Published in final edited form as:

Biometrics. 2014 December ; 70(4): 981–992. doi:10.1111/biom.12198.

A Bayesian Nonparametric Model for Spatially Distributed Multivariate Binary Data with Application to a Multidrug-Resistant Tuberculosis (MDR-TB) Study

Jian Kang^{1,†}, Nanhua Zhang^{2,†}, and Ran Shi¹

Jian Kang: jian.kang@emory.edu

¹Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, U.S.A

²Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, U.S.A

Summary

There has been an increasing interest in the analysis of spatially distributed multivariate binary data motivated by a wide range of research problems. Two types of correlations are usually involved: the correlation between the multiple outcomes at one location and the spatial correlation between the locations for one particular outcome. The commonly used regression models only consider one type of correlations while ignoring or modeling inappropriately the other one. To address this limitation, we adopt a Bayesian nonparametric approach to jointly modeling multivariate spatial binary data by integrating both types of correlations. A multivariate probit model is employed to link the binary outcomes to Gaussian latent variables; and Gaussian processes are applied to specify the spatially correlated random effects. We develop an efficient Markov chain Monte Carlo algorithm for the posterior computation. We illustrate the proposed model on simulation studies and a multidrug-resistant tuberculosis case study.

Keywords

Bayesian methods; Drug resistance; Gaussian processes; Spatially distributed multivariate binary data

1. Introduction

It is common to observe spatially distributed multivariate binary data in many research areas, such as dental research (Bandyopadhyay, Reich, and Slate, 2009), toxicology (Davidov and Peddada, 2011), ecology (Dormann, 2007) and environmental research (Wall and Liu, 2009), where multiple binary outcomes are observed for subjects from different

© 2014, The International Biometric Society

Correspondence to: Jian Kang, jian.kang@emory.edu.

[†]Equally contributed.

The web supplementary materials including Web Appendices A and B referenced in the article are available at the *Biometrics* website on Wiley Online Library. The R and C++ source code along with example data are available at the webpage <http://web1.sph.emory.edu/users/jkang30/software/BayesSpatMultiBinary.html>. The code provides R and C++ functions to perform posterior computations for the proposed Bayesian model for spatially distributed multivariate binary data.

spatial locations. It usually involves two types of correlations: the correlation between the multiple outcomes within the same subject (at one location) and the spatial correlation between subjects (multiple locations) for one outcome. Most existing statistical methods (Cox, 1972; Liang et al., 1992; Carey et al., 1993; Chib and Greenberg, 1998; Bandyopadhyay et al., 2009; Franzese and Hays, 2009; Davidov and Peddada, 2011) dealing with multivariate binary data only consider one type of correlation while ignoring or inappropriately modeling the other one. To address this limitation, in this article, we propose a Bayesian nonparametric model for spatially distributed multivariate binary outcomes motivated by the analysis of a dataset collected from San Juan de Lurigancho (SJL), Peru to study the drug resistance in the treatment of multidrug-resistant tuberculosis (MDR-TB).

1.1. Multidrug-Resistant Tuberculosis (MDR-TB)

Tuberculosis (TB) is a common infectious disease that claims an estimated 1.7 million lives each year and the number of new cases was more than nine million in 2011 (Lawn and Zumla, 2011). It creates huge burdens across the globe, especially in developing countries due to higher incidence rates (Kumar, Abbas, and Aster, 2012). The recommended treatment for new onset of tuberculosis is to use a combination of antibiotic containing rifampin (RIF) along with isoniazid (INH), pyrazinamide (PZA), and ethambutol (EMB) (Lawn and Zumla, 2011). However, drug resistance is very common in tuberculosis treatment (Dye et al., 2002). According to the World Health Organization (WHO, 2010), 3.6 % of all TB cases are estimated to have MDR-TB. Although there have been a few studies on the mechanism of drug resistance in tuberculosis (Al-Orainey, 1990; Crofton et al., 1997; Rodrigues, Gomes, and Rebelo, 2007), why tuberculosis is resistant to a certain treatment is largely unknown. It is well known that drug resistance of TB is unevenly distributed and therefore MDR-TB is perceived as problems of local rather than global importance (Dye et al., 2002). Studying the distribution of drug resistance over different regions is particularly important in guiding treatment of MDR-TB in a certain region. By correctly choosing an effective drug for a certain region, we expect to treat the patients more efficiently and save the cost of applying drugs that patients are more likely to be resistant to. Spatial modeling and prediction of drug resistance can be helpful for treatment decision-making in some high MDR-TB burden regions where funding for universal resistance testing is not available (Resch et al., 2006).

The motivating data were collected from a study of a cohort of patients diagnosed with pulmonary TB and MDR-TB over an 18-month period in San Juan de Lurigancho (SJL), Peru (Jacob et al., 2010). Only patients with no prior history of TB were included in this study. In Peru, patients with newly diagnosed TB were usually treated with first-lines drugs administered under directly observed therapy (DOTS) (WHO, 2010; Resch et al., 2006). DOTS-Plus, which entailed the addition of second-line drugs, was suggested on patients with long-standing disease due to highly resistant strains of *Mycobacterium tuberculosis* (Mitnick et al., 2003). Community-based therapy for MDR-TB could lead to variation in drug resistance pattern over different regions (Mitnick et al., 2003). Eligible subjects received an explanation of the study and provided written consents to participate. Initial data from the screening included past medical history, demographic information including age, gender, occupation, address etc. Drug susceptibility testing for isoniazid (INH), rifampin (RIF), ethambutol (EMB), and streptomycin (SM) was performed on the initial sputum

culture isolates of all enrolled subjects. Subjects with initial drug-resistant tuberculosis isolates were confirmed and treated using a treatment regimen with duration deemed appropriate by the Committee of the National Tuberculosis Control Programme (NTCP) and the Committee for Evaluation of Retreatment (CER) in Peru.

Geocoordinates, including latitude, longitude, were recorded for subjects in this dataset. All subjects started anti-TB chemotherapy after collection of baseline samples and completion of initial measurements. The outcomes of interest are the drug sensitivity to four first-line drugs: INH, RIF, EMB, and SM in Lowenstein–Jensen (LJ) medium. For each subject, the demographic information is also collected including age, gender, marital status, family size (number of persons living in the house), and whether the subject works in a health care center (yes, or no).

The goal of the analysis is to study the spatial distribution and dependence of the MDR-TB and identify the important demographic effects on the MDR-TB. There are two sources of dependence that need to be considered. First, patients in the same region might share similar drug resistance profiles. This is in part due to the infectious nature of MDR-TB, and patients with MDR-TB might be infected from the same sources. Also, patients in the same region share many other characteristics that affect their immune systems. These suggest the drug resistance should be spatially correlated. Second, the resistance to one drug might be correlated with that of another drug because some chromosomal mutations might lead to resistance to multiple drugs (Wade and Zhang, 2004).

