysis.

To complete this classical approach, we also ran a stepwise forward analysis starting from all the available variables. Unfortunately, two of them ("estimated number of others 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 [9]. This left a total of 26 available variables. We thus carried out a stepwise procedure starting with all the 26 variables ; this led to 10 of them being included in the final model, including "Years from menarche to first intercourse". Table 8 presents the logistic regression coefficients and standard errors estimated in the "final" model.

4.2 Bayesian Model Averaging

For comparability, we used the same variable definitions as in [9], we did not consider interactions, and we analyzed 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. This is then pared down to the models in Occam's window, namely the most likely model *a posteriori* and the models whose posterior probabilities are within a factor of 20 of that of the most likely one. To do this, highly unlikely models are first excluded using the adapted leaps and bounds algorithm as implemented in the `bic.logit` or `bic.glm` S-PLUS function (available at www.research.att.com/˜volinsky/bma.html or at lib.stat.cmu.edu/S), and then the posterior probabilities of the remaining models are computed using the Laplace approximation, as implemented in the `glib` S-PLUS function[1] (see Appendix for details).

The prior distribution of β_i is a centered 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%. For a dichotomous variable, this OR is e^{β_i}, and the interval [1/7,7] corresponds to choosing $\beta_i \sim$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%. 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

[1] 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 analysis are performed on models with no intercept term, comparisons were made with the full model, *i.e.* the model including all the 26 variables.

Table 8: Cervical cancer study analyses

	Classical Variable Selection			Bayesian Model Averaging		
	$\hat{\beta}$	SE	p-value	$E(\beta\|D)$	$sd(\beta\|D)$	$Pr(\beta \neq 0\|D)$
Sexual partners before age 20	0.70	0.25	$< .01$	0.61	0.23	100%
Years using barrier contraceptives	-0.15	0.046	$< .001$	-0.12	0.044	100%
Episodes of genital warts	1.48	0.48	$< .01$	0.93	0.44	97%
Years of education	-0.26	0.079	$< .001$	-0.24	0.076	100%
Years since last Pap smear (log)	0.99	0.20	$< .001$	0.71	0.18	100 %
Cumulative smoking exposure	0.0029	0.00073	$< .001$	0.0024	0.0007	100%
Cumulative douche use	0.0089	0.0023	$< .001$	0.0078	0.0023	100%
Yrs. from menarche to first intercourse	-0.078	0.051	0.13	-0.074	0.05	39%
Genital herpes (yes/no)			NS	0.31	0.69	13%
Age at first intercourse			NS	-0.047	0.065	12%
Income			NS	-0.085	0.12	9%
Partner with genital warts			NS	0.11	0.51	7%
Age at first regular intercourse			NS	-0.03	0.04	6%
Gonorrhea or syphilis (yes/no)			NS	-0.88	0.61	66%
Other cervicitis (yes/no)	1.23	0.54	$< .05$	0.73	0.46	82%
Intra-uterine devices	-0.14	0.074	$< .10$	-0.11	0.073	13%

estimates of the regression coefficients (see the right hand side of Table 8). The posterior probability of being a risk factor, given by $\Pr(\beta_i \neq 0|D)$, is also shown. Table 8 summarizes the estimation of the logistic coefficients as well as the associated probabilities: the p-values and the $\Pr(\beta_i \neq 0|D)$ for the "classical" and Bayesian analyses, respectively. The variables which do not figure in Table 8 had both a nonsignificant p-value and a posterior probability lower than 5%.

We now compare the results of the classical analysis with those of BMA. Six variables had p-values below .01, and these all had posterior probabilities of 100%[2], so that for these variables the two approaches were in agreement.

Two variables had p-values between .01 and .05. For one of these, the number of genital warts episodes, the posterior probability was 97%, so the two analyses agreed. For the other one, "other cervicitis", the posterior probability was 82%, so that the p-value seems too decisive once model uncertainty is taken into account. An interesting point about the "other

[2] Note that a reported posterior probability of 100% 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 be close to, but not exactly equal to, 100%.

cervicitis" variable is that it did not figure in the published results (Table 7), because the authors initially excluded variables whose p-values was greater than .10 in a univariate analysis, which was the case for "other cervicitis"[3]. 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%, in most cases well below. There was one notable exception, however, namely "gonorrhea or syphilis", whose posterior probability was 66%. Thus the classical analysis missed the (weak) evidence in the data for an effect of this variable. This high posterior probability seems to indicate that this variable is a good marker of past sexual risk factors. Note nevertheless that only 14% of patients were exposed. In this context of low exposure frequency, the choice of ϕ is more influential and might have contributed to the difference between classical and Bayesian analyses concerning the selection of this variable.

