Lung Cancer Pathological Image Analysis Using a Hidden Potts Model

Cancer Informatics
Volume 16: 1–9
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176935117711910

Qianyun Li[1], Faliu Yi[2], Tao Wang[3], Guanghua Xiao[3] and Faming Liang[1]

[1]Department of Biostatistics, University of Florida, Gainesville, FL, USA. [2]Image Analysis, UT Southwestern Medical Center, Dallas, TX, USA. [3]Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, USA.

ABSTRACT: Nowadays, many biological data are acquired via images. In this article, we study the pathological images scanned from 205 patients with lung cancer with the goal to find out the relationship between the survival time and the spatial distribution of different types of cells, including lymphocyte, stroma, and tumor cells. Toward this goal, we model the spatial distribution of different types of cells using a modified Potts model for which the parameters represent interactions between different types of cells and estimate the parameters of the Potts model using the double Metropolis-Hastings algorithm. The double Metropolis-Hastings algorithm allows us to simulate samples approximately from a distribution with an intractable normalizing constant. Our numerical results indicate that the spatial interaction between the lymphocyte and tumor cells is significantly associated with the patient's survival time, and it can be used together with the cell count information to predict the survival of the patients.

KEYWORDS: Potts model, double Metropolis-Hastings, intractable normalizing constant, survival analysis

RECEIVED: December 10, 2016. **ACCEPTED:** March 28, 2017.

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 2243 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Liang's research was partially supported by grants DMS-1612924 and R01-GM117597. The latter is awarded as part of the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences.

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

CORRESPONDING AUTHOR: Faming Liang, Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA. Email: faliang@ufl.edu

Introduction

Lung cancer is the most common human cancer and the deadliest in the United States and globally. Non–small-cell lung cancer (NSCLC) is the most common cause of lung cancer death, accounting for up to 85% of deaths from lung cancer. Within NSCLC, adenocarcinoma and squamous cell carcinoma are the 2 major subtypes, with distinct prognoses and therapeutic remedies.[1,2] Current guidelines for treating lung cancer are largely based on clinical and pathological staging systems. However, the outcome varies widely.[1,2] Identifying the biomarkers that are responsible for patient risk prediction could help with treatment options, as well as understanding features of the tumor. Pathological examination of tumor tissue slides is routine in lung cancer diagnosis. The pathological images are widely available from routine clinical practices. Recent studies[3–5] have shown that the growth patterns of lung tumors are associated with patient survival outcomes. The pathological image features derived from the computer-aided pathological analysis have been used to predict the survival of patients with breast cancer[6,7] and complement cancer genomic profiling.[7,8] These studies demonstrate the feasibility of using digital pathological image analysis for objective and unbiased clinical prognosis. However, there still lacks such an analysis for lung cancer due to the complexity and heterogeneity of the disease.

Modeling spatial correlations in images is fundamental for pathological image analysis. In statistics, Markov random field models, such as the Ising model and Potts model, have been widely used to extract spatial correlation information for imaging data,[9–11] The major difficulty with the Ising and Potts models is their intractable normalizing constant, which makes their parameters hard to be estimated.

In this article, we model the spatial distribution of different types of cells, namely, lymphocyte, stroma, and tumor cells, using a modified Potts model. The parameters of the Potts model are often called interaction parameters, which characterize the clustering behavior of the same types of spins (ie, cells in the context of this article). We estimate the parameters of the Potts model using the double Metropolis-Hastings (DMH) algorithm[12] under a Bayesian framework. The DMH algorithm is very efficient for sampling from distributions with intractable normalizing constants, especially for the distributions defined on a large-scale lattice. We found that the interaction between lymphocytes and tumor cells is significantly associated with the patient's survival time, and furthermore, it can be used together with the cell count information to improve the prediction of the patient's survival time.

The remainder of this article is organized as follows. Section "Potts model" describes the modified Potts model and gives the details on how the DMH algorithm can be used to estimate its parameters. Section "Lung cancer imaging data" proposes a hidden, modified Potts model and presents our findings for lung cancer pathological image data. Section "Discussion" concludes the article with a brief discussion.

Potts Model

A modified Potts model

The q-state Potts model[13] consists of a 2-dimensional lattice of spins, where each spin takes values from a set of q different elements. The energy function of the model is given by

$$H(\boldsymbol{x}) = -\sum_{(i,j)\sim(i',j')} J_{ij,i'j'}\delta\left(x_{ij}, x_{i'j'}\right), \tag{1}$$

where $\boldsymbol{x} = \{x_{ij}\}$ denotes the collection of all spins of the model, $(i,j)\sim(i',j')$ denotes the neighboring pairs of the spins, $J_{ij,i'j'}$ denotes the interaction parameter between the spin x_{ij} and the spin $x_{i'j'}$, $\delta(x_{ij}, x_{i'j'})$ is an indicator function which is equal to 1 if $x_{ij} = x_{i'j'}$ and 0 otherwise, and the sum takes over all neighboring pairs of spins. If $q = 2$, the Potts model is reduced to the Ising model.[14] The Potts model is also related to, and generalized by, several other models, such as the XY model,[15] the Heisenberg model[16] and the N-vector model.[17] Generalizations of the Potts model have been used to model grain growth in metals and coarsening in foams.[18,19] A further generalization of the model, known as the cellular Potts model,[20] has been used to simulate static and kinetic phenomena in foam and biological morphogenesis.

In the standard form of Potts model (1), $J_{ij,i'j'}$ represents the strength of interaction between the same type of neighboring spins as the function $\delta(x_{ij}, x_{i'j'})$ takes a non-zero value if and only if $x_{ij} = x_{i'j'}$, where the value of x_{ij} indicates the type of the spin. However, for the lung cancer pathological imaging data, we would like to study the interactions between different types of cells, which are coded as 1 for lymphocyte cells, 2 for stroma cells, and 3 for tumor cells. For this reason, we consider a modified Potts model with the energy function given by

$$H(\boldsymbol{x}) = -\sum_{(i,j)\sim(i',j')} \theta_{kl}\left\{1-\delta\left(x_{ij} = k, x_{i'j'} = l\right)\right\}, \tag{2}$$

where k and l take values from the set {1,2,3}, θ_{12} represents the interaction between lymphocyte and stroma cells, θ_{13} represents the interaction between lymphocyte and tumor cells, and θ_{23} represents the interaction between stroma and tumor cells. This modified Potts model can also be viewed as a simplified cellular Potts model without the volume and surface constraints.

