Published in final edited form as:
Stat Med. 2008 December 20; 27(29): 6158–6174. doi:10.1002/sim.3434.

Finding Factors Influencing Risk: Comparing Variable Selection Methods Applied to Logistic Regression Models of Cases and Controls

Michael D. Swartz[1,*], **Robert K. Yu**[1], and **Sanjay Shete**[1]

[1] Department of Epidemiology, Unit 1340, U. T. M. D. Anderson Cancer Center, PO Box 301430, Houston, TX, 77230-1439

Abstract

When modeling the risk of a disease, the very act of selecting the factors to include can heavily impact the results. This study compares the performance of several variable selection techniques applied to logistic regression. We performed realistic simulation studies to compare five methods of variable selection: (1) a confidence interval approach for significant coefficients (CI), (2) backward selection, (3) forward selection, (4) stepwise selection, and (5) Bayesian stochastic search variable selection (SSVS) using both informed and uniformed priors. We defined our simulated diseases mimicking odds ratios for cancer risk found in the literature for environmental factors, such as smoking; dietary risk factors, such as fiber; genetic risk factors such as XPD; and interactions. We modeled the distribution of our covariates, including correlation, after the reported empirical distributions of these risk factors. We also used a null data set to calibrate the priors of the Bayesian method and evaluate its sensitivity. Of the standard methods (95% CI, backward, forward and stepwise selection) the CI approach resulted in the highest average percent of correct associations and lowest average percent of incorrect associations. SSVS with an informed prior had higher average percent of correct associations and lower average percent of incorrect associations than did the CI approach. This study shows that Bayesian methods offer a way to use prior information to both increase power and decrease false-positive results when selecting factors to model complex disease risk.

Keywords

Bayesian logistic regression; Case-control analyses; Logistic regression; Prior calibration; Variable selection

Introduction

To discover and describe the etiology of complex diseases requires simultaneously examining the association of multiple factors with the disease. The risk factors included in the final model of disease association can heavily impact the results of the study. If a researcher includes too many covariates, or not enough data, the coefficients in a logistic regression or conditional logistic regression model can be biased. The coefficients of various factors would then be overestimated, which would lead to more false-positive results for association of the risk factors [1,2]. Therefore, the selection of factors associated with a given disease is a very important part of model building. Researchers commonly use logistic or conditional logistic regression with standard stepwise statistical methods to select

[*]Corresponding author, mdswartz@mdanderson.org.

associated factors for a particular disease. However, standard stepwise statistical methods for variable selection tend to select the inflated coefficients and, at the end of the analysis, report those inflated coefficients [3]. This is because the standard selection methods, which involve statistically testing a variable's significant contribution to a model in a stepwise fashion, ignore the uncertainty introduced by the model selection process [4–6]. Recent research suggests that hierarchical shrinkage estimation and/or Bayesian methods offer a solution to reducing the false-positives and overestimation of the coefficients [3,7–11]. In particular, some authors propose using informed priors in a Bayesian analysis to reduce these errors [3,10,11].

One approach that accounts for the uncertainty introduced by model selection, Bayesian model averaging (BMA), has recently been applied to logistic regression for case-control studies [12,13]. BMA more accurately models the variation in the estimates for the coefficients by accounting for the uncertainty surrounding model selection. Specifically, it incorporates prior model information as the prior probability of a model appropriately modeling the association [6,14–16]. Thus, BMA methods create a comprehensive mathematical model of the associations in the data that contains both the uncertainty in estimating the associations in the data through the model coefficients and the uncertainty in selecting the model itself.

An alternative type of Bayesian model averaging, known as stochastic search variable selection (SSVS), focuses on variable selection rather than model selection [17–20]. Like BMA, SSVS captures the uncertainty surrounding model selection, but it is framed in terms of the variables, rather than the model as a whole. With SSVS, priors can be defined in terms of the prior probability of a variable's inclusion in the model, which is interpreted as the prior probability of a specific variable's association with the outcome. The model probability can then be defined as the joint probability of all the variables that compose that particular model. SSVS allows researchers to think in terms of specific covariates—or in the epidemiology setting, specific risk factors— both as they construct prior model probabilities, and as they make posterior inferences. Therefore, SSVS provides an alternative to BMA that may facilitate prior elicitation and posterior interpretation.

Previous research has shown that BMA outperforms the standard selection methods [12,13] when using uninformed priors. The purpose of this study was to compare the performance on a logistic regression model of standard variable selection techniques with that of the Bayesian SSVS technique, using both informed and uninformed priors. We simulated data from distributions with parameters chosen to reflect reality to compare the average percent correct and percent incorrect associations detected by the selection methods. The scenarios we present here are in the context of hypothesis driven studies, where the number of covariates is smaller than the sample size (e.g. a candidate gene association study).

Methods

Simulating Covariates

In order to compare model selection methods, we simulated realistic data by defining the parameters for the distributions, from which we simulated our covariates using values found in the literature for lung or colorectal cancer. The studies we used consisted of predominantly non-Hispanic Caucasians, and further details can be found in the individual references. We selected factors that drastically increase risk, moderately increase risk, and moderately decrease risk. For continuous covariates, such as nutritional factors, we collected means, variances, and some relevant covariances to simulate covariates from univariate or multivariate normal distributions. For categorical covariates, such as smoking status, we first simulated values from a standard univariate normal distribution or a multivariate normal

distribution with mean 0, variance 1, and relevant covariances from the literature and then categorized the distributions using appropriate values that correspond to prevalences reported in healthy populations (usually control samples from a case-control study).