1.2. Multivariate Binary Data Analysis

A wide range of statistical methods have been proposed for the modeling of multivariate binary data. Carey, Zeger, and Diggle (1993) proposes alternating logistic regression that permits simultaneously regressing the response on explanatory variables and modeling the association among responses, using generalized estimating equations (Liang and Zeger, 1986). Alternatively, multivariate probit models (Ashford and Sowden, 1970; Amemiya, 1974) have been proposed. They characterize the multivariate binary response using a correlated Gaussian distribution for underlying latent variables that are manifested as discrete variables through a threshold specification. The Bayesian analysis of multivariate probit model becomes increasingly popular due to the development of efficient posterior computational algorithms for the model fitting (Chib and Greenberg, 1998) through Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990; Albert and Chib, 1993). Those methods are not directly applicable to the MDR-TB problem, since they are not designed to take into account of the two types of correlations.

Many techniques have been developed for modeling spatially dependent multivariate binary data. These methods include autologistic models (Besag, 1975; Dormann, 2007; Bandyopadhyay et al., 2009), the generalized linear mixed-effects model with spatial random errors (Diggle, Tawn, and Moyeed, 1998), the multivariate probit model with spatial random errors (Weir and Pettitt, 1999; Franzese and Hays, 2009). Autologistic models establish a correspondence between the binary response and the explanatory variables through a logistic regression and account for spatial correlation using autoregression. However, autologistic models are associated with complications due to intractable

normalizing factor in a fully Bayesian framework (Bandyopadhyay et al., 2009). Therefore, some models resort to pseudo-likelihood (Weir and Pettitt, 1999) and may give biased estimates (Dormann, 2007). Many of these methods are developed for areal data, which are defined as a district or a region (such as zip code). Wall and Liu (2009) proposed a spatial latent class model for spatially distributed multivariate binary data. It extends the classical latent class model by adding spatial structure to the latent class distribution through the use of the multinomial probit model. This model did not adjust for the covariate effects and used computationally intensive cross-validation procedures to choose the number of mixture components.

To overcome the limitations of the current methods, we propose a Bayesian nonparametric regression model for the spatially distributed multivariate binary data based on Gaussian processes (GPs). GPs have received attentions from both statistics and machine-learning communities (Higdon, Swall, and Kern, 1999; Rasmussen and Williams, 2006). They have been widely used in nonparametric regression, classification problem and spatio-temporal modeling, especially in modeling spatial random effects (Banerjee, Gelfand, and Carlin, 2003; Banerjee et al., 2008) due to its attractive theoretical properties and the availability of efficient computational algorithms to fit these models. The GP model fitting for spatial statistics may suffer from large computational burden when the number of spatial locations is large. A common approach to tackling this problem is to seek approximations to the spatial process by kernel convolutions, low rank splines or basis functions (Rasmussen and Williams, 2006; Banerjee et al., 2008) so that the original GP is replaced by a stochastic process in a lower dimensional subspace. Other methods proposed for this “large n ” problem include the use of approximation to the likelihood or using a Markov random field (Rue and Tjelmeland, 2002) to approximate the random-field model. The approximation to the likelihood approach could suffer from inadequacy, especially for multivariate processes. The use of the Markov random field is best suited for points on a regular grid, and could introduce unquantifiable errors in precision with irregular locations. In this article, we take the kernel convolution to reduce the dimension. This approach provides good accuracy and only requires a moderate computational cost, given a set of kernels.

Our proposed model has the following notable features: (1) it models the spatial effects through a stationary GP on the multiple drug resistance of TB and is able to adjust for covariate effects; (2) it takes into account the spatially distributed between-drug correlations that provide richer information for understanding the multiple drug resistance in different regions; (3) the computational cost of our model fitting is relatively small for a large spatial data set.

The rest of this article is organized as follows. In Section 2, we introduce a Bayesian nonparametric model for the spatially distributed multivariate binary data. We discuss the model properties and the posterior computation strategy. Then, we analyze the MDR-TB data using the proposed method in Section 3. Section 4 presents simulation studies that demonstrate the performance of the proposed method. Section 5 concludes the article with some discussions.

2. Model

Suppose our dataset consists of n subjects. For each subject i , we have q binary outcomes, denoted by $\boldsymbol{\delta}(\mathbf{s}_i) = (\delta_1(\mathbf{s}_i), \dots, \delta_q(\mathbf{s}_i))'$, observed at a spatial location $\mathbf{s}_i \in \mathbb{R}^d$, where $\mathbb{R}^d$ is a d -dimensional Euclidean space and $\delta_j(\mathbf{s}_i) \in \{0, 1\}$. We also have p covariates denoted by $\mathbf{w}_i = (w_{i1}, \dots, w_{ip})'$. We denote the entire observed data by $\mathcal{D} = \{\boldsymbol{\delta}(\mathbf{s}_i), \mathbf{w}_i\}_{i=1}^n$.

2.1. A Spatial Multivariate Probit Model

For each subject $i = 1, \dots, n$, we model the conditional probability mass function of $\boldsymbol{\delta}(\mathbf{s}_i)$ given the covariates effect $\mathbf{g}(\mathbf{w}_i)$, by introducing spatial effects $\mathbf{f}(\mathbf{s}_i)$ and subject effect $\boldsymbol{\eta}_i$. Specifically, we have

$$\Pr\{\boldsymbol{\delta}(\mathbf{s}_i) = \mathbf{k} | \mathbf{g}(\mathbf{w}_i)\} = \mathbb{E}[\Pr\{\boldsymbol{\delta}(\mathbf{s}_i) = \mathbf{k} | \mathbf{f}(\mathbf{s}_i), \boldsymbol{\eta}_i, \mathbf{g}(\mathbf{w}_i)\}], \quad (1)$$

where $\mathbf{k} = (k_1, k_2, \dots, k_q)'$ with $k_j \in \{0, 1\}$, $\mathbf{g}(\mathbf{w}_i) = (g_1(\mathbf{w}_i), \dots, g_q(\mathbf{w}_i))'$, $\mathbf{f}(\mathbf{s}) = (f_1(\mathbf{s}), \dots, f_q(\mathbf{s}))'$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{iq})'$. The expectation $\mathbb{E}[\cdot]$ in (1) is taken with respect to the joint distribution of $\mathbf{f}(\mathbf{s}_i)$ and $\boldsymbol{\eta}_i$. In this model, $f_j(\mathbf{s}_i)$, $g_j(\mathbf{w}_i)$ and η_{ji} are the spatial effects, covariate effects and subject effects on binary outcome $\delta_j(\mathbf{s}_i)$ respectively. Note that the covariate effects $\mathbf{g}(\mathbf{w}_i)$ are spatially independent and the subject effects $\boldsymbol{\eta}_i$ is introduced to characterize the marginal correlation between $\delta_1(\mathbf{s}_i)$, $\delta_2(\mathbf{s}_i)$, $\dots$, and $\delta_q(\mathbf{s}_i)$.

