Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280217747311
journals.sagepub.com/home/smm
SAGE

Identifying the source of food-borne disease outbreaks: An application of Bayesian variable selection

Rianne Jacobs,[1] Emmanuel Lesaffre,[2] Peter FM Teunis,[3,4] Michael Höhle[5] and Jan van de Kassteele[1]

Abstract

Early identification of contaminated food products is crucial in reducing health burdens of food-borne disease outbreaks. Analytic case-control studies are primarily used in this identification stage by comparing exposures in cases and controls using logistic regression. Standard epidemiological analysis practice is not formally defined and the combination of currently applied methods is subject to issues such as response misclassification, missing values, multiple testing problems and small sample estimation problems resulting in biased and possibly misleading results. In this paper, we develop a formal Bayesian variable selection method to account for misclassified responses and missing covariates, which are common complications in food-borne outbreak investigations. We illustrate the implementation and performance of our method on a *Salmonella* Thompson outbreak in the Netherlands in 2012. Our method is shown to perform better than the standard logistic regression approach with respect to earlier identification of contaminated food products. It also allows relatively easy implementation of otherwise complex methodological issues.

Keywords

Bayesian variable selection, food-borne disease outbreaks, misclassification, missing value imputation, spike and slab prior

1 Introduction

With food chains becoming increasingly complex and food products being transported across the globe with increasing ease, contaminated food products can rapidly cause food-borne disease outbreaks.[1] Such outbreaks constitute a large health burden on society.[2] Examples include the *Salmonella* Thompson 2012 outbreak in the Netherlands with 1149 laboratory-confirmed cases (including 4 deaths)[3] and the *Escherichia coli* O104:H4 2011 outbreak in Germany with 3816 reported cases (including 54 deaths).[4] Early detection of such outbreaks and the subsequent identification of the contaminated food products is crucial in reducing the disease burden of such outbreaks. The aim of this paper is to develop a methodologically sound procedure to assist epidemiologists in identifying contaminated food products.

In current practice, the identification of the contaminated food product is a long and cumbersome process. The process involves several steps which are not clear cut, much like a criminal investigation. Information is incomplete, delayed, uncertain and continuously updated.[5] Outbreak investigations are, therefore, not fixed designs, but constantly evolving studies. Often they also involve many different authorities, e.g. veterinary and food-safety agencies responsible for knowing the distributional networks and health departments responsible for the registration of the human cases affected by the outbreak.[6] This paper focuses on the efforts conducted by such

[1]Department of Statistics, Informatics and Modelling, RIVM, Bilthoven, Netherlands
[2]L-Biostat, KU Leuven, Leuven, Belgium
[3]Centre for Zoonoses and Environmental Microbiology, RIVM, Bilthoven, Netherlands
[4]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA
[5]Department of Mathematics, Stockholm University, Stockholm, Sweden

Corresponding author:
Jan van de Kassteele, RIVM, PO Box 1, Bilthoven 3720 BA, Netherlands.
Email: jan.van.de.kassteele@rivm.nl

health agencies which use the identified outbreak cases and the likely pathogen of the outbreak in order to infer about the food product responsible for the outbreak using epidemiological methods.

Analytic case-control studies are the primary epidemiological tool in this process of identification used by epidemiologists. The basic concept of case-control outbreak investigations is the comparison of exposures in cases and controls.[5] Once an outbreak has been detected, cases which are known to belong to the outbreak are usually already available, e.g., because of specific symptoms, laboratory confirmation, etc. Based on this, a formal case definition is usually drafted. To start a case-control study, controls need to be selected in such a way that they are comparable to the cases and without symptoms. Wacholder et al.[7] provide some guidelines on minimizing the bias in control selection. One such source of bias may be confounding. Variables such as age, gender and geographical region may be important confounders when comparing cases and controls and various methods are available to deal with such confounding.[8]

In the first stage of data collection, extensive questionnaires, known as trawling or shotgun questionnaires, are used to obtain information on a wide range of exposures. Cases and controls fill out the questionnaire on what products they consumed in a specified period: the start and length of this period are determined by when the cases became ill and by the incubation period of the disease (ranging from a few hours for norovirus up to a week for most *Salmonella* serovars). During the outbreak, as new insights and information become available, the questionnaire is updated and subsequently becomes more focused as food products are excluded or more detailed questions are added. This questionnaire dynamic is one of the complications of case-control outbreak investigations. In addition, one can very well imagine the practical difficulties that subjects have trying to recall their dietary consumption in the given time period, resulting in many data being missing or being reported erroneously. There may be recall bias in that cases and controls remember their food consumption differently: cases may ''over-remember'' a positive exposure, while controls may have forgotten exposures.[5] Finally, despite that controls are questioned on their symptoms, it is impossible to confirm whether they are indeed true controls (i.e. not infected) or rather asymptomatic infections (i.e. infected but not ill). This may then result in misclassification of the response.

The statistical analysis of the questionnaires typically involves classical (conditional) logistic regression to investigate exposure effects while correcting for confounders. Due to the large number of different food products that people may have consumed, one often has a variable selection problem, where one attempts to identify relevant exposures. Moreover, in the beginning of an outbreak, the number of covariates (i.e. food products) may be greater than the number of observations. Classical variable selection procedures, i.e. a combination of univariable analysis and stepwise, forward or backward selection based on p-values,[9] are most employed, thereby ignoring small sample bias and problems of multiple testing. When searching for the causative agent of a food-borne disease outbreak, we, therefore, need a far more sophisticated variable selection procedure.

We argue that the Bayesian approach offers powerful tools to deal with variable selection problems complicated by some of the above-mentioned issues of outbreak investigations. Bayesian methods allow us to use external information (in the form of prior distributions) to aid the modelling when data are scarce. This is crucial in the analysis of our case-control data especially in the light of early identification when very few questionnaires have yet been returned. In addition, such methods provide us with the flexibility to account for the problems of missing covariates and misclassified responses in a unified framework which is hard to solve in a formal frequentist setting. Moreover, the methods are not afflicted with the typical frequentist problems associated with multiple testing, such as unreliable p-values and biased estimation results.[10] In this paper, we, therefore, develop a formal Bayesian variable selection method, based on the stochastic search variable selection (SSVS) procedure,[11] which accounts for misclassified responses and deals with the problem of missing covariates. We illustrate our method on the Dutch *Salmonella* Thompson outbreak data.

The paper is structured in the following way. In Section 2, we present the outbreak data motivating the work. In Section 3, we present the Bayesian variable selection method and its implementation. In Section 4, we present the data analysis results of the implemented method on the *Salmonella* Thompson case study data, compare the Bayesian analysis with the standard and Lasso logistic regression approaches and present the results of the sensitivity analysis. In Section 5, we discuss the method and results. In Section 6, we summarize the paper and discuss the impact of our methodology on epidemiological practice.

2 Data

We motivate our methodological developments by data obtained from a series of case-control studies performed by the Dutch National Institute for Public Health and the Environment (RIVM) as part of the outbreak

investigation of a large nationwide *Salmonella* Thompson outbreak.[3] The case-control studies ran from 16 August 2012 to 28 September 2012, when smoked salmon was identified as the source. During the study, various potential sources were identified by the investigation, namely minced meat, ready-to-eat raw vegetables, ice cream and finally smoked fish.

During the outbreak investigation, the food-consumption questionnaire was updated. The first version of the questionnaire was a very broad trawling questionnaire containing 178 items. In the second version, ambiguous products and those with a very low consumption frequency were removed from the questionnaire. In further updates of the questionnaire, respondents were asked to give more details about specific food products. All the food products which were removed from the first version of the questionnaire, were also removed from the dataset. Only food products that were available for all versions of the questionnaire have been included in the dataset. This results in 108 covariates, which include age, gender, 95 food products and 11 supermarket covariates. Age is a continuous covariate and standardized in the analysis. All the other covariates are binary-valued.

