Research Article

Statistics in Medicine

Received 22 June 2011, Accepted 8 March 2012 Published online 24 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5388

Point source modeling of matched case–control data with multiple disease subtypes

Shi Li,[a] Bhramar Mukherjee[a*†] and Stuart Batterman[b]

In this paper, we propose nonlinear distance-odds models investigating elevated odds around point sources of exposure, under a matched case-control design where there are subtypes within cases. We consider models analogous to the polychotomous logit models and adjacent-category logit models for categorical outcomes and extend them to the nonlinear distance-odds context. We consider multiple point sources as well as covariate adjustments. We evaluate maximum likelihood, profile likelihood, iteratively reweighted least squares, and a hierarchical Bayesian approach using Markov chain Monte Carlo techniques under these distance-odds models. We compare these methods using an extensive simulation study and show that with multiple parameters and a nonlinear model, Bayesian methods have advantages in terms of estimation stability, precision, and interpretation. We illustrate the methods by analyzing Medicaid claims data corresponding to the pediatric asthma population in Detroit, Michigan, from 2004 to 2006. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: asthma cases; conditional likelihood; disease subclassification; iteratively reweighted least square; Markov chain Monte Carlo; matched case–control; point source modeling

1. Introduction

In case–control designs, matching is commonly implemented to avoid bias due to potential confounders. In an individually matched case–control study, effects of potential risk factors are typically ascertained through a conditional likelihood approach such as conditional logistic regression (CLR) [1]. Extension of CLR to situations with multiple subtypes of cases or controls has been made through polychotomous CLR (PCLR), which is more efficient than carrying out separate CLRs for subgroups [2]. Liang and Stewart [2], Becher and Jockel [3], and Becher [4] applied PCLR models to matched case–control studies with two control groups, typically hospital and population controls. Thomas *et al*. [5] and Durbin and Pasternack [6] applied PCLR models to analyze multiple disease groups with one set of controls. Sinha *et al*. [7] considered a Bayesian semiparametric model for analyzing matched case–control data with multiple disease states and missing exposure values. Mukherjee *et al*. [8] considered cases having multiple disease states with a natural ordering in matched case–control studies. Mukherjee *et al*. [9] proposed a methodology to fit stratified proportional odds models by amalgamating conditional likelihoods obtained from all possible binary collapsing of the ordinal scale.

Studies since the 1990s [10–13] have investigated elevated risk of respiratory diseases around putative point sources of environmental pollution. Diggle *et al*. [14] described an extension to matched case–control designs of the parametric modeling framework in [10, 12], using a conditional likelihood approach. Asthma and chronic obstructive airways disease were associated with proximity of residence

[a] *Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[b] *Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
**Correspondence to: Bhramar Mukherjee, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
[†] *E-mail: bhramar@umich.edu*

to major roads in East London. The possibility of residual spatial variation always exists in such environmental epidemiology studies. Diggle *et al*. [15] modeled the residual spatial variability as a Gaussian random field and proposed a Bayesian inferential approach using Markov chain Monte Carlo (MCMC) methods. Recently, there has been an increasing interest in modeling disease risk in relation to point sources of pollution in a Bayesian framework [16–18]. Wakefield and Morris [16] described a Bayesian hierarchical modeling of disease risk around a point source, embedding models proposed by Diggle *et al*. [13]. They discussed issues of the sensitivity to prior specification for this class of models. Dreassi *et al*. performed a sensitivity analysis to investigate how the specification of the distance-odds functions and the choice of prior distributions affect results under case–control studies [19]. Rodrigues *et al*. [20] provided a semiparametric approach for point process modeling using generalized additive model and illustrated the flexibility of this approach with applications in epidemiology and criminology. All of the aforementioned spatial environmental epidemiology studies considered only the standard binary case–control states.

The purpose of this article is to incorporate the distance-odds model around point sources into the analysis of matched case–control data with *multiple* disease or control states. We extend the idea of the polychotomous logit model and the adjacent-category logit model from the standard categorical data literature [21] to the nonlinear distance-odds model framework. The extensions with nonlinear odds function lead to some unique observations specific to the distance odds model. We evaluate maximum likelihood, profile likelihood, iteratively reweighted least squares (IRLS), and a hierarchical Bayesian approach using MCMC under the proposed models. We compare inference methods and various types of point source models using an extensive simulation study. Simulation studies that compare the frequentist properties (such as bias, mean squared error (MSE), and coverage probability) of the proposed methods and models are not available in the literature, not even for binary case–control states.

We organize the rest of the paper as follows. Section 2 describes the general model formulation. Section 2.1 reviews the distance-odds model with binary outcomes as proposed by Diggle *et al.*; Section 2.2 considers the extension of the distance-odds model with polychotomous outcomes under matched case–control data and considers various inference approaches. Section 3 explores the performance of the proposed models and inference methods using extensive simulation studies. We consider Analysis of the Detroit Asthma Morbidity, Air Quality and Traffic (DAMAT) study as a case study in Section 4. Section 5 concludes with a discussion.

2. Model formulation

2.1. *Review of distance-odds model with binary outcome by Diggle et al.* [14]

Diggle *et al*. [10, 12] proposed the distance-odds model for characterizing elevated risk around putative point sources of environmental pollution in case–control studies. The model assumes that the odds of disease, $r(x)$ as a function of distance x from the point source, is proportional to the decay function $f(x)$, as given in the following:

$$\begin{aligned} &\frac{P(Y=1|x)}{P(Y=0|x)} = \frac{p(x)}{1-p(x)} = r(x) = \rho f(x) \quad \text{and} \\ &f(x) = 1 + \alpha \exp\left(-(x/\beta)^2\right), \qquad (\alpha, \beta) \in (-1, \infty) \times (0, \infty), \end{aligned} \tag{1}$$

where Y is the disease status ($Y = 1$ for case; $Y = 0$ for control), x is the distance from the point source, and ρ is the background odds of disease in the case–control population. (For a case–control study that is embedded in a cohort study, ρ is typically given by $\rho = (q_1/q_2)\kappa$, where κ is the background odds of disease in the study cohort and q_1 and q_2 are the proportions of cases and controls sampled from the cohort respectively.) The parameters (α, β) in model (1) have a natural interpretation: α is proportional to the disease odds at the point source ($\alpha = [r(0)/\rho] - 1$); β measures the rate of decay with increasing distance from the point source, in the unit of distance x. Under this model setting, as $x \to \infty$, we have $f(x) \to 1$ and the risk function $p(x) = P(Y = 1|x) = \rho f(x)/(1 + \rho f(x)) \to \rho/(1 + \rho)$, that is, the background risk in the case–control population [14]. Note that, if $f(x) = \exp(\beta x)$ is chosen with $r(x) = \rho f(x)$ in model (1), then one would have that $\log(r(x)) = \log(\rho) + \beta x$, which becomes the usual logistic regression model that assumes a linear distance-odds relationship with log odds ratio β and intercept $\log(\rho)$. However, usually the odds of disease changes nonlinearly with increasing distance from the point source, for example, with increasing distance to an industrial park, the odds of asthma might decrease much faster within 0–200 m than within 1000–1200 m. Another possible disadvantage of

the log-linear model is that for $\beta < 0$ (that implies increasing odds with decreasing distance), $r(x) \to 0$ and $p(x) \to 0$ as $x \to \infty$, but these do not converge to background odds or risk, which would be a desirable property. For non-rare diseases such as asthma, the log-linear distance-odds model is questionable. These disadvantages of log-linear model lead us to focus on the nonlinear distance-odds model (1) proposed by Diggle *et al*. [10].

As an extension to model (1), Diggle and Rowlingson [12] assumed multiplicative risk factors for the combined effects of S point sources and allowed for covariate adjustment via additional log-linear terms. In the presence of S point sources and W spatially referenced covariates $Z_w(x), w = 1, \ldots, W$, the resulting distance-odds model takes the form

$$r(\mathbf{x}) = \rho f(\mathbf{x}) \quad \text{and} \quad f(\mathbf{x}) = \exp\left(\sum_{w=1}^{W} \phi_w Z_w(\mathbf{x})\right) \prod_{s=1}^{S} f_s(x_s), \tag{2}$$

where $\mathbf{x} = (x_1, \ldots, x_S)$ and x_s and $f_s(x_s)$ are the distance and the decay function for the sth point source, respectively. Here, each $f_s(x_s)$ takes the same functional form as in model (1), that is, $f_s(x_s) = 1 + \alpha_s \exp(-(x_s/\beta_s)^2)$.

For a 1:M matched case–control study with N matched pairs, the risk of disease for an individual at distance x in the ith stratum can be expressed as [14]

$$P_i(Y = 1|x) = \frac{r_i(x)}{1 + r_i(x)} = \frac{\rho_i f(x)}{1 + \rho_i f(x)}, \quad i = 1, \ldots, N,$$

where the baseline odds ρ_i for the ith stratum can potentially vary across matched pairs under the matched case–control design. The conditional likelihood, given the exposure vector at distance $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i(M+1)})$ for the ith stratum, that the case is at distance x_{i1} is

$$\begin{aligned} L_i(\alpha, \beta) &= P(Y_{i1} = 1, Y_{i2} = \cdots = Y_{i(M+1)} = 0 | Y_{i1} + Y_{i2} + \cdots + Y_{i(M+1)} = 1, \mathbf{x_i}) \\ &= \frac{\frac{\rho_i f(x_{i1})}{\prod_{j=1}^{M+1}(1+\rho_i f(x_{ij}))}}{\frac{\rho_i f(x_{i1})}{\prod_{j=1}^{M+1}(1+\rho_i f(x_{ij}))} + \frac{\rho_i f(x_{i2})}{\prod_{j=1}^{M+1}(1+\rho_i f(x_{ij}))} + \cdots + \frac{\rho_i f(x_{i(M+1)})}{\prod_{j=1}^{M+1}(1+\rho_i f(x_{ij}))}} \\ &= \frac{f(x_{i1})}{\sum_{j=1}^{M+1} f(x_{ij})}, \quad i = 1, \ldots, N, \end{aligned} \tag{3}$$

where Y_{ij} and x_{ij} are the disease status and distance for the jth individual in the ith stratum respectively, $i = 1, \ldots, N; j = 1, \ldots, M+1$. The general form of the conditional likelihood is (3). For one point source binary model, $f(x)$ is as given in (1), where as for multiple point sources, binary model (with possible covariate adjustment) $f(x)$ is as given in (2).