Corresponding to the energy function (2), the probability mass function of the modified Potts model is given by

$$\begin{aligned} f(\boldsymbol{x}\,|\,\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})}\exp\left\{\sum_{(i,j)\sim(i',j')} \theta_{kl}\left\{1-\delta\left(x_{ij} = k, x_{i'j'} = l\right)\right\}\right\} \\ &\triangleq \frac{\varphi(\boldsymbol{x},\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}, \end{aligned} \tag{3}$$

where $\boldsymbol{\theta} = \{\theta_{12}, \theta_{13}, \theta_{23}\}$, and $Z(\boldsymbol{\theta})$ is the normalizing constant. As an exact evaluation of $Z(\boldsymbol{\theta})$ needs to sum over the entire space of $\boldsymbol{x}$, which consists of 3^N different elements with N denoting the total number of spins, $Z(\boldsymbol{\theta})$ is intractable even for a small size model, say $N = 100$. How to estimate the parameters for such a model has been studied in the recent literature.[9,12,21–23]

DMH algorithm for the modified Potts model

Suppose that we are interested in estimating $\boldsymbol{\theta}$ for the modified Potts model under a Bayesian framework. Let $\pi(\boldsymbol{\theta})$ denote the prior density function of $\boldsymbol{\theta}$. Then, the posterior density function is given by

$$\pi\left(\boldsymbol{\theta}|\boldsymbol{x}\right) \propto \frac{\varphi(\boldsymbol{x},\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}\pi(\boldsymbol{\theta}). \tag{4}$$

It is easy to see that the Metropolis-Hastings (MH) algorithm cannot be directly applied to simulate from this posterior as the acceptance probability would involve an unknown normalizing constant ratio $Z(\boldsymbol{\theta})/Z(\boldsymbol{\theta}')$, where $\boldsymbol{\theta}'$ denotes the proposed value.

To address this issue, some auxiliary variable Markov chain Monte Carlo (MCMC) algorithms have been proposed, which aim to have the normalizing constant ratio $Z(\boldsymbol{\theta})/Z(\boldsymbol{\theta}')$ canceled in simulations by augmenting appropriate auxiliary variables to the target distribution and/or the proposal distribution. Along this direction, Møller et al[21] proposed an algorithm which augments both the target and proposal distributions, and Murray et al[22] proposed the so-called exchange algorithm which arguments only the proposal distribution. Although the underlying idea is very attractive, these 2 algorithms require the auxiliary variables to be drawn using a perfect sampler.[24] As perfect sampling can be very expensive or impossible for many models with intractable normalizing constants, the applications of these algorithms are highly hindered. To overcome this difficulty, Liang[12] proposed the DMH algorithm, and Liang et al[23] proposed an adaptive exchange algorithm. The adaptive exchange algorithm generates auxiliary variables via an importance sampling procedure from a Markov chain running in parallel, and the DMH algorithm generates auxiliary variables through a short run of the MH algorithm initialized with the original observation. As noted in Liang,[12] initializing the auxiliary MH chain with the original observation improves convergence of the algorithm. Other than the auxiliary variable MCMC algorithm, some approximation-based algorithms have been proposed in the literature, such as maximum pseudo-likelihood estimation,[25] Monte Carlo maximum likelihood estimation,[26] and adaptive kernel smoothing,[27,28] which approximate the likelihood function, the normalizing constant $Z(\boldsymbol{\theta})$, or the normalizing constant ratio $Z(\boldsymbol{\theta})/Z(\boldsymbol{\theta}')$. A recent comparative review[29] concludes that, compared with other algorithms, the DMH algorithm is very efficient for complex models with intractable normalizing constants, although the estimates are only asymptotically correct. The DMH algorithm is adopted in this article for estimating the parameters of the modified Potts model. The DMH algorithm can be described as follows.

Suppose that we are interested in simulating a sample $\boldsymbol{y}$ from $f(\cdot\,|\,\boldsymbol{\theta}')$ using the MH algorithm. If starting with the current state x, the transition probability, $P_{\boldsymbol{\theta}'}^{(m)}(\boldsymbol{y}\,|\,\boldsymbol{x})$, is given by

Table 1. Parameter estimates for the simulated data by the double Metropolis-Hastings algorithm, where SE(·) denotes the standard error of the corresponding estimate.

$(\theta_{12}, \theta_{13}, \theta_{23})$	$\hat{\theta}$	SE $(\hat{\theta}_{12})$	$\hat{\theta}_{13}$	SE $(\hat{\theta}_{13})$	$\hat{\theta}_{23}$	SE $(\hat{\theta}_{23})$
(0, 0.1, 0.1)	0.0360	0.0032	0.0975	0.0069	0.1016	0.0068
(0, 0.3, 0.3)	0.0403	0.0035	0.3099	0.0053	0.3108	0.0078
(0, 0.3, 0.7)	0.0444	0.0039	0.2980	0.0092	0.7058	0.0093
(0.1, 0.1. 0.1)	0.1005	0.0059	0.0988	0.0071	0.0996	0.0070
(0.3, 0.3, 0.3)	0.2981	0.0083	0.2926	0.0079	0.2994	0.0098
(0.1, 0.3, 0.5)	0.0971	0.0090	0.3069	0.0119	0.5024	0.0116

$$P_{\theta'}^{(m)}\left(\boldsymbol{y}|\boldsymbol{x}\right) = K_{\theta'}\left(\boldsymbol{x}\to\boldsymbol{x}_1\right)\ldots K_{\theta'}\left(\boldsymbol{x}_{m-1}\to\boldsymbol{y}\right), \qquad (5)$$

where $K(\cdot\to\cdot)$ is the MH transition kernel. Provided that the Markov chain has reached equilibrium states, then, by the detailed balance condition, we have

$$\frac{P_{\theta'}^{(m)}\left(\boldsymbol{x}|\boldsymbol{y}\right)}{P_{\theta'}^{(m)}\left(\boldsymbol{y}|\boldsymbol{x}\right)} = \frac{f(\boldsymbol{x}\,|\,\boldsymbol{\theta}')}{f(\boldsymbol{y}\,|\,\boldsymbol{\theta}')}, \qquad (6)$$