The 95 food product covariates indicate whether a person did (1) or did not (0) eat that product. The 11 supermarket covariates indicate whether a person does (1) or does not (0) buy most of their groceries at that shop. Food covariates that were not filled in are assumed to be zero. This is a reasonable assumption because it is often the case that respondents only mark the food products that they have consumed. Respondents also had the opportunity to respond with ''maybe''. In the analysis in 2012, these products were assumed to be consumed for persons with a ''maybe'' answer and set to 1, thereby possibly over-estimating food consumption. In our dataset, we instead deal with these covariates as being missing. Following this definition of missing values, the percentage of missing covariates per respondent is up to 40% for the cases and 67% for the controls.

The case-control study was designed as an individually matched case-control study. Because of expected non-response among the controls and to ensure at least one control per case, four controls per case were sampled from the general Dutch population from the same or neighbouring municipality with similar age and same gender.[3] Both cases and controls got their questionnaire sent by mail. The final dataset has 302 observations of which 109 are cases and 193 are controls. Age has a bimodal distribution with the highest frequencies in the age groups 10–19 and 60–69. Non-response in the controls was higher among males than among females. Female controls in the 60–69 age group had the highest response rate, almost 100%. For the other age groups and for the males, the response rate of the controls was close to 25%.

3 Methods

3.1 Logistic regression model with misclassification

Misclassified responses in a case-control study result when one cannot confirm whether cases and controls truly reflect the underlying true disease status. Let Y_i be, a possibly misclassified, observed disease status, and T_i a true disease status for person $i = 1, 2, \ldots, n$. Then the misclassification model is given by

$$\begin{aligned} P(Y_i|T_i, \boldsymbol{X_i}) &= P(Y_i|T_i) \\ P(Y_i = 1|T_i = 1) &= \text{Se} \\ P(Y_i = 0|T_i = 0) &= \text{Sp} \end{aligned} \tag{1}$$

The first line in equation (1) indicates that we assume nondifferential misclassification, i.e. the misclassification does not depend on the covariates $\boldsymbol{X_i}$. The last two equations define the sensitivity (Se) and specificity (Sp). We can then write the logistic regression model, corrected for misclassification, as

$$\begin{aligned} Y_i &\sim \text{Bernoulli}\ (\mu_i) \\ \mu_i &= (\text{Se})\pi_i + (1 - \text{Sp})(1 - \pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + \boldsymbol{x}_i^{'}\boldsymbol{\beta} \end{aligned} \tag{2}$$

where $\boldsymbol{x}_i' \in \mathbb{R}^{1\times p}$ denotes the ith row of the design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ denotes the vector of unknown regression coefficients and $\pi_i = P(T_i = 1|\boldsymbol{X_i})$ denotes the probability of having a true case. Note that the model in equation (2) simplifies to the classical logistic regression model for $\text{Se} = \text{Sp} = 1$ which corresponds to the situation of no misclassification, i.e. $T_i \equiv Y_i$ for $i = 1, 2, \ldots, n$. Equation (2) is widely used when one needs to correct for misclassification in the response in a logistic regression.[12–17]It can also be seen as a generalization of the logistic regression model. Rousseeuw and Christmann[18] illustrate this model as in Figure 1 and refer to it as the hidden

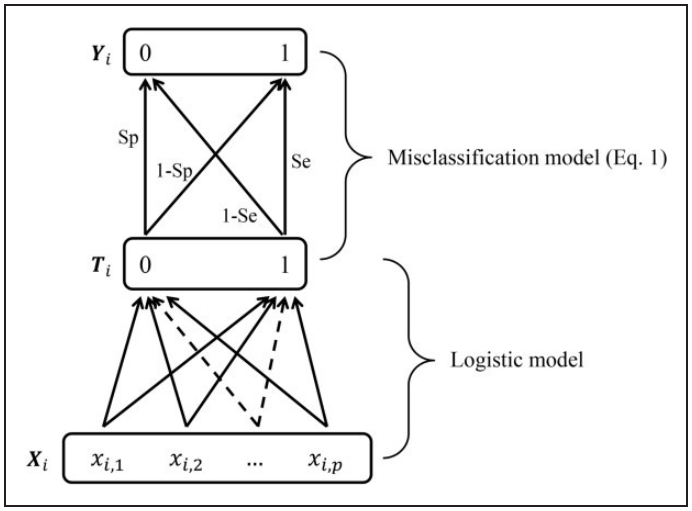

Figure 1. Generalization of logistic regression as illustrated by Rousseeuw and Christmann.[18]

logistic regression model because the true response T_i is hidden by the misclassification model in the top part of Figure 1.

A simplification of the model in equation (2) is obtained when either one or both of the misclassification errors do not occur, i.e. when either $\mathrm{Se} = 1$ or $\mathrm{Sp} = 1$ or both. For our data, we have $\mathrm{Sp} = 1$. A case only entered the dataset if it was twice laboratory-confirmed. We can, therefore, safely assume that no non-infected person incorrectly entered the dataset as a case, implying that the specificity is one. On the other hand, it is well known that food-related pathogen infections, such as *Salmonella* and *Campylobacter*, are often asymptomatic.[5,19–21] An infected person, therefore, may not have become ill, incorrectly entering the dataset as a control, implying $P(Y_i = 0|T_i = 1) > 0$ which leads to a sensitivity of less than one.

The model in equation (2) simplifies to $\mu_i = (\mathrm{Se})\pi_i$. This model, however, is still unidentifiable without extra information. This extra information can come from a validation dataset. In a Bayesian setting, one may, in addition, also use historical information to provide a prior distribution, $\mathrm{Beta}(a_1, b_1)$, on the sensitivity.

Although we are dealing with a matched case control study, we have not yet taken account of this data structure in our model. Moreover, we only consider a main effects model with linear terms.

3.2 Bayesian variable selection

To incorporate Bayesian variable selection in the model, we apply the SSVS procedure[11] in which a mixture prior on the parameters, β_j, consisting of one spike and one slab Gaussian component, is constructed (see Figure 2). The variance of the spike component is given by $\tau^2 > 0$ and the variance of the slab component is given by $c^2\tau^2 > 0$. The mathematical formulation of the SSVS prior for $j = 1, 2, \ldots, , p$ is given by

$$\beta_j\,|\tau^2, c^2 \sim \gamma_j \mathrm{N}(0, \tau^2 c^2) + (1 - \gamma_j)\mathrm{N}(0, \tau^2) \tag{3}$$

$$\gamma_j\,|\omega_j \sim \text{Bernoulli}\ (\omega_j) \tag{4}$$

$$\omega_j \sim \text{Beta}\ (a_{j,0}, b_{j,0}) \tag{5}$$

where γ_j is the indicator variable for inclusion of β_j into the model with ω_j the inclusion probability of the jth covariate.

The choice of the parameters τ and c can be guided by noting that the spike and slab components intersect at $\epsilon = \tau\sqrt{2\log(c)c^2/(c^2 - 1)}$. The point ϵ can be seen as a threshold for ''practical significance'' in that all coefficients $\beta_j \in [-\epsilon, \epsilon]$ can be interpreted as zero.[22] For a fixed c, the standard deviation τ can be selected to reflect our chosen value of practical significance. The posterior choice of covariates can then be based on the posterior probability of obtaining a significantly large β_j, i.e. the posterior inclusion probability $P(\beta_j \notin [-\epsilon, \epsilon]|\mathrm{Data})$. In the context of food-borne disease outbreaks, due to the encoding of the exposure, we are only interested in positive regression coefficients and consequently in the one-sided posterior inclusion probability, $P(\beta_j > \epsilon|\mathrm{Data})$.