Denote the conditional likelihood by L, the corresponding log-likelihood by l $\left(l = \log(L) = \sum_{i=1}^{N} \log(L_i) = \sum_{i=1}^{N} l_i\right)$, and the parameters to be estimated by $\boldsymbol{\theta}$. The maximum likelihood estimates (MLEs) of $\boldsymbol{\theta} = (\alpha, \beta)$ in the one point source binary outcome model can be obtained by maximizing the logarithm of the conditional likelihood

$$l(\alpha, \beta) = \sum_{i=1}^{N} \log\left(\frac{f(x_{i1})}{\sum_{j=1}^{M+1} f(x_{ij})}\right) = \sum_{i=1}^{N} \log\left(\frac{1 + \alpha \exp(-(x_{i1}/\beta)^2)}{\sum_{j=1}^{M+1}\left[1 + \alpha \exp(-(x_{ij}/\beta)^2)\right]}\right).$$

Similarly, the MLEs of $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}) = (\alpha_1, \ldots, \alpha_S, \beta_1, \ldots, \beta_S, \phi_1, \ldots, \phi_W)$ in the S point sources binary outcome model with W covariates can be obtained by maximizing

$$\begin{aligned} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}) &= \sum_{i=1}^{N} \log\left(\frac{f(\boldsymbol{x_{i1}})}{\sum_{j=1}^{M+1} f(\boldsymbol{x_{ij}})}\right) \\ &= \sum_{i=1}^{N} \log\left(\frac{\exp\left(\sum_{w=1}^{W} \phi_w Z_w(\boldsymbol{x_{i1}})\right) \prod_{s=1}^{S}\left(1 + \alpha_s \exp\left(-(x_{i1s}/\beta_s)^2\right)\right)}{\sum_{j=1}^{M+1}\left[\exp\left(\sum_{w=1}^{W} \phi_w Z_w(\boldsymbol{x_{ij}})\right) \prod_{s=1}^{S}\left(1 + \alpha_s \exp\left(-(x_{ijs}/\beta_s)^2\right)\right)\right]}\right), \end{aligned}$$

where $\boldsymbol{x_{ij}} = (x_{ij1}, \ldots, x_{ijS})$ and x_{ijs} is the distance of the jth individual in the ith stratum from the sth point source. We can find more detailed discussion of parameter estimation and inference for the models with binary outcomes in [14].

2.2. *Distance-odds model with polychotomous outcome*

In this section, we extend the distance-odds model reviewed in Section 2.1 to situations where cases can have multiple disease states. Without loss of generality, we illustrate the methods and formulation in the following sections for a 1:M matched case–control data set with N matched pairs, where outcomes can belong to one of the K disease categories (for example, with $K = 2$; poor prognosis: $Y = 2$; fair prognosis: $Y = 1$) and one control group ($Y = 0$). These methods can be readily applied to situations with multiple control states and to situations with variable matching ratios. The distance-odds model is adapted to both polychotomous-category model (PCM) and adjacent-category model (ACM) setting (Remark 1). The PCMs are considered when one tries to distinguish nominal disease subtypes to the controls. The ACMs are more appropriate when there is a natural ordering of the disease subclassifications.

2.2.1. Polychotomous-category distance-odds model. For the PCM setting, the odds of disease for the jth individual in the ith stratum at distance x_{ij} is modeled as

$$r^k(x_{ij}) = \frac{P(Y_{ij} = k|x_{ij})}{P(Y_{ij} = 0|x_{ij})} = \rho_{ik} f_k(x_{ij}), \qquad i = 1, \ldots, N; \quad j = 1, \ldots, M+1; \quad k = 1, \ldots, K, \tag{4}$$

where the baseline odds ρ_{ik} can potentially vary across matched pairs i and disease categories k and the distance-odds function $f_k(x)$ can also vary among disease categories. Note that, if $f_k(x) = \exp(\beta_k x)$ is chosen in model (4) with multiplicative nuisance parameters $\rho_{ik} = \gamma_i \times \lambda_k$, one would have that

$$\log(r^k(x_{ij})) = \log\left(\frac{P(Y_{ij} = k|x_{ij})}{P(Y_{ij} = 0|x_{ij})}\right) = \log(\gamma_i) + \log(\lambda_k) + \beta_k x_{ij}, \tag{5}$$

which becomes the polychotomous logistic regression models [21] that assumes a linear distance-odds relationship. Nonlinear distance-odds models such as (1) are desired, with advantages over log-linear models as discussed in Section 2.1. With the use of the K equations in (4) along with one more constraint that $\sum_{k=0}^{K} P(Y_{ij} = k|x_{ij}) = 1$, the risk of disease can be written in terms of ρ_{ik} and f_k for the corresponding individual, that is,

$$P(Y_{ij} = 0|x_{ij}) = \frac{1}{1 + \sum_{k=1}^{K}[\rho_{ik} f_k(x_{ij})]},$$
$$P(Y_{ij} = k|x_{ij}) = \frac{\rho_{ik} f_k(x_{ij})}{1 + \sum_{k=1}^{K}[\rho_{ik} f_k(x_{ij})]}, \quad k = 1, \ldots, K.$$

Let k_i denote the disease states of the case subject in matched set i, $k_i \in (1, \ldots, K)$. The conditional likelihood for the ith stratum, given a matched case–control pair at distance $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{i(M+1)})$, that the case (in category k_i) is at distance x_{i1} is

$$\begin{aligned} L_i^{k_i} &= P(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{i(M+1)} = 0 | Y_{i1} + Y_{i2} + \cdots + Y_{i(M+1)} = k_i, \mathbf{x_i}) \\ &= \frac{\rho_{ik_i} f_{k_i}(x_{i1}) / \prod_{j=1}^{M+1}\left[1 + \sum_{k=1}^{K} \rho_{ik} f_k(x_{ij})\right]}{\sum_{j=1}^{M+1} \rho_{ik_i} f_{k_i}(x_{ij}) / \prod_{j=1}^{M+1}\left[1 + \sum_{k=1}^{K} \rho_{ik} f_k(x_{ij})\right]} \\ &= \frac{f_{k_i}(x_{i1})}{\sum_{j=1}^{M+1} f_{k_i}(x_{ij})}. \end{aligned} \tag{6}$$

The general form of the conditional likelihood is (6). For one point source PCM, $f_k(x)$ is given as $f_k(x) = 1 + \alpha_k \exp\left(-(x/\beta_k)^2\right)$; for multiple point sources PCM, $f_k(x)$ is given as $f_k(\boldsymbol{x}) = \exp\left(\sum_{t=1}^{T} \phi_{kt} Z_{kt}(\boldsymbol{x})\right) \prod_{s=1}^{S} f_{ks}(x_s)$, where $f_{ks}(x_s) = 1 + \alpha_{ks} \exp\left(-(x_s/\beta_{ks})^2\right)$.

2.2.2. Adjacent-category distance-odds model. For the ACM setting, the adjacent odds of disease between category K versus $K-1$ for the jth individual in the ith stratum can be modeled as

$$r^k(x_{ij}) = \frac{P(Y_{ij}=k|x_{ij})}{P(Y_{ij}=k-1|x_{ij})} = \rho_{ik} f_k(x_{ij}), \qquad i=1,\ldots,N; \quad j=1,\ldots,M+1; \quad k=1,\ldots,K. \tag{7}$$

Again, the baseline odds ρ_{ik} can vary across matched pairs i and disease categories k, and the distance-odds function $f_k(x)$ can vary across disease categories. One point source ACM and multiple point sources ACM (with possible covariate adjustment) can be formulated similarly as PCM with different choices of f_k. For these nonlinear settings, ACM cannot be represented as a reparameterization of PCM as in log-linear models (Remark 1). Thus, both ACM and PCM are needed for ordered and nominal disease subclassifications, respectively. Note that if $f_k(x) = \exp\left(\beta_k^* x\right)$ is chosen in model (7) with multiplicative nuisance parameters $\rho_{ik} = \gamma_i^* \times \lambda_k^*$, one would have that

$$\log(r^k(x_{ij})) = \log\left(\frac{P(Y_{ij}=k|x_{ij})}{P(Y_{ij}=k-1|x_{ij})}\right) = \log\left(\gamma_i^*\right) + \log\left(\lambda_k^*\right) + \beta_k^* x_{ij}, \tag{8}$$

which reduces to the polychotomous logistic regression models in adjacent category setting [21] that assumes a linear distance-odds relationship. The risk of disease can be represented in terms of ρ_{ik} and f_k as

$$P(Y_{ij}=0|x_{ij}) = \frac{1}{1+\sum_{k=1}^{K}\left[\prod_{h=1}^{k}\rho_{ih}f_h(x_{ij})\right]},$$
$$P(Y_{ij}=k|x_{ij}) = \frac{\prod_{h=1}^{k}\rho_{ih}f_h(x_{ij})}{1+\sum_{k=1}^{K}\left[\prod_{h=1}^{k}\rho_{ih}f_h(x_{ij})\right]}, \quad k=1,\ldots,K.$$

It follows that the conditional likelihood for the ith stratum is

$$\begin{aligned} L_i^{k_i} &= P(Y_{i1}=k_i, Y_{i2}=\cdots=Y_{i(M+1)}=0|Y_{i1}+Y_{i2}+\cdots+Y_{i(M+1)}=k_i, \mathbf{x_i}) \\ &= \frac{\prod_{h=1}^{k_i}\rho_{ih}f_h(x_{i1}) / \prod_{j=1}^{M+1}\left[1+\sum_{k=1}^{K}\left[\prod_{h=1}^{k}\rho_{ih}f_h(x_{ij})\right]\right]}{\sum_{j=1}^{M+1}\left[\prod_{h=1}^{k_i}\rho_{ih}f_h(x_{ij})\right] / \prod_{j=1}^{M+1}\left[1+\sum_{k=1}^{K}\left[\prod_{h=1}^{k}\rho_{ih}f_h(x_{ij})\right]\right]} \\ &= \frac{\prod_{h=1}^{k_i} f_h(x_{i1})}{\sum_{j=1}^{M+1}\left[\prod_{h=1}^{k_i} f_h(x_{ij})\right]}. \end{aligned} \tag{9}$$

One special case of interest is the homogeneity of the adjacent odds ratios with one unit increase in distance across case categories, that is,

$$\frac{r^K(x+1)}{r^K(x)} = \frac{r^{K-1}(x+1)}{r^{K-1}(x)} = \cdots = \frac{r^1(x+1)}{r^1(x)}, \quad \forall x \quad \Leftrightarrow \quad \alpha_1=\alpha_2=\cdots=\alpha_K \quad \text{and} \quad \beta_1=\beta_2=\ldots=\beta_K. \tag{10}$$

We call this special case in (10) the homogeneous ACM.