The following are examples of the covariate simulations we performed. Fiber intake, a continuous covariate, is correlated with folate intake in colorectal cancer cases and controls. Therefore, we simulated fiber and folate from a multivariate normal distribution, setting the correlations for fiber and folate to be .35 [21] and using a mean of 17.8 g/day and standard deviation of 7.84 g for fiber [22] and a mean of 331 mcg/day and standard deviation of 137 mcg for folate [23]. From these values, we constructed a multivariate covariance matrix and mean vector. For the categorical covariates of smoking status and alcohol intake, we simulated smoking and alcohol intake as dichotomous, (smoker or non-smoker; drinker or non-drinker). We first simulated a multivariate normal distribution with mean 0, variance 1, and covariance 0.54 from [24]. We then used a cut point of −0.8416212 for smoking to match the reported prevalence of .20 [25], and a cut point of −0.1509692 for alcohol to match the prevalence of .44 [26]. For other covariates, we used parameters on the order of those found in the literature. For the full list of covariate distributions, see Table 1.

Simulated Diseases

We simulated all diseases using a logistic regression model, defining the coefficients using odds ratios found in the literature as a guide. For the complete list of odds ratios from the literature and the corresponding coefficients, see Table 2. We simulated 3 diseases, with each disease having its own model and analysis to highlight different risk mechanisms. We also simulated a null data set to calibrate our priors. The first disease (Simulated Disease 1) we simulated using a model containing only main effects. We defined cases and controls by defining the probability of a disease as

$$\text{logit}(P(Disease))=2.708smoke+0.7419alcohol+0.61XPD-0.6349folate-.5108MTHFR. \quad (1)$$

This model yielded about a 30% disease prevalence on average in the simulated population. When we analyzed this disease, we included all true covariates and 7 unassociated covariates (e.g., fiber, X6–X10, X15) not originally in the simulation model (see Table 3). The second disease model (Simulated Disease 2) was also simulated from a model using only main effects, but we used more covariates:

$$\begin{aligned}\text{logit}(P(Disease))=&2.708smoke\\&+0.7419alcohol\\&+0.6098XPD\\&-0.6349folate\\&-0.5108MTHFR\\&+0.4055X6\\&-0.5878X7\\&-0.6913X8-0.6419X9+0.5306X10.\end{aligned} \quad (2)$$

This model yielded about a 2.5% disease prevalence on average in the simulated population. When we analyzed Simulated Disease 2, we omitted covariates X8 and X9 (which were included in the simulation model, equation (2)) but included all other associated covariates and an additional 4 unassociated covariates (X11–X14, see Table 4) to examine how each selection method (described below) performed when not all causal covariates were included

in the analysis. The third disease model (Simulated Disease 3) was simulated including the main effects and a non-nested interaction term:

$$\text{logit}(P(Disease))=2.708smoke+0.7419alcohol+0.6098XPD-0.6349folate-0.5108MTHFR+0.9163X15*X7. \quad (3)$$

This disease yielded about a 30% disease prevalence on average in the simulated population. When analyzing Simulated Disease 3, we included all the associated covariates and the interaction and analyzed an additional 7 unassociated covariates and 7 unassociated interactions (see Table 5).

For each disease scenario, we simulated 500 replicates, with 500 cases and 500 controls. To accurately simulate the disease, for each scenario, we simulated 200,000 individuals with disease penetrance defined by the appropriate logistic regression model. Then we sampled 500 cases and 500 controls from this simulated population. For each replicate, we simulated a new population. For the null data set, we randomly selected one replicate from Simulated Disease 2 and independently assigned case status. To simulate the case status, we simulated a binary random variable with probability parameter 0.5. If the binary variable was 1, the status was changed; if the binary variable was 0, the status did not change. This method can be used on any collected data set, and on average keeps the number of cases and controls equal. We created 500 null data sets from the selected replicate.

Variable selection methods

Using the simulated disease models described above, we compared the performance of five different model selection methods: a confidence interval approach for significant coefficients (CI), backward selection, forward selection, stepwise selection, and Bayesian stochastic search variable selection (SSVS) using both informed and uniformed priors. The first four methods we consider standard methods, easily implemented in a software package. The first method consists of calculating the 95% confidence interval for the β coefficient. If the confidence interval does not include 0, we select the variable as important to the model. The next methods, backward, forward, and stepwise selection are all well known selection methods [27]. Backward selection first estimates the coefficients for each covariate in the model and then calculates the Wald statistic for each coefficient. The covariate whose coefficient has the largest p-value greater than a threshold is removed and never re-enters the model. The process is repeated until all covariates are significant. For forward selection, SAS calculates the chi-squared statistic for each covariate not yet entered in the model, and the largest statistic with a p-value less than an entry value enters in the model and remains in the model. Stepwise selection in SAS is calculated like forward selection, with the flexibility of removing a variable if its p-value becomes larger than a threshold [28]. Once each standard selection method was complete, we screened the remaining covariates with the 95% confidence interval method described above to decide on a final model. For this study, we investigated the performance of these standard selection methods using 3 different significance thresholds: 0.01, 0.05, and 0.20.

SSVS, which was recently extended to conditional logistic regression [29], is applicable to generalized linear models [18,30–32], and here we apply it to logistic regression. We begin with the familiar logistic regression likelihood:

$$\text{logit}(P(Y=1))=\sum_{i=1}^{p}\beta_i x_i. \quad (4)$$

Then, we model the model selection uncertainty through a mixture prior assigned to each coefficient:

$$\beta_i|\gamma_i \sim \gamma_i N(0, c\tau)+(1-\gamma_i)N(0,\tau), \tag{5}$$

where γ_i is a Bernoulli indicator with probability w_i, and we describe c and τ next.

The variable selection mechanics come through the Bernoulli indicator and the values for c and τ. We choose τ to be a small value so that when $\gamma_i = 0$, the prior on β_i focuses the probability on values of β_i close enough to 0 that we can consider the covariate as having a negligible contribution to the model. And we choose c to be large enough such that when $\gamma_i = 1$, the product of $c\ \tau$ spreads the probability distribution of β_i to cover reasonable values of the coefficient for the covariate that is selected. To select covariates, we use the median model decision rule [33]: select all covariates where $P(\gamma_i=1|\text{data}) \geq 50\%$.