Let $q(\boldsymbol{\theta}'\,|\,\boldsymbol{\theta}_t,\boldsymbol{x})$ denote the proposal distribution for drawing a new parameter vector $\boldsymbol{\theta}'$. The DMH algorithm iterates between the following steps:

(a) Draw a new sample $\boldsymbol{\theta}'$ from the proposal density function $q(\boldsymbol{\theta}'\,|\,\boldsymbol{\theta}_t,\boldsymbol{x})$.
(b) Given , ′ ,simulate an auxiliary variable $\boldsymbol{y}\sim P_{\theta'}^{(m)}(\boldsymbol{y}\,|\,\boldsymbol{x})$ starting from the observation $\boldsymbol{x}$ and calculate the MH ratio:

$$\begin{aligned} r\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}',\boldsymbol{y}|\boldsymbol{x}\right) &= \frac{\pi\left(\boldsymbol{\theta}'\right)}{\pi\left(\boldsymbol{\theta}_t\right)}\frac{f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_t)}{f(\boldsymbol{x}\,|\,\boldsymbol{\theta}_t)}\frac{P_{\theta'}^{(m)}\left(\boldsymbol{x}|\boldsymbol{y}\right)}{P_{\theta'}^{(m)}\left(\boldsymbol{y}|\boldsymbol{x}\right)} \\ &= \frac{\pi\left(\boldsymbol{\theta}'\right)}{\pi\left(\boldsymbol{\theta}_t\right)}\frac{f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_t)}{f(\boldsymbol{x}\,|\,\boldsymbol{\theta}_t)}\frac{f(\boldsymbol{x}\,|\,\boldsymbol{\theta}')}{f(\boldsymbol{y}\,|\,\boldsymbol{\theta}')}. \end{aligned} \qquad (7)$$

(c) Set $\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}'$ with probability $\min\{1,r(\boldsymbol{\theta}_t,\boldsymbol{\theta}',\boldsymbol{y}|\boldsymbol{x})\}$; set $\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t$ otherwise.

Suppose that a sequence of samples $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n$ has been collected from a run of DMH. An approximate Bayesian estimator of $\boldsymbol{\theta}$ can then be obtained by averaging over the samples:

$$\bar{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}_i.$$

As implied by equation (6), the DMH sampler is almost exact for those parameters around the true value of $\boldsymbol{\theta}$. Hence, the estimator $\bar{\boldsymbol{\theta}}$ can be rather accurate even when m is not very large.

Simulation examples

We tested the performance of the DMH algorithm on the modified Potts model using simulated examples. We considered a variety of values of $\boldsymbol{\theta}$ as given in Table 1. For each setting of $\boldsymbol{\theta}$, we simulated 30 data sets independently using the Gibbs sampler on a 50×50 lattice. To simulate each data set, the Gibbs sampler was run for 10^5 iterations with random starting configurations.

To conduct a Bayesian analysis for the simulated data, we let $\boldsymbol{\theta}$ be subject to a uniform prior distribution, ie, $\pi(\boldsymbol{\theta})\propto1$ for $\boldsymbol{\theta}\in[0,1]^3$. For each of the simulated data sets, DMH was run for 6000 iterations, where the first 1000 iterations were discarded for the burn-in process and the samples generated in the remaining iterations were used for inference. At each iteration, the auxiliary sample was simulated using the Gibbs sampler with a single sweep for all elements. Each run costs about 18 seconds central processing unit (CPU) time on a Dell OptiPlex 9020 computer. The estimates of $\boldsymbol{\theta}$ are summarized in Table 1, where each estimate is obtained by averaging more than 30 independent data sets. Table 1 indicates that the DMH algorithm works well in parameter estimation for the modified Potts model.

Lung Cancer Imaging Data

The data set and accessibility

The pathological images and survival status from 205 patients with NSCLC in the National Lung Screening Trial (NLST) study[30] were collected. The characteristics of the participants are summarized in Table 2. The Kaplan-Meier curve for survival of the whole set of patients is shown in Figure 1. For each patient, 1, 2, or 3 tissue slides were first taken, where the number of slides depends on the size of tumor. For patients with a large size of tumor, more slides are needed to have a more comprehensive characterization of the tumor, and vice versa. Then, each tissue slide was examined by a lung cancer pathologist, the regions of interest (ROIs) within the tumor region(s) were determined, and 5 ROIs were randomly selected from each patient for further analysis. In total, we had 1585 ROI images.

Table 2. NSCLC patient characteristics (N = 205).

COHORT		NLST
No. of patients		N = 205
Age at diagnosis, y Median [LQ-HQ]		64 [60-68]
Follow-up, y Median [LQ-HQ]		6.6 [5.4-6.9]
Vital status, %	Alive	136
	Deceased	64
	NA	5
Sex, %	M	112
	F	89
	NA	4
Cancer stage, %	I	135
	II	20
	III	33
	IV	13
	NA	4
Smoking status, %	Smoker	110
	Nonsmoker	91
	NA	4

Abbreviations: HQ, high quartile; LQ, low quartile; NLST, National Lung Screening Trial.