Remark 1 (Connection between the ACM and the PCM)
For the log-linear case of the ACM and the PCM as given in Equations (5) and (8), respectively, the logarithm of the polychotomous odds can be rewritten as the sum of the logarithm of the adjacent-category odds, that is,

$$\log\left(\frac{P(Y_{ij}=k|x_{ij})}{P(Y_{ij}=0|x_{ij})}\right) = \sum_{h=1}^{k}\log\left(\frac{P(Y_{ij}=h|x_{ij})}{P(Y_{ij}=h-1|x_{ij})}\right) = k\log\left(\gamma_i^*\right) + \sum_{h=1}^{k}\log\left(\lambda_h^*\right) + \sum_{h=1}^{k}\beta_h^* x_{ij}. \tag{11}$$

Comparing Equation (11) with (5), one would have the well-known one-to-one mapping between the polychotomous odds ratio and the adjacent-category odds ratio, that is, $\beta_k = \sum_{h=1}^{k}\beta_h^*, k=1,\ldots,K$.

However, similar mapping between PCM and ACM for the nonlinear distance-odds model cannot be established even for the simplest case with $K = 2$. For example,

$$\frac{P(Y_{ij} = k|x_{ij})}{P(Y_{ij} = 0|x_{ij})} = \prod_{h=1}^{k} \frac{P(Y_{ij} = h|x_{ij})}{P(Y_{ij} = h-1|x_{ij})}, \quad k = 1, 2$$

$$\Rightarrow 1 + \alpha_k \exp\left(-(x_{ij}/\beta_k)^2\right) = \prod_{h=1}^{k} \left(1 + \alpha_h^* \exp\left(-\left(x_{ij}/\beta_h^*\right)^2\right)\right), \quad k = 1, 2.$$

When $k = 1$, $\alpha_1 = \alpha_1^*$ and $\beta_1 = \beta_1^*$; when $k = 2$, the aforementioned equation does not have closed-form solutions for (α_2, β_2) in terms of $\left(\alpha_1^*, \beta_1^*, \alpha_2^*, \beta_2^*\right)$. Therefore, PCM is not a natural reparameterization of ACM as in the log-linear model case. Consequently, ACM or homogeneous ACM cannot be fitted as a special case of the PCM setting.

2.3. *Estimation and inference*

2.3.1. Maximum likelihood approach. Without loss of generality, the first subject in each stratum is always considered as the case when deriving the likelihood and fitting the models, that is, $Y_{i1} = k_i, k_i \in (1, \ldots, K)$. Thus, the actual contribution of the ith stratum to the conditional likelihood is $L_i^{k_i}$ as given in (6) for PCM or as given in (9) for ACM, respectively. For example, the MLEs for ACM can be obtained by maximizing the logarithm of the conditional likelihood

$$\begin{aligned}\sum_{i=1}^{N} \log\left(L_i^{k_i}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right) &= \sum_{i=1}^{N} \log\left(\frac{\prod_{h=1}^{k_i} f_h(x_{i1})}{\sum_{j=1}^{M+1} \prod_{h=1}^{k_i} f_h(x_{ij})}\right) \\ &= \sum_{i=1}^{N} \log\left(\frac{\prod_{h=1}^{k_i} \left(1 + \alpha_h \exp\left(-(x_{i1}/\beta_h)^2\right)\right)}{\sum_{j=1}^{M+1} \prod_{h=1}^{k_i} \left(1 + \alpha_h \exp\left(-(x_{ij}/\beta_h)^2\right)\right)}\right),\end{aligned} \tag{12}$$

or the following in the most general case with multiple sources and covariate adjustment

$$\begin{aligned}&\sum_{i=1}^{N} \log\left(L_i^{k_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi})\right) = \sum_{i=1}^{N} \log\left(\frac{\prod_{h=1}^{k_i} f_h(\boldsymbol{x_{i1}})}{\sum_{j=1}^{M+1} \prod_{h=1}^{k_i} f_h(\boldsymbol{x_{ij}})}\right) \\ &= \sum_{i=1}^{N} \log\left(\frac{\prod_{h=1}^{k_i} \left[\exp\left(\sum_{w=1}^{W} \phi_{hw} Z_{hw}(\boldsymbol{x_{i1}})\right) \prod_{s=1}^{S} \left(1 + \alpha_{hs} \exp\left(-(x_{i1s}/\beta_{hs})^2\right)\right)\right]}{\sum_{j=1}^{M+1} \prod_{h=1}^{k_i} \left[\exp\left(\sum_{w=1}^{W} \phi_{hw} Z_{hw}(\boldsymbol{x_{ij}})\right) \prod_{s=1}^{S} \left(1 + \alpha_{hs} \exp\left(-(x_{ijs}/\beta_{hs})^2\right)\right)\right]}\right).\end{aligned} \tag{13}$$

Under the homogeneity assumption in (10), maximizing (12) or (13) would be reduced to the constrained optimization problem with restriction $(\alpha_1 = \cdots = \alpha_K, \beta_1 = \cdots = \beta_K)$ or $(\alpha_{1s} = \cdots = \alpha_{Ks}, \beta_{1s} = \cdots = \beta_{Ks}, \forall s)$, respectively. The MLEs of PCMs can be obtained similarly. Standard errors of the parameter estimates can be calculated from the square root of the diagonal elements of the inverse of the Hessian matrix of the corresponding conditional likelihood, and then the 95% Wald-type confidence intervals (CI) can be constructed.

2.3.2. Profile likelihood approach. Parameter estimates and CIs can also be obtained using the profile likelihood. For the one point source homogeneous ACM, the simplest case with two parameters, the profile likelihood method reduces $l(\alpha, \beta)$ to a function of a single-parameter β, by treating α as nuisance parameter and maximizing over it. The profile likelihood for β is defined as

$$\tilde{l}(\beta) = \max_{\alpha} l(\alpha, \beta).$$

Suppose that the maximum of the function $\tilde{l}(\beta)$ is located at $\tilde{\beta}$ and the corresponding optimizer over α is $\tilde{\alpha}(\tilde{\beta})$. Thus, $(\tilde{\alpha}(\tilde{\beta}), \tilde{\beta})$ would be the MLE based on the profile likelihood. The CI based on profile likelihood for β is defined as

$$\left\{\beta : 2[l(\tilde{\alpha}(\tilde{\beta}), \tilde{\beta}) - \tilde{l}(\beta)] \leq \chi^2_{1,0.95}\right\},$$

where $\chi^2_{1,0.95}$ is the 95th upper quantile of the χ^2 distribution with one degree of freedom. This approach reduces the number of independent parameters by expressing some of them as functions of the others, instead of dealing with all the parameters simultaneously. It is helpful in the presence of many parameters, such as in (12) and (13).

Remark 2 (Identifiability and Monte Carlo tests)
The likelihood-based inference described in Sections 2.3.1 and 2.3.2 assumes that usual regularity conditions hold [22]. Under these regularity conditions, approximate CIs for the MLEs can be derived from the asymptotic multivariate normality of the MLEs and the estimated Hessian matrix. The likelihood ratio statistics for testing $H_0 : f(x) = 1$ has an asymptotic chi-squared distribution under the same regularity conditions. Diggle *et al*. [14] pointed out that with an insufficient sample size, the log-likelihood surface of (α, β) may be far from quadratic and standard likelihood-based asymptotics are unreliable. Moreover, these models have an irregularity at the null hypothesis of $H_0 : f(x) = 1$, because $f(x) = 1$ corresponds to one of the two parameters of (α, β) equal to 0 with the other indeterminate, in the situation where there is no covariate adjustment. Monte Carlo tests can be used as an alternative. One thousand data sets can be simulated under the null, and the observed values of the likelihood ratio statistics $LR = 2 \times (l(\hat{\alpha}, \hat{\beta}) - l(\alpha = 0 \text{ or } \beta = 0)) = 2\left(l(\hat{\alpha}, \hat{\beta}) - N \log\left(\frac{1}{M+1}\right)\right)$ can be ranked among the 1000 simulated LR values. If the observed LR ranks kth largest among 1000 simulated values, the p-value of the Monte Carlo test is $k/1001$ and the test is exact [14, 23].

2.3.3. Iteratively reweighted least square regression. Another alternative approach is IRLS regression. As the strata are mutually independent under the matched case–control design, it is not necessary to further consider the correlation between the residuals from different strata. Typically, one can write the nonlinear regression model with binary response Y_i as

$$Y_i = p_i(\boldsymbol{x_{i,\theta}}) + \varepsilon_i,$$

where Y_i is the observed binary response, $p_i(\boldsymbol{x_{i,\theta}})$ is the predicted probability from the model for subject i, and $\varepsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed random errors, $i = 1, \ldots, N$. Under the conditional framework given there being a matched case–control pair at distance $\mathbf{x_i}$, we can treat each stratum as a single 'subject' with response $\sum_{j=1}^{M+1} I(Y_{ij} = k_i) = I(Y_{i1} = k_i)$ (assumed the first subject to be the case) and predicted probability $L_i^{k_i}$ as given in Section 2.2. The sum of squared error (*SSE*) is given by

$$SSE(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(I(Y_{i1} = k_i) - L_i^{k_i}(\boldsymbol{x_i}, \boldsymbol{\theta})\right)^2.$$

One can further assume that the variance structure of the errors to be $\varepsilon_i \sim N\left(0, \sigma_k^2\right)$ for $\{i : k_i = k\}$, that is, for all the strata where case response equals to k. Then, the IRLS estimation can be realized by iteratively minimizing the weighted *SSE*

$$SSE(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \left[\sum_{i:k_i=k} \left(I(Y_{i1} = k) - L_i^k(\boldsymbol{x_i}, \boldsymbol{\theta})\right) \Sigma_k^{-1} \left(I(Y_{i1} = k) - L_i^k(\boldsymbol{x_i}, \boldsymbol{\theta})\right) \right], \qquad (14)$$

where Σ_k is the pooled variance of errors from all strata where the case response equals k. In the initial step of IRLS, $\boldsymbol{\theta}$ is estimated by minimizing the weighted *SSE* with all $\Sigma_k^{(0)}$ set to identity. An estimate for $\Sigma_k^{(1)}$ is then calculated by $(1/df_k) \sum_{i:k_i=k} r_i^{(0)2}$, where the residuals $r_i^{(0)} = I(Y_{i1} = k_i) - L_{i1}^{k_i}\left(\boldsymbol{x_i}, \hat{\boldsymbol{\theta}}^{(\mathbf{0})}\right)$ and df_k is the degree of freedom (the size of the set $\{i : k_i = k\}$ minus the number of parameters in the model). The estimated $\hat{\Sigma}_k^{(1)}$ are used as the weights in the next step of IRLS to minimize the weighted *SSE*. Parameter estimation is simply realized by iterating this process further, calculating updated estimates for Σ_k's, estimating the model parameters $\boldsymbol{\theta}$ with updated weights,

and iterating until convergence. The standard errors can be calculated from the Hessian matrix of the corresponding log-likelihood

$$\sum_{k=1}^{K}\sum_{i:k_i=k}\left[-\frac{1}{2}\log\left(2\pi\sigma_k^2\right)-\frac{1}{2}\left(\frac{I(Y_{i1}=k)-L_i^k(\boldsymbol{x_i},\boldsymbol{\theta})}{\sigma_k}\right)^2\right].$$

IRLS estimate and MLE were shown to be consistent and asymptotically normal under the assumption that the errors are normally distributed as $\varepsilon_i \sim N\left(0,\sigma_k^2\right)$ for $\{i : k_i = k\}$ [24].