Comparing Analyses

For each simulated scenario, and the null set, we used available software to implement all forms of variable selection. We used SAS 9.1 (Cary, NC) to implement the standard variable selection methods. For forward, backward, and stepwise selection, after the covariates were selected, we applied the 95% confidence interval rule described above for further selection. We implemented SSVS using WinBUGS 1.4.2 [34], assigning c and τ using a rationale similar to that used previously [29]. Since the coefficients have the interpretation of the log-odds ratio, we chose c to be 1000 and τ to be 0.001. This covered the range from −3 to 3, with a little over 99% of the probability mass, when $\gamma = 1$ and restricted the range from about -0.003 to 0.003 when $\gamma = 0$. We also assigned different types of priors. We used non-differentiating prior probabilities of inclusion, where all the prior w_i's were given the same value, either 0.5, matching those used in [29], or 0.25, matching those used in [35]. Additionally, we used a differentiating prior where we set w_i to 0.5 for some variables and 0.25 for others. This models the idea of using informative priors for those variables with information in the literature. We used a $w = 0.5$ for MTHFR, smoking, alcohol, XPD and folate to represent upweighting the prior according to the literature. We also use a $w = 0.5$ for fiber to reflect using colon cancer literature to inform our risk prior for lung cancer, as a result misspecifying our prior. For the remaining covariates we used $w_i = 0.25$ to represent skepticism of association for those remaining covariates. We also created a second type of informative prior by setting all the w_i's to 0.25 and setting the mean of the beta distribution to the value for the logistic regression model simulation. One caveat, if this were a real study on real data, we would spend more time eliciting expert opinion regarding the importance of each risk factor, rather than weighing the priors according to our simulation.

Once all the methods were run on each of the 500 replicates, we calculated the percentage of replicates where each covariate was selected by each method. We used these values to calculate the average correct association percentage and incorrect association percentage for each method. To calculate the average correct association percentage, we added up the percentages of both the associated variables selected and the unassociated variables not selected for a model and divided by the total number of variables (equation (6))

$$\%correct\ association = \frac{\sum_{associated} P(selected|associated) + \sum_{unassociated} [1 - P(selected|unassociated)]}{Total\ number\ of\ variables} \tag{6}$$

Likewise, to calculate average percentage of incorrect association, we added up the percentages of those associated variables not selected, and those unassociated variables selected, and divided by the total number of variables (equation (7)).

$$\%incorrect\ association = \frac{\sum_{associated} [1 - P(selected|associated)] + \sum_{unassociated} P(selected|unassociated)}{Total\ number\ of\ variables} \quad (7)$$

For simulated disease 3, since the disease was simulated with the interaction term alone, without the component main effects, we do not count the component main effects as true risk factors, and instead count them as incorrect associations when they are included in the model, with or without the interaction term. We count the main effects this way since this paper examines how variable selection methods recover the true model.

For an additional measure of comparison, we also present the ratio of the average percent correct to average percent incorrect. This can be interpreted as the Bayes factor describing support of correct association when the prior probability of correct association is equal to the prior probability of incorrect association. We caution over-interpreting these ratios however, because in all cases of selection using standard or Bayesian methods (except SSVS when w =0.5), this assumption may not be accurate.

Results

The percentages of all replicates for the simulated diseases for which each method selected each covariate is presented in Tables 3, 4, and 5. Each column of the tables corresponds to a method used. 95% CI lists the percentage of replicates where the 95% confidence interval for that covariate did not include 0. Backward, forward, and stepwise columns refer to the corresponding selection methods, and the p values for sub-headings refer to the threshold values (p=0.01, 0.05 and 0.20). Then the tables present the columns for SSVS with subheadings for the different priors. First is w=0.25 for all covariates considered, followed by SSVS with w=0.5 for all covariates considered. The next column, w=0.25 or 0.5, refers to selecting the prior probability of inclusion based on the literature. And the final column reports the performance of SSVS using w=0.25 for all covariates, and centering the prior for β at the simulated log odds ratio.

Results of Simulated Disease 1

We began our comparisons with a simple model: 5 main effects, with some factors that increased risk and some factors that decreased risk. For this simple model, all methods had comparable power to select the true variables (see Table 3). However, the standard methods had higher incorrect association percentages than the SSVS methods. Of the standard methods, the 95% confidence interval has the lowest average percentage of incorrect associations at 5.9%. Backward, forward, and stepwise all had average incorrect association percentages of 7.9% or higher for every threshold. As to be expected, as the threshold increased, the incorrect associations increased. For SSVS, using w = 0.5 gave the largest average percentage of incorrect associations (3%), but the percentage was still less than that of the 95% CI method. Next largest was centering the priors for β at the simulation value followed by SSVS using w = 0.25. Choosing the prior probability for inclusion to be 0.5 for the associated covariates (and for fiber) and 0.25 for the unassociated covariates (except fiber) gave the lowest incorrect association percentage of 1.1%. Interpreting the ratios on a Bayes Factor scale [5], only SSVS shows strong evidence for correct associations (ratio greater than 20), with the differentiating prior yielding the largest ratio.

Results of Simulated Disease 2

Next, we analyzed Simulated Disease 2, where we simulated the disease from main effects only, but omitted some of the associated main effects from consideration in the final analysis. The average ratios were lower across all methods than those for Simulated Disease 1 (see Table 4). The 95% confidence interval approach had the lowest average incorrect association percentage rate of the standard methods and the highest average correct association percentage (19.4% and 80.6%, respectively) among the standard methods. It also had the highest average ratio of the standard methods (4.1). The other standard selection methods performed very similarly with each other. As the threshold increased, so did the average correct association percentages. However, the incorrect association percentages decreased as the threshold increased. The average ratio ranged between 2.5 and 3.5 for all three stepwise selection methods, across all thresholds, which borders between barely worth a mention and positive evidence for selecting correct associations on the Bayes factor scale [5]. Meanwhile, SSVS had the lowest average incorrect association rates, the maximum being 17.9% when centering the priors for β, and the minimum being 9.6% for w=0.5. SSVS also has the highest average ratio of all the methods, with the prior w = 0.5 giving the highest average ratio of 9.4, while the priors centered on the simulation values for β has the lowest average ratio of 4.6, almost as low as the 95%CI method of selection. However, all SSVS methods had ratios in the positive evidence range of a Bayes Factor scale (ratios greater than 3.2) [5].