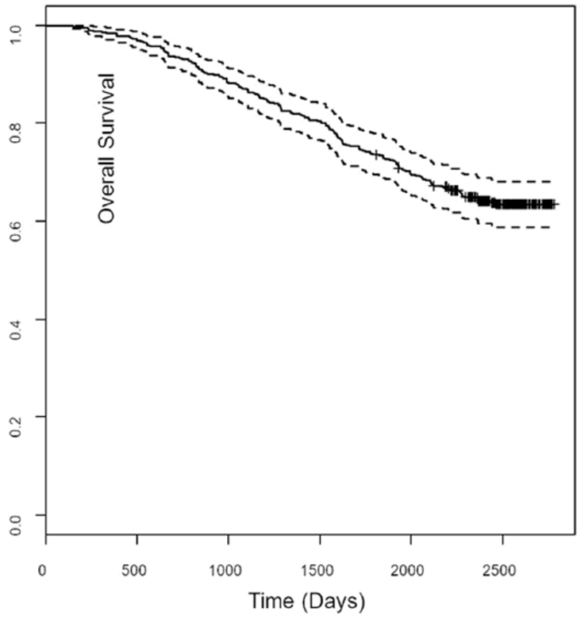

Figure 1. Kaplan-Meier plot for 205 patients with non–small-cell lung cancer.

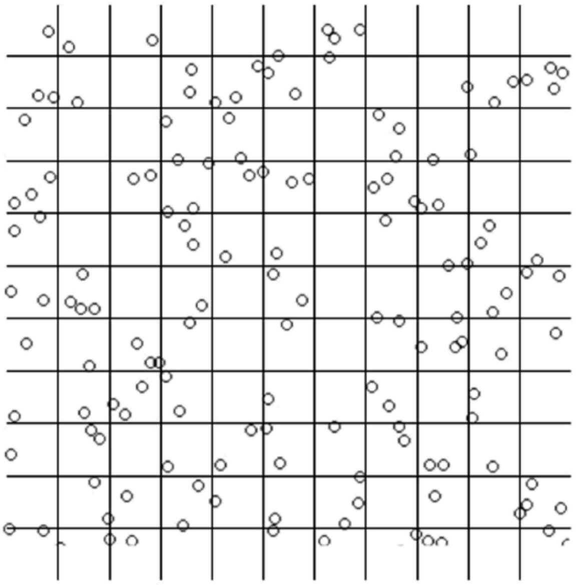

Figure 2. Illustration of the auxiliary lattice, where the circles represent cells. In real data sets, the location of a cell is given by a point so that a cell cannot belong to more than 1 square.

In the ROIs, the nucleus of each cell and the cell boundary were determined using a watershed method.[31] For each cell, a 160 pixels by 160 pixels image patch was extracted around the center, and the cell type was predicted using a convolutional neural network[32] developed from another study. The prediction was evaluated by the lung cancer pathologist, and the accuracy was more than 94%. The cell locations (ie, the coordinates of the cell on the ROI image) and the predicted cell types were used as inputs of the proposed method in this article. We are making the code publicly available. Once the paper is accepted, the code will be linked to the published version of the article.

DMH for a hidden Potts model

As the cell locations are irregular, it is difficult to model the pathological image by a Potts model. In particular, it is difficult to identify the neighboring cells for each cell. For this reason, we model the image by a hidden Potts model by introducing an auxiliary lattice to the image, which is illustrated by Figure 2. We note that the idea of modeling spatial data via an auxiliary lattice has been explored in the literature.[33,34]

Consider a pathological image with n cells located at $s_1,\ldots,s_n$. Let y_k denote the type of the cell located at s_k, and it takes value 1 for lymphocyte cells, 2 for stroma cells, and 3 for tumor cells. Let $\boldsymbol{y}=\{y_k\}$ denote the collection of observed cell types in the image. Let $W=\{(i,j):i=1,\ldots M,j=1,\ldots,N\}$ denote the auxiliary lattice, which partitions the image into $(M+1)(N+1)$ squares. Denote the squares by $C_1,C_2,\ldots,C_{(M+1)(N+1)}$. Let C_k denote the square that s_k belongs to. For convenience, we let each C_k be compact, ie, including all the boundary points of the square, while assuming that there are no cells belonging to 2 squares. If a cell is exactly on the boundary of some squares, then we randomly assign the cell to one of them. Let $X_{ij},(i,j)\epsilon W$, denote the hidden cell types at the auxiliary lattice. Conditional on X, we model the distribution of $\{Y_k\}$ by

$$P\left(Y_k = q|\boldsymbol{X},\gamma\right) = \frac{\exp\left(\gamma\sum_{(i,j)\epsilon C_k}\delta\left(X_{ij}=q\right)\right)}{\sum_{q'=1}^{3}\exp\left(\gamma\sum_{(i,j)\epsilon C_k}\delta\left(X_{ij}=q'\right)\right)}, \quad q=1,2,3, \tag{8}$$

where γ is a projection parameter with a prespecified value. The larger γ is, the more similar the original and imputed images are.

Let $f(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, as specified in equation (3), denote the likelihood function of the hidden Potts model. Let $\pi(\boldsymbol{\theta})$ denote the prior density function of $\boldsymbol{\theta}$. Assume that all Y_k's are independent conditional on $\boldsymbol{X}$, then we have

$$f\left(\boldsymbol{x},\theta|\boldsymbol{y},\gamma\right) \propto \left[\prod_{k=1}^{n} P\left(Y_k = y_k|\boldsymbol{x},\gamma\right)\right] f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Therefore, the full conditional posterior of X_{ij} is given by

$$\begin{aligned} &P\left(X_{ij} = q|\gamma,\boldsymbol{y},\boldsymbol{x}_{\partial_{ij}},\boldsymbol{\theta}\right) \\ &\propto \prod_{y_k\in C^{ij}} P\left(Y_k = y_k|X_{ij} = q,\boldsymbol{x}_{\partial_{ij}},\gamma\right)\exp \\ &\left(\sum_{(i',j')\in\partial_{ij}} \theta_{qx_{i'j'}}\left(1-\delta\left(q,x_{i'j'}\right)\right)\right), \end{aligned} \tag{9}$$

