Original Article

Spatial modeling for correlated cancers using bivariate directed graphs

Leiwen Gao[1], Sudipto Banerjee[1], Abhirup Datta[2]

[1]University of California, Los Angeles, CA, USA; [2]Johns Hopkins University, Baltimore, MD, USA

Contributions: (I) Conception and design: All authors; (II) Administrative support: S Banerjee; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: L Gao; (V) Data analysis and interpretation: L Gao; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Sudipto Banerjee. University of California, Los Angeles, CA, USA. Email: sudipto@ucla.edu.

Background: Disease maps are an important tool in cancer epidemiology used for the analysis of geographical variations in disease rates and the investigation of environmental risk factors underlying spatial patterns. Cancer maps help epidemiologists highlight geographic areas with high and low prevalence, incidence, or mortality rates of cancers, and the variability of such rates over a spatial domain. They can also be used to detect "hot-spots" or spatial clusters which may arise due to common environmental, demographic, or cultural effects shared by neighboring regions. Statistical methods for spatial data formulate models to capture spatial autocorrelation and produce cancer maps to better detect clustering and hotspots. When more than one cancer is of interest, the models must also capture the inherent or endemic association between the diseases in addition to the spatial association. This article develops interpretable and easily implementable spatial autocorrelation models for two or more cancers.

Methods: The article builds upon recent developments in univariate disease mapping that have shown the use of mathematical structures such as directed acyclic graphs (DAGs) to capture spatial association for a single cancer. The advantage of using DAGs over other existing models is the easier interpretation of spatial association. The current manuscript extends this family of directed acyclic graphical models to estimate inherent or endemic association for two cancers in addition to the association over space (clustering) for each of the cancers. The method builds a Bayesian hierarchical model where the spatial effects are introduced as latent random effects for each cancer. A valid joint probability model is constructed by first modeling the marginal distribution of one disease followed by the second disease conditional on the first. This approach ensures easier interpretation of model parameters and helps to separate the spatial autocorrelation for each cancer from the association between the two cancers.

Results: We analyze the relationship between esophagus and lung cancer extracted from the Surveillance, Epidemiology, and End Results (SEER) Program for their incidence rates in the years 2012–2016 across 58 counties in California. Our analysis shows statistically significant association between the county-wide incidence rates of lung and esophagus cancer across California. After accounting for explanatory variables (smoking, age, education, employment, sex, race, health insurance and poverty), esophagus cancer rates exhibit weaker spatial association than lung cancer rates for data counties in California.

Conclusions: The bivariate directed acyclic graphical model performs better than competing bivariate spatial models in the existing literature. This improvement is seen both in terms of the model's fit to the data and complexity of the model.

Keywords: Bayesian hierarchical models; directed acyclic graphs (DAGs); disease mapping; Surveillance, Epidemiology, and End Results database (SEER database); spatial statistics

Received: 20 November 2019; Accepted: 09 June 2020; Published: 20 November 2020.

doi: 10.21037/ace-19-41

View this article at: http://dx.doi.org/10.21037/ace-19-41

Introduction

Disease mapping, which refers to techniques for mapping and analysis of geographical variations in disease rates and the investigation of environmental risk factors underlying these patterns, has long been an important tool in cancer epidemiology (1). Disease maps are used to highlight geographic areas with high and low prevalence, incidence, or mortality rates of cancers, and the variability of such rates over a spatial domain (2). They can also be used to detect "hot-spots" or spatial clusters which may arise due to common environmental, demographic, or cultural effects shared by neighboring regions (3). Maps of crude incidence or mortality rates can be misleading when the population sizes for some of the units are small, which results in large variability in the estimated rates, and makes it difficult to distinguish chance variability from genuine differences. The correct geographic allocation of health care resources can be greatly enhanced by deployment of statistical models that allow a more accurate depiction of true disease rates and their relation to explanatory variables (covariates). Many tasks critical for successful cancer surveillance and control require new inferential methods to handle these complex and often spatially indexed data sets. Since local sample sizes within each spatial region are too low for design-based solutions to attain desired levels of statistical precision (4), much recent work in disease-mapping has been carried out within the context of Bayesian hierarchical models (5). The body of scientific literature on modern methods for geographic disease mapping is too vast to be reviewed here. Comprehensive reviews of prevalent statistical disease mapping methods and their implementation using available software can be found, among several other sources (6-9).