We assume that given $\mathbf{f}(\mathbf{s}_i)$, $\mathbf{g}(\mathbf{w}_i)$ and $\boldsymbol{\eta}_i$, $\delta_1(\mathbf{s}_i)$, $\dots$, $\delta_q(\mathbf{s}_i)$ are conditionally independent of one another. Furthermore, we assume

$$\Pr\{\boldsymbol{\delta}(\mathbf{s}_i) = \mathbf{k} | \mathbf{f}(\mathbf{s}_i), \boldsymbol{\eta}_i, \mathbf{g}(\mathbf{w}_i)\} = \prod_{j=1}^q \Pr\{\delta_j(\mathbf{s}_i) = k_j | f_j(\mathbf{s}_i), \eta_{ji}, \mathbf{g}_j(\mathbf{w}_i)\}. \quad (2)$$

For each $j = 1, \dots, q$, we model $\delta_j(\mathbf{s}_i)$ using a generalized linear model through a probit link function given $g_j(\mathbf{w}_i)$, $f_j(\mathbf{s}_i)$ and η_{ji} ,

$$\Phi^{-1}[\Pr\{\delta_j(\mathbf{s}_i) = k_j | f_j(\mathbf{s}_i), \eta_{ji}, \mathbf{g}_j(\mathbf{w}_i)\}] = f_j(\mathbf{s}_i) + \mathbf{g}_j(\mathbf{w}_i) + \eta_{ji}, \quad (3)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The observed data likelihood based on (1)–(3) is computationally intractable for a large spatial dataset. To address this issue, for each subject i , we introduce a set of latent independent normal variables $\mathbf{z}_i = (z_{i1}, \dots, z_{iq})^T$ to resolve model (1)–(3). We have for $i = 1, \dots, n$ and $j = 1, \dots, q$,

$$[\delta_j(\mathbf{s}_i) | \mathbf{z}_{ij}] = \mathbf{I}[\mathbf{z}_{ij} > 0], [\mathbf{z}_{ij} | \mathbf{f}_j(\mathbf{s}_i), \mathbf{g}_j(\mathbf{w}_i), \eta_{ji}] \sim \text{Normal}(f_j(\mathbf{s}_i) + \mathbf{g}_j(\mathbf{w}_i) + \eta_{ji}, \mathbf{1}), \quad (4)$$

where indicator function $\mathbf{I}[\mathcal{A}] = 1$ if event $\mathcal{A}$ occurs, $\mathbf{I}[\mathcal{A}] = 0$ otherwise. By integrating out $\mathbf{z}_{ij}$, model (4) reduces to (3).

2.2. Prior Specification

2.2.1. Gaussian processes—We start from a brief overview of the Gaussian processes (Rasmussen and Williams, 2006). A Gaussian process is defined as a collection of random variables, any finite collection of which have a joint Gaussian distribution. A Gaussian process $f(\mathbf{x})$ on space $\mathbb{R}^d$ is completely specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$, denoted $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. By *Mercer's theorem*, (Higdon et al., 1999; Rasmussen and Williams, 2006), the covariance function can be decomposed as

$k(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{\infty} v_l \psi_l(\mathbf{x}) \psi_l(\mathbf{x}')$, where $\{v_l\}_{l=1}^{\infty}$ are the eigenvalues and $\{\psi_l(\mathbf{x})\}_{l=1}^{\infty}$ are the eigenfunctions. They satisfy $\int k(\mathbf{x}, \mathbf{x}') \phi_l(\mathbf{x}) d\mathbf{x} = v_l \psi_l(\mathbf{x}')$ and $v_l > 0$ for $l = 1, 2, \dots$. This decomposition leads to an equivalent model representation of the Gaussian process:

$f(\mathbf{x}) = \sum_{l=1}^{\infty} \zeta_l \psi_l(\mathbf{x})$, where $\zeta_l \sim \text{Normal}(0, v_l)$. This representation involves an infinite number of parameters that has a nature of the nonparametric model. In practice, we are interested in approximating $f(\mathbf{x})$ on a finite number of points $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$, which is given by

$$f(\mathbf{x}_i) \approx \sum_{j=1}^L \zeta_j \tilde{\psi}_j(\mathbf{x}_i) \text{ with } \zeta_j \sim \text{Normal}(0, \tilde{v}_j), \quad (5)$$

where $\{\tilde{\psi}_j\}_{j=1}^L$ with $\tilde{\boldsymbol{\psi}} = (\psi_1(\mathbf{x}_1), \dots, \psi_1(\mathbf{x}_n))^T$ and $\{\tilde{\psi}_j\}_{j=1}^L$ are respectively the eigenvectors and the eigenvalues of the covariance matrix $\{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$. The number of components L is usually chosen such that $\sum_{j=1}^L \tilde{v}_j / \sum_{j=1}^n \tilde{v}_j > \alpha$ for a given $\alpha \in (0, 1)$. We refer to (5) as an *eigen decomposition approximation* to $f(\mathbf{x}_i)$ for $i = 1, \dots, n$. Next, we discuss the prior specifications for spatial effects in the model.

2.2.2. Priors for spatial effects—We assume that the spatial effects $f_1(\mathbf{s}), \dots, f_q(\mathbf{s})$ are mutually independent at a location $\mathbf{s} \in \mathbb{R}^d$. And for $j = 1, \dots, q$, $f_j(\mathbf{s})$ is a Gaussian process defined on $\mathbb{R}^d$, that is,

$$[f_j(\mathbf{s}) | \sigma_j^2] \sim \mathcal{GP}\{\mathbf{f}_0(\mathbf{s}), \sigma_j^2 \kappa(\mathbf{s}, \mathbf{s}')\}, \quad (6)$$

where $f_0(\mathbf{s})$ is a mean function shared by different $f_j(\mathbf{s})$'s and σ_j^2 is the variance of $f_j(\mathbf{s})$. The kernel correlation function $\kappa(\mathbf{s}, \mathbf{s}')$ is set to be the same for different $f_j(\cdot)$'s and it characterizes the smoothness of the GPs. Based on (5), we consider the eigen decomposition approximation to $f_j(\mathbf{s}_i)$, for $j = 1, \dots, q$ and $i = 1, \dots, n$, that is,

$$f_j(\mathbf{s}_i) \approx \sum_{l=1}^L \alpha_{jl} \phi_l(\mathbf{s}_i) = \boldsymbol{\alpha}_j^T \boldsymbol{\phi}_i \text{ with } \boldsymbol{\alpha}_j \sim \text{Normal}(\boldsymbol{\alpha}_0, \sigma_j^2 \boldsymbol{\Lambda}), \quad (7)$$

where $\alpha_j = (\alpha_{j1}, \dots, \alpha_{jn})^T$ for $j = 1, \dots, q$, $\phi_i = (\phi_1(\mathbf{s}_i), \dots, \phi_L(\mathbf{s}_i))^T$ and $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_L\}$. $\{\lambda_l\}_{l=1}^L$ and $\{(\phi_l(\mathbf{s}_1), \dots, \phi_l(\mathbf{s}_n))^T\}_{l=1}^L$ are respectively the eigen values and eigen vectors of the covariance matrix $\{\kappa(\mathbf{s}_i, \mathbf{s}_j)\}_{1 \leq i, j \leq n}$ with $\lambda_l = \lambda_{l+1}$ for $l = 1, \dots, L$.