where $q\epsilon\left\{1,2,3\right\}$, C^{ij} denotes the union of the squares that the spin (i,j) belongs to, and $\partial_{\boldsymbol{ij}}$ denotes the neighboring spins of the spin (i,j). In this article, a free boundary condition is assumed for the model, under which the boundary points have fewer neighboring spins than the interior spins. After the normalization for equation (9), we have

$$P\left(X_{ij} = q|\gamma,\boldsymbol{y},\boldsymbol{x}_{\partial_{ij}},\boldsymbol{\theta}\right) \frac{\prod_{y_k\in C^{ij}} P\left(Y_k = y_k|X_{ij} = q,\boldsymbol{x}_{\partial_{ij}},\gamma\right) \exp\left(\sum_{(i',j')\in\partial_{ij}} \theta_{qx_{i'j'}}\left(1-\delta\left(q,x_{i'j'}\right)\right)\right)}{\sum_{q=1}^{3}\prod_{y_k\in C^{ij}} P\left(Y_k = y_k|X_{ij} = q,\boldsymbol{x}_{\partial_{ij}},\gamma\right) \exp\left(\sum_{(i',j')\in\partial_{ij}} \theta_{qx_{i'j'}}\left(1-\delta\left(q,x_{i'j'}\right)\right)\right)}. \tag{10}$$

Therefore, given $\boldsymbol{y}$, the projection parameter, and the model parameter, the hidden Potts model can be imputed using the Gibbs sampler through iteratively drawing X_{ij} s from distribution (10).

Similarly, we can get the full conditional posterior of $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y},\gamma) \propto f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{11}$$

As γ is a prespecified constant, the posterior can be simulated by iterating between the following 2 steps:

DMH algorithm for hidden Potts models

- Impute the hidden Potts model by simulating from equation (10) for all $(i,j)\in W$.
- Simulate $\boldsymbol{\theta}$ from distribution (11) using the DMH algorithm.

This algorithm consists of a few tunable parameters, including γ, M, and N. As mentioned previously, determine the similarity of the observed and imputed images. To make the 2 images more similar, we set γ to a large value. In all examples of this article, we set $\gamma=10$. The parameters M and N determine the size of the auxiliary lattice. Following the suggestion of Park and Liang,[33] we choose M and N such that $n\approx MN$. To be precise, we set M and N in the following way for each image: Let d_1 *and* d_2 denote the ranges of the cell locations at x-axis and y-axis, respectively. Then, we set the side length of each square to $l=\sqrt{d_1d_2/n}$, and set $M=[d_1/l]+1$ and $N=[d_2/l]+1$, where $[z]$ denotes the integer part of z.

Numerical results

The algorithm described above was applied to the 1585 ROI images. For each image, the algorithm was run for 6000 iterations, where the first 1000 iterations were discarded for the burn-in process and the samples generated in the remaining iterations were used for inference, and it cost about 14 minutes of CPU time on a Dell OptiPlex 9020 computer. The CPU time may vary slightly according to the values of M and N. Figure 3 shows the observed (left panels) and imputed (right panels) images for 3 ROIs, where each of the imputed images is obtained by averaging over the samples generated in a single run of the DMH algorithm. As it takes a large value, the imputed image is almost the same at each iteration after the simulation has reached equilibrium. As shown in Figure 3, the imputed images are very similar to the observed ones.

To assess the association between the spatial distributions of cells in pathological images and patients' survival status, we fitted a Cox proportional hazards model on the survival time with respect to the estimates of the interaction parameters of the hidden Potts model. Here, the survival time is defined as the time from diagnosis (of lung cancer) to death from all causes; right-censored cases exist. A patient with censored survival time means the patient is still alive at the last follow-up or lost to follow-up. In the Cox regression model, the hazard function has the form

$$h(t;Z) = h_o(t)\exp\left(\beta_1 Z_1 + \cdots + \beta_k Z_k\right),$$

where $h_o(t)$ is the baseline hazard function, and $Z_1,\ldots,Z_k$ are covariates having a multiplicative effect on the hazard function. For our model, we have $k=3$ and the covariates $Z_1=\hat{\theta}_{12}$,