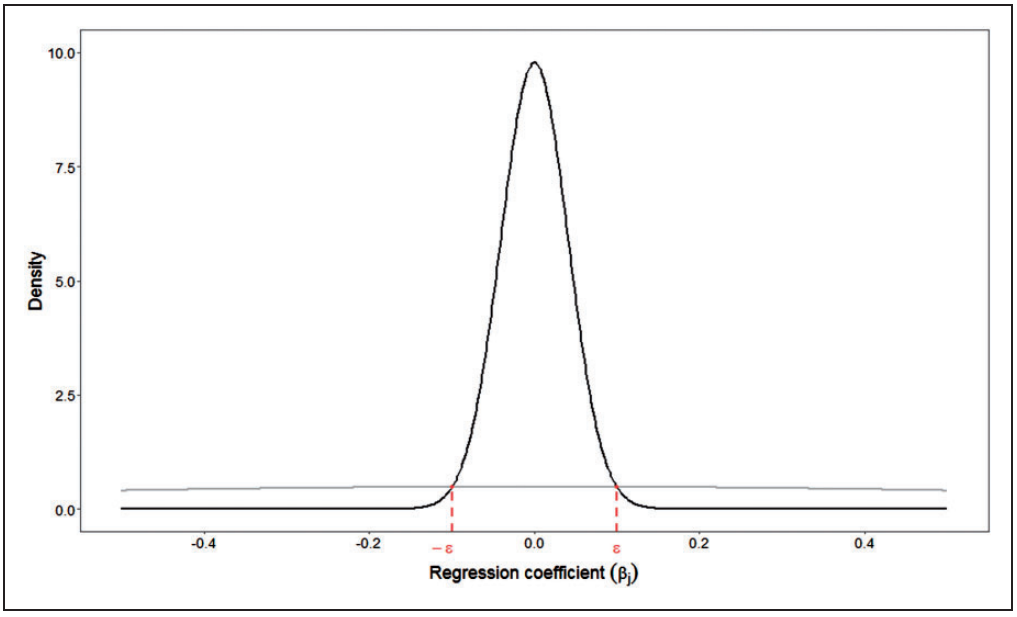

Figure 2. Spike (black curve) and slab (grey curve) prior distribution as used in the stochastic search variable selection procedure. ϵ indicates the threshold for practical significance.

The parameters $a_{j,0}$ and $b_{j,0}$ in equation (5) are chosen to reflect prior knowledge about the probability that a covariate should be in the model. As this may be a somewhat abstract exercise, we illustrate the mathematical relationship between the choice of $a_{j,0}$ and $b_{j,0}$ and the prior inclusion probability $P(\beta_j > \epsilon)$. From equations (3) to (5), it is possible to derive the prior marginal distribution of β_j, namely $\beta_j \sim a_{j,0}/(a_{j,0}+b_{j,0})N(0,\tau^2c^2)+ b_{j,0}/(a_{j,0}+b_{j,0})N(0,\tau^2)$. Large $b_{j,0}$ and small $a_{j,0}$ values result in prior inclusion probabilities close to zero and small $b_{j,0}$ values and large $a_{j,0}$ values result in prior inclusion probabilities close to 0.5. For similar $a_{j,0}$ and $b_{j,0}$ values, the one-sided prior inclusion probability is around 0.25 with increasing values of $a_{j,0}$ and $b_{j,0}$ resulting in more informative priors. In this paper, $a_{j,0}$ and $b_{j,0}$ are currently taken equal for all j covariates.

Alternatively, one can specify $\omega_j = 0.5$ which substantially reduces computational time and often provides sensible results.[23] In this context, ω_j denotes the prior fraction of covariates in the model with low values favouring parsimonious models.

3.3 Missing covariates

In a standard Bayesian setting with complete data, we are interested in the posterior distribution $p(\boldsymbol{\theta}_{Y|X}|\, y_i, \boldsymbol{x}_i)$, where $\boldsymbol{\theta}_{Y|X}$ is the vector of parameters associated with the likelihood of the response model (for example (Se, $\beta_0, \boldsymbol{\beta}'$)$'$ in equation (2)). In a missing data setting, however, $\boldsymbol{X}$ consists of two parts, the completely observed variables, $\boldsymbol{X}_{\text{obs}} \in \mathbb{R}^{n\times q}$, and the incompletely observed variables, $\boldsymbol{X}_{\text{mis}} \in \mathbb{R}^{n\times r}$. The relevant posterior distribution then becomes $p(\boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}_X, \boldsymbol{x}_{i,\text{mis}}|\, y_i, \boldsymbol{x}_{i,\text{obs}})$ with $\boldsymbol{x}_{i,\text{mis}} = (x_{i,\text{mis}_1}, x_{i,\text{mis}_2}, \ldots, , x_{i,\text{mis}_r})'$, $\boldsymbol{x}_{i,\text{obs}} = (x_{i,\text{obs}_1}, x_{i,\text{obs}_2}, \ldots, , x_{i,\text{obs}_q})'$ and where $\boldsymbol{\theta}_X$ denotes the parameters associated with the likelihood of the incompletely observed variables, $\boldsymbol{X}_{\text{mis}}$. This posterior can be written as

$$p(\boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}_X, \boldsymbol{x}_{i,\text{mis}}|\, y_i, \boldsymbol{x}_{i,\text{obs}}) \propto p(\, y_i|\boldsymbol{x}_{i,\text{mis}}, \boldsymbol{x}_{i,\text{obs}}, \boldsymbol{\theta}_{Y|X})\, p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}, \boldsymbol{\theta}_X)\pi(\boldsymbol{\theta}_{Y|X})\pi(\boldsymbol{\theta}_X)$$

where $\pi(\boldsymbol{\theta}_{Y|X})$ and $\pi(\boldsymbol{\theta}_X)$ denote prior distributions.[24] The joint likelihood of the missing covariates can conveniently be written as the product of conditional distributions

$$p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}, \boldsymbol{\theta}_X) = p(x_{i,\text{mis}_1}|\boldsymbol{x}_{i,\text{obs}}, \boldsymbol{\theta}_{X_1}) \prod_{j=2}^{r} p(x_{i,\text{mis}_j}|x_{i,\text{mis}_1}, \ldots, , x_{i,\text{mis}_{(j-1)}}, \boldsymbol{x}_{i,\text{obs}}, \boldsymbol{\theta}_{X_j}) \tag{6}$$

with $\boldsymbol{\theta}_X = (\boldsymbol{\theta}_{X_1}, \boldsymbol{\theta}_{X_2}, \ldots, , \boldsymbol{\theta}_{X_r})$.[24,25] We assume a model-based approach here in which each of the probability distribution functions in equation (6) is chosen from the exponential family according to the type of the

respective covariate with the dependence on previous variables modelled by a generalized linear model with regression coefficients $\boldsymbol{\theta}_{X_j} = (\alpha_{0,j}, \alpha_{1,j}, \alpha_{2,j} \ldots \alpha_{j-1,j})'$. In our application, because all covariates are binary, we assume a Bernoulli response with a logistic regression model.

Some words on the ordering of the imputation models in equation (6) are in order here. It is not obvious what the order of the imputation models should be[26] and this may influence the results. In the case of continuous and categorical missing covariates, Chen and Ibrahim[27] suggest to condition the categorical imputation models on the continuous covariates first. Erler et al.[24] ordered the imputation models according to the number of missing values, starting with the covariate with the least missing values, suggesting a possible gain in computational time. It has been shown, however, that the sequential specification as in equation (6) is quite robust against changes in the ordering.[27,28] In our analysis, we used the covariates in the order as they appeared in the dataset.