2.2.3. Priors for covariate effects—The covariate effects g_j can take any form, however, in this application, we particularly consider a linear model and assign a normal prior to the coefficient, that is,

$$g_j(\mathbf{w}_i) = \beta_j^T \mathbf{w}_i, \beta_j \sim \text{Normal}(\beta_0, \Omega_j), \quad (8)$$

where β_0 represents the a priori common covariate effect on the multiple binary outcomes. The covariance matrix Ω_j measures the deviation of the effect on $\delta_j(\mathbf{s}_i)$ from the common effect. Although the prior specification for η can be very flexible, we assume the subject effects are identical to different binary outcomes and assign a normal prior, that is,

$$\eta_i = \mathbf{1}_q \eta_i, \eta_i \sim \text{Normal}(0, \sigma_\eta^2), \quad (9)$$

where $\mathbf{1}_q$ is a q dimensional column vector with all elements being 1's and σ^2 characterizes the variability of the subject effects. For the hyperpriors, we assume

$$\begin{aligned} \beta_0 &\sim \text{Normal}(\mathbf{0}, \Omega_0), \alpha_0 \sim \text{Normal}(\mathbf{0}, \sigma_0^2 \mathbf{I}_L), \\ \sigma_\eta^2 &\sim \text{Inv-Gamma}(a_\eta, b_\eta), \sigma_j^2 \sim \text{Inv-Gamma}(a, b), \\ \text{and } \Omega_j &\sim \text{Inv-Wishart}(\mathbf{S}_\Omega, \nu_\Omega), \text{ for } j=0, 1, \dots, q, \end{aligned}$$

where $\mathbf{I}_L$ is an L dimensional identity matrix.

2.3. Model Properties and Posterior Inference

Our model introduces the *spatial correlations* for each outcome, that is, for $j = 1, \dots, q$, and $i, i' \in \{1, \dots, n\}$,

$$\text{cov}[\delta_j(\mathbf{s}_i), \delta_j(\mathbf{s}_{i'})] = \text{cov}[\Phi(\alpha_j^T \phi_i + \beta_j^T \mathbf{w}_i + \eta_i), \Phi(\alpha_j^T \phi_{i'} + \beta_j^T \mathbf{w}_{i'} + \eta_{i'})], \quad (10)$$

where α_j and β_j introduce the correlation, since η_i and $\eta_{i'}$ are independent. The subject-specific variable ϕ_i and $\mathbf{w}_i$ mainly affect the values of correlations which vary for different binary outcomes.

In addition, our model is able to characterize the correlations between different binary responses over space. We refer them as *between-outcome correlations*, that is, a priori marginal correlation structure between $\delta_j(\mathbf{s}_i)$'s. Specifically, given a spatial location $\mathbf{s}_i$, for $i = 1, \dots, n$, we have for any $j, k \in \{1, \dots, q\}$,

$$\text{cor}[\delta_j(\mathbf{s}_i), \delta_k(\mathbf{s}_i)] = \text{cov}[\Phi(\alpha_j^T \phi_i + \beta_j^T \mathbf{w}_i + \eta_i), \Phi(\alpha_k^T \phi_i + \beta_k^T \mathbf{w}_i + \eta_i)]. \quad (11)$$

By integrating out all other parameters in the hierarchical model, α_j and α_k are correlated as well as β_j and β_k . Also the subject effect η_i contributes to the outcome correlations. This further implies that these correlations vary over space.

For the posterior computation, we resort to Gibbs sampling to simulate the joint posterior distribution of $\{\mathbf{z}_i\}_{i=1}^n, \{\beta_j\}_{j=0}^q, \{\alpha_j\}_{j=0}^q, \{\eta_i\}_{i=1}^n, \{\Omega_j\}_{j=0}^q, \{\sigma_j^2\}_{j=0}^q$ and σ^2 given the spatial smoothing parameter ρ in the covariance kernel $\kappa(\mathbf{s}, \mathbf{s}')$. The full conditionals of the parameters and the corresponding sampling schemes are provided in Web Appendix A in the Web Supplementary Materials. The spatial correlation ρ is chosen by maximizing the marginal likelihood profile of ρ which is estimated by the Monte Carlo simulation.

3. Analysis of the MDR-TB Dataset

In this section, we illustrate our method on the analysis of the multi drug resistant tuberculosis (MDR-TB) data from San Juan de Lurigancho (SJL), Peru, as described in Section 1. The spatial multivariate binary responses are the drug susceptibility test results (0 = sensitive and 1 = resistant) of four first-line drugs (isoniazid, rifampin, ethambutol, and streptomycin) on the initial sputum culture isolates of all 780 enrolled subjects, with latitude and longitude recorded for each subject. The covariates considered in this article include: gender (male/female), age at enrollment (years), marriage (married or living with partner/otherwise), family size (number of persons living in the house ranging from 1 to 20), and whether the subject works in a health care center (yes/no). The questions of interest include: (1) what do the spatial distributions of the drug resistance look like for these four drugs? (2) are the four drug resistance correlated with each other? (3) do the MDR-TB outcomes depend on any other factors?

To address the above questions, we apply our model by choosing $L = 36$ in (7) such that $\sum_{l=1}^{36} \lambda_l / \sum_{l=1}^{780} \lambda_l > 0.9$. We ran the posterior computation algorithm with 60,000 iterations and a burn-in period of 15,000. For each model parameter, the posterior convergence was assessed using trace plots, auto-correlation plots, as well as the Gelman-Rubin convergence diagnostics (Gelman and Rubin, 1992) by running multiple chains. All chains of fixed effect parameters became stable after 15,000 iterations. The potential scale reduction factors are all close to 1. We used the remaining 45,000 to make posterior inference on the parameters of our interest. The posterior predictive distribution of the spatial random effects, as well as the pairwise spatial correlations were also obtained. Table 1(A) presents the mean predicted probabilities of drug resistance for each district in SJL, Peru. Table 1(B) presents the estimates of key parameters in the model. Figures 1 and 2 present the spatial random effects and the pairwise correlations for the four drug resistance outcomes over space, respectively.

3.1. Spatial Effects

Figure 1 shows the spatial effects for four drug resistance outcomes. As the three drugs (INH, RIF, and EMB) have strong spatial correlations, the spatial effects of the three drugs

are similar across regions. Large spatial effects were found in the West side boundary (mainly District Santa Rosa and District Ayacucho) for these three drugs, implying increased probability of drug resistance to these three drugs in these regions. Patients in the south of this region (District Caja de Agua) are less likely to have resistance to the drug RIF.

Table 1(A) shows the mean predicted probabilities of drug resistance across different districts. For drug INH, RIF, and EMB, high probabilities of drug resistance were found in District Campoy, Azcarruz Alto, Sta Fe de Totoritas, Santa Rosa, and Ayachucho, while low probabilities of drug resistance were found in the following Districts: Caja de Agua, Piedra Liza, Bayovar, Mangamarca, 15 de Enero. For the drug SM, District Sta Fe de Totoritas has the highest mean probability of resistance. This is consistent with estimated spatial effects.

3.2. Dependence of Multi Drug Resistance

Figure 2 presents the pairwise correlations for the four drug resistance outcomes over space. We identify a strong correlation among INH, RIF, and EMB over space, implying that the drug resistance to these three drugs are similar over regions. Meanwhile, we see a relatively small spatial correlation of resistance to SM with the other three drugs. Identifying drugs that have resistance profiles different from other common drugs is very useful from the clinical perspective, because those are more likely to be the effective alternative when the patients have resistance to common drugs.