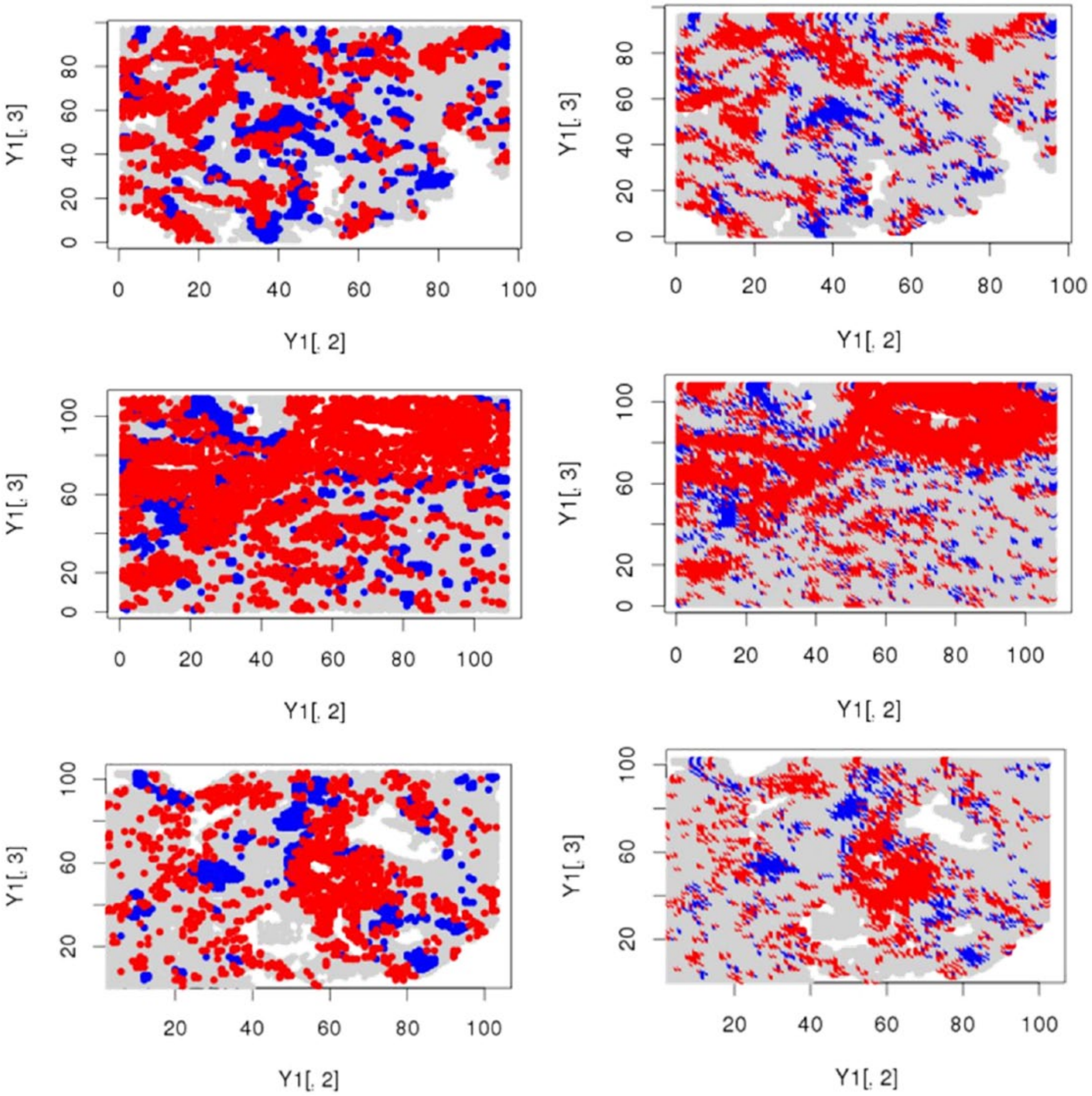

Figure 3. Comparison of the observed (left panels) and imputed (right panels) images for 3 regions of interest which are all from different patients: lymphocyte cells are in blue, stroma cells are in red, and tumor cells are in light gray.

$Z_2 = \hat{\theta}_{13}$, and $Z_3 = \hat{\theta}_{23}$, which were produced by the proposed method with the projection parameter $\gamma = 10$. The parameters $\beta_1, \ldots, \beta_k$ are estimated using the R package "survival."[35] As the data set contains multiple observations per patient, the generalized estimating equation method was used to compute a robust variance for each parameter estimate.[36]

Table 3 summarizes the estimates of the parameters of the Cox regression model. The overall *P* value for the significance of the model is .03374. The parameter estimates indicate that a higher value of the interaction between lymphocytes and tumor cells is significantly associated with a higher risk of death (at a significance level of .05); the hazard coefficient shows a negative correlation between the survival time and the value of the interaction between lymphocyte and tumor cells. In other words, widespread tumor cells indicate severity of the disease. Table 3 also shows that the interaction between stroma and tumor cells is also weekly associated with the risk of death (at a significance level of .1). However, there is no evidence suggesting any

Table 3. Survival analysis for lung cancer pathological images with the cell spatial interaction information ($\gamma = 10$).

PARAMETER	COEF	EXP(COEF)	SE	*P* VALUE
θ_{12}	−0.2631	0.7687	7.0305	.968
θ_{13}	2.5395	12.6739	0.9767	**.0043**
θ_{23}	0.7078	2.0296	0.4049	.0848

Abbreviation: SE, standard error.
The significance of the bold value(s) is identified at a level of 0.05.

association between the interaction of lymphocytes and stroma cells and the risk of death. In summary, we may conclude that the proposed hidden Potts model is able to extract some useful information about the status of the disease.

To assess the sensitivity of the results to the projection parameter γ, different values of γ were tried. The results are reported in Tables 4 and 5, which indicate that our results are not sensitive to the value of γ.

We also assessed the proportional hazards assumption for the Cox regression model on this data set.[37] The *P* value for the whole model is .219, and the *P* values of the respective parameters are all greater than .10, suggesting that the proportional hazards assumption is valid and the regression coefficients $\boldsymbol{\theta} = \{\theta_{12}, \theta_{13}, \theta_{23}\}$ remain constant over time. Figure 4 shows the plot of scaled Schoenfeld residuals versus time.

Finally, we conducted a survival analysis based on the cell count information only. In particular, we considered the cell count ratios, lymphocyte/stroma and tumor/stroma. The results, which are shown in Table 6, indicate that the cell count information is also very useful in predicting patient survival; the hazard coefficient implies a negative correlation between the survival time and the ratio tumor/stroma. It is very interesting to point out that the spatial interaction information learned by the proposed method for different types of cells is complementary to the cell count information. As indicated in Table 7, the prediction for the survival can be further improved by using both of them. With both information, the overall *P* value (in likelihood ratio test) of the Cox regression model has been improved to .00539 from .02367 (with the cell count information only). In addition, compared with Tables 3 and 6, the significance levels of θ_{13}, θ_{23}, and tumor/stroma ratio are also improved.

Table 4. Survival analysis for lung cancer pathological images with the cell spatial interaction information ($\gamma = 8$).

PARAMETER	COEF	EXP(COEF)	SE	*P* VALUE
θ_{12}	−1.5196	0.2188	1.5611	.2864
θ_{13}	2.4410	11.4846	1.1122	**.0058**
θ_{23}	−0.0294	0.9734	0.3782	.9424

Abbreviation: SE, standard error.
The significance of the bold value(s) is identified at a level of 0.05.

Table 5. Survival analysis for lung cancer pathological images with the cell spatial interaction information ($\gamma = 12$).

PARAMETER	COEF	EXP(COEF)	SE	*P* VALUE
θ_{12}	−1.6388	0.1942	1.5536	.2461
θ_{13}	2.5323	12.5911	1.1047	**.0037**
θ_{23}	−0.0101	0.9900	0.3765	.9782

Abbreviation: SE, standard error.
The significance of the bold value(s) is identified at a level of 0.05.