Results of Simulated Disease 3

The results for Simulated Disease 3 show how the selection methods perform when analyzing an interaction term. All associated covariates and interaction terms were included in the analysis (see Table 5). Again, of the standard methods, using the 95% confidence interval to detect associated covariates and interactions had the lowest average incorrect association percentage and the highest correct association percentage (6.7% and 93.3%, respectively). Comparing the stepwise selection methods, backward selection had the lowest average incorrect association percentages, and the best average ratios for each threshold value. SSVS, regardless of prior specification, had the smallest average percentage of incorrect associations across the replicates, and the highest average percentage of correct associations. Considering the average ratio, the differentiating prior of using w = 0.25 or 0.5 depending on the covariate w = 0.25 gave the strongest evidence for finding true associations (average ratio = 136), followed by using w = 0.25 which yielded an average ratio of 99. The lowest average ratio was 18.2 for the prior w = 0.5, which is still positive, and almost strong evidence for detecting correct associations. Centering the prior for β actually reduced the average ratio from simply using w = 0.25, but still provides much stronger evidence for detecting correct associations than using w=0.5 without centering the priors (44.9 versus 18.2 respectively).

Analysis of Null Replicates

We present the analysis of the null replicates slightly differently. We report the average and the maximum false-positive frequencies for each selection method, along with the percentage of covariates selected by each method, and do not calculate any ratios (see Table 6). Also since we are simulating a null set, we don't have any prior information about the importance of particular covariates. Therefore, we used four different values of the prior probability of inclusion, rather than the previous differentiating priors. The standard methods all had similar average false-positive frequencies. Of the standard methods (holding the threshold constant for comparisons), stepwise selection and forward selection had the smallest false positive percentage compared to backwards selection. Using 95% CI has a false positive percentage that is not as low as the p = 0.01 threshold, but is lower than the p = 0.05 and 0.20 thresholds for forward and stepwise selection methods. Backwards selection

had the highest false positive percentages for each threshold and was larger than the 95% CI method for thresholds greater than or equal to $p = 0.05$. Using a prior value of $w = 0.25$ and 0.4 for SSVS had a false positive percentage similar to the selection methods with $p = 0.01$ threshold. Increasing w to 0.5 or 0.6 increased the false positive percentage greater than stepwise or forward selection with $p = 0.01$ threshold, but not as high as using thresholds greater than or equal to $p = 0.05$.

Discussion

We compared Bayesian SSVS with informed and non-informed priors and standard variable selection methods. We compared their performance on 3 different simulated diseases: main effects only; main effects only, but some covariates were not considered in the analysis; and main effects with an interaction. Across all analyses, the Bayesian SSVS methods had the lowest average incorrect association percentages and had average correct association percentages and average ratios greater than those of the standard selection methods, regardless of how much information was modeled in the priors. Thus, this study agrees with previous studies [12,13], showing that methods accounting for model selection uncertainty tend to be more powerful in finding the true model than methods that do not account for model selection uncertainty.

This study also shows some advantages to including prior information when it is present. In most scenarios, the average ratio was highest when we used SSVS with a discriminatory prior probability of inclusion ($w = 0.25$ or 0.50, depending on the covariate). These results also show that informatively increasing the prior probability of inclusion for only those variables believed *a priori* to be associated with the disease controls the average incorrect associations and both increases the average correct association, and increases their ratio. However, we see that when we inform the prior distribution of the log odds ratios (β coefficients), SSVS did not perform as well as using discriminating prior probabilities of inclusion (informing the w_i's). Therefore, informing the prior probability of inclusion has a stronger impact on controlling the average incorrect associations and increasing the correct associations than informing the prior distributions on the coefficients.

We also see that this method exhibits robustness in some situations. When we increased the prior probability of inclusion for fiber to 0.5, even though it wasn't truly simulated to be associated with the disease, SSVS correctly determined that this variable was not associated with the simulated disease across all scenarios. Looking across different prior values, we also see that the method is very robust for finding the true associated risk factors when all factors were investigated, such as in Simulated Diseases 1 and 3. However, the method became more sensitive to the prior probabilities for inclusion of associated variables when some of the associated variables were omitted from the analysis, as in Simulated Disease 2.

Our analysis of null data sets serves two purposes. First, it evaluates the sensitivity to different prior probabilities of inclusion. Second, it provides a model for calibrating the prior probability of inclusion for this and future studies. We specifically chose to generate disease status unassociated with a fixed set of covariates to model using one data set, as would be the case with a real data set. This calibration study we performed here can be done for any Bayesian study, allowing researchers to calibrate uninformative priors for individual data sets. For this set (Simulated Disease 2), we see that increasing the prior probability of inclusion did not increase average percentage of incorrect associations appreciably.