Remark 3
For the three methods discussed in Section 2.3.1–2.3.3, instead of working directly on (α_{ks},β_{ks}) with a range of $(-1,\infty)\times(0,\infty)$, we performed unrestricted optimizations on the one-to-one transformed parameters $(u_{ks},v_{ks}) = (\log(1+\alpha_{ks}),\log(\beta_{ks}))$ that span the whole real plane and then transformed the results back in terms of the original parameters (α_{ks},β_{ks}).

2.3.4. Bayesian approach. The Bayesian approach provides an alternative to the frequentist inferential strategies described in Section 2.3.1–2.3.3. A proper Bayesian approach would be to use the full likelihood and specify a prior distribution on the nuisance parameters $\boldsymbol{\rho} = (\rho_1,\ldots,\rho_N)$. However, the full likelihood approach would encounter the difficulty of prior specification and estimation of $\boldsymbol{\rho}$. One can use a marginal likelihood instead, which integrates out the nuisance parameters with respect to a random distribution. Rice [25, 26] discussed the equivalence between the use of conditional and marginal likelihoods for matched case–control study. Diggle *et al*. pointed out that the conditional likelihood approach is consistent with the full likelihood approach for the binary outcome model with independent priors for $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$ [14]. Therefore, we proceed with the conditional likelihood as the basis for Bayesian inference.

Prior specification. We primarily considered in this paper the following sets of mutually independent prior distributions on $(\boldsymbol{u},\boldsymbol{v}) = (u_{11},\ldots,u_{KS},v_{11},\ldots,v_{KS})$,

$$\log(1+\alpha_{ks}) = u_{ks} \sim N\left(\mu_{u_{ks}},\sigma_{u_{ks}}^2\right),$$
$$\log(\beta_{ks}) = v_{ks} \sim N\left(\mu_{v_{ks}},\sigma_{v_{ks}}^2\right), \quad k=1,\ldots K; \quad s=1,\ldots S,$$

where the mean and variance of α_{ks} are $\mu_{\alpha_{ks}} = \exp\left(\mu_{u_{ks}}+\frac{1}{2}\sigma_{u_{ks}}^2\right)-1$ and $\sigma_{\alpha_{ks}}^2 = \left(\exp\left(\sigma_{u_{ks}}^2\right)-1\right)\exp\left(2\mu_{u_{ks}}+\sigma_{u_{ks}}^2\right)$, respectively. Similarly, $\mu_{\beta_{ks}} = \exp\left(\mu_{v_{ks}}+\frac{1}{2}\sigma_{v_{ks}}^2\right)$ and $\sigma_{\beta_{ks}}^2 = \left(\exp\left(\sigma_{v_{ks}}^2\right)-1\right)\exp\left(2\mu_{v_{ks}}+\sigma_{v_{ks}}^2\right)$. We considered both informative and noninformative (or vague) prior distributions. For informative priors, with our knowledge of roadway effects on asthma and the literature reviewed in Section 1, the prior distribution of α_{ks} was set with mean $\mu_{\alpha_{ks}} = 0.5$ and variance $\sigma_{\alpha_{ks}}^2 = 0.25$ (thus, $P(0.1 < \alpha_{ks} < 1.0) \approx 0.95$). For other types of health outcomes or pollution sources, different informative priors could be used. Given the fact that the point source effects on health outcomes (e.g., roadway effects on asthma) last only for a few hundred meters in most of the literature, prior distributions of β_{ks} were set with means $\mu_{\beta_{ks}} = 400$ and variance $\sigma_{\beta_{ks}}^2 = 150$ (thus, $P(50 < \beta_{ks} < 750) \approx 0.95$). For noninformative priors, the same mean $(\mu_{\alpha_{ks}},\mu_{\beta_{ks}}) = (0.5, 400)$ with large variance $\left(\sigma_{\alpha_{ks}}^2,\sigma_{\beta_{ks}}^2\right) = (0.5, 400)$ were used for (α_{ks},β_{ks}). It follows that $P(-0.2 < \alpha_{ks} < 2.0) \approx 0.95$ and $P(50 < \beta_{ks} < 1500) \approx 0.95$, which should contain the prior knowledge about (α,β). For the rest of the paper, we focus on $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and primarily proceed using models without covariate adjustment.

We perform a sensitivity analysis by comparing the posterior distributions derived from various normal priors with the same means of $(\mu_{u_{ks}},\mu_{v_{ks}})$ but different choices of $\left(\sigma_{u_{ks}}^2,\sigma_{v_{ks}}^2\right)$. Wakefield and Morris [16] suggested using independent Uniform prior distribution on (α,β) on the range of $(-1,\alpha_{max})\times(0,\beta_{max})$ for the one point source binary model (1), where α_{max} and β_{max} are the maximum plausible values based on current epidemiological knowledge. We also consider this Uniform prior distribution on (α_{ks},β_{ks}) with different choices of α_{max} and β_{max} as part of the sensitivity analysis.

Sampling algorithm. The joint posterior distribution can be expressed as

$$\pi(\boldsymbol{u},\boldsymbol{v}|X,Y) \propto \pi(\boldsymbol{u},\boldsymbol{v}) \times L(\boldsymbol{u},\boldsymbol{v}).$$

where $\pi(\boldsymbol{u},\boldsymbol{v})$ is the prior distribution and $L(\boldsymbol{u},\boldsymbol{v})$ is the conditional likelihood in terms of the transformed parameters $(u_{ks},v_{ks}) = (\log(1+\alpha_{ks}),\log(\beta_{ks}))$. Because the full conditional distributions of

the parameters do not follow a standard distributional form, the MCMC method is used to generate random draws from the posterior distributions. For two-parameter models such as the one point source homogeneous ACM, the random walk Metropolis–Hastings algorithm is used to generate a Markov chain that has the limit distribution equal to the target posterior distribution. For four(or more)-parameter models such as one point source ACM, computationally it is hard to draw simultaneously from the joint distribution using Metropolis–Hastings algorithm. Instead, we use a componentwise Metropolis–Hastings within Gibbs algorithm. We discuss the computational strategy corresponding to these MCMC algorithms in Appendices B and C (Supporting information[‡]). The convergence of these Markov chains are examined using Gelman and Rubin's convergence diagnostic [27]. In this study, the random walk Metropolis–Hastings or Metropolis–Hastings within Gibbs algorithm for the proposed models converge to their limit distributions after 2000–4000 runs. The chains have autocorrelations up to 20. Therefore, the chains are refined by choosing a common burn-in period of 5000 and a common thinning frequency of 20. We performed these MCMC algorithms for a length of $T = 45000$. After burn-in and thinning, the resulting Markov chains of length 2000 are treated as random draws from the target posterior distribution.

As a Bayesian counterpart to the Monte Carlo test discussed in Remark 2, Bayes factors [28] are considered to test the null hypothesis that $H_0 : f(x) = 1$. The Bayes factor for comparing the current model M_1 to the null model M_0 is defined as the ratio of the posterior probability to the prior probability, which is given by

$$B = \frac{P(M_1|Y)/P(M_0|Y)}{P(M_1)/P(M_0)} = \frac{\int_\theta \pi(Y|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)\mathrm{d}\boldsymbol{\theta}}{\int_\phi \pi(Y|\boldsymbol{\phi}, M_0)\pi(\boldsymbol{\phi}|M_0)\mathrm{d}\boldsymbol{\phi}} = \frac{P(Y|M_1)}{P(Y|M_0)}.$$

The calculation of the Bayes factor B is not straightforward using MCMC. We used the importance sampling estimator $\frac{1}{T}\sum_{t=1}^{T}\left[l(\boldsymbol{\theta^t})\pi(\boldsymbol{\theta^t})/g(\boldsymbol{\theta^t})\right]$ as suggested by Diggle *et al.* [14], where the prior distribution on $\boldsymbol{\theta}$ is used as the importance distribution $g(\boldsymbol{\theta})$ and $\boldsymbol{\theta^t}$ are sampled from $g(\boldsymbol{\theta})$. Kass and Raftery [28] suggested calculating $2\log(B)$ as a Bayesian analogue of a log-likelihood ratio statistics or deviance. Values greater than 2 indicate increasing evidence against M_0: between 2 and 6 is 'positive' evidence, 6 to 10 is 'strong', and over 10 is 'very strong' evidence against M_0 [14, 28]. We can find a number of alternatives in [29].

3. Simulation study

We consider two case subgroups ($K = 2$) and one control group and up to two point sources in the following simulation study. Specifically, we conduct four different settings of simulations where the true models are as follows: (1) one point source PCM; (2) one point source ACM; (3) one point source homogeneous ACM; and (4) two point sources homogeneous ACM.

3.1. Simulation design

We generate a large cohort of $L = 1,000,000$ people initially. We include two independent risk factors, age and gender, for this cohort, of which we set the distributions similar to those for the pediatric population of the Detroit Medicaid data source. Specifically, we generate gender from a Bernoulli distribution with probability 0.55 for being a male; we generate age from a piecewise Uniform distribution with a range of 2–18 and then rounded to integer values. We generate the exposure variable, distance to the point source, from a mixture distribution of Uniform and Gamma. Specifically, we generate distances (in meters) from the first and second sources from $0.15 \cdot \text{Uniform}(0, 500) + 0.85 \cdot \text{Gamma}(\text{shape} = 3, \text{rate} = 0.003)$ and $0.2 \cdot \text{Uniform}(0, 500) + 0.8 \cdot \text{Gamma}(\text{shape} = 3, \text{rate} = 0.005)$, respectively. Simulation studies are based on this fixed cohort with mutually independent covariates of age, gender, and distances with distributions described previously.

The disease status for the cohort would be different for different choices of distance-odds model or true parameter settings. For example, for one point source ACM, the disease states ($k = 0, 1, 2$) are generated using the subject-specific risk functions $\boldsymbol{p(x)}$ in (15) with certain fixed values of $(\alpha_1, \beta_1, \alpha_2, \beta_2)$.