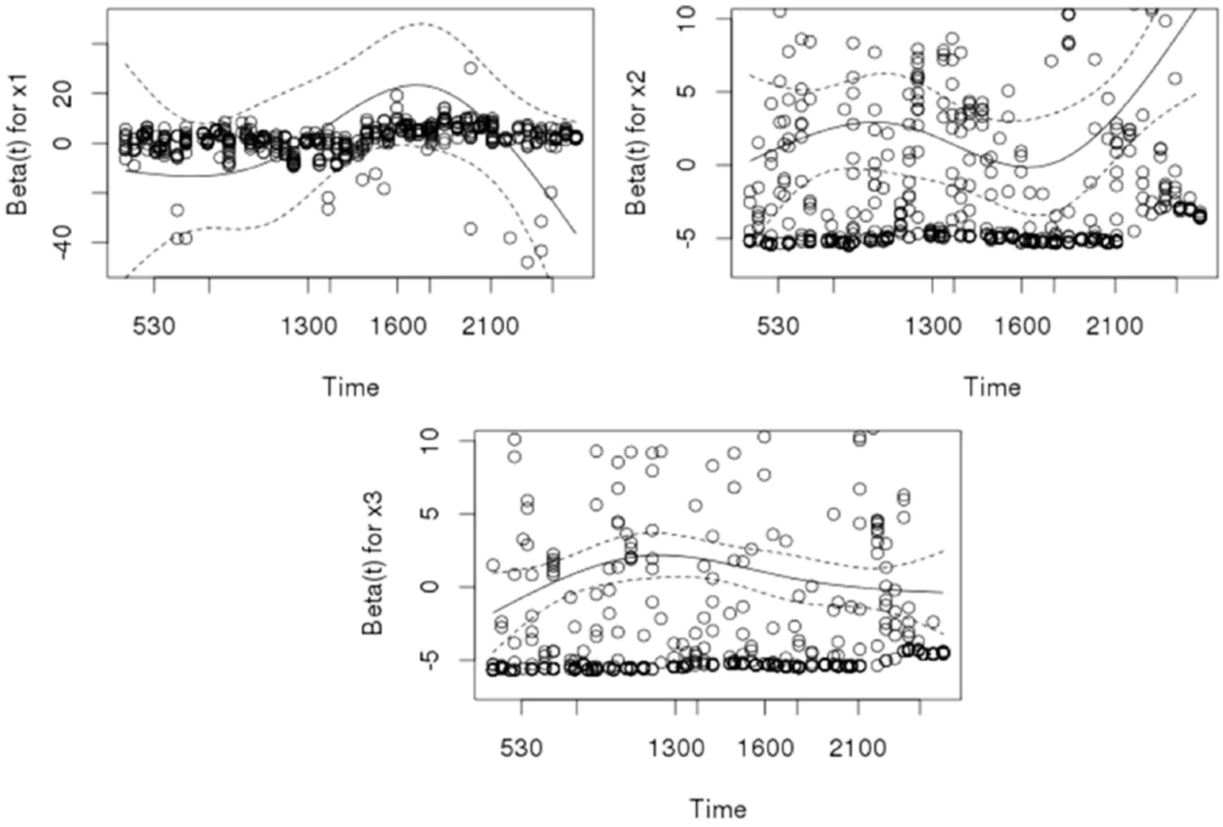

Figure 4. Scaled Schoenfeld residuals versus time for $\theta = \{\theta_{12}, \theta_{13}, \theta_{23}\}$.

Table 6. Survival analysis for lung cancer pathological images with cell count information only.

PARAMETER	COEF	EXP(COEF)	SE	*P* VALUE
Lymphocyte/ stroma ratio	9.999e−01	1.050e−04	1.5611	.233
Tumor/stroma ratio	1.381e−03	1.001e+00	4.192e−04	**1.59e−07**

Abbreviation: SE, standard error.
Likelihood ratio test = 7.49 on 2 df, *P* = .02367; Wald test = 27.55 on 2 df, *P* = 1.039e−06.
Score (log-rank) test = 11.87 on 2 df, *P* = .002651; Robust = 11.12, *P* = .003856.
The significance of the bold value(s) is identified at a level of 0.05.

Table 7. Survival analysis for lung cancer pathological images with both the cell count information and the cell spatial interaction information.

PARAMETER	COEF	EXP(COEF)	SE	*P* VALUE
Lymphocyte/ stroma ratio	−9.417e−05	9.999e−01	1.099e−04	.20458
Tumor/ stroma ratio	1.445e−03	1.001e+00	4.243e−04	**4.94e−08**
θ_{12}	3.451e−01	1.412e+00	6.935e+00	.95539
θ_{13}	2.400e+00	1.103e+01	9.714e−01	**.00511**
θ_{23}	8.215e−01	2.274e+00	4.059e−01	**.04495**

Likelihood ratio test = 16.57 on 5 df, *P* = .00539; Wald test = 40.34 on 5 df, *P* = 1.277e−07.
Score (log-rank) test = 21.99 on 5 df, *P* = .0005265; Robust = 18.84, *P* = .002063.
The significance of the bold value(s) is identified at a level of 0.05.

Discussion

In this article, we have proposed to model pathological images using a hidden Potts model and applied the DMH algorithm to estimate the model parameters. The introduction of auxiliary lattice makes the proposed method very general, which can be used for any type of imaging data with or without regular observations. The auxiliary lattice also helps reduce the complexity of imaging data and defines a concise and explicit neighborhood for each spin of the hidden Potts model. Other auxiliary variable MCMC algorithms, eg, the adaptive exchange algorithm,[23] can potentially be applied to this problem. However, it would be more time-consuming given the hidden structure of the proposed Potts model.

For the lung cancer pathological imaging data, our study shows that the survival time of NSCLC patients might be significantly associated with the strength of interactions between lymphocyte and tumor cells. The spatial interaction parameter together with the cell count information can potentially be used as a biomarker for prognosis and personalized treatments of patients with NSCLC. It would be of great interest to extend the proposed method to other pathological imaging data.