The calibration study, combined with the results across the different priors, has important implications for using informed priors, and Bayesian methods in general, for epidemiological research. The fact that using a differentiated prior probability of inclusion

performed the best in most cases suggests that informing the prior can be useful. Even in Simulated Disease 2, where we did not include all of the true associated covariates in the analysis, although the non-differentiated prior with w=0.5 performed best, the differentiated prior was still close, and controlled the incorrect associations well. Therefore, a differentiated prior is recommended. In the absence of prior knowledge on how to inform the prior, these results recommend using the same value between 0.25 and 0.5 for all the w_i. These prior values, combined with the median model decision rule [33] leads to the same decision rule as including variables whose Bayes factor is greater than 3 or 1 respectively. Since a Bayes Factor greater than 3 leads to "substantial evidience" this supports previous findings recommending using $w = 0.25$ [35]. Also, since the incorrect association rates do not increase appreciably, and the method is more powerful when all the true covariates are present in the pool of searchable variables, it is better to include more, rather than fewer, variables in the initial pool of searchable variables, to a point. Although informing the prior improves variable selection performance, we caution the readers that, as discussed in previous reports, these methods are not to replace careful consideration of available scientific information or data analysis [12,13]. The epidemiologist should approach the study with a preconceived set of candidate variables to investigate as potential risk factors or confounders. However, these results imply that if there is debate about whether or not to include a specific factor as a potential, it can be included, using an appropriately calibrated prior. However, we remind the reader that in this study we used a relatively small number of covariates, representing a typical hypothesis driven approach focusing on specific covariates and interactions. Theoretically, SSVS can be applied to problems with many more covariates (see [18,36]), but this would increase computational burden, and often requires custom computer code to implement. As the number of variables increase, identifiability has to be addressed as well, [37–40] which is beyond the scope of this paper.

This study also compared standard methods using the same data. We see that for all simulated diseases, the standard stepwise selection methods did not offer any improvement in the average percentage of correct or incorrect associations, or their ratio over simply calculating the 95% confidence interval for the log odds ratio and selecting those variables whose confidence intervals did not include 0. In all three simulated diseases, by our measures, the 95% confidence interval method performed better than the standard selection methods, regardless of the threshold value. Even looking at the individual variables, there were only a few variables where either backward selection or forward selection performed better than constructing the 95% confidence intervals. As for within the stepwise selection methods, on average, they performed similarly across the scenarios without interactions. However when investigating specific interactions, backward selection with a stringent threshold performed the best for the non-nested models.

This study supports using Bayesian variable selection methods with informed priors when looking for factors associated with disease, both risk factors and protective factors. With appropriately informed priors, SSVS had the highest average percentages of correct associations and lowest average percentages of incorrect associations of all the methods examined. Using a calibration study with the data at hand helps to calibrate the priors to some optimum setting, and also helps evaluate robustness to the prior for certain analyses. Note that the methods used in this study are model selection methods, which are applicable to many types of generalized linear models. On the basis of results in the literature regarding performance with a continuous response [5,6,18], we conjecture that these results generalize to a continuous response. Finally, this study showed that of the standard selection methods examined, none provided improvement over the classical method of calculating the 95% confidence intervals and evaluating whether or not the interval for the log odds ratio includes 0. With this study, we see that Bayesian methods offer a way to bring prior information into the analysis in such a way as to increase detection of true associations.

Acknowledgments

We would like to acknowledge our sources of funding: Michael Swartz was funded by NCI grant number 1K07CA123109-01A1. We also wish to thank the reviewers for their constructive comments that helped improve this paper.

References

1. Greenland S. When should epidemiologic regressions use random coefficients? Biometrics. 2000; 56:915–921. [PubMed: 10985237]
2. Jewell NP. Small-Sample Bias of Point Estimators of the Odds Ratio from Matched Sets. Biometrics. 1984; 40:421–435. [PubMed: 6487726]
3. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. International Journal of Epidemiology. 2007; 36:195–202. [PubMed: 17329317]
4. George EI. The variable selection problem. Journal of the American Statistical Association. 2000; 95:1304–1308.
5. Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995; 90:773–795.
6. Raftery AE. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika. 1996; 83:251–266.
7. Budtz-Jorgensen E, et al. Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. Annals of Epidemiology. 2007; 17:27–35. [PubMed: 17027287]
8. Greenland S. Bayesian perspectives for epidemiological research: I. foundations and basic methods. International Journal of Epidemiology. 2006; 35:765–775. [PubMed: 16446352]
9. Gustafson P. Sample size implications when biases are modelled rather than ignored. Journal of the Royal Statistical Society, Series A. 2006; 169:865–881.
10. Thomas DC, Witte JS, Greenland S. Dissecting effects of complex mixtures: who's afraid of informative priors? Epidemiology. 2007; 18:186–190. [PubMed: 17301703]
11. Young J. Statistical errors in medical research--a chronic disease? Swiss Med Wkly. 2007; 137:41–43. [PubMed: 17299668]
12. Viallefont V, Raftery AE, Richardson S. Variable selection and Bayesian model averaging in case-control studies. Statistics in Medicine. 2001; 20:3215–3230. [PubMed: 11746314]
13. Wang D, Zhang W, Bakhai A. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. Statistics in Medicine. 2004; 23:3451–3467. [PubMed: 15505893]
14. Madigan D, Raftery AE. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occams Window. Journal of the American Statistical Association. 1994; 89:1535–1546.
15. Raftery AE. Bayesian Model Selection in Social Research. Sociological Methodology. 1995; 25:111–163.
16. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. Journal of the American Statistical Association. 1997; 92:179–191.
17. George, EI. Discussion of Bayesian model averaging and model search strategies by M.A. Clyde. In: Bernardo, J., et al., editors. Bayesian Statistics. Vol. 6. Oxford University Press; London: 1999.
18. George EI, McCulloch RE. Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association. 1993; 88:881–889.
19. George EI, McCulloch RE. Approaches for Bayesian variable selection. Statistica Sinica. 1997; 7:339–373.
20. Mitchell TJ, Beauchamp JJ. Bayesian Variable Selection in Linear Regression. Jounral of the American Statistical Association. 1988; 83:1023–1032.
21. Bingham SA, et al. Is the association with fiber from foods in colorectal cancer confounded by folate intake? Cancer Epidemiology, Biomarkers and Prevention. 2005; 14:1552–1556.