[‡]*Supporting information may be found in the online version of this article.*

Specifically, the outcome for the lth patient Y_l is generated from the multinomial distribution with probabilities

$$\begin{aligned} P(Y_l = 0|x_l) &= \frac{1}{1+\rho_{l1}f_1(x_l)+\rho_{l1}\rho_{l2}f_1(x_l)f_2(x_l)}, \\ P(Y_l = 1|x_l) &= \frac{\rho_{l1}f_1(x_l)}{1+\rho_{l1}f_1(x_l)+\rho_{l1}\rho_{l2}f_1(x_l)f_2(x_l)}, \\ P(Y_l = 2|x_l) &= \frac{\rho_{l1}\rho_{l2}f_1(x_l)f_2(x_l)}{1+\rho_{l1}f_1(x_l)+\rho_{l1}\rho_{l2}f_1(x_l)f_2(x_l)}, \quad l = 1,\ldots,L. \end{aligned} \tag{15}$$

The subject-specific nuisance parameter for the lth patient can be generated using $\rho_{lk} = \exp(b_{0k} + b_1 \times \text{age}_l + b_2 \times \text{gender}_l)$, $k = 1, 2$. The parameters $(b_{01}, b_{02}, b_1, b_2)$ can be obtained from the Detroit Medicaid data. Here, we use $b_1 = -0.05$ and $b_2 = 0.3$. The intercepts b_{01} and b_{02} can be varied within a range of $(-2.0, -0.5)$ to generate different desired disease prevalence. Typically, about 20% of subjects of the cohort are generated as cases, of which all disease subcategories have roughly the same proportion ($k = 1, \approx 10\%; k = 2, \approx 10\%$). After the disease status is generated for the cohort, $R = 500$ matched case–control data sets are then generated, each with N 1:1 matched pairs. We also consider different sample sizes $N = 500$, 1000, and 2000. Specifically, for each of the R matched case–control data sets, N cases are randomly drawn from the cohort, and then they are randomly matched with controls by age (within 2 years) and gender. We did not consider covariate adjustment in the simulation study because both covariates of age and gender are matched.

Under each model setting, we calculate parameter estimates with 95% CIs by using MLE, profile likelihood, and IRLS described in Section 2.3.1–2.3.3. Because of the identifiability problem of the likelihoods for the proposed models, there are a few runs ($< 5\%$) that fail to converge or converge for the point estimates but can not obtain CIs (for example, failure to invert the Hessian matrix using maximum likelihood method). We removed the nonconverged data sets among the $R = 500$ ones. We summarize the simulation results on the remaining R' data sets where all three frequentist methods converge. We summarize the R' estimates in terms of relative bias (e.g., relative bias for a parameter θ is $\left(\frac{1}{R'}\sum_{i=1}^{R'}\hat{\theta}_{(i)} - \theta_{\text{true}}\right)/\theta_{\text{true}} \times 100\%$), MSE (e.g., $MSE = \frac{1}{R'}\sum_{i=1}^{R'}(\hat{\theta}_{(i)} - \theta_{\text{true}})^2$), and coverage probability (the proportion that the 95% CIs cover the true value is calculated as an *ad hoc* estimate of the true coverage probability among these R' runs). For the Bayesian approach, the posterior mode as well as 95% highest posterior density (HPD) interval are estimated based on 2000 draws (after burn-in and thinning) from the posterior distribution. Because the posterior distributions of α and β are both positively skewed (a heavy right tail for β), the posterior mean is not used. To compare with the frequentist results such as MLE, we use the posterior mode instead of the median, because the posterior mode asymptotically converges to MLE. We summarize the R' posterior modes in terms of relative bias and MSE for the same R' data sets. We calculate the coverage probability as the proportion of times that the 95% HPD intervals cover the true value.

3.2. *Simulation results*

Table I shows a summary of the simulation results comparing convergence rate, relative bias, and coverage probability by different methods and by different sample sizes for the four distance-odds models (i.e., one point source PCM, ACM and homogeneous ACM, and two point sources homogeneous ACM). We summarize the MSE comparison in Figure 1. Because the three frequentist methods of MLE, profile likelihood, and IRLS regression provide very similar and consistent results, we primarily focus on the difference between the broad class of frequentist and Bayesian approaches, which is described in the following text in terms of convergence, relative bias, MSE, and coverage probability separately. Additionally, the following results hold for $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}$'s. The complete numerical simulation results are shown in Tables A.1–A.6 (Supporting information).

Convergence. For all four distance-odds models with a large sample size such as $N = 2000$, the frequentist methods perform well in terms of convergence with a joint convergence rate $R'/R > 90\%$. Typically, less than 5% of runs failed to converge for each of the three frequentist methods. With a decreased sample size of $N = 500$, the 90% joint convergence rate remains for the two homogeneous models. However, failures increase to 30% for one point source PCM and ACM using frequentist methods. Thus, we performed and presented the simulations for these two models for a sample size of $N = 1000$ in Table I,

Table I. Summary of the simulation results in terms of convergence rate, relative bias, and coverage probability comparing frequentist and Bayesian methods using different sample sizes.

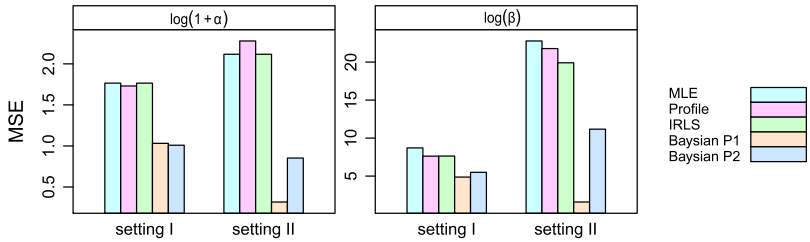
		Frequentist method				Bayesian method			
		Lack of	RB[a] (%)			Lack of	RB (%)		
N = 2000		convergence[a] (%)	α	β	CP[a](%)	convergence	α	β	CP (%)
One point source	ACM (H)[b]	3	(−0.5, 6.9),	(−0.4, 1.3)	93.8(91, 97)	None noted	(−8.3, 7.6),	(−0.4, 3.9)	93.2(81, 98)
	ACM	8	(−3.4, 9.6),	(−1.9, 1.7)	95.2(90, 98)	None noted	(−9.2, 7.7),	(−1.1, 3.2)	94.1(92, 98)
	PCM	9	(−0.2, 4.6),	(−0.9, 2.1)	93.7(91, 98)	None noted	(−4.2, 4.9),	(−0.9, 4.2)	93.6(89, 97)
Two point source	ACM (H)	5	(0.4, 10.7),	(−0.3, 0.6)	94.6(89, 97)	None noted	(−9.2, 9.2),	(−2.9, 2.8)	94.9(84, 99)
N = 500 or 1000[c]									
One point source	ACM (H)	8	(9.0, 25.6),	(−0.8, 1.7)	93.2(84, 99)	None noted	(−15.8, 17.6),	(−0.8, 6.0)	94.3(67, 99)
	ACM	15	(3.3, 17.5),	(−0.6, 1.4)	93.6(80, 99)	None noted	(−20.1, 15.7),	(−4.4, 8.3)	95.5(86, 100)
	PCM	13	(1.2, 14.9),	(−0.2, 0.8)	93.6(89, 100	None noted	(−14.8, 14.1),	(−3.4, 4.1)	95.5(88, 99)
Two point source	ACM (H)	10	(10.0, 23.1),	(−0.4, 0.8)	94.5(88, 99)	None noted	(−24.3, 18.9),	(−4.6, 4.2)	95.6(81, 100)

[a]Lack of convergence: mean of the none-convergence rates $(R - R')/R$ across parameter settings; RB: range of the relative biases across parameter settings; CP: mean and the range of the coverage probabilities across parameter settings.

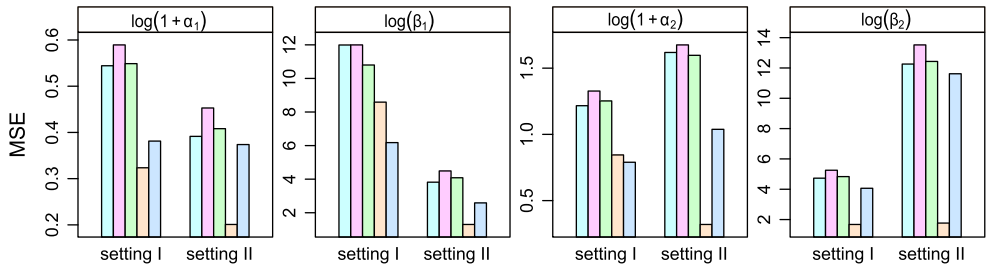
[b]Homogeneous adjacent-category model.

[c]$N = 500$ for the two homogeneous models; $N = 1000$ for one point source PCM and ACM.

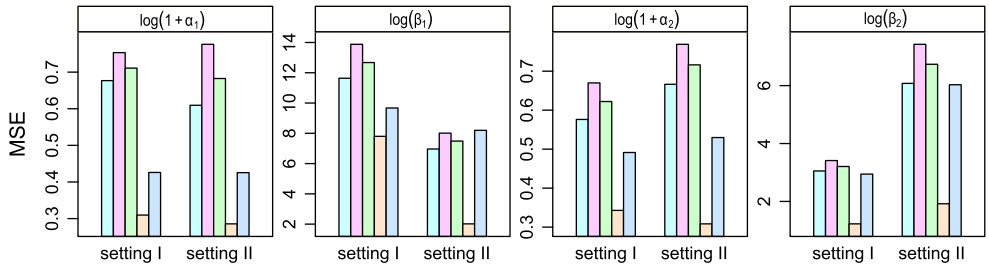
(a) one point source homogeneous adjacent category model. Sample size $N = 500$; Settings I and II refer to $(\alpha, \beta) = (0.7, 300)$ and $(0.4, 500)$ respectively.



(b) one point source adjacent category model. Sample size $N = 1000$; Settings I and II refer to $(\alpha_1, \beta_1, \alpha_2, \beta_2) = (0.3, 300, 0.7, 500)$ and $(0.4, 500, 0.4, 500)$ respectively.



(c) one point source polychotomous category model. Sample size $N = 1000$; Settings I and II refer to $(\alpha_1, \beta_1, \alpha_2, \beta_2) = (0.3, 300, 0.7, 500)$ and $(0.4, 500, 0.4, 500)$ respectively.



(d) two point sources homogeneous adjacent category model. Sample size $N = 500$; Settings I and II refer to $(\alpha_{11}, \beta_{11}, \alpha_{12}, \beta_{12}) = (0.5, 500, 0.3, 300)$ and $(0.4, 500, 0.4, 500)$ respectively.