The only other non-significant variable for which the posterior probability approached 50% was "years from menarche to first intercourse" (39%). 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.

Note that the regression logistic parameter estimates from the Bayesian analysis are shrunk towards zero in comparison with the classical analyses. This is a usual phenomenon in Bayesian estimation and is related to the choice of prior variances for these coefficients which will be discussed later.

5 DISCUSSION

In observational 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

[3] The results published in [9] were based on stepwise variable selection starting from 18 rather than 26 variables.

selection. Our simulation study, designed to resemble typical case-control studies, has shown that among variables with a given range of p-values, for example the range [.01,.05] commonly declared to be "significant", the proportion that actually are risk factors tends to be much lower than one minus the p-value (only 49% actually were risk factors in our simulated case-control studies with 1,000 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 grossly misleading in this context.

If Bonferroni-corrected p-values had been used instead, the correspondence between nominal and observed levels would have been better, but the Bonferroni correction assumes uncorrelated variables and is rarely used in case-control studies. It was not used once in our sample of 49 case-control studies from the AJE.

We have proposed Bayesian model averaging as a formal way of accounting for model uncertainty in case-control studies analyzed 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. But the model uncertainty that remains should be taken into account when final conclusions are drawn.

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 there we have taken all 2^q models to be equally likely *a priori*. Our experience has been that the results tend to be relatively insensitive to deviations from this specification. There are important ways in which this prior distribution may need to deviate from equiprobability should be noted, following [1]. If a variable C is known from other studies to be undoubtedly related to the disease, and if this association is not subsidiary to a possible exposure/disease association, 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. Moreover, if a factor is thought to be sufficiently important to be used as a matching or balancing factor in the study design, it should be included in all the models considered.

There are various possible ways other than equal probability to assign prior model probabilities. One approach that seems promising is to elicit prior model probabilities from health professionals [32]. In the results reported in [32], this gave better predictive performance than BMA with equal prior model probabilities.

Regression estimates also depend on the prior distribution of β_i, and more precisely on the prior scale parameter ϕ of the centered 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% 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 OR's as high as 14 (instead of 7) for variables in case-control studies similar to the one being analyzed, we might double the value of ϕ, yielding a 95% 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 (*i.e.* OR in the interval [1/14, 14]). The posterior probabilities and estimates of the coefficients were essentially unchanged from these of Table 8 (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”, “Gonorrhea 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 (3) 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 often be cast in this framework, since it can be thought of as the approximate log-odds ratio in a stratum from a hypothetical future large sample exchangeable with the current one. The relevant stratum would be one defined by stratifying by *all* the potential confounders $X_2, \ldots, X_q$.

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, i.e. 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 validly interpreted in the usual way. One possible objection to such a prior is that we would never believe that a coefficient could be exactly zero, although we might well expect it to be small. This concern is alleviated by the fact that the results from this prior are very similar to those from a different prior in which the probability is spread over a moderately small interval about zero instead of being concentrated precisely at zero; this interval can be as wide as half a standard error [33].

One strong result to emerge from our simulation study is the difficulty of interpretation of 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.

Acknowledgments: The authors are grateful to R. Peters for providing us with the data from the cervical cancer case-control study reanalyzed 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 Office of Naval Research Grant N00014-96-1-0192.

Appendix

Reducing the set of models $\{M_k, k = 1, \ldots, K\}$ With q explanatory variables, and no interaction, 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 following principle (also referred to as Occam's window by Madigan and Raftery [11]), which consists of:

- calculating the posterior probabilities of all models, using a workable fast approximation,

- identifying the "best" model (i.e. the one with the highest posterior probability) : M_b,

- eliminating the models which are more than δ times less probable than the best one.

More precisely, we keep the models M_k, satisfying :

$$\frac{P(M_b|D)}{P(M_k|D)} < \delta \tag{9}$$

We actually run this algorithm twice:

- 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 $\delta = 100$;

- 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 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.

References

[1] N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research. Vol 1 : The analysis of case-control studies.* Lyon : IARC scientific publication no.32, 1980.

[2] N.E. Breslow. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91:14–28, 1996.