3.3. Demographic Variable Effects

Table 1(B) presents the posterior mean, and 95% credible intervals of key fixed-effects parameters. Age is found to be a significant predictor of drug resistance for all four drugs, which implies that young patients are associated with a high chance of drug resistance. Being married significantly increases the chance of having resistance to the drug INH; large effects of being married were also found for the drug RIF and EMB though the effects were not significant. Working in a health care center decreases the chance of drug resistance but it was not statistically significant. The effect of age is consistent with previous studies (Faustini, Hall, and Perucci, 2006).

3.4. Model Assessments

We perform the posterior predictive checking using the χ^2 discrepancy (Gelman, Meng, and Stern, 1996). In our model, it is defined as

$$\chi^2(\delta, \theta) = \sum_{i=1}^n \sum_{j=1}^q \frac{\{\delta_j(s_i) - \Phi(\alpha_j^T \phi_i + \beta_j^T \mathbf{w}_i + \eta_i)\}^2}{\Phi(\alpha_j^T \phi_i + \beta_j^T \mathbf{w}_i + \eta_i) [1 - \Phi(\alpha_j^T \phi_i + \beta_j^T \mathbf{w}_i + \eta_i)]},$$

where $\delta = \{\delta(s_i)\}_{i=1}^n$ denotes the observed data and θ is a collection of all the parameters in model (1), including $\{\alpha_j\}_{j=1}^q$, $\{\beta_j\}_{j=1}^q$ and $\{\eta_i\}_{i=1}^n$. Let $\{\theta^{(k)}\}_{k=1}^K$ be the set of posterior samples of parameters we obtained from the MCMC algorithm. For each $k = 1, \dots, K$, we sample the data from model (1), denoted $\delta^{(k)}$, then we obtain the posterior predictive

p -value estimated by $\hat{p} = \frac{1}{K} \sum_{k=1}^K I[\chi^2(\delta^{(k)}, \theta^{(k)}) > \chi^2(\delta, \theta^{(k)})] = 0.347$. This implies that our model fits data well.

3.5. Model Comparisons

To demonstrate the superiority of our proposed model for the analysis of MDR-TB data, we also analyze the data using other models. We compare the fitting results and evaluate the drug resistance prediction accuracy for different models. To be more specific, we consider four probit regression models with or without taking into account different types of correlations in the data. We refer to our proposed model as **Model 1**, and we define three other simpler models below:

- **Model 2:** Multiple univariate probit regression models (one for each drug) including the spatial random effects but ignoring the between-drug correlations;
- **Model 3:** A multivariate probit regression model incorporating the between-drug correlations but ignoring spatial random effects;
- **Model 4:** Multiple univariate probit regression models (one for each drug) ignoring both spatial correlations and between-drug correlations;

where Models 2–4 are the simpler models nested within our proposed model (Model 1) by ignoring one or two types of correlations.

We perform a comparison of the posterior inference on covariate effects from four models (see Web Tables 2–4 in the Web Supplementary Materials for the results by the simpler models). We find that the significant effects of some demographic variables on the drug resistance is only detected by the proposed model and cannot be identified by the other simpler models. For example, the proposed model fitting implies that “Age” and “Family Size” are strongly associated with the SM resistance, and “Marital Status” is significantly associated with the INH resistance, while all the simpler models do not detect those associations.

In Section 4, we conduct additional simulation studies on assessing the parameter estimates of covariate effects for all models, where we show that our proposed model achieves the best performance among the four. For the simpler models, we also check the goodness of fit using the same approach for Model 1 in Section 3.4, the posterior predictive p -values for Models 2, 3, and 4 are, respectively, 0.366, 0.058, and 0.004, suggesting that Models 3 and 4 do not well fit the data as both of them ignore the spatial effects. Although the posterior predictive p -value suggests that Model 2 also have a good fit of the data, we further demonstrate the superiority of the proposed model (Model 1) by computing the drug resistance prediction accuracy.

Specifically, a prediction on the drug resistance can be obtained by thresholding the posterior predictive probability at 0.5. We use the ten-fold cross validation by splitting the data into ten groups with nine groups as the training data of model fitting and one group as the test data for model validation. The accuracy for each drug fitted by Models 1–4 are listed in Table 1(C). The overall accuracies for the four models are respectively 0.88, 0.83, 0.78,

and 0.76, where Model 1 (the proposed model) has the highest prediction accuracy, Model 2 has a better accuracy than Model 3, and Model 4 has the lowest prediction accuracy. Thus, incorporating both spatial correlations and between-drug correlations into the model can substantially improve the model fitting and model prediction accuracy. This demonstrates that the proposed model is advantageous over the simpler models.

Moreover, we validate the assumption that the spatial correlation parameter ρ is homogeneous across outcomes by comparing the proposed model (Model 1) with an extension by introducing different correlation parameters for different outcomes. We refer to this model as Model 0. Models 0 and 1 provide a similar posterior inference on the demographic effects (Please refer to Web Table 1 in Web Supplementary Materials). Both of them demonstrate a good fit of the MDR-TB data with a similar posterior predictive p -values (0.347 vs. 0.356). Also, they have the same overall prediction accuracy of 0.88 on the drug resistance (see Table 1(C) for more details). Thus, the above model fitting and checking results are not sensitive to the homogeneous spatial correlation assumption, implying that the proposed model is valid for the MDR-TB data analysis.

4. Simulation Studies

In this section, we conduct simulation studies to evaluate model fitting performance.

4.1. Set Up

To generate simulated datasets, we randomly select n locations in $[0, 1]^2$. At each location, four binary outcomes are simulated from the model (4) given covariate effects and spatial effects. Three covariates are continuous, drawn independently from normal distributions with variance 1 and mean 1, 2, and 3, respectively. The other covariate is binary, following a Bernoulli distribution with probability 0.5. The covariate coefficients are set as $(-1, 0.5, 0, 0.5)$, $(-0.5, 0.25, 0, -1)$, $(-1, 0.25, 0, 1)$ and $(0.5, -0.25, 0, -0.5)$. For $i = 1, \dots, n$ and $j = 1, 2, 3, 4$, the spatial effects in the model are designed in the following two cases:

- Case I: Deterministic functions are specified, that is,

$$f_j(\mathbf{s}_i) = 6 \sin \left(2\pi \left(\mathbf{x}_{i1} + \frac{\pi \mathbf{j}}{4} \right) \right) \cos \left(2\pi \left(\mathbf{x}_{i2} + \frac{\pi \mathbf{j}}{4} \right) \right),$$

where $\mathbf{s}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2})$.

- Case II: Truncated Gaussian processes are drawn from (7), that is,

$$f_j(\mathbf{s}_i) = \sum_{l=1}^{50} \alpha_{jl} \phi_l(\mathbf{s}_i),$$

where $\alpha_{1l} = 3 \sin(l/8) + 3 \sin(l/10)$, $\alpha_{2l} = 3 \cos(l/8) + 3 \cos(l/10)$, $\alpha_{3l} = 3 \sin(l/8) + 3 \cos(l/10)$ and $\alpha_{4l} = 3 \cos(l/8) + 3 \sin(l/10)$.