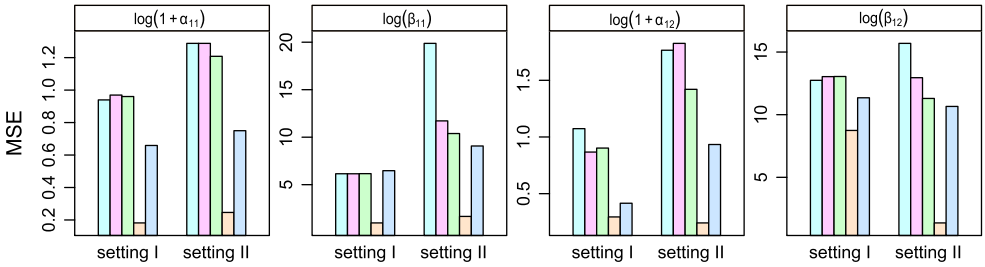


Figure 1. Mean squared errors for two settings of true parameter values under various distance-odds models, using MLE, profile likelihood, IRLS, and Bayesian methods with $R = 500$ simulations. Bayesian P1 and P2 refer to two choices of prior distributions; Prior 1: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.25, 150)$; Prior 2: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.5, 400)$. Y-axis (MSE values) is scaled by a multiplier of 100.

where a joint convergence rate of 85% occurs using frequentist methods. In the Bayesian approach, we numerically assessed the convergence of the posterior chains by the Gelman–Rubin convergence diagnostic [27]. We detected no problems either numerically or via examining the trace plots in our limited simulation study. The MCMC method does not require the usual regularity conditions [22] or any asymptotic normality assumption, and it yields exact posterior distributions for all sample sizes. It also avoids the identifiability issue but needs a careful choice of the covariance matrix of the proposal distribution because of the strong correlations among the model parameters.

Relative bias. When $N = 2000$, we observe low relative biases (with range $(-9.2, 10.7)\%$ for $\boldsymbol{\alpha}$'s and $(-2.9, 4.2)\%$ for $\boldsymbol{\beta}$'s) for both frequentist and Bayesian methods for all models with different choices of true parameter settings (shown in Table I; numerical details shown in Tables A.1 and A.2 (Supporting information)). Thus, both methods have performed well with large sample size in terms of relative bias. For smaller sample sizes ($N = 500$ for the two homogeneous models; $N = 1000$ for one point source PCM and ACM), relative biases of $\boldsymbol{\alpha}$ are usually as high as 25%, whereas relative biases of $\boldsymbol{\beta}$ are still well controlled ($< 5\%$, except few extreme setting). Note that estimates of $\boldsymbol{\alpha}$ are biased upwards (Table I) using frequentist methods with these small sample sizes, whereas Bayesian methods do not suffer as much. The aforementioned results are consistent across inference methods for each model as shown in Table I (numerical details shown in Tables A.3–A.6 (Supporting information)).

Mean squared error. When the sample size $N = 2000$, the MSEs are consistent across methods for each distance-odds models with different true parameters. Figure 1 shows the MSEs corresponding to each method with smaller sample sizes of $N = 500$ or 1000. The three frequentist approaches using MLE, profile, and IRLS method show very similar MSE values, whereas the Bayesian approach shows consistently lower MSEs than frequentist approach for each distance-odds model regardless of true parameters values. Note that, for the Bayesian approach, the MSEs derived from informative priors are much lower than those from noninformative (vague) priors for each setting as expected. Thus, if prior knowledge is available, it should be used to enhance precision for these distance-odds models.

Coverage probability. In Table I, when $N = 2000$, the coverage probabilities are around 95% for all the models and methods in our simulation study. For smaller sample sizes of $N = 500$ or 1000, the coverage probabilities fall below the nominal level for some parameter settings; however, they are still around 95% on average (shown in Table I; numerical details shown in Tables A.3–A.6 (Supporting information)). Note that these percentages are estimated based on the R' data sets where all three frequentist methods converge. In addition, the Bayesian approach provides comparable percentages based on all $R = 500$ data sets. Therefore, it is more stable than the frequentist methods in terms of coverage probability and convergence.

In summary, Bayesian methods, especially incorporated with prior knowledge, have advantages in terms of estimation stability and precision for the proposed nonlinear distance-odds models with multiple disease subtypes.

4. A case study: the Detroit asthma morbidity, air quality and traffic study

The present study describes a population-based matched case–control analysis investigating associations between acute asthma outcomes and proximity of residence to major roads in Detroit, MI.

4.1. Study design: health data and distance measurements

We examined the pediatric population ($2 - 18$ years of age) served by Medicaid for the study period from 2004 through 2006. The Medicaid data provide the most complete and readily available source of healthcare utilization across Detroit. The population consists mainly of African American children from lower income families and is considered a high-risk population for asthma-related events [30]. The data included an encrypted Medicaid identifier, age, sex, race/ethnicity, utilization dates, and diagnostic codes for inpatient admissions and emergency department visits, and geocoded home residence at the time of each healthcare visit. To ensure a full claims history, the study population was restricted to those with continuous Medicaid enrollment (more than 11 months in each year), full Medicaid coverage, and no other insurance. Asthma cases were identified as all children who made at least one asthma claim during the 3-year study period, indicated by primary diagnostic code 493.X (International Classification of Diseases, 9th Revision, Clinical Modification). Controls were defined as children whose primary diagnosis was injury or poisoning. Each asthma case was matched with one control on the basis of gender, race, and age (within 2 years). Asthma cases were further grouped into multiple disease categories ($K = 2$), based on the frequency of acute asthma outcomes ($Y = 2$ for claimants with two or more asthma claims; $Y = 1$ for claimants with exactly one asthma claim). We can find details on the descriptive analysis of this data set in [31].

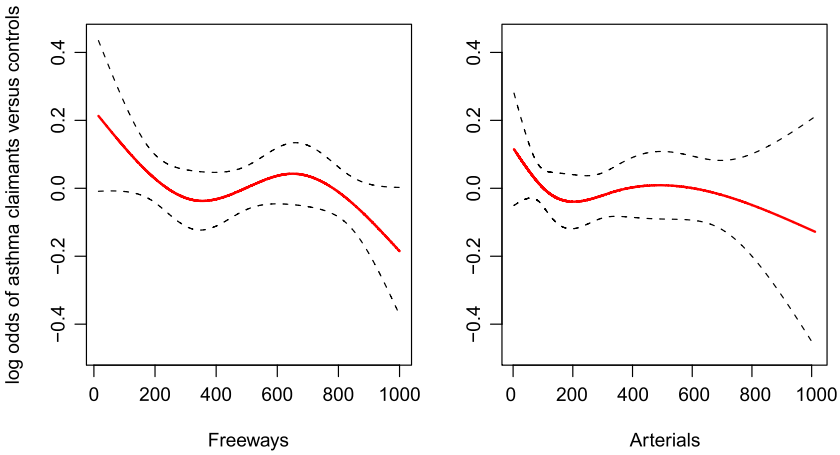


Figure 2. Estimated natural spline terms of distance showing the distance-odds relationships for asthma claimants versus controls, using (binary) conditional logistic regression model with spline of distance as its argument. The solid lines show the point estimates; the dashed lines show the 95% confidence bands.

The geocoded residence information was used to estimate the distance to major roads in Detroit, defined as state and interstate freeways and major arterials with annual average daily traffic flows exceeding 50,000 and 20,000 vehicles per day, respectively. The freeways and the arterials are considered as the first and second point sources, respectively. Shape files providing coordinates of road centerlines were obtained from the Southeast Michigan Council of Governments. These files and the geocoded claim data were merged into ARCGIS 9.3 (Environmental Systems Research Institute, Redlands, CA, USA) to determine the proximity to each major road. Because of confidentiality concerns, claim locations were reported only to the closest 10 m. The road centerline does not account for the width of the highway and median strip, if any, which can exceed 30 m for sections of some freeways. Taken together, these factors suggested that differences on the order of at least 20 to 50 m would be meaningful.

4.2. Results and discussion

We performed separate analyses for one and two point source(s) models. For one point source (freeways) models, the study region was restricted to 1000 m buffer of freeways, which consisted of 2669 1:1 matched case–control pairs. For two point sources (freeways and arterials) models, the study region was restricted to 1000 m buffer of freeways or arterials, which consisted of 4081 1:1 matched case–control pairs. Figure 2 illustrates the natural spline fit and 95% confidence band for the relationship between distance to roadways and odds of being an asthma claimant, using a CLR model with only spline of distance as its argument. These plots provide an exploratory analysis of the data, which indicate increasing risk with proximity to both types of roads, where the freeways appear to have stronger effects. There may be a threshold distance beyond which the roadway effect vanishes. The increase of odds at 600 m of freeways is not statistically significant, which could be an artifact of the smoothing parameter ($df = 3$ in the natural spline).

Method comparison. The frequentist methods of MLE, profile likelihood, and IRLS provide similar point estimates and CIs with essentially the same AIC values for each distance-odds model (Tables A.7 and A.8 (Supporting information)). Thus, we primarily discuss results as frequentist method (MLE as demonstration) versus Bayesian method in the main text. Table II shows the parameter estimates and 95% CIs using likelihood method and posterior modes with 95% HPD intervals using Bayesian methods, for one point source models. Additionally, the corresponding contour plots of the conditional log-likelihood surfaces for these one point source models are shown in Figure A.1 (Supporting information). Note that these log-likelihood surfaces are not far from quadratic in shape given the large sample size of 2669 asthma cases in the DAMAT study. Note also that the contour lines near $u = 0$ (or equivalently $\alpha = 0$) are almost vertical, which implies the identifiability issue that a wide range of β can provide the same value of likelihood values. Fortunately, the peaks of the likelihood surfaces are not close to the null for these one point source models. For the Bayesian method, estimated marginal posterior densities for one

point source models are shown in Figures A.2 and A.3 (Supporting information), where the locations of the posterior modes are close to each other for the two prior choices for each parameter under each model. Posterior densities of $\boldsymbol{\beta}$ are highly right skewed, especially for noninformative prior distribution with much wider HPDs than those derived from informative priors (shown in Table II). Thus, the frequentist likelihood-based inference method or a noninformative Bayesian method should be avoided for these distance-odds models in presence of well-elicited prior knowledge.

Model selection. Generally, the distance-odds models are selected *a priori* in the study design stage. For example, different choices of the numbers of point sources would provide different study regions with different sample sizes. As discussed in Section 2.2, the choice between PCMs and ACMs can also be considered *a priori* on the basis of the interest of nominal or ordered disease subclassifications. Model selection can also be based on AICs for frequentist method or DICs for Bayesian method. For example, ACM (homogeneous) has the smallest AIC value among the four one point source models as shown in Table II. However, the differences among these AICs are very small and of little practical concern. In this case, all these one point source models fit almost equally well. For both informative and noninformative priors, one point source PCM and ACM have similar and relatively lower DIC values than the other two models. There is evidence that the more sophisticated models that allow different functional forms of odds between case subtypes are preferred even after penalizing for the additional number of parameters using the Bayesian approach. Therefore, a PCM (smallest DIC) with informative priors is the preferred approach among all one point source models for the DAMAT study (different numbers of point sources with different sample sizes are not directly comparable). Similarly, Table III shows the corresponding results for the two point sources binary model and homogeneous ACM, where the latter with an informative prior Bayesian approach is preferred.