In the case of many covariates, as in our data, it is reasonable to assume that some of the parameters $(\alpha_{0,j}, \alpha_{1,j}, \alpha_{2,j} \ldots \alpha_{j-1,j})'$ are zero due to sparse relationships among the covariates. Similar to the variable selection of the response model (equation (3)), we perform variable selection in each of the conditional regression models of the covariate probability model in equation (6). This variable selection is implemented by providing not only the regression parameters of the response model $(\beta_1, \beta_2, \ldots, , \beta_j)'$ with the spike and slab prior distribution (equations (3) to (5)), but also those of the covariate models $(\alpha_{0,j}, \alpha_{1,j}, \alpha_{2,j} \ldots \alpha_{j-1,j})'$. The resulting two-level variable selection model was developed by Mitra and Dunson.[29]

3.4 Prior specification

The variance parameters of the spike and slab prior distribution need to be specified. On the basis of expert knowledge, large β's are very unlikely in practice. Moreover, in our experience, allowing large values of β a priori may hamper the convergence of the MCMC algorithm. With this in mind and choosing a practical significance level of $\epsilon = 0.05$, and $c = 100$, we obtain $\tau = 0.0165$ which results in a slab distribution of $\mathrm{N}(0, 1.65^2)$. This slab distribution results in a prior median odds ratio of one with 2.5th and 97.5th percentile given by 0.04 and 25.25, respectively.

In order to make the model identifiable, an informative prior for the sensitivity is required. Based on some preliminary expert knowledge about the sensitivity, we used a $\mathrm{Se} \sim \mathrm{Beta}(33, 4)$ prior which assumes a median sensitivity of 90% and 5th percentile of 80%.

The prior distribution for the ω_j's (equation (5)) was set to $\omega_j \sim \mathrm{Beta}(1, 2)$. This results in a one-sided prior inclusion probability of $P(\beta_j > \epsilon) = 0.16$. The $\mathrm{Beta}(1, 2)$ distribution is a positively skewed distribution giving more weight to small probabilities. The distribution, therefore, favours more parsimonious models among those covariates available from the questionnaires. The ω_j's for the variable selection of the covariate models were also given a $\mathrm{Beta}(1, 2)$ distribution.

The remaining parameters, i.e. intercept terms for response and covariates models as well as the regression coefficients for age and gender, were given a diffuse normal prior distribution, $\mathrm{N}(0, 1000)$.

3.5 Implementation

Our final model combines the two-level SSVS procedure with the likelihood of the logistic regression model from Section 3.1. Our model is easily implemented in the R Software[30] using JAGS[31] for implementing the MCMC sampling. The model statement for the likelihood and the SSVS priors is given in Supplemental Material.

We ran five chains with a burn-in of 1000 iterations and then a further 4000 iterations per chain, resulting in a posterior sample of size 20000. Trace plots of the inclusion probabilities, ω_j's, and the sensitivity, Se, were used to assess mixing. Visual inspection of trace plots indicated good mixing.

3.6 Model performance

We study the performance of our model in two ways: (i) how our model performs during the outbreak (Figure 3) and (ii) how the different parts of the model compare to standard logistic regression and Lasso logistic regression (Figure 4).

i. During the actual outbreak in 2012, the cases and controls entered the dataset as the questionnaires were returned. To mimic this dynamic, we sorted our dataset according to the return date of the questionnaires. This return date was constructed from two existing variables, namely questionnaire fill-in date and

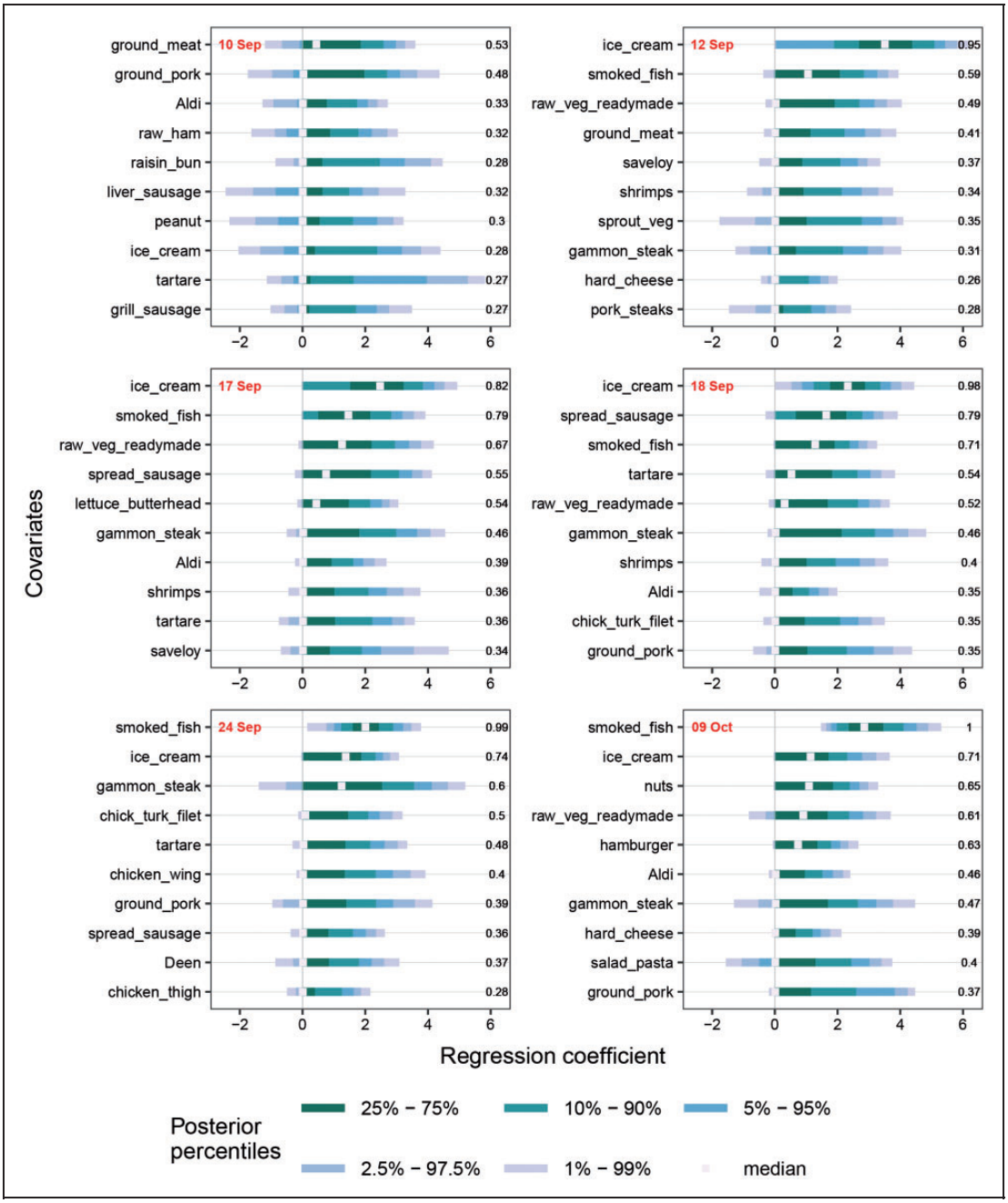

Figure 3. Posterior percentiles of regression coefficients and corresponding one-sided posterior inclusion probabilities, $P(\beta_j > 0.05|\text{Data})$, in the analysis of subsets of the *Salmonella* Thompson data mimicking the available data at different time points during the outbreak.

questionnaire return date. For those observations with a missing return date, the date was taken to be three working days after the fill-in date which was the average duration of questionnaire return for the questionnaires for which we had a return date and fill-in date. If the fill-in date was also missing, the return date was taken to be two weeks after the questionnaire was sent out, which was determined in consultation with the epidemiologist who worked the *Salmonella* outbreak case in 2012. We fit our model to various subsets of the data which mimic the available data at a certain date during the outbreak. The posterior distributions of the regression coefficients of the top ten covariates, sorted according to their posterior median, are plotted (Figure 3). The distributions are visualized by plotting various percentiles