For the above two cases, random effects η are independently drawn from Normal(0, 0.5) at each location. To determine the hyper-parameter ρ , we compute the marginal likelihood profile at a set of different values of ρ based on simulations. For Case II, the maximum value of log likelihood is achieved at $\rho = 0.19$, which is close to the truth $\rho = 0.2$. To perform the posterior inference, for both cases, we chose L such that $\sum_{l=1}^L \lambda_l > 0.85 \sum_{l=1}^n \lambda_l$. We set $\nu_{\Omega} = 6$ and $\mathbf{S}_{\Omega} = \mathbf{I}_4$ in the priors of the covariance matrices Ω_j . The priors for the variances σ^2 , σ_0^2 and σ_0^2 , $j = 1, 2, \dots, q$ are all set to be an inverse gamma distribution with shape 0.0001 and scale 0.0001. The initial values for all elements of $\{\mathbf{z}_i\}_{i=1}^n$, $\{\beta_j\}_{j=0}^q$, $\{\alpha_j\}_{j=0}^q$ and $\{\eta_i\}_{i=1}^n$ are set as zeros. The initial values of all Ω_j , for $j = 0, 1, \dots, 4$, are $0.6\mathbf{I}$. The initial values for all other variance elements are ones. All MCMC chains are updated for 10,000 iterations with a 2,000 burn-in.

4.2. Model Fitting

Figure 3 shows the estimated posterior mean of spatial effects over the entire space using cubic spline interpolations compared with the true spatial effects when the data sample size is 800. We can see that the posterior mean of the spatial effects map is quite close to the true spatial map in each case for each outcome.

We quantify the posterior inference accuracy on the model parameters by computing the difference between the posterior sample and the truth. We refer to this difference as the posterior sample bias. A good posterior inference on a parameter is reflected by the fact that the mean of posterior sample bias should be close to zero and its 95% credible intervals should cover zero. For spatial effects, we focus on the spatial map at locations where the true values achieve the maximum and the minimum. Table 2(A) presents the results for the two cases with different data sample size: 800, 1400 and 2000. The results show that the 95% credible intervals cover the truth at the extreme value locations. The widths of the credible intervals become smaller as the sample size increases. To evaluate the overall accuracy of the spatial effects for the entire space, we compute the estimated mean square error (MSE) for 8,000 posterior samples in Table 2(B). It is defined as

$$\hat{\text{MSE}}(\{\mathbf{F}_j^{(k)}\}_{k=1}^{8000}) = \frac{1}{nK} \sum_{k=1}^{8000} \|\mathbf{F}_j^{(k)} - \mathbf{F}_j\|_{\mathcal{F}}^2, j=1, 2, 3, 4,$$

where $\mathbf{F}_j^{(k)}$ is a 100×100 matrix representing the two-dimensional interpolated spatial map for outcome j based on the posterior sample k , for $k = 1, \dots, 8000$, and $\mathbf{F}_j$ represents the true spatial map. The matrix norm $\|\cdot\|_{\mathcal{F}}$ is the Frobenius norm. In Table 2(C), we summarize the posterior inference accuracy on the covariate effects with different data sample size. In particular, we focus on three types of effects: (1) small positive effects ($\beta_{22} = 0.25$); (2) small negative effects ($\beta_{42} = -0.25$); (3) no effects ($\beta_{33} = 0$). The results show that the estimated posterior means of the coefficients are very close to true values and all the 95% credible intervals cover the true values respectively. None of the 95% credible intervals for

the small positive effects and small negative effects cover zero. This implies that our proposed method has the power to detect those effects.

In summary, the simulation studies suggest that our methods provide very good estimates and inferences on the spatial effects and covariates effects for both cases I and II.

4.3. Model Comparisons

To demonstrate the advantages of the proposed model, we also fit the simulated datasets using the three simpler models (Models 2–4) defined in Section 3.5. Since not all the simpler models can make inference on the spatial correlations and between-outcome correlations, we primarily concentrate on the posterior inference on the covariate effects by comparing the squared bias and the MSE of the posterior sample for the covariate coefficients. Table 3 summarizes the comparisons for both cases I and II when the sample size is 800. The results suggest that our proposed model has the smallest MSE and the squared bias in both cases. The simplest model by ignoring both spatial correlations and between-outcome correlations has the largest MSE and squared bias. This implies that the modeling of both types of correlations is necessary for the posterior inference on the covariate effects, and it can substantially improve the accuracy of the posterior inference.

5. Discussion

In this article, we propose a Bayesian nonparametric model using Gaussian processes for the analysis of spatially distributed multivariate binary outcomes motivated by a multi-drug resistance tuberculosis problem. Using the MDR-TB incidence data from SJL, Peru, we demonstrate that our model takes into account two sources of dependence (1) the spatial dependence of one subject on other neighboring subjects, and (2) the correlation among the multiple binary responses from the same subject. Our model provides a good prediction for the spatial pattern of the drug resistance profiles over different regions. Our analysis shows a strong correlation among INH, RIF, and EMB over space, implying that the resistance profiles of these three drugs are similar across the study region. The correlations of SM with the other three drugs are low. This suggests that the SM could be an alternative when patients having high resistance to other three drugs. The estimated spatial effects for four drug resistance profiles also provide insights into the choice of a more effective drug for patients in that region. The fixed effect estimates show that age was a significant predictor of drug resistance for all four drugs, with younger age associated with increased chance of drug resistance. This result supports previous findings (Faustini et al., 2006).

In the motivating MDR-TB data, we completely collected all four outcomes at all locations of interest, however, we expect the proposed model to have even greater benefits for fitting the unbalanced data (i.e., some locations provide only a partial vector of the binary outcomes) since leveraging both between-drug and spatial correlations will yield better predictions for the missing outcomes at locations where not all binary outcomes are observed.

There are two possible future directions of this work. First, our modeling approach can readily be extended for the analysis of spatially distributed multiple discrete outcomes with

more than two categories or ordinal variables by using a different link function. Second, in this work, we analyzed subject-level data but it is also possible to group these patients in districts and model the multivariate responses at district level through a conditionally autoregressive (CAR) model. Similar to the model proposed in this article, the new model needs to account for both the spatial dependence of neighboring districts and the correlation among multivariate responses on the same individual.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are thankful to the Editor, the Associate Editor and an anonymous reviewer for their helpful and constructive comments that have led to a substantial improvement on this manuscript.