Estimation and interpretation. Table II shows the parameter estimates and 95% CIs using MLE, and posterior modes with 95% HPD intervals using Bayesian methods, for the one point source models (binary/ACM/PCM). Generally, the point estimates of $\hat{\alpha}$ and $\hat{\beta}$ lay within 0.1–0.4 and 100–300 respectively for the one point source models, which implies that the roadway effect on asthma only lasts up to a few hundred meters and that the increase in risk is modest. Take the one point source PCM that has the smallest DIC as an example, the MLE (or posterior mode) $\hat{\alpha}_2 = 0.39(0.32)$ is slightly larger than $\hat{\alpha}_1 = 0.21(0.25)$ as shown in Table II. It implies that, at the point source, the odds of asthma for claimants with two or more claims ($k = 2$) versus controls is slightly higher than the odds for claimant with exactly one claim ($k = 1$) versus controls. Table III shows the results for two point sources models. In general, we have $\hat{\alpha}_{11} > \hat{\alpha}_{12}$ and $\hat{\beta}_{11} > \hat{\beta}_{12}$, which implies that the odds of asthma at freeways is higher than the odds at arterials and the freeways effects last longer than arterials. Figure 3 shows the estimated distance-odds functions $\hat{\boldsymbol{f}}_{\boldsymbol{k}}$ for the one point source PCM, using MLE and Bayesian method with informative priors. Note that the Bayesian method with prior knowledge provides consistently higher estimates of $\boldsymbol{f}_{\boldsymbol{k}}$ than MLE. For both case subgroups, $\hat{\boldsymbol{f}}_{\boldsymbol{k}}$ deceases rapidly within 0–300 m, and then the roadway effect on asthma lasts up to 400 m off freeways using MLE method and 600 m using Bayesian method, respectively. The 95% credible regions are above unity up to a distance of 350 m. Note that the MLE of $\boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is estimated by plugging in the MLE of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ using their invariant property; the posterior distribution of $\boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is estimated by draws from the posterior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for fixed grid values of distance x (every 0.5 m). Note also that, for interval estimates of a function of parameters, the 95% Bayesian credible region can be directly obtained from the draws; however, the calculation of the frequentist confidence bands for the MLE of $\boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is not straight forward. This requires the Delta theorem (calculation of the first and second derivatives of the complex likelihood function) and relies on asymptotic properties needing a large sample size.

Table IV shows the p-values of the Monte Carlo test and the Bayes factors for testing $H_0 : f(x) = 1$ for one and two point source(s) distance-odds models. Evidence of associations ($H_1 : f(x) > 1$) is found for most models using the MC test (p-value < 0.05) or Bayes factors ($B > 2$). Strongest associations are found for PCM among one point source models and for homogeneous ACM among two point sources models respectively, which is consistent with the results shown in Tables II and III.

Sensitivity analysis. The results in Tables II–IV show consistency for different choices of the distance-odds models under a matched case–control study. Similar conclusion can be drawn using these models that there is evidence of the roadway effect on asthma and that the effect is modest and only lasts up to

Table II. Parameter estimates with 95% confidence intervals for one point source models using MLE, profile likelihood, and IRLS methods; and posterior modes with 95% highest posterior density (HPD) credible intervals using MCMC.

MLE[a]	Binary model	α	β			AIC
	Estimate	0.258	174.1			3699.9
	CI[a]	(−0.042, 0.558)	(55.7, 292.4)			
	ACM (homogeneous)	α_1	β_1			
	Estimate	0.188	168.8			3699.8
	CI	(−0.023, 0.398)	(57.8, 279.8)			
	ACM (general)	α_1	β_1	α_2	β_2	
	Estimate	0.215	176.0	0.130	153.4	3703.7
	CI	(−0.126, 0.557)	(41.6, 310.4)	(−0.484, 0.744)	(87.5, 394.2)	
	PCM	α_1	β_1	α_2	β_2	
	Estimate	0.208	191.8	0.392	154.1	3703.6
	CI	(−0.118, 0.534)	(9.1, 374.6)	(−0.242, 1.025)	(26.5, 281.7)	

Bayesian P1[a]	Binary model	α	β			DIC
	Posterior mode	0.247	228.6			3686.2
	Posterior median	0.289	290.7			
	CI (HPD)[a]	(0.034, 0.487)	(121.0, 592.1)			
	ACM (homogeneous)	α_1	β_1			
	Posterior mode	0.177	182.7			3686.8
	Posterior median	0.156	202.5			
	CI (HPD)	(0.025, 0.361)	(118.1, 550.0)			
	ACM (general)	α_1	β_1	α_2	β_2	
	Posterior mode	0.194	192.5	0.242	222.5	3675.3
	Posterior median	0.244	287.2	0.261	326.8	
	CI (HPD)	(0.004, 0.461)	(125.5, 667.8)	(−0.072, 0.505)	(116.7, 623.3)	
	PCM	α_1	β_1	α_2	β_2	
	Posterior mode	0.246	231.3	0.320	259.0	3674.9
	Posterior median	0.298	256.8	0.366	269.7	
	CI (HPD)	(0.028, 0.514)	(113.3, 737.8)	(0.049, 0.649)	(115.3, 602.6)	
Bayesian P2[a]	Binary model	α	β			DIC
	Posterior mode	0.285	152.7			3682.7
	Posterior median	0.203	398.2			
	CI (HPD)	(0.027, 1.308)	(154.6, 1401.2)			
	ACM (homogeneous)	α_1	β_1			
	Posterior mode	0.192	160.5			3683.2
	Posterior median	0.216	395.2			
	CI (HPD)	(0.005, 1.086)	(79.3, 1263.2)			
	ACM (general)	α_1	β_1	α_2	β_2	
	Posterior mode	0.212	177.5	0.162	172.5	3670.4
	Posterior median	0.312	425.7	0.188	256.7	
	CI (HPD)	(−0.039, 0.896)	(96.5, 1347.8)	(−0.181, 0.840)	(39.5, 1123.9)	
	PCM	α_1	β_1	α_2	β_2	
	Posterior mode	0.258	243.2	0.286	147.0	3669.4
	Posterior median	0.346	566.4	0.367	218.9	
	CI (HPD)	(−0.056, 0.822)	(76.5, 1490.9)	(−0.080, 0.818)	(44.9, 1267.8)	

[a]MLE, maximum likelihood estimate; CI, confidence/credible interval; HPD, highest posterior density; Bayesian P1 and P2 refer to two settings of prior choice; Prior 1: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.25, 150)$; Prior 2: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.5, 400)$.

Statistics in Medicine

Table III. Parameter estimates with 95% confidence intervals for two point sources models using MLE, profile likelihood, and IRLS methods; and posterior modes with 95% highest posterior density (HPD) credible intervals using MCMC.

		First point source		Second point source		
MLE[a]	Binary model	α_{11}	β_{11}	α_{12}	β_{12}	AIC
	Estimate	0.228	309.2	−0.098	180.5	5657.5
	CI	(−0.177, 0.663)	(97.3, 575.8)	(−0.420, 0.223)	(17.4, 376.2)	
	ACM (homogeneous)	α_{11}	β_{11}	α_{12}	β_{12}	
	Estimate	0.179	283.6	−0.134	114.8	5656.0
	CI	(0.001, 0.360)	(68.4, 535.1)	(−0.357, 0.093)	(6.5, 233.2)	
Bayesian P1[a]	Binary model	α_{11}	β_{11}	α_{12}	β_{12}	DIC
	Posterior mode	0.280	257.9	0.061	270.0	5609.1
	Posterior median	0.304	302.1	0.089	300.1	
	CI (HPD)[a]	(0.127, 0.462)	(171.9, 533.2)	(−0.072, 0.200)	(132.1, 543.7)	
	ACM (homogeneous)	α_{11}	β_{11}	α_{12}	β_{12}	
	Posterior mode	0.205	294.0	0.019	261.4	5593.8
	Posterior median	0.212	360.2	0.021	340.4	
	CI (HPD)	(0.075, 0.354)	(155.0, 509.8)	(−0.083, 0.122)	(143.1, 633.2)	
Bayesian P2[a]	Binary model	α_{11}	β_{11}	α_{12}	β_{12}	DIC
	Posterior mode	0.248	327.5	0.007	182.3	5604.6
	Posterior median	0.303	430.2	0.011	434.6	
	CI (HPD)	(0.069, 0.474)	(122.9, 627.2)	(−0.131, 0.134)	(75.1, 1225.7)	
	ACM (homogeneous)	α_{11}	β_{11}	α_{12}	β_{12}	
	Posterior mode	0.186	228.0	−0.006	149.4	5595.2
	Posterior median	0.222	340.4	0.011	480.9	
	CI (HPD)	(0.051, 0.354)	(129.0, 645.8)	(−0.120, 0.108)	(70.1, 1243.2)	

[a]MLE, maximum likelihood estimate; CI, confidence/credible interval; HPD, highest posterior density; Bayesian P1 and P2 refer to two settings of prior choice; Prior 1: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.25, 150)$; Prior 2: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.5, 400)$.

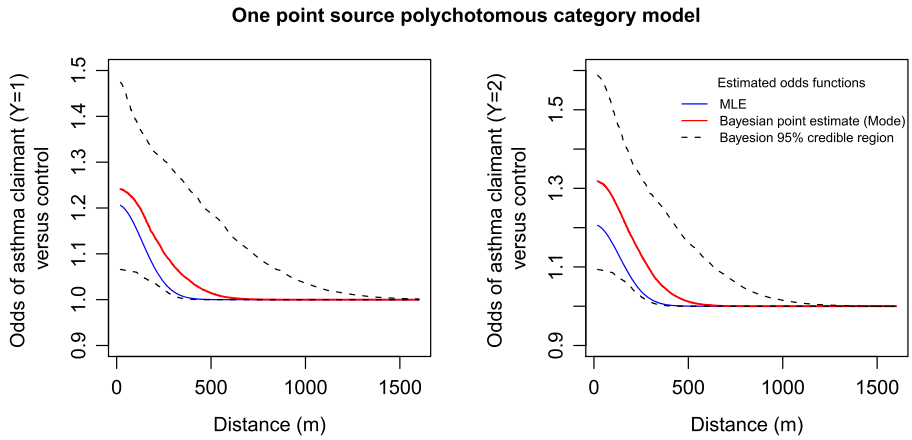


Figure 3. Estimated distance-odds functions for the one point source polychotomous-category model. The solid blue line shows the MLE of the odds function; the solid red line shows the Bayesian posterior mode estimate with 95% credible region (dashed lines). Parameters of prior distribution used are $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.25, 150)$.

Table IV. Monte Carlo test p-values and Bayes factors $2\log(B)$ for the null hypothesis that $H_0 : f(x) = 1$ for various point source(s) models.