Acknowledgements

The authors thank the editor, associate editor, and 4 referees for their constructive comments, which have led to significant improvement of this paper.

Author Contributions

QL conducted simulations and drafted the paper; FY and TW preprocessed the image data and made the data ready to be used by QL for statistical analysis; GX identified the problem and discussed with FL to initiate the project; FL initiated and directed the project, proposed the model, and wrote the paper with QL.

REFERENCES

1. Cross MD. Advances in NSCLC: histologic distinction between adenocarcinoma and squamous cell carcinoma. *Med Lab Obs*. 2012;44:40–42.
2. Maeda H, Matsumura A, Kawabata T, et al. Adenosquamous carcinoma of the lung: surgical results as compared with squamous cell and adenocarcinoma cases. *Eur J Cardiothorac Surg*. 2012;41:357–361.
3. Amin MB, Tamboli P, Merchant SH, et al. Micropapillary component in lung adenocarcinoma: a distinctive histologic feature with possible prognostic significance. *Am J Surg Pathol*. 2002;26:358–364.
4. Barletta JA, Yeap BY, Chirieac LR. Prognostic significance of grading in lung adenocarcinoma. *Cancer*. 2010;116:659–669.
5. Borczuk AC, Qian F, Kazeros A, et al. Invasive size is an independent predictor of survival in pulmonary adenocarcinoma. *Am J Surg Pathol*. 2009;33:462–469.
6. Tabesh A, Teverovskiy M, Pang H-Y, et al. Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans Med Imaging*. 2007;26:1366–1378.
7. Beck AH, Sangoi AR, Leung S, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*. 2011;3:108ra13.
8. Yuan Y, Failmezger H, Rueda OM, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med*. 2012;4:157ra43.
9. Li S. *Markov Random Field Modeling in Image Analysis*. London, England: Springer; 2009.
10. Ayasso H, Mohammad-Djafari A. Joint NDT image restoration and segmentation using Gauss-Markov-Potts prior models and variational Bayesian computation. *IEEE Trans Image Process*. 2010;19:2265–2277.
11. Green PJ, Richardson S. Hidden Markov models and disease mapping. *J Am Statist Assoc*. 2002;97:1055–1070.
12. Liang F. A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *J Stat Comput Sim*. 2010;80:1007–1022.
13. Wu F. The Potts model. *Rev Mod Phys*. 1982;54:235–268.
14. Ising E. Beitrag zur Theorie des Ferromagnetismus. *Z Phys*. 1925;31:253–258.
15. Kosterlitz JM. The critical properties of the two-dimensional xy model. *J Phys C: Solid State Phys*. 1974;7:1046–1060.
16. Polyakov AM. Interaction of goldstone particles in two dimensions: applications to ferromagnets and massive Yang-Mills fields. *Phys Lett B*. 1975;59:79–81.
17. Stanley HE. Dependence of critical properties upon dimensionality of spins. *Phys Rev Lett*. 1968;20:589–592.
18. Sahni PS, Grest GS, Anderson MP, Srolovitz DJ. Kinetics of the q-state Potts model in two dimensions. *Phys Rev Lett*. 1983;50:263–266.
19. Glazier JA, Anderson MP, Grest GS. Coarsening in the two-dimensional soap forth and the large-q Potts model: a detailed comparison. *Philos Mag B*. 1990;62:615–646.
20. Graner F, Glazier JA. Simulation of biological cell sorting using a two-dimensional extended Potts model. *Phys Rev Lett*. 1992;69:2013–2016.
21. Møller J, Pettitt AN, Reeves R, Berthelsen KK. An efficient Markov Chain Monte Carlo method for distributions with intractable normalizing constants. *Biometrika*. 2006;93:451–458.
22. Murray I, Ghahramani Z, MacKay DJC. MCMC for doubly-intractable distributions. In: Proceedings of 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI); 2006. https://dslpitt.org/uai/papers/06/p359-murray.pdf.

23. Liang F, Jin I-H, Song Q, Liu JS. An adaptive exchange algorithm for sampling from distribution with intractable normalizing constants. *J Am Statist Assoc*. 2016;111:377–393.
24. Propp J, Wilson D. Exact sampling with coupled Markov Chains and applications to statistical mechanics. *Random Struct Algor*. 1996;9:223–252.
25. Besag J. Spatial interaction and the statistical analysis of lattice systems. *J Roy Stat Soc B*. 1974;36:192–236.
26. Geyer CJ, Thompson EA. Constrained Monte Carlo maximum likelihood for dependent data. *J Roy Stat Soc B*. 1992;54:657–699.
27. Liang F. Continuous Contour Monte Carlo for marginal density estimation with an application to a spatial statistical model. *J Comput Graph Stat*. 2007;16:608–632.
28. Atchade YF, Lartillot N, Robert CP. Bayesian computation for intractable normalizing constants. *Braz J Stat*. 2013;27:416–436.
29. Park J, Haran M. *Bayesian Inference in the Presence of Intractable Normalizing Functions: A Comparative Review* (Manuscript). State College, PA: Department of Statistics, Penn State University; 2016.
30. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365:395–409.
31. Gonzalez RC, Woods RE. *Digital Imaging Processing*. New York, NY: Prentice Hall; 2002.
32. Shin H, Roth H, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35:1285–1298.
33. Park J, Liang F. Bayesian analysis of geostatistical models with an auxiliary lattice. *J Comput Graph Stat*. 2012;21:453–475.
34. Xu G, Liang F, Genton M. A Bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. *Stat Sinica*. 2015;25:61–79.
35. Therneau TM. Package survival. https://cran.r-project.org/web/packages/survival/survival.pdf. Accessed December 1, 2016.
36. Therneau T, Grambsch P, Pankratz VS. Penalized survival models and frailty. *J Comput Graph Stat*. 2003;12:156–175.
37. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81:515–526.