[3] Bishop Y.M.M., Fienberg S.E., and Holland P.W. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass.: MIT Press, 1975.

[4] R.M. Mickey and S. Greenland. The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, 129(1):125–137, 1989.

[5] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* John Wiley and Sons, 1989.

[6] D.A. Savitz, K-A. Tolo, and C. Poole. Statistical significance testing in the american journal of epidemiology, 1970-1990. *American Journal of Epidemiology*, 139(10):1047–1052, 1994.

[7] D.A. Freedman. A note on screening regression equations. *The American Statistician*, 37:152–155, 1983.

[8] A.J.Miller. *Subset Selection in Regrerssion.* Chapman and Hall, 1990.

[9] R.K. Peters, D. Thomas, D.G. Hagan, T.M. Mack, and B.E. Henderson. Risk factors for invasive cervical cancer among latinas and non-latinas in los angeles country. *Journal of the National Cancer Institute*, 77(5):1063–1077, 1986.

[10] E.E. Leamer. *Specification Searches: Ad Hoc Inference With Nonexperimental Data.* New York: Wiley, 1978.

[11] D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1535–1546, 1994.

[12] A.E. Raftery. *Bayesian model selection in social research (with Discussion). In* Sociological Methodology 1995 *(edited by P.V. Marsden)*, pages 111–163. Cambridge, Mass.: Blackwell Publishers, 1995.

[13] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[14] D. Draper. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, series B*, 57:45–98, 1995.

[15] A.E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83:251–266, 1996.

[16] C.C. Chatfield. Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, series A*, 158:419–466, 1995.

[17] A.E. Raftery, D. Madigan, , and J.A. Hoeting. Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997.

[18] P.M. Lee. *Bayesian Statistics: An Introduction.* Oxford, U.K.: Oxford University Press, 1989.

[19] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* New York: Wiley, 1994.

[20] A. Gelman, J.B. Carlin, H.S. Stern, , and D.B. Rubin. *Bayesian Data Analysis.* London: Chapman and Hall, 1995.

[21] D.B. Rubin and N. Schenker. Efficiently simulating the coverage properties of interval estimates. *Applied Statistics*, 35:159–167, 1986.

[22] H. Jeffreys. *Theory of Probability.* Oxford, U.K.: Oxford University Press, 3rd edition, 1961.

[23] A. Racine, A.P. Grieve, H. Fluhler, and A.F.M. Smith. Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statictics*, 35:93–150, 1986.

[24] S. Richardson, M. Gerber, and S. Cénée. The role of fat, animal protein and some vitamin consumption in breast cancer: a case-control study in southern france. *International Journal of Cancer*, 48:1–9, 1991.

[25] A.E. Raftery and S. Richardson. *Model Selection for Generalized Linear Models via GLIB : Application to Nutrition and Breast Cancer. in Bayesian Biostatistics*, chapter 12, pages 321–353. Oxford University Press, 1996.

[26] A.E. Raftery, D. Madigan, and C.T. Volinsky. *Accounting for model uncertainty in survival analysis improves predictive performance (with Discussion). In* Bayesian Statistics 5 *(J.M. Bernardo* et al.*, eds.)*, pages 323–349. Oxford, U.K.: Oxford University Press, 1995.

[27] G.M. Furnival and Jr. Wilson, R.W. Regression by leaps and bounds. *Technometrics*, 16:499–511, 1974.

[28] C.T. Volinsky, D. Madigan, A.E. Raftery, and R.A. Kronmal. Bayesian model averaging in proportional hazards models: Assessing the risk of a stroke. *Applied Statistics*, 46:433–448, 1997.

[29] D.G. Kleinbaum, L.L. Kupper, and H. Morgenstern. *Epidemiologic Research : Principles and Quantitative Methods*. Belmont, CA : Lifetime Learning Publications, 1982.

[30] P. Armitage. *Statistical Methods in Medical Research.* Blackwell Scientific Publications, 1971.

[31] K.J. Rothman and S.Greenland. *Modern Epidemiology, Second Edition.* Lippincott-Raven, 1998.

[32] D. Madigan, J. Gavrin, and A.E. Raftery. Enhancing the predictive performance of Bayesian graphical models. *Communications in Statistics - Theory and Methods*, 24:2271–2292, 1995.

[33] J.O. Berger and M. Delampady. Testing precise hypotheses (with discussion). *Statistical Science*, 3:317–352, 1987.