22. Hudson TS, et al. Dietary fiber intake: assessing the degree of agreement between food frequency questionnaires and 4-day food records. Journal of the American College of Nutrition. 2006; 25:370–381. [PubMed: 17031005]
23. Shen H, et al. Dietary Folate Intake and Lung Cancer Risk in Former Smokers: A Case-control analysis. Cancer Epidemiology, Biomarkers and Prevention. 2003; 12:980–986.
24. Batel P, et al. Relationship between alcohol and tobacco dependencies among alcoholics who smoke. Addiction. 1995; 90:977–980. [PubMed: 7663320]
25. Centers for Disease Control and Prevention. Tobacco Use Among Adults --United States, 2006. Morbidity and Mortality Weekly Report [Serial Online]. 2006; 55:1145–1148.
26. Encyclopedia of Public Health Online. 2002. http://www.answers.com/topic/alcohol-use-and-abuse?cat=health
27. Neter, J., et al. Applied Linear Statistical Models. 4. Irwin; Chicago: 1996.
28. SAS Institute Inc. SAS OnlineDoc® 9.1.3. SAS Institute Inc; Cary, NC: 2004.
29. Swartz MD, et al. Stochastic Search Gene Suggestion: A Bayesian Hierarchical Model for Gene Mapping. Biometrics. 2006; 62:495–503. [PubMed: 16918914]
30. Chipman, H.; George, EI.; McCulloch, RE. The Practical Implementation of Bayesian Model Selection. In: Lahiri, P., editor. Model Selection. Institute of Mathematical Statistics; Beachwood, Ohio: 2001. p. 65-134.
31. Dellaportas P, Smith AFM. Bayesian-Inference for Generalized Linear and Proportional Hazards Models Via Gibbs Sampling. Applied Statistics-Journal of the Royal Statistical Society Series C. 1993; 42:443–459.
32. Ntzoufras I, Forster JJ, Dellaportas P. Stochastic search variable selection for log-linear models. Journal of Statistical Computation and Simulation. 2000; 68:23–37.
33. Barbieri MM, Berger JO. Optimal predictive model selection. Annals of Statistics. 2004; 32:870–897.
34. Spiegelhalter, DJ., et al. WinBUGS. Cambridge; 2007.
35. Swartz MD, Shete S. The Null Distribution of Stochastic Search Gene Suggestion: A Bayesian Approach to Gene Mapping. BMC Proceedings. 2007; 1:S113–S118. [PubMed: 18466454]
36. Swartz MD, et al. Model selection and Bayesian methods in statistical genetics: summary of group 11 contributions to Genetic Analysis Workshop 15. Genetic Epidemiology. 2007; 31(Suppl 1):S96–102. [PubMed: 18046760]
37. Eberly LE, Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. Statistics in Medicine. 2000; 19:2279–2294. [PubMed: 10960853]
38. Gelfand AE, Sahu K. Identifiability, improper priors, and Gibbs sampling for generalized linear models. Journal of the American Statistical Association. 1999; 94:247–253.
39. Lindley, DV. Bayesian statistics; a review. Regional conference series in applied mathematics 2; Philadelphia: Society for Industrial and Applied Mathematics; 1971. p. 83
40. Rannala B. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Systematic Biology. 2002; 51:754–760. [PubMed: 12396589]
41. Centers for Disease Control. Cigarette Smoking among Adults - United States, 2006. Morbidity and Mortality Weekly Report. 2007; 56:1157–1161. [PubMed: 17989644]
42. Spitz MR, et al. Modulation of nucleotide excision repair capacity by XPD polymorphisms in lung cancer patients. Cancer Research. 2001; 61:1354–1357. [PubMed: 11245433]
43. Shen H, et al. Polymorphisms of Methylene-tetrahydrofolate Reductase and Risk of Lung Cancer: A Case-Control Study. Cancer Epidemiology, Biomarkers and Prevention. 2001; 10:397–401.
44. Shi Q, et al. Polymorphisms of methionine synthase and methionine synthase reductase and risk of lung cancer: a case-control analysis. Pharmacogenet Genomics. 2005; 15:547–555. [PubMed: 16006998]

Table 1

Simulation parameters for covariate distributions

Covariate	Distribution	Population Prevalence	Reference
Smoking	Multivariate normal with Alcohol. Mean 0, variance 1, correlation 0.54	0.2	Centers for Disease Control and Prevention, 2006 [41]
Alcohol	Multivariate normal with Smoking. Mean 0, variance 1, correlation 0.54	0.44	Encyclopedia of Public Health, 2002 [26], Correlation from Batel, 1995 [24]
XPD (Lys751Gln)	N(0,1)	LL = 0.442 LG = 0.45 GG=0.108	Spitz, M.R. et al., 2001 [42]
Fiber	Multivariate normal with Folate. Mean 77.8 g, variance 7.84^2, correlation 0.35	Continuous Covariate	Hudson, TS et al., 2006 [22]
Folate*	Multivariate normal with Fiber. Mean 331 mcg, variance 137^2, correlation 0.35	Continuous Covariate	Shen H. et al, 2001 [43] Correlation: Bingham et al, 2005 [21]
MTHFR (C677T)	N(0,1)	TT = 0.42 TC = 0.455 CC=0.105	Shi, Q. et al, 2005 [44]
X6†	N(74.47, 25.53)	Continuous	Arbitrary
X7†	N(0, 33.55)	Continuous	Arbitrary
X8*	N(237.25,7.57)	Continuous	Arbitrary
X9	N(15, 5)	Continuous	Arbitrary
X10	N(0,1)	AA = 0.5 AB=0.45 BB=0.05	Arbitrary
X11	N(17, 5)	Continuous	Arbitrary
X12	N(22, 6)	Continuous	Arbitrary
X13	N(25, 7)	Continuous	Arbitrary
X14	N(12, 5)	Continuous	Arbitrary
X15	N(0, 1)	1 = 95% 2= 5%	Arbitrary-Similar to recessive coding for a SNP**

*Covariates transformed by dividing by 100 before simulating the disease and analysis to facilitate computations of the probabilities.