Statistical models for mapping a single disease have employed probability distributions such as Markov random fields or MRFs (10) that introduce dependence using the adjacency information among the different regions on a map. Two conspicuous examples are the conditional autoregression (CAR) and simultaneous autoregression (SAR) models (11-14) for further discussions on CAR and SAR models. More recently, directed acyclic graphical autoregressive (DAGAR) models that employ directed acyclic graphs (DAGs) have been developed as an alternative to CAR or SAR models (15). A specific motivation for DAGAR models is that they impart greater interpretability to the spatial autocorrelation parameter.

In this article, we will perform joint spatial mapping of two different types of cancers. Joint modeling is appropriate when different diseases have been observed over the same spatial units and when the diseases themselves are related to each other, say because they share the same set of spatially distributed risk factors or the presence of one disease in a spatial unit may encourage or inhibit the presence of the second disease in the same spatial unit. In other words, we seek models to capture the spatial association for each disease as well as the association between the diseases. There is a substantial literature on multivariate disease mapping that has demonstrated, theoretically and empirically, the benefits of jointly modeling several potentially related cancers, as opposed to modeling them independently (16-20). While it has been assertively demonstrated that independent models for cancers can lead to biased results because of unaccounted associations among the cancers, the current literature is largely based on using CAR models for spatial mapping (21-23). For example, a bivariate CAR model has been proposed for modeling two associated diseases (21). Extensions such as a generalized multivariate CAR model (GMCAR) have been developed and compared with other multivariate CAR models (24,25) revealing strong correlation of mortality rates for lung and esophageal cancer (26). Our proposed bivariate DAGAR (BDAGAR) model for modeling two diseases over the same spatial region will help epidemiologists and spatial analysts better interpret the association among the cancers.

The incidence of adenocarcinoma of lung and esophageal cancer have been found to share common risk factors including gastroesophageal reflux disease (GERD), obesity and its associated metabolic syndrome (diabetes, hypertension and hyperlipidemia) (27). In terms of metabolic mechanisms, it has also been reported that cytochrome P450 2C19 (CYP2C19) may participate in the activation of procarcinogen of both lung and esophageal cancer, and CYP2C19 poor metabolizers (PMs) have higher incidence of two cancers (28). Given the potential association between the incidence of lung and esophageal cancer, the remainder of this article proceeds by developing a class of BDAGAR models, conducting some disease mapping for these two different cancers, and summarizing with some concluding remarks.

Methods

Our approach will be to construct a probability model for each disease using the distribution specified by DAGAR. We will extend the univariate DAGAR to a bivariate model by modeling the distribution of one disease as a univariate

DAGAR and the conditional distribution of the second disease given the first also as a DAGAR. In this sense, our BDAGAR is analogous to the bivariate CAR models (26). We develop notations and briefly discuss the univariate DAGAR in the following section.

DAGAR for modeling a single disease

We consider a geographic map of our region of interest (e.g., a particular state) delineated by k distinct administrative regions (e.g., counties or ZIP codes) with clear non-overlapping boundaries separating them. Let $w = (w_1, w_2, ..., w_k)^T$ be a $k \times 1$ vector consisting of spatially associated random effects corresponding to each region. We develop a spatially correlated model using a DAG. The geographic map provides us with a list of neighbors for each region. Neighbors can be defined by the user. Common definitions include when two regions share a common boundary or if their centers are within a certain fixed distance, although the model and resulting distribution theory hold for any fixed set of neighbors. The data structure for the geographic map and its neighbors is defined as a graph, denoted $G = \{V, E\}$, where the regions are indexed by an ordered set $V = \{1, 2, ..., k\}$ and form the vertices of the graph and E is the collection of edges between the vertices, i.e., the collection of ordered pairs (j, j') such that j and j' are geographic neighbors based upon some specified definition.