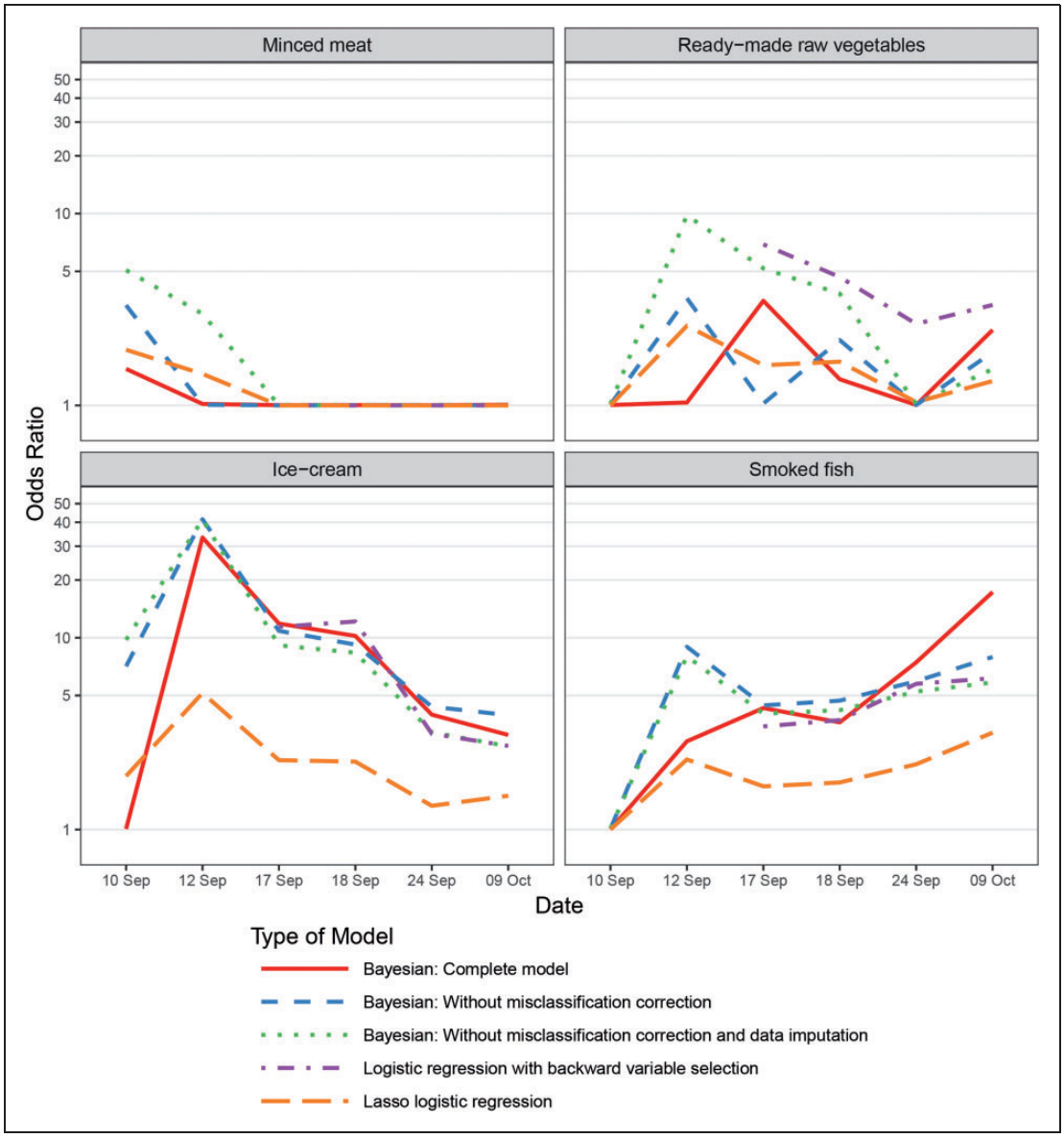

Figure 4. Odds ratios plotted over time comparing three Bayesian methods with standard and Lasso logistic regression for four potential sources of the Dutch *Salmonella* Thompson 2012 outbreak.

of the distribution. The posterior distributions indicate our updated believe about the value of the regression coefficient and consequently the amount of association with the response: the further the distribution is removed from zero, the larger the association. We also include the posterior inclusion probabilities $P(\beta_j > \epsilon|\text{Data}) = P(\beta_j > 0.05|\text{Data})$ for each covariate. The posterior inclusion probability expresses how often a covariate appears in the model and is a transparent way of expressing the uncertainty surrounding the effect of a covariate on the outcome.[32] This probability reflects our updated belief that a particular food culprit might be a relevant culprit as compared with the prior probability of $P(\beta_j > \epsilon) = 0.16$. One can categorize these posterior probabilities as: ≥ 0.50, ≥ 0.75, ≥ 0.95 and ≥ 0.99, corresponding to weak, positive, strong and very strong evidence for an association with the response.[32,33]

ii. Our model consists of three distinct parts, namely Bayesian variable selection, misclassification correction and missing value imputation. Each of these parts contributes to the overall performance of our model. We study this contribution by analysing our data using three different analyses: (i) Bayesian variable selection, missing data imputation and misclassification correction, (ii) Bayesian variable selection, missing data imputation and

(iii) Bayesian variable selection only. The last analysis is run on the exact same data as the standard and Lasso logistic regression analyses, with the products filled in as ''maybe'' set to ''consumed''.

For the standard logistic regression analysis, we used the classical variable selection procedures as mentioned in Section 1. First we fitted a univariable logistic model for each covariate from which we selected the covariates with a p-value less than 0.2. This selection of covariates was then entered into a multivariable logistic model on which we performed backward variable selection based on the AIC.

For the Lasso approach, we fitted a Lasso logistic regression using the glmnet package[34] in R. With the Lasso approach, the likelihood function is penalized for the size of the regression coefficients, thereby shrinking the regression coefficients towards zero, i.e. the odds ratios towards one. This penalization allows us to obtain maximum likelihood estimates, even in small samples when standard logistic regression fails. The penalization parameter λ is estimated by cross-validation, choosing the value of λ which minimizes the mean cross-validated error.

The three Bayesian methods and the standard and Lasso logistic regression approach are compared by plotting the odds ratios of the five models over time (Figure 4). For the Bayesian methods, we plot the posterior median odds ratio, for the standard logistic regression method, we plot the odds ratios obtained from the final multivariable model and for the Lasso approach, we plot the odds ratios obtained from the model with the optimal cross-validated λ as described above.

4 Results

4.1 Data analysis

In this section, we analyse the Dutch *Salmonella* Thompson data using our proposed model. We analysed five subsets of the data with our model to retrospectively determine when our model would have identified the smoked fish as a probable source. The five subsets contained data up to 10 September, 12 September, 17 September, 18 September and 24 September. The posterior distributions of the regression coefficients and the posterior inclusion probabilities for the five subsets and the complete dataset (9 October) are illustrated in Figure 3.

We see that on 10 September, although ground meat has a slightly elevated posterior median value, there is very little evidence in the data for any of the covariates to be included in the model. The posterior inclusion probabilities are all less than 50%, except for ground meat (53%) for which there is only very weak evidence for an association with the response. This may partly be due to the still very limited number of controls present in the data at this point (41 cases and 20 controls). Between 10 and 12 September, the dataset increases to 115 observations with 42 cases and 73 controls. Note the sudden increase in the number of controls. And now the data contain enough information for the variable selection procedure to provide some evidence for including covariates. The posterior inclusion probability for ice-cream is 0.95, providing strong evidence for an association with the response, making it a possible source. During the actual outbreak, ice-cream was investigated, but it did not turn out to be the source. Also, smoked fish has a relatively large posterior median compared to the other covariates, but its posterior inclusion probability (0.59) still only provides weak evidence.