†Covariates transformed by dividing by 10 before simulating the disease and analysis to facilitate computations of the probabilities.

**SNP = single nucleotide polymorphism.

Table 2

Covariates and odds ratios extracted from the literature

Covariate	Odds Ratio	β-coefficient[+]	Reference
Smoking	15	2.708	Le Marchand et al., 2002
Alcohol	2.1	0.7419	Bandera, E. V., et al. 2001
XPD	1.84	0.6098	Spitz et al., 2001
Fiber	1	0	
Folate	.53	−0.6349	Shen et al. 2003
MTHFR	.6	−0.5108	Shi et al. 2005[*]

[*] The odds ratio for MTHFR is actually the lower endpoint of the 95% confidence interval in this source.

[+] β-coefficient = log(odds ratio)

Table 3

Percent each covariate was selected, average percentage of correct associations and average percentage of incorrect associations of the 500 replicates of Simulated Disease 1 for each selection method

Covariate	95% CI	Backward			Forward			Stepwise			SSVS			
		p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	w=0.25	w=0.5	w=0.5 or 0.25	informed β-prior means
Smoking^{+}	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Alcohol^{+}	100	99.8	100	100	99.8	100	100	99.8	100	100	100	100	100	99.6
XPD^{+}	100	100	100	100	100	100	100	100	100	100	99.8	100	100	99.8
Fiber	7.4	4.8	11.6	10.2	5.0	11.4	10.2	4.4	11.4	10.2	0	0	0	0
Folate^{+}	100	100	100	100	100	100	100	100	100	100	100	100	100	100
MTHFR^{+}	97.6	87.0	96.2	97.6	84.0	95.0	97.4	83.4	95	97.4	89.2	95.4	93.6	87.2
X6	14.0	23.4	30.8	26.4	22.0	30.4	26.4	21.8	30.4	26.4	0	0	0	0
X7	5.2	1.2	5.2	5.6	1.2	5.0	5.6	1.2	5	27.8	0	0	0	0
X8	16.4	26.0	34.6	27.8	24.6	33.6	27.8	24.2	33.6	25.2	4.2	12.8	3.8	3.6
X9	14.8	22.6	29.8	25.0	22.4	30.2	25.2	22.4	30.2	25.2	0	0	0	0
X10	5.4	2.6	8.0	6.8	3.0	7.4	6.8	3.0	7.4	6.8	1.2	5.6	0.6	0.6
X15	5.4	1.0	6.0	5.6	1.5	5.6	5.6	1.0	5.6	5.6	3.0	13.2	2.8	2.8
Avg. correct Associations	94.1	92.1	89.2	90.9	92.0	89.3	90.8	92.1	89.3	89.2	98.4	97.0	98.9	98.3
Avg. incorrect associations	5.9	7.9	10.8	9.2	8.0	10.7	9.2	7.9	10.7	10.8	1.6	3.0	1.1	1.7
Avg. Ratio*	15.9	11.7	8.2	9.9	11.5	8.3	9.9	11.7	8.3	8.2	60.6	32.1	87.2	57.8

$^{+}$denotes covariate in true model

*Avg. Ratio is calculated as the Avg. Correct Associations/Avg. Incorrect Associations

Table 4

Percent each covariate was selected, average percentage of correct associations and average percentage of incorrect associations of the 500 replicates of Simulated Disease 2 for each selection method.

Covariate	95% CI	Backward			Forward			Stepwise			SSVS			
		p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	w=0.25	w=0.5	w=0.5 or 0.25	informed β-prior means
Smoking^{+}	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Alcohol^{+}	100	35.6	60.8	83.6	33.8	59.6	83.6	33.0	59.4	83.6	49.2	68.6	60.0	38.4
XPD^{+}	76.8	46.8	73.6	76.6	43.8	73.0	76.6	43.4	73.0	76.6	53.8	71.0	64.4	49.2
Fiber	10.4	5.0	13.4	11.4	5.2	13.2	21.4	4.6	13.2	11.4	0	0	0	6.8
Folate^{+}	100	100	100	100	100	100	100	100	100	100	100	100	100	100
MTHFR^{+}	89.0	74.2	90.8	89.0	73.8	90.4	89.2	73.4	90.4	89.2	68.4	86.4	83.2	63.2
X6^{+}	100	100	100	100	99.2	100	100	95.0	100	100	99.4	99.8	99.8	99.2
X7^{+}	100	100	100	100	100	100	100	100	100	100	100	100	100	100
X10^{+}	53.2	22.2	49.6	52.0	20.6	49.2	52.0	19.8	49.2	52.0	29.0	49.6	44.0	24.2
X11	42.4	38.8	53.6	47.6	36.8	53.2	47.6	36.4	53.0	47.6	0	0	0	0
X12	50.4	48.6	57.6	54.2	46.0	56.4	54.4	45.8	56.4	54.4	0	0	0	0
X13	46.8	43.0	58.4	53.0	42.4	58.2	53.0	41.6	58.2	53.0	0	0	0	0
X14	21.6	15.8	29.0	26.0	15.4	29.0	25.8	15.4	29.0	26.0	0	0	0	0
Avg. Correct Associations	80.6	71.4	74.0	77.6	71.2	74.0	76.8	70.8	74.0	77.6	84.6	90.4	88.6	82.1
Avg. Incorrect Associations	19.4	28.6	25.9	22.4	28.8	26.0	23.1	29.2	26.0	22.4	15.4	9.6	11.4	17.9
Avg. Ratio*	4.1	2.5	2.9	3.5	2.5	2.8	3.3	2.4	2.8	3.5	5.5	9.4	7.74	4.6

$^{+}$Denotes covariate in the true model

*Avg. Ratio is calculated as the Avg. Correct Associations/Avg. Incorrect Associations.