References

- Al-Orainey I. Drug resistance in tuberculosis. *Journal of Chemotherapy*. 1990; 2:147. [PubMed: 2116504]
- Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*. 1993; 88:669–679.
- Amemiya T. Bivariate probit analysis: Minimum chi-square methods. *Journal of the American Statistical Association*. 1974; 69:940–944.
- Ashford J, Sowden R. Multi-variate probit analysis. *Biometrics*. 1970; 26:535–546. [PubMed: 5480663]
- Bandyopadhyay D, Reich BJ, Slate EH. Bayesian modeling of multivariate spatial binary data with applications to dental caries. *Statistics in Medicine*. 2009; 28:3492–3508. [PubMed: 19902498]
- Banerjee, S.; Gelfand, AE.; Carlin, BP. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC; 2003.
- Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*. 2008; 70:825–848.
- Besag J. Statistical analysis of non-lattice data. *The Statistician*. 1975; 24:179–195.
- Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*. 1993; 80:517–526.
- Chib S, Greenberg E. Analysis of multivariate probit models. *Biometrika*. 1998; 85:347–361.
- Cox DR. The analysis of multivariate binary data. *Applied Statistics*. 1972; 21:113–120.
- Crofton, SJ.; Chaulet, P.; Maher, D.; Grosset, J.; Harris, W.; Horne, N.; Iseman, M.; Watt, B. *Guidelines for the Management of Dug-Resistant Tuberculosis*. World Health Organization; Geneva: 1997.
- Davidov O, Peddada S. Order-restricted inference for multivariate binary data with application to toxicology. *Journal of the American Statistical Association*. 2011; 106:1394–1404. [PubMed: 22973069]
- Diggle PJ, Tawn J, Moyeed R. Model-based geostatistics. *Journal of the Royal Statistical Society, Series C*. 1998; 47:299–350.
- Dormann CF. Assessing the validity of autologistic regression. *Ecological Modelling*. 2007; 207:234–242.
- Dye C, Williams BG, Espinal MA, Raviglione MC. Erasing the world's slow stain: Strategies to beat multidrug-resistant tuberculosis. *Science*. 2002; 295:2042–2046. [PubMed: 11896268]
- Faustini A, Hall AJ, Perucci CA. Risk factors for multidrug resistant tuberculosis in Europe: A systematic review. *Thorax*. 2006; 61:158–163. [PubMed: 16254056]

- Franzese R, Hays J. The spatial probit model of interdependent binary outcomes: Estimation, interpretation, and presentation. 2009 Working Paper.
- Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*. 1990; 85:398–409.
- Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*. 1996; 6:733–760.
- Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7:457–472.
- Higdon D, Swall J, Kern J. Non-stationary spatial modeling. *Bayesian Statistics*. 1999; 6:761–768.
- Jacob BG, Krapp F, Ponce M, Gotuzzo E, Griffith DA, Novak RJ. Accounting for autocorrelation in multi-drug resistant tuberculosis predictors using a set of parsimonious orthogonal eigenvectors aggregated in geographic space. *Geospatial Health*. 2010; 4:201–217. [PubMed: 20503189]
- Kumar, V.; Abbas, AK.; Aster, JC. Robbins Basic Pathology. Saunders: Elsevier Health Sciences; 2012.
- Lawn SD, Zumla AI. Tuberculosis. *The Lancet*. 2011; 378:57–73.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*. 1992; 54:3–40.
- Mitnick C, Bayona J, Palacios E, Shin S, Furin J, Alcántara F, Sánchez E, Sarria M, Becerra M, Fawzi MCS, Kapiga S, Neuberg D, Maguire JH, Kim JY, Farmer P. Community-based therapy for multidrug-resistant tuberculosis in lima, peru. *New England Journal of Medicine*. 2003; 348:119–128. [PubMed: 12519922]
- Rasmussen, CE.; Williams, C. Gaussian Processes for Machine Learning. The MIT press; 2006.
- Resch SC, Salomon JA, Murray M, Weinstein MC. Cost-effectiveness of treating multidrug-resistant tuberculosis. *PLoS Medicine*. 2006; 3:e241. [PubMed: 16796403]
- Rodrigues P, Gomes MGM, Rebelo C. Drug resistance in tuberculosis reinfection model. *Theoretical Population Biology*. 2007; 71:196–212. [PubMed: 17174368]
- Rue H, Tjelmeland H. Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*. 2002; 29:31–49.
- Wade MM, Zhang Y. Mechanisms of drug resistance in mycobacterium tuberculosis. *Frontiers in Bioscience: A Journal and Virtual Library*. 2004; 9:975–994. [PubMed: 14766424]
- Wall MM, Liu X. Spatial latent class analysis model for spatially distributed multivariate binary data. *Computational Statistics and Data Analysis*. 2009; 53:3057–3069. [PubMed: 20161235]
- Weir I, Pettitt A. Spatial modelling for binary data using a hidden conditional autoregressive Gaussian process: A multivariate extension of the probit model. *Statistics and Computing*. 1999; 9:77–86.
- WHO S. T. I. Treatment of Tuberculosis: Guidelines. World Health Organization; Geneva: 2010.

Figure 1.
Posterior mean of spatial effects for the resistance of the four drugs at each location.

Figure 2.
Estimated between-drug correlation over space.

Figure 3. Top four pairs show the estimated posterior mean of spatial effects for Case I and the bottom four are for Case II, where the sample size $n = 800$.

Table 1

Model fitting results for the MDR-TB dataset

(A) Estimated posterior probabilities of drug resistance for different regions									
District	INH	RIF	EMB	SM	District	INH	RIF	EMB	SM
San Fernando, CS	0.17	0.10	0.23	0.04	La Libertad, CS	0.11	0.06	0.17	0.02
La Huayrona, CS	0.11	0.07	0.16	0.03	Juan Pablo II, CS	0.16	0.11	0.22	0.04
Canto Grande, CS	0.15	0.10	0.21	0.05	Azcarruz Alto, CS	0.19	0.13	0.27	0.06
Jose C Mariategui, CS	0.15	0.09	0.24	0.03	de Octubre, PS	0.15	0.10	0.22	0.04
Huascar XV, CS	0.16	0.11	0.22	0.05	Sta Fe de Totoritas, PS	0.21	0.16	0.27	0.11
Huascar II, CS	0.11	0.08	0.15	0.03	Proyectos Especiales, PS	0.12	0.07	0.23	0.02
Ganimedes, CS	0.13	0.09	0.19	0.05	Santa Rosa, PS	0.33	0.24	0.40	0.11
Cruz de Motupe, CS	0.10	0.05	0.17	0.01	Ayacucho, PS	0.26	0.16	0.35	0.05
Piedra Liza, CS	0.10	0.05	0.17	0.02	Zarate, PS	0.15	0.09	0.21	0.05
Bayovar, CS	0.10	0.06	0.17	0.01	Medalla Milagrosa, PS	0.14	0.09	0.21	0.03
Jaime Zubieta, CS	0.15	0.08	0.23	0.03	Campoy Alto, CS	0.03	0.02	0.05	0.01
San Juan, CS	0.11	0.07	0.18	0.05	Montenegro, PS	0.15	0.07	0.21	0.02
Mangamarca, CS	0.09	0.04	0.17	0.01	Santa Maria, PS	0.17	0.09	0.27	0.04
San Hilarion, CS	0.17	0.11	0.25	0.03	Tupac Amaru II, PS	0.14	0.09	0.22	0.03
Campoy, CS	0.21	0.13	0.31	0.05	Caja de Agua, PS	0.11	0.05	0.19	0.02
de Enero, CS	0.06	0.02	0.11	0.01	San Benito	0.15	0.09	0.23	0.04