	MC test	Bayes factors	
Model	p-value	P1	P2
One point source			
Binary model	0.04	3.52	2.89
ACM (homogeneous)	0.06	4.32	3.41
ACM (general)	0.02	6.29	6.16
PCM	<0.01	7.12	6.04
Two point source			
Binary model	0.04	3.11	2.57
ACM (homogeneous)	<0.01	6.69	5.98

Bayesian P1 and P2 refer to two settings of prior choice; Prior 1: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.25, 150)$; Prior 2: $(\mu_\alpha, \mu_\beta) = (0.5, 400)$ and $\left(\sigma_\alpha^2, \sigma_\beta^2\right) = (0.5, 400)$.

a few hundred meters. As a sensitivity analysis of the prior specification, posterior densities are derived and compared from different choices of prior distributions for the one point source PCM. For normal priors on (u, v) with different variances $\left(\sigma_u^2, \sigma_v^2\right)$, the posterior modes are close to each other for each parameter under each model shown in Figure 4. However, the posterior modes are sensitive to the choice of $\alpha_{\max}$ and $\beta_{\max}$ using Uniform priors on $(-1, \alpha_{\max})$ and $(0, \beta_{\max})$. When $\alpha_{\max}$ and $\beta_{\max}$ are large, these Uniform priors still put equal weights on the whole range of $(-1, \alpha_{\max})$ and $(0, \beta_{\max})$ that may overly weight the upper extreme values. Wakefield and Morris [16] have also pointed out the influence of the Uniform priors, which reflects the fact that there is little information in the likelihood as a result of sparsity of data in the upper extremes. Thus, the parameterization (u, v) with normal priors appear to be more robust.

5. Discussion

In this paper, we extended the distance-odds model of Diggle *et al.* [14] to models where there are subtypes within cases under a matched case–control design. The extension to subclassification within cases is nontrivial with these nonlinear odds functions under a matched design. Maximum likelihood, profile likelihood, IRLS, and a Bayesian approach using MCMC methods were evaluated under the proposed models. We compared these methods via an extensive simulation study evaluating frequentist properties, such as relative bias, MSE ,and coverage probability, and showed that Bayesian methods have

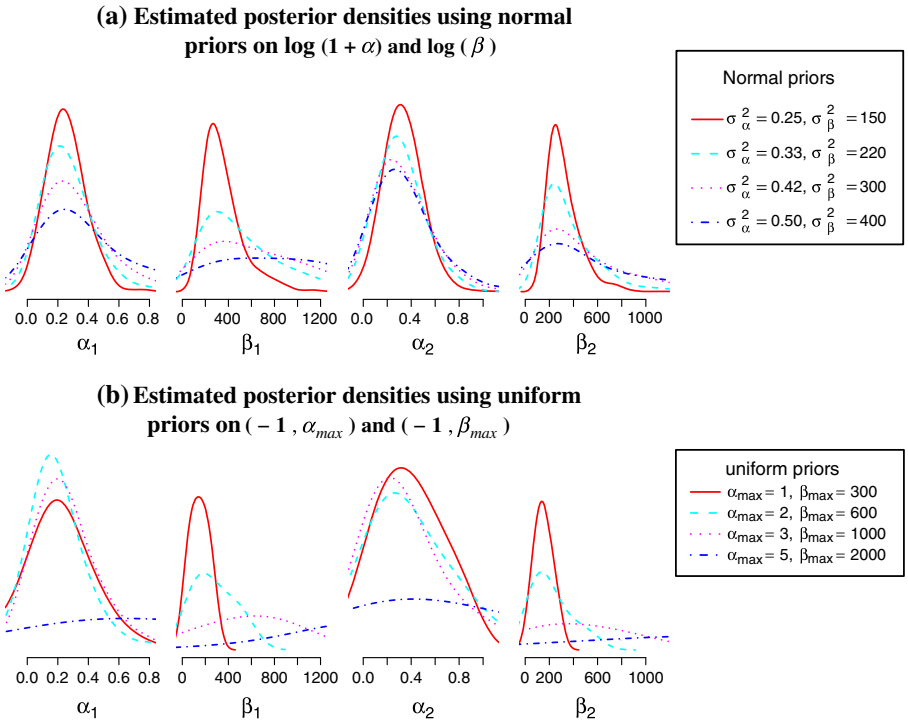


Figure 4. Estimated posterior densities for different settings of prior choices for the one point source polychotomous-category model for the Detroit Medicaid data, as a sensitivity analysis.

advantages in terms of estimation stability, precision, and interpretation. The Bayesian methods are able to yield direct HPD for complex nonlinear distance-odds functions and does not require large sample approximation. There is no simulation study in the literature that compares the convergence, relative bias, MSE, or coverage probability for these point source models, even for the basic binary outcome model. We apply the proposed models and methods to a population-based matched case–control study investigating associations between acute asthma outcomes and proximity of residence to major roads by analyzing Medicaid claims data for the pediatric asthma population in Detroit, MI, from 2004 to 2006. We also perform a sensitivity analysis to investigate how the choice of distance-odds models and specification of the prior distributions affect the results. Typically, the results were consistent for different choices of models and normal prior distributions on the transformed parameters for the DAMAT study.

We did not consider the extension of the nonlinear distance-odds model to the proportional odds model setting in the study, which is most commonly used for ordered data. We realize that the conditional likelihood does not apply to this model because of the nuisance parameters remaining in the nonlinear odds functions. Moreover, the prospective–retrospective conversion for case–control data is only valid for a multiplicative intercept model. In addition, the residual spatial correlations can be modeled either parametrically or semiparametrically. These issues remain to be explored in future research.

Acknowledgements

We appreciate the help from Robert Wahl, Elizabeth Wasilevich, and Erika Garcia who contributed to the overall DAMAT study design and the Medicaid data ascertainment and use and the help from Huda Elasaad who provided the distance measurement using ARCGIS 9.3 desktop software. Although portions of the research described in this article have been funded in part by the United States Environmental Protection Agency through grant EPA-G2007-STAR-A1 to Science to Achieve Results (STAR) Program: Development of Environmental Health Outcome Indicators, it has not been subjected to the agency's required peer and policy review and therefore does not necessarily reflect the views of the agency and no official endorsement should be inferred. The research of Bhramar Mukherjee was partially supported by the NSF grant DMS 1007494.

References

1. Breslow NE, Day NE, Halvorsen KT, Prentice RL, Sabai C. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* 1978; **108**:299–307.
2. Liang KY, Stewart W. Polychotomous logistic regression methods for matched case-control studies with multiple case or control groups. *American Journal of Epidemiology* 1987; **125**:720–730.
3. Becher J, Jockel KH. Bias adjustment with polychotomous logistic regression in matched case-control studies with two control groups. *Biometrical Journal* 1990; **7**:801–816.
4. Becher H. Alternative parameterization of polychotomous models: theory and application to matched case-control studies. *Statistics in Medicine* 1991; **10**:375–382.
5. Thomas DC, Goldberg M, Dewar R, Siemiatycki J. Statistical methods relating several exposure factors to several diseases in case-heterogeneity studies. *Statistics in Medicine* 1986; **5**:49–60.
6. Durbin N, Pasternack BS. Risk assessment for case-control subgroups by polychotomous logistic regression. *American Journal of Epidemiology* 1986; **6**:1101–1117.
7. Sinha S, Mukherjee B, Ghosh M. Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics* 2004; **60**:41–49.
8. Mukherjee B, Liu I, Sinha S. Analysis of matched case-control data with multiple ordered disease states: possible choices and comparisons. *Statistics in Medicine* 2007; **26**:3240–3257.
9. Mukherjee B, Ahn J, Liu I, Sanchez BN. On elimination of nuisance parameters in stratified proportional odds model by amalgamating conditional likelihoods. *Statistics in Medicine* 2008; **27**:4950–4971.
10. Diggle PJ. A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *Journal of the Royal Statistical Society A* 1990; **153**:349–362.
11. Lawson AB. On the analysis of mortality events associated with a pre-specified fixed point. *Journal of the Royal Statistical Society A* 1993; **156**:363–377.
12. Diggle PJ, Rowlingson BS. A conditional approach to point process modeling of raised incidence. *Journal of the Royal Statistical Society A* 1994; **157**:433–440.
13. Diggle PJ, Elliott P, Morris SE, Shaddick G. Regression modeling of disease risk in relation to point sources. *Journal of the Royal Statistical Society A* 1997; **160**:491–505.
14. Diggle PJ, Morris SE, Wakefield J. Point-source modeling using matched case-control data. *Biostatistics* 2000; **1**:89–105.
15. Diggle PJ, Moyeed RA, Tawn JA. Model-based geostatistics (with discussion). *Applied Statistics* 1998; **47**:299–350.
16. Wakefield JC, Morris SE. The Bayesian modelling of disease risk in relation to a point source. *Journal of the American Statistical Association* 2001; **96**:77–91.
17. Lawson AB, Browne WJ, Vidal Rodeiro CL. *Disease Mapping with WinBugs and MlWin*. Wiley: New York, 2003.
18. Congdon P. *Applied Bayesian Modelling*. Wiley: New York, 2003.
19. Dreassi E, Lagazio C, Maule M, Magnani C, Biggeri A. Sensitivity analysis of the relationship between disease occurrence and distance from a putative source of pollution. *Geospatial Health* 2008; **2**:263–271.
20. Rodrigues A, Diggle PJ, Assuncao R. Semi-parametric approach to point source modeling in epidemiology and criminology. *Journal of the Royal Statistical Society C* 2010; **59**:533–542.
21. Agresti A. *Categorical Data Analysis*. Wiley: New York, 2002.
22. Breslow NE, Day NE. Statistical methods in cancer research. Volume I – The analysis of case-control studies. *IARC Scientific Publications* 1980; **32**:335–338.
23. Barnard GA. Contribution to the discussion of Professor Bartlett's paper. *Journal of the Royal Statistical Society B* 1963; **25**:294.
24. Gallant AR. *Nonlinear Statistical Models*. Wiley: New York, 1987.
25. Rice KM. Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association* 2004; **99**:510–522.
26. Rice KM. Equivalence between conditional and random-effects likelihoods for pair-matched case-control studies. *Journal of the American Statistical Association* 2008; **103**:385–396.
27. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–511.
28. Kass RE, Raftery AE, Bayes factors. *Journal of the American Statistical Association* 1995; **90**:773–795.
29. Diciccio TJ, Kass RE, Raftery AE, Wasserman L. Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* 1997; **92**:903–915.
30. Wu YC, Batterman S. Proximity of schools in Detroit, Michigan to automobile and truck traffic. *Journal of Exposure Science and Environmental Epidemiology* 2006; **16**:457–470.
31. Li S, Batterman S, Wasilevich E, Elasaad H, Wahl R, Mukherjee B. Asthma exacerbation and proximity of residence to major roads: a population-based matched case-control study among the pediatric Medicaid population in Detroit, Michigan. *Environmental Health* 2011; **10**:34. DOI: 10.1186/1476-069X-10-34.