Moving on in time to 17 and 18 September, more covariates start to have a positive posterior median. Ice-cream is still a likely suspect on 18 September with a posterior distribution detached from zero and a posterior inclusion probability (0.96) providing strong evidence. Arriving now at 24 September, we see a change in the variable selection. By now we have 74 cases and 174 controls. Smoked fish has now moved to the top with a high posterior median and a posterior inclusion probability (0.99) indicating very strong evidence for an association with the response. As mentioned in Section 2, smoked fish, more specifically smoked salmon, was indeed found to be the source of the outbreak after laboratory confirmation.

In Figure 4, we compare the results of the three Bayesian analyses (variable selection, misclassification correction and missing value imputation) with one another and with the standard and Lasso logistic regression analyses. We plot the odds ratios for the four food products that came up as likely suspects during the outbreak analysis in 2012, namely minced meat, read-made raw vegetables, ice-cream and smoked fish.[3]

For the standard logistic regression approach of 10 September, the number of covariates was, even after the univariable preselection, still too large to be able to fit a multivariable model. Thus, no odds ratios are available for this date. Also, for the 12 September analysis, we could not fit a multivariable model. Although the estimation algorithm did converge, some of the odds ratios were unrealistically large (≥ 1000). This may be due to a specific problem such as partial or complete separation.[35,36] In general, it is a sign that our data contain too little information to give reliable maximum likelihood estimates of the regression coefficients.

For the Bayesian analysis, we first consider the analysis which is most comparable to the standard logistic regression analysis, namely the model in which we only have the Bayesian variable selection part (dotted line in Figure 4). We see that for the food products that were not contaminated, the Bayesian odds ratios were consistently smaller than those of the frequentist analysis. Because these food products were not the contaminated food products, we want their odds ratios to be as close to one as possible. For the smoked fish, the Bayesian model performs similarly to the frequentist analysis, resulting in similar odds ratios.

In the next comparison, we add the missing data imputation (dashed line in Figure 4). For ice-cream, this model performs similarly to the previous Bayesian model. For minced meat and ready-made vegetables, it performs better in that the odds ratios are even lower. Also for the smoked fish, the Bayesian variable selection with missing data imputation performs better than both the frequentist and the Bayesian variable selection only model, resulting in higher odds ratios.

Finally, we arrive at the complete Bayesian variable selection model (solid line in Figure 4) in which we, in addition to the missing data imputation, also include the misclassification correction. Here we clearly see a large gain in performance for smoked fish: our complete Bayesian model has an odds ratio of up to three times as large as the frequentist odds ratio. Comparing the graphs for ice-cream and smoked fish, we also note that on 24 September, the Bayesian models and especially the complete model return a higher odds ratio for smoked fish than for ice-cream. Although this is also the case for the frequentist analysis in Figure 4, in the complete multivariable model (not shown), there were still covariates with higher odds ratios than smoked fish. Only after 29 September, did the frequentist multivariable model estimate smoked fish to have the highest odds ratio. Our Bayesian model, therefore, identifies smoked fish as a likely suspect earlier in the outbreak than the standard frequentist method.

For the Lasso approach, we clearly see the shrinkage effect for all four food products and all time points. Adding shrinkage to the logistic regression ensures that we get estimates even with little data. This is an advantage over the standard logistic regression for which we could not fit a multivariable model in the beginning. Compared to the Bayesian analysis, however, the Lasso approach does not perform as well. First, due to the shrinkage, the odds ratios for the various food products are very similar, making it difficult to identify the most probable suspects as none of them stand out. Second, the Lasso approach shrinks both the weak and the large effects. This is seen by considering the slope of the smoked fish odds ratios. As the evidence in the data increased for smoked fish, the full Bayesian model shows a steep increase in odds ratio reflecting this evidence. For the Lasso approach, however, there is only a slight increase as the odds ratios are kept small due to the shrinkage. In conclusion, although the shrinkage of the Lasso approach is an advantage compared to the standard logistic regression analysis, it has the unwanted side effect of also shrinking the large effects, making easy identification of likely suspects more difficult than in the Bayesian case.

We, therefore, see that in comparing our Bayesian variable selection model with the standard logistic backward variable selection model, our model is able to come up with the correct food product earlier in the outbreak, i.e. based on less data.

4.2 Sensitivity analysis

In order to investigate the effect of our informed prior distributions, we performed a small sensitivity analysis. The values of the spike and slab parameters, ϵ, τ and c, have been found to influence the model results in Bayesian variable selection. The performance of the variable selection (in terms of the number of correctly and falsely selected variables) increases with decreasing slab variance.[37] This is due to the fact that for lower slab variance, smaller effects are penalized less than larger effects.[37] The marginal posterior summary statistics, such as the posterior median or mean, of the regression coefficients, however, tend to remain stable across different values of the parameters.[22]

We found that smaller values of ϵ have a larger shrinkage effect resulting in the posterior distributions of the regression coefficients to be narrower and shrunk towards zero. We also considered the effect of the prior distribution of the inclusion probabilities on the results. Using a uniform prior distribution allows more variables into the model by resulting in higher posterior medians for more of the regression coefficients. On the other hand, a strong positively skewed distribution such as a Beta$(1, 10)$ prior distribution forced a very parsimonious model with only one regression coefficient, namely smoked fish, having a positive posterior median. Finally, we considered the prior distribution for the sensitivity parameter. We allowed a slighter wider distribution, but with the same median of 0.9, namely Beta$(16, 2)$. This prior distribution resulted in very wide posterior distributions with longer negative tails than for the case of SE $\sim$ Beta$(33, 4)$. In all of the above scenario's, the regression coefficient of smoked fish still had the highest posterior median. There was some variation in which variables ended up in the top ten of highest regression coefficient values.

5 Discussion

This paper is an attempt at improving the current analytical methods for source identification in the epidemiological investigation of food-borne disease outbreaks. Identifying the contaminated food product among the many food products from a trawling questionnaire is very much like a criminal investigation. One needs to keep an open mind so as not to exclude any possible suspects, but simultaneously one needs enough focus and detail to find the culprit as quickly as possible.

This balance between open mindedness and focus is especially apparent when specifying the prior distributions for the inclusions probabilities, ω_j's, of our model. In our analysis of the data, all ω_j's were given the same prior distribution, namely Beta(1, 2). This is a relatively uninformative distribution giving slightly more weight to smaller inclusion probabilities thereby enforcing a preference for a more parsimonious model. Applying a non-informative prior to all the covariates is very much an open-mind approach. It will, however, not contribute much to finding a likely suspect early in the outbreak. For this we need to focus on likely suspects by providing the prior distributions with external information about such likely suspects. The Bayesian variable selection model lends itself especially well to include information on individual or groups of inclusion probabilities. Literature, historical outbreak data and expert knowledge are rich information sources that can be used to inform the prior distributions for the inclusion probabilities. Given the pathogen, which during a food-borne disease outbreak is usually known, certain food products or groups of food products are more likely suspects than others. This higher likelihood can be quantified by a distribution and is used together with the data to guide the variable selection process. As a word of caution, however, when informing prior distributions, one should keep in mind that we are dealing with an outbreak which usually occurs due to some unusual circumstance, where some unexpected food product is contaminated.[3] The *Salmonella* Thompson 2012 outbreak is such an example. *Salmonella* is known to occur to a large extent in poultry, eggs, pigs and bovine and not in fish.[38,39] Also, very few *Salmonella* outbreaks associated with fish have been reported in literature.[40] Indeed, during the 2012 outbreak, the *Salmonella* was *on* the fish due to reusable dishes used in the processing line and not *in* the fish. If the prior information of low *Salmonella* occurance in fish had been incorporated in the analysis, it would possibly have taken even longer to identify the smoked fish as the culprit. More data are then needed to counteract the ''wrong'' information in the prior distribution.