Table 5

Percent each covariate was selected, average percentage of correct associations and average percentage of incorrect associations of the 500 replicates of Simulated Disease 3 for each selection method

Covariates	95% CI	Backward			Forward			Stepwise			SSVS			
		p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	w=0.25	w=0.5	w=0.5 or 0.25	informed β-prior means
Smoking^{+}	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Alcohol^{+}	100	99.0	99.8	100	99.6	99.8	100	99.6	99.8	100	99.4	100	100	98.8
XPD^{+}	100	99.6	100	100	99.8	100	100	99.8	100	100	99.8	100	100	99.6
Fiber	25.0	25.6	43.0	42.4	5.2	12.0	13.0	4.8	12.0	13.2	0	0	0	0
Folate^{+}	100	100	100	100	100	100	100	100	100	100	100	100	100	100
MTHFR^{+}	98.2	88.6	97.0	98.0	81.4	95.0	98.2	81.0	95.0	98.0	91.2	96.4	96.6	89.6
X6	31.8	43.2	58.8	57.4	18.2	28.4	26.2	18.2	28.0	27.0	1.2	4.0	1.6	0.6
X7	5.2	5.8	8.8	8.8	13.0	15.2	10.2	12.8	15.2	10.2	0	0.2	0	0
X8	7.0	14.0	18.6	15.2	20.8	30.8	26.6	20.4	30.8	27.0	2.8	7.6	3.2	2.4
X9	6.4	10.6	14.4	11.0	21	28.6	23.7	20.6	28.6	24.2	0	0	0	19.8
X10	6.4	2.0	7.8	7.0	2.6	8.0	8.0	2.4	8.0	7.8	1.8	5.8	1.8	1.6
X15	3.8	8.4	15.4	17.0	33.8	37.2	22.0	34.2	37.8	22.8	0.4	13.2	0.4	0.8
Smoke* X15	2.0	0.2	2.8	3.6	8.8	16.4	14.4	7.2	16.2	14.6	0	28.8	0	1.0
Alcoh* X15	2.0	0.2	4.4	4.8	1.6	6.2	6.6	1.4	6.2	6.8	1.6	20.4	1.6	2.4
XPD* X15	3.2	0.4	4.2	5.2	1.6	6.0	7.4	1.0	5.8	7.4	2.4	18.0	2.4	2.6
Fiber* X6	21.2	22.0	39.2	35.6	0	0.8	2.4	0	0.8	3.0	0	0	0	0
Folate* X15	01.8	0.6	4.8	5.6	3.4	10.4	12.8	2.4	10.6	13.4	0.2	2.2	0.2	0.2
Fiber* X15	5.6	1.6	6.2	7.0	0	0.2	1.6	0	0.2	18.0	0	0.2	0	0
Folate* X7	5.0	1.0	5.4	5.6	0.2	0.1	1.8	1.0	1.0	1.8	0	0	0	0
X7* X15^{+}	93.6	99.2	100	99.8	3.6	17.2	40.8	3.2	17.0	41.6	99.8	100	100	99.8
Avg. Correct Associations	93.3	92.5	88.2	88.6	87.7	85.6	88.1	79.7	85.5	87.1	99.0	94.8	99.2	97.8
Avg. Incorrect Associations	6.7	7.5	11.9	11.4	12.3	14.4	11.9	20.3	14.4	12.9	1.0	5.2	0.7	2.18
Avg. Ratio*	13.9	12.4	7.4	7.8	7.1	5.9	7.4	3.9	5.9	6.8	99.0	18.2	136.0	44.9

$^{+}$denotes covariate in the true model

* Avg. Ratio is calculated as the Avg. Correct Associations/Avg. Incorrect Associations

Table 6

Percentage of covariates selected and average false positives out of the 500 null replicates

Covariate	95% CI	Backward			Forward			Stepwise			SSVS			
		p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	p=0.01	p=0.05	p=0.20	w=0.25	w=0.4	w=0.5	w=0.6
Smoking	6.2	1.0	7.2	7.8	0.2	4.0	8.8	0.2	4.0	8.8	0.8	2.8	4.6	8.0
Alcohol	5.6	1.2	6.4	8.6	0.6	2.8	7.2	0.6	2.4	7.4	0.4	1.8	3.2	6.6
XPD	5.8	1.6	8.4	7.0	0.2	3.8	7.2	0.4	3.8	7.2	1.0	2.4	3.0	5.2
Fiber	4.4	2.0	8.2	9.6	0.2	2.8	7.0	0.4	2.8	7.0	0	0	0	0
Folate	5.4	2.2	7.4	8.8	0.8	2.6	5.8	0.8	2.6	6.0	0.4	0.4	0.4	1.0
MTHFR	6.0	1.2	7.4	7.6	0.8	4.4	8.4	0.8	4.4	8.4	0.4	2.0	2.6	2.0
X6	4.4	2.6	9.8	9.6	0.2	1.4	4.2	0.2	1.2	4.6	0	0	3.2	5.0
X7	4.4	1.4	6.6	5.0	1.0	4.8	6.2	1.0	4.8	6.2	0	0	0	0
X10	5.0	1.8	6.2	6.4	0.6	4.4	7.2	0.6	4.4	7.0	1.2	2.2	0	0
X11	4.4	2.6	9.8	12.2	0.4	1.4	7.6	0.4	1.4	5.6	0	0	3.2	4.6
X12	7.8	3.2	12.4	14.8	0.2	1.6	7.2	0.2	1.4	7.2	0	0	0	0
X13	5.2	1.8	8.4	13.2	0	1.6	8.0	0	1.6	8.2	0	0	0	0
X14	4.6	3.2	1.0	10.4	0.4	3.0	7.2	0.4	2.8	7.2	0	0	0	0
Avg. False Positives	5.3	2.0	7.6	9.3	0.4	3.0	7.1	0.5	2.9	7.0	0.3	0.9	1.6	2.5
Max. False Positives	7.8	3.2	12.4	14.8	1.0	4.8	8.8	1	4.8	8.8	1.2	2.8	4.6	8.0