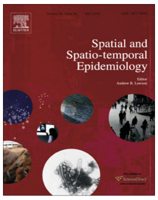
Spatial and Spatio-temporal Epidemiology 9 (2014) 1–12

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



Original Research

Spatial correlation in Bayesian logistic regression with misclassification

Kristine Bihrmann [a,*], Nils Toft [a], Søren Saxmose Nielsen [a], Annette Kjær Ersbøll [b]

[a] *Faculty of Medical and Health Sciences, University of Copenhagen, Grønnegårdsvej 8, DK-1870 Frederiksberg C, Denmark*
[b] *National Institute of Public Health, University of Southern Denmark, Øster Farimagsgade 5A, 2, DK-1353 Copenhagen K, Denmark*

ARTICLE INFO

Article history:
Received 7 January 2013
Revised 14 February 2014
Accepted 20 February 2014
Available online 1 March 2014

Keywords:
Bayesian logistic regression
Misclassification
Spatial correlation
Conditional autoregressive models
Slice sampling

ABSTRACT

Standard logistic regression assumes that the outcome is measured perfectly. In practice, this is often not the case, which could lead to biased estimates if not accounted for.

This study presents Bayesian logistic regression with adjustment for misclassification of the outcome applied to data with spatial correlation. The models assessed include a fixed effects model, an independent random effects model, and models with spatially correlated random effects modelled using conditional autoregressive prior distributions (ICAR and ICAR(ρ)). Performance of these models was evaluated in a simulation study.

Parameters were estimated by Markov Chain Monte Carlo methods, using slice sampling to improve convergence.

The results demonstrated that adjustment for misclassification must be included to produce unbiased regression estimates. With strong correlation the ICAR model performed best. With weak or moderate correlation the ICAR(ρ) performed best. With unknown spatial correlation the recommended model would be the ICAR(ρ), assuming convergence can be obtained.

1. Introduction

Logistic regression is often used in epidemiology for estimating the effect of observed covariates on a binary outcome. The standard model does, however, assume that the outcome is measured perfectly, i.e. sensitivity and specificity are both 1. In practice, this is often not the case. Many diagnostic tests for infectious diseases, for example, are not perfect which can result in a misclassified outcome. This could lead to biased estimates (Copeland et al., 1977) if not accounted for in the model.

Magder and Hughes (1997) introduced the use of an EM algorithm to incorporate information on misclassification in fixed effects logistic regression in a frequentist setting. In their approach, sensitivity and specificity were assumed fixed at predetermined values. This assumption was relaxed by e.g. Lyles et al. (2011) using validation-data based adjustment for misclassification. McInturff et al. (2004) presented a method to account for misclassification in a Bayesian setting. The Bayesian approach offers the possibility of including uncertainty in sensitivity and specificity through prior distributions. Here we will focus solely on a Bayesian approach.

If study subjects are clustered, e.g. animals are located in a herd or repeated measures are made on a subject, random effects can be included in the model to account for correlation within these clusters. Paulino et al. (2005) considered Bayesian logistic regression with correlated misclassified data and included independent random effects in the model. Random effects can also be included to account for spatial correlation in data. These random effects will be correlated according to a given spatial structure and will in this paper, as in many other studies,

[*] Corresponding author. Tel.: +45 65 50 77 36.
E-mail address: krbi@sund.ku.dk (K. Bihrmann).

http://dx.doi.org/10.1016/j.sste.2014.02.002
1877-5845/© 2014 Elsevier Ltd. All rights reserved.

be modelled by a conditional autoregressive (CAR) prior distribution. A simple CAR model was proposed by Besag et al. (1991) and expanded by e.g. Cressie (1993) and Leroux et al. (1999). However, accounting for misclassification in models including these spatially structured random effects has, to our knowledge, not been studied.

In this paper we present a Bayesian logistic regression model with adjustment for misclassification of the outcome and including spatially structured random effects modelled by CAR priors. The objectives were to (1) estimate parameters using Markov Chain Monte Carlo methods which posed some challenges in terms of convergence, (2) evaluate the effect of misclassification and compare the performance of two CAR priors as well as two models without spatial structure when applied to data with misclassification and spatial correlation among observations, and (3) study model fit, in particular with regards to inclusion of spatial structure. The objectives were addressed in a simulation study.

2. Statistical model

Consider a population of subjects clustered in known geographic locations, e.g. inhabitants in municipalities or animals within herds. All subjects are classified in two groups according to the outcome of a test for a given condition. Let Y_{ij} denote the test status (positive test = 1, negative test = 0) of subject: j, location: $i, j = 1, \ldots, n_i$, $i = 1, \ldots, n$. Then $Y_{ij} \sim \text{Bernoulli}\ (p_{ij})$, where $p_{ij} = \Pr$ (positive test), $j = 1, \ldots, n_i, i = 1, \ldots, n$. A traditional logistic regression model with random effect $\boldsymbol{V}^t = (V_1, \ldots, V_n)$ is given by

$$\text{logit}(p_{ij}) = a + \boldsymbol{bx} + V_i, \tag{1}$$

where a is the intercept, $\boldsymbol{b}$ is a vector of regression parameters, and $\boldsymbol{x}^t = (x_1, \ldots, x_m)$ is a set of covariates.