Although it is well-known that *Salmonella* infections are often asymptomatic, literature on estimating the incidence of asymptomatic infections is very sparse. Jertborn et al.[41] studied asymptomatic infections among a group of Swedish travelers. In this group, the sensitivity was $\text{Se} = P(Y = 1|T = 1) = 0.41$. This may suggest that our sensitivity of 0.9 is way too large. This study, however, is hardly comparable to our situation. There have also been attempts to calculate so-called ''multipliers'' which estimate the *Salmonella* infections in the population from known culture-confirmed cases.[19,42–44] It is, however, not easy to isolate the asymptomatic infections from these calculations. Constructing a scientifically substantiated informative prior distribution for the sensitivity requires an even bigger literature review combined with expert elicitation and is outside the scope of this paper.

The missing data imputation method described in Section 3.3 may look similar to the chained equations approach as used in multiple imputation by chained equations (MICE), also known as regression switching.[45,46] MICE, however, does not explicitly specify a joint distribution for the covariates. It considers each variable separately, imputing it using all the other variables as predictors. This may give convergence problems if the individual models are not compatible with each other or with a multivariate distribution.[29,47] In addition, MICE is not embedded in a Bayesian setting, making it incompatible within a full Bayesian variable selection procedure.

As mentioned in the Introduction, the current methodology in analysing case-control studies for source identification in food-borne disease outbreaks includes a combination of univariable and multivariable logistic regression together with backward and forward selection. This is an ill-defined and to some extent ad hoc way in searching for contaminated food products. Some of the problems associated with this method include well known model selection problems with determining strength of effects based on p-values, number of variables close to or even exceeding the number of observations and no standard or well-defined way of dealing with the large percentage of missing values.

In a study comparing stepwise and backward variable selection methods with Bayesian model averaging in case-control studies, Viallefont et al.[32] found that p-values are no longer reliable in determining risk factors after stepwise or backward variable selection has been applied. In a simulation study, the authors showed that of those variables whose regression coefficient had a p-value between 0.01 and 0.05 (typically assumed to be significant in epidemiological analyses) after stepwise variable selection, only 49% were actual risk factors by design. This was 57% when backward variable selection was applied. These proportions are well below one minus the nominal significance level. This concern about stepwise variable selection and p-values is also discussed by Harrell.[10] With Bayesian variable selection, we do not have these problems because there is no reliance on p-values.

Fitting a multivariable model using standard logistic regression is problematic early in an outbreak. At this point, data are still scarce resulting in more covariates than observations ($p > n$). In this case, it is not possible to obtain maximum likelihood estimators for the regression coefficients as was seen in Section 4.1. Both the Lasso logistic regression and the Bayesian analyses do not have this problem, because they introduce some form of constraint on the size of the regression coefficients.

Furthermore, using a Bayesian model allows automation of the analyses as new data enters the case-control study. Currently, when new case-control data enter the study, the analyses have to be rerun over the cumulative dataset at that point. In a Bayesian setting, however, the model can be transformed into a state-of-the-art learning algorithm whereby the posterior output of an earlier analysis is used as prior input for the analysis with the new data. These ''new'' priors can also be combined with other new insights obtained during the outbreak investigation, thereby reflecting the dynamics of the investigation. Such a learning algorithm has the potential to greatly improve the efficiency of the identification process. Further research is required to obtain full advantage of such dynamic modelling.

In the context of model choice, we have not yet made full use of all the modelling options. First, we have not extensively examined all analysis options for matched data. In this context, one could analyse the data using a conditional logistic regression model. Alternatively, one might use a random effects model, inserting a random effect for the groups of matched case-controls. In the extreme case, where one has 1-to-1 matching, one should use conditional logistic regression, as using unconditional logistic regression is biased and produces inflated odds ratio estimates.[48] This, however, is not the case in our case study. In fact, using an unconditional analysis on matched data may even increase the precision of the estimates. This is due to the fact that, in an unconditional analysis, cases with identical values for their matched variables and cases without corresponding controls can still be included in the analysis.[48] It is, however, still necessary to control for the matched variables, as we do, by including age and gender into our model. If we were to have used conditional logistic regression, we would expect less precision in the estimation, resulting in more difficult detection of significant variables. Second, the choice of model terms can still be optimized. One might want to include interaction terms and possibly non-linear terms for, e.g., age. Finally, we have not yet dealt with the misclassification in the covariates. Such misclassification can be dealt with by a misclassification model.[49] Dealing properly with the matched design of the data, the choice of model terms and misclassification in the covariates may lead to substantial improvements of the model. This is a topic for future research.

SSVS is a standard Bayesian variable selection method and is over 20 years old. It has many advantages over other Bayesian variable selection methods. The most obvious alternatives are model selection procedures in which a model choice criteria is calculated, such as the deviance information criterion (DIC).[50] These, however, are only feasible to use when the number of possible models is limited.[51] With over 100 possible variables in our case study, there are just too many possibilities. Moreover, for model selection procedures, the possible models are usually chosen from some theoretical basis to limit the choice of models. In our case, we do not have that. All combinations of variables are possible and the SSVS method can handle that especially well. Reversible Jump algorithms,[52] another method for model selection, are difficult to implement in practice, partially due to the fact that the dimension of the model in each run changes as different number of variables may be selected in each run. SSVS, in contrast to the other methods, allows for simple and complex models which gives a lot of flexibility. Computational challenges are minimized because the dimension of the model stays the same during the whole analysis and the actual variable selection is only performed during the interpretation of the results after the algorithm has converged. This also provides flexibility in the selection process by allowing the researcher to decide on the best selection criteria relevant to the application.

The main limitation of our model is the computation time. The large computational burden comes from the MCMC sampling of the two-level variable selection process. In our application, the MCMC sampling took approximately 7 h with parallel computing on five cores. Although this is a long time to wait when one is developing and studying the methodology and performing sensitivity analyses, it is less of a problem in practice. During an outbreak, new data may enter the study on a daily basis and the model can be run overnight to have the new results ready each morning.

6 Summary and impact on epidemiological practice

In this paper, we developed a method that deals with the problems of variable selection, missing covariates and misclassified responses in the context of source identification in the investigation of food-borne disease outbreaks. We have shown how a Bayesian analysis allows a relatively easy implementation of these concepts in the

re-analysis of the Dutch *Salmonella* Thompson 2012 outbreak data. Moreover, the Bayesian analysis performed better than both the standard and Lasso logistic regression models in identifying the responsible food source.

The method presented in this paper constitutes a first attempt at formalizing the methodology necessary for the analytic part of food-borne disease outbreak investigations. Current procedures are very much ad hoc in nature. In the interest of public health and for the task of lowering the disease burden of food-borne disease outbreaks, outbreak investigations are desperately in need of sound statistical methodology – methodology which can not only deal with the many challenges in case-control studies, but also exploit the structures in the data and the dynamics of the outbreak in order to identify the contaminated food product as quickly as possible. The Bayesian variable selection method presented in this paper is an example of such methodology which can provide epidemiologists with a streamlined statistical tool to aid the outbreak investigations. It only seems natural that such more formal decision support tools become part of standard epidemiological practice.

Acknowledgements

The authors thank Agnetha Hofhuis from the National Institute for Public Health and the Environment for providing the data of the Dutch *Salmonella* Thompson 2012 outbreak. The authors thank Eelco Franz, Maarten Schipper and the reviewer for their cricital reading of the paper and their comments which improved the paper.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the National Institute for Public Health and the Environment (RIVM) and through their Strategic Programme (SPR) which contributes to solutions to societal challenges through interdisciplinary research and by supporting innovation and capacity building at RIVM.