(B) Posterior means, posterior standard deviations and 95% credible intervals for the covariate effects on the multiple drug resistance

Effect	INH			RIF		
	Mean	S.D.	95% C.I.	Mean	S.D.	S.D.
Intercept	-1.293	0.470	(-2.227,-0.330)	-1.707	0.476	0.476
Gender (male/female)	0.189	0.244	(-0.289, 0.675)	0.102	0.255	0.255
Age (years)	-0.040	0.012	(-0.063,-0.016)	-0.035	0.012	0.012
Marital status (yes/no)	0.593	0.225	(0.165, 1.044)	0.371	0.245	0.245
Family size (1 – 20)	-0.014	0.031	(-0.079, 0.043)	-0.034	0.039	0.039
Work for health care (Yes/No)	-0.658	0.528	(-1.674, 0.347)	-0.759	0.580	0.580

(B) Posterior means, posterior standard deviations and 95% credible intervals for the covariate effects on the multiple drug resistance							
Effect	INH			RIF			SM
	Mean	S.D.	95% C.I.	Mean	S.D.	95% C.I.	
Effect	Mean	S.D.	95% C.I.	Mean	S.D.	95% C.I.	
Intercept	-1.223	0.449	(-2.132,-0.390)	-2.008	0.607	(-3.150,-0.734)	
Gender (male/female)	0.434	0.212	(0.050, 0.860)	-0.140	0.338	(-0.765, 0.529)	
Age (years)	-0.032	0.011	(-0.051,-0.011)	-0.034	0.018	(-0.070,-0.001)	
Marital status (yes/no)	0.350	0.220	(-0.089, 0.758)	0.082	0.315	(-0.536, 0.687)	
Family size (1 – 20)	0.018	0.032	(-0.044, 0.076)	-0.106	0.052	(-0.216,-0.015)	
Work for health care (yes/no)	-0.521	0.487	(-1.500, 0.383)	-0.062	0.616	(-1.204, 1.078)	

(C) Ten-fold cross-validation prediction accuracy on drug resistance between five different models

	Model 0	Model 1	Model 2	Model 3	Model 4
INH	0.86	0.86	0.82	0.76	0.75
RIF	0.90	0.90	0.84	0.79	0.78
EMB	0.80	0.80	0.75	0.68	0.66
SM	0.97	0.97	0.92	0.87	0.83
Overall	0.88	0.88	0.83	0.78	0.76

Table 2
Simulation results for posterior inference on the spatial effects and covariate coefficients

(A) Bias of posterior samples for spatial effects at the locations with maximum and minimum values and their 95% credible intervals

Outcome	Case I			Case II		
	$n = 800$	$n = 1400$	$n = 2000$	$n = 800$	$n = 1400$	$n = 2000$
Max						
1	-0.57 (-2.21, 1.40)	-0.60 (-2.03, 0.99)	-0.58 (-1.84, 0.69)	-0.54 (-1.09, 0.13)	-0.28 (-0.78, 0.25)	-0.31 (-0.75, 0.19)
2	0.06 (-1.42, 1.82)	-0.49 (-1.89, 1.04)	0.08 (-0.97, 1.21)	-0.34 (-0.89, 0.25)	-0.21 (-0.25, 0.68)	0.04 (-0.39, 0.33)
3	-0.45 (-2.04, 1.31)	-0.42 (-1.93, 1.26)	-0.34 (-1.35, 0.77)	-0.30 (-0.84, 0.23)	-0.01 (-0.46, 0.44)	-0.19 (-0.23, 0.59)
4	-0.64 (-2.13, 1.03)	-0.03 (-1.43, 1.61)	0.09 (-0.90, 1.25)	0.13 (-0.37, 0.60)	-0.05 (-0.36, 0.48)	0.01 (-0.34, 0.38)
Min						
1	0.71 (-1.48, 2.73)	0.42 (-1.04, 1.75)	0.32 (-0.65, 1.21)	0.49 (-0.19, 1.08)	0.38 (-0.17, 0.91)	0.30 (-0.19, 0.78)
2	1.00 (-0.69, 2.50)	-0.32 (-1.88, 1.31)	-0.15 (-1.56, 1.26)	0.44 (-0.14, 1.00)	0.05 (-0.52, 0.58)	0.34 (-0.09, 0.69)
3	0.92 (-1.04, 2.66)	0.45 (-1.36, 2.12)	-0.03 (-1.24, 1.13)	0.05 (-0.42, 0.46)	0.40 (-0.12, 0.82)	0.03 (-0.41, 0.43)
4	0.60 (-1.35, 2.20)	0.18 (-1.38, 1.61)	-0.22 (-1.37, 0.87)	0.02 (-0.45, 0.46)	0.07 (-0.32, 0.48)	-0.03 (-0.47, 0.38)

(B) Mean square errors of the posterior sample of spatial effects maps

Outcome	Case I			Case II		
	$n = 800$	$n = 1400$	$n = 2000$	$n = 800$	$n = 1400$	$n = 2000$
1	0.192	0.123	0.070	0.112	0.071	0.035
2	0.169	0.104	0.075	0.096	0.083	0.045
3	0.110	0.092	0.074	0.118	0.087	0.039
4	0.082	0.071	0.047	0.081	0.076	0.035

(C) Estimated posterior mean of covariate coefficients and their 95% credible intervals for different types of effects (small positive effects: $\beta_{21} = 0.25$, small negative effects: $\beta_{42} = -0.25$, and no effects $\beta_{33} = 0$)

β_{ij}	$n = 800$	$n = 1400$	$n = 2000$
Case I			
$\beta_{22} = 0.25$	0.26 (0.13, 0.43)	0.27 (0.13, 0.31)	0.23 (0.14, 0.31)

(C) Estimated posterior mean of covariate coefficients and their 95% credible intervals for different types of effects (small positive effects: $\beta_{22} = 0.25$, small negative effects: $\beta_{42} = -0.25$, and no effects $\beta_{33} = 0$)

β_{ij}	$n = 800$	$n = 1400$	$n = 2000$
$\beta_{42} = -0.25$	-0.29 (-0.42, -0.17)	-0.32 (-0.42, -0.23)	-0.26 (-0.34, -0.18)
$\beta_{33} = 0$	0.06 (-0.06, 0.18)	0.02 (-0.07, 0.11)	0.06 (-0.02, 0.13)
Case II			
$\beta_{22} = 0.25$	0.32 (0.21, 0.45)	0.26 (0.18, 0.35)	0.24 (0.16, 0.32)
$\beta_{42} = -0.25$	-0.22 (-0.33, -0.12)	-0.27 (-0.34, -0.19)	-0.28 (-0.36, -0.21)
$\beta_{33} = 0$	0.02 (-0.07, 0.12)	0.02 (-0.06, 0.09)	0.003 (-0.06, 0.08)

Table 3
Simulation results for comparisons of posterior inference accuracy on the covariate effects

	Model 1	Model 2	Model 3	Model 4
Case I				
MSE	0.197	0.218	0.402	0.645
Squared bias	0.023	0.029	0.143	0.182
Case II				
MSE	0.315	0.492	0.762	0.816
Squared bias	0.042	0.055	0.160	0.214