Given sensitivity *se* and specificity *sp* of the test, the probability of a positive test can be written as

$$p_{ij} = se \times \pi_{ij} + (1 - sp) \times (1 - \pi_{ij}), \tag{2}$$

where $\pi_{ij} = \Pr$ (condition truly present) (Rogan and Gladen, 1978). A logistic regression model adjusted for misclassification is then given by

$$logit(\pi_{ij}) = \alpha + \boldsymbol{\beta x} + U_i \tag{3}$$

The random effect $\boldsymbol{U}^{\,t} = (U_1, \ldots, U_n)$ is included to account for residual spatial correlation. The spatial correlation is modelled by an $n \times n$ matrix $\boldsymbol{W}$ describing the neighbourhood relations. Neighbours are correlated whereas non-neighbours are conditionally independent given all other random effects. In this study, neighbourhood relations are defined in terms of distance: locations within a predefined distance D of each other are considered neighbours and correlation is related to their inverse distance. Hence, the (r, s) element of W is given by

$$W_{\text{rs}} = \begin{cases} \frac{1}{d_{\text{rs}}} & \text{if } 0 < d_{\text{rs}} \leqslant D \\ 0 & \text{otherwise} \end{cases}$$

where d_{rs} is the distance between location r and s.

The random effect $\boldsymbol{U}$ is modelled by a CAR model for $U_i|\boldsymbol{U}_{-i}, i = 1, \ldots, n$, where $\boldsymbol{U}_{-i}$ denotes all the elements in $\boldsymbol{U}$ except U_i. A simple example of such a model is the intrinsic conditional autoregressive (ICAR) model proposed by Besag et al. (1991) and given by

$$U_i|\boldsymbol{U}_{-i} \sim N\left(\sum_{k=1}^{n} \frac{W_{\text{ik}}}{W_{i+}} U_k, \frac{\sigma^2}{W_{i+}}\right), \quad i = 1, \ldots, n, \tag{4}$$

where $W_{i+} = \sum_{k=1}^{n} W_{ik}$. Hence, the conditional mean of the random effect U_i is a weighted average of the neighbouring random effects and the precision depends on the amount of information supplied by neighbours (many neighbours located nearby means a large precision). The joint distribution corresponding to (4) is improper since it specifies only differences in U_i's. This can, however, be fixed by the constraint $\sum_{i=1}^{n} U_i = 0$. Another way to assure propriety is to include a correlation parameter ρ which leads to the ICAR(ρ) model

$$U_i|\boldsymbol{U}_{-i} \sim N\left(\rho \sum_{k=1}^{n} \frac{W_{\text{ik}}}{W_{i+}} U_k, \frac{\sigma^2}{W_{i+}}\right), \quad i = 1, \ldots, n, \tag{5}$$

suggested by e.g. Sun et al. (1999). The joint distribution corresponding to (5) is $\text{N}_\text{n}(0, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \sigma^2 \left(\widetilde{\mathbf{W}} - \rho W\right)^{-1} \tag{6}$$

with $\widetilde{\mathbf{W}} = \text{diag}(W_{1+}, \ldots, W_{n+})$. To ensure the covariance matrix $\boldsymbol{\Sigma}$ to be positive definite, the correlation parameter ρ is bounded by -1 and 1. Hence, the ICAR model (4) is just the special case assuming maximum correlation. A point of criticism of the ICAR(ρ) model is that even with no spatial correlation present ($\rho = 0$) the variance still depends on the neighbours. With no residual spatial correlation between clusters the random effects should, however, be modelled as independent and identically Normal distributed, i.e. $\boldsymbol{U} \sim N_n(0, \sigma^2 \boldsymbol{I})$ – in the following termed the random effects model. The logistic regression model could be further simplified by ignoring any clusters and omitting the random effect U in (1) and (3) – in the following termed the fixed effects model.

3. Simulation study

3.1. Motivating example: paratuberculosis

Paratuberculosis is a chronic infection in cattle caused by *Mycobacterium avium* subsp. *paratuberculosis* (MAP) (Harris and Barletta, 2001). A voluntary control programme on paratuberculosis is offered to Danish dairy farmers. On January 1, 2009, 28% of the herds participated in the programme (Bihrmann et al., 2012). Geographic location of these herds are given as UTM coordinates representing the largest building on the premises. Participating herds are tested for MAP four times a year using a milk antibody ELISA (ID-Screen®, ID-Vet, Montpellier, France). Tests are performed and recorded on individual cows. One date of testing within each herd tested between October 2008 and June 2009 was included in this study. A total of 194,465 cows in 1503 herds were tested. For more details on these data, please refer to Bihrmann et al. (2012).