Supplementary Material

Supplementary material is available for this article online.

References

1. Bernard H, Faber M, Wilking H, et al. Large multistate outbreak of norovirus gastroenteritis associated with frozen strawberries, Germany, 2012. *Euro Surveill* 2014; **19**: 20719.
2. Hald T, Aspinall W, Devleesschauwer B, et al. World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. *PloS One* 2016; **11**: e0145839.
3. Friesema I, de Jong A, Hofhuis A, et al. Large outbreak of Salmonella Thompson related to smoked salmon in the Netherlands, August to December 2012. *Euro Surveill* 2014; **19**: 20918.
4. Frank C, Werber D, Cramer JP, et al. Epidemic profile of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. *N Engl J Med* 2011; **365**: 1771–1780.
5. Dwyer DM, Strickler H, Goodman RA, et al. Use of case-control studies in outbreak investigations. *Epidemiol Rev* 1994; **16**: 109–123.
6. Buchholz U, Bernard H, Werber D, et al. German outbreak of Escherichia coli O104:H4 associated with sprouts. *N Engl J Med* 2011; **365**: 1763–1770.
7. Wacholder S, McLaughlin JK, Silverman DT, et al. Selection of controls in case-control studies: I. Principles. *Am J Epidemiol* 1992; **135**: 1019–1028.
8. Korn EL. Estimating the utility of matching in case-control studies. *J Chronic Dis* 1984; **37**: 765–772.
9. Hosmer DW, Lemeshow S and Sturdivant RX. *Applied logistic regression*, 3rd ed. Hoboken: John Wiley & Sons, 2013. [ISBN 9780470582473].
10. Harrell FE. *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd ed. New York: Springer, 2015. [ISBN 9783319194240].
11. George EI and McCulloch RE. Variable selection via Gibbs sampling. *J Am Stat Assoc* 1993; **88**: 881–889.

12. Ekholm A and Palmgren J. A model for a binary response with misclassifications. In: Gilchrist R (ed.) *Proceedings of the international conference on generalised linear models*. New York: Springer-Verlag, 1982, pp.128–143.
13. Gerlach R and Stamey J. Bayesian model selection for logistic regression with misclassified outcomes. *Stat Modell* 2007; **7**: 255–274.
14. Liu J, Gustafson P and Huo D. Bayesian adjustment for the misclassification in both dependent and independent variables with application to a breast cancer study. *Stat Med* 2016; **35**: 4252–4263.
15. Luan X, Pan W, Gerberich SG, et al. Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Stat Med* 2005; **24**: 2221–2234.
16. Magder LS and Hughes JP. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol* 1997; **146**: 195–203.
17. Roy S, Banerjee T and Maiti T. Measurement error model for misclassified binary responses. *Stat Med* 2005; **24**: 269–283.
18. Rousseeuw PJ and Christmann A. Robustness against separation and outliers in logistic regression. *Comput Stat Data Anal* 2003; **43**: 315–332.
19. Simonsen J, Strid M, Molbak K, et al. Sero-epidemiology as a tool to study the incidence of Salmonella infections in humans. *Epidemiol Infect* 2008; **136**: 895–902.
20. Simonsen J, Teunis PF, van Pelt W, et al. Usefulness of seroconversion rates for comparing infection pressures between countries. *Epidemiol Infect* 2011; **139**: 636–643.
21. Teunis PF, Falkenhorst G, Ang C, et al. Campylobacter seroconversion rates in selected countries in the European Union. *Epidemiol Infect* 2013; **141**: 2051–2057.
22. Lesaffre E and Lawson AB. *Bayesian biostatistics*. Chichester: John Wiley & Sons, 2012.
23. George EI and McCulloch RE. Approaches for Bayesian variable selection. *Stat Sin* 1997; **7**: 339–373.
24. Erler NS, Rizopoulos D, van Rosmalen J, et al. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Stat Med* 2016; **35**: 2955–2974.
25. Ibrahim JG, Chen MH and Lipsitz SR. Bayesian methods for generalized linear models with covariates missing at random. *Canad J Stat* 2002; **30**: 55–78.
26. Bartlett JW, Seaman SR, White IR, et al. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Meth Med Res* 2015; **24**: 462–487.
27. Chen MH and Ibrahim JG. Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* 2001; **57**: 43–52.
28. Zhu J and Raghunathan TE. Convergence properties of a sequential regression multiple imputation algorithm. *J Am Stat Assoc* 2015; **110**: 1112–1124.
29. Mitra R and Dunson D. Two-level stochastic search variable selection in GLMs with missing predictors. *Int J Biostat* 2010; **6**: 33.
30. R Core Team. R: a language and environment for statistical computing, www.r-project.org/ (2016, accessed 27 November 2017).
31. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik K, Leisch F and Zeileis A (eds) *Paper presented at 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria, 20–22 March 2003. ISSN 1609-395X, 2003.
32. Viallefont V, Raftery AE and Richardson S. Variable selection and Bayesian model averaging in case-control studies. *Stat Med* 2001; **20**: 3215–3230.
33. Kass RE and Raftery AE. Bayes factors. *J Am Stat Assoc* 1995; **90**: 773–795.
34. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Software* 2010; **33**: 1–22.
35. Albert A and Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984; **71**: 1–10.
36. Lesaffre E and Albert A. Partial separation in logistic discrimination. *J R Stat Soc B* 1989; **51**: 109–116.
37. Rockova V, Lesaffre E, Luime J, et al. Hierarchical Bayesian formulations for selecting variables in regression models. *Stat Med* 2012; **31**: 1221–1237.
38. EFSA and ECDC. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2014. *EFSA J* 2015; **13**: 4329.
39. Mølbak K, Simonsen J, Jørgensen CS, et al. Seroincidence of human infections with nontyphoid salmonella compared with data from public health surveillance and food animals in 13 European countries. *Food Safety* 2014; **59**: 1599–1606.
40. Heinitz ML, Ruble RD, Wagner DE, et al. Incidence of Salmonella in fish and seafood. *J Food Protect* 2000; **63**: 579–592.
41. Jertborn M, Haglind P, Iwarson S, et al. Estimation of symptomatic and asymptomatic Salmonella infections. *Scand J Infect Dis* 1990; **22**: 451–455.

42. Chalker RB and Blaser MJ. A review of human salmonellosis: III. Magnitude of salmonella infection in the United States. *Rev Infect Dis* 1988; **10**: 111–124.
43. Simonsen J, Mølbak K, Falkenhorst G, et al. Estimation of incidences of infectious diseases based on antibody measurements. *Stat Med* 2009; **28**: 1882–1895.
44. Voetsch AC, van Gilder TJ, Angula FJ, et al. FoodNet estimate of the burden of illness caused by nontyphoidal Salmonella infections in the United States. *Clin Infect Dis* 2004; **38**(Suppl. 3): S127–S134.
45. Van Buuren S, Boshuizen HC and Knook DL. Multiple imputation of missing blood presure covariates in survival analysis. *Stat Med* 1999; **18**: 681–694.
46. Van Buuren S and Oudshoorn CGM. Multivariate imputation by chained equations: MICE V1. 0 user's manual No. PG/VGZ/00.38. Technical report, TNO, 2000.
47. Horton NJ and Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 2007; **61**: 79–90.
48. Pearce N. Analysis of matched case-control studies. *Br Med J* 2016; **352**: i969.
49. Gustafson P. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Boca Raton: Chapman & Hall/CRC, 2004. [ISBN 1584883359].
50. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. *J R Stat Soc B* 2002; **64**: 583–639.
51. O'Hara RB and Sillanpää MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal* 2009; **4**: 85–118.
52. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**: 711–732.