The DAGAR model specifies $w \sim N(0, \tau Q(\rho))$, where $Q(\rho)$ is a spatial precision matrix that depends only upon a spatial autocorrelation parameter ρ and τ is a positive scale parameter. To describe $Q(\rho)$, we define neighbor sets $N(i) = \{j < i : j \sim i\}$, where $i \in V \backslash \{1\}$, i.e., the set V excluding the region indexed by 1, and $j \in V$. Thus, $N(i)$ includes geographic neighbors of region j that precede i in the ordered set V. The precision matrix $Q(\rho) = (I - B)^T F (I - B)$, where B is a $k \times k$ strictly lower-triangular matrix with entries b_{ij} and F is a $k \times k$ diagonal matrix with diagonal elements f_{ii} such that

$$b_{ij} = \begin{cases} 0 & \text{if } j \notin N(i) \\ \dfrac{\rho}{1 + (n_{<i} - 1)\rho^2} & \text{if } i = 2, 3, ..., k, j \in N(i) \end{cases} \quad \text{and} \quad f_{ii} = \frac{1 + (n_{<i} - 1)\rho^2}{1 - \rho^2} \qquad [1]$$

where $n_{<i}$ is the number of members in $N(i)$. The above definition of b_{ij} is consistent with the lower-triangular structure of B because $j \notin N(i)$ for any $j \geq i$. The derivation of B and F as functions of a spatial correlation parameter ρ is based upon forming local autoregressive models on embedded spanning trees of subgraphs of G (15).

A BDAGAR model

We now extend the DAGAR to the bivariate case, where we jointly model two cancers across regions. Let $w_i = (w_{i1}, w_{i2}, ..., w_{ik})^T$ be the spatial random effect vector for disease i, where w_{ij} refers to the spatial random effect for disease i in region j. We will build a hierarchical model:

$$p(w_1, w_2) = N(w_1 \mid 0, \tau_1 Q_1(\rho_1)) \times N(w_2 \mid A_{21} w_1, \tau_2 Q_2(\rho_2)) \qquad [2]$$

where $N(\cdot \mid \mu, Q)$ denotes a normal density with mean μ and precision matrix Q. The precision matrices $\tau_i Q_i(\rho_i)$ for $i = 1, 2$ are the DAGAR precision matrices formed with the entries of B and F described in Eq. [1] with ρ_i. Therefore, in Eq. [2] we model w_1 as a univariate DAGAR and w_2 conditional on w_1 also as a DAGAR. Each disease has its own distribution and there are two spatial autocorrelation parameters (ρ_1 and ρ_2) corresponding to the two diseases. This ensures that spatial associations specific to each disease will be captured.

The matrix A_{21} models the association between the two diseases. We use a parametric form $A_{21} = \eta_0 I_k + \eta_1 M$, where M is the binary adjacency matrix of the geographic map, i.e., $m_{ij} = 1$ if $i \sim j$ and 0 otherwise. The joint distribution of $w = (w_1^T, w_2^T)^T$ is now derived from Eq. [2] as $w \sim N(0, Q_w)$, where the precision matrix Q_w is

$$Q_w = \begin{bmatrix} \tau_1 Q_1(\rho_1) + \tau_2 A_{21}^T Q_2(\rho_2) A_{21} & \tau_2 A_{21}^T Q_2(\rho_2) \\ \tau_2 Q_2(\rho_2) A_{21} & \tau_2 Q_2(\rho_2) \end{bmatrix} \qquad [3]$$

and the covariance matrix Q_w^{-1} is

$$Q_w^{-1} = \begin{bmatrix} \tau_1^{-1} Q_1^{-1}(\rho_1) & \tau_1^{-1} Q_1^{-1}(\rho_1) A_{21}^T \\ \tau_1^{-1} A_{21} Q_1^{-1}(\rho_1) & \tau_1^{-1} A_{21} Q_1^{-1}(\rho_1) A_{21}^T + \tau_2^{-1} Q_2^{-1}(\rho_2) \end{bmatrix} \qquad [4]$$

We call a normal distribution with the above precision, or covariance, matrix, the BDAGAR model. The interpretation of ρ_1 and ρ_2 is clear: ρ_1 measures the spatial association for the first cancer, while ρ_2 is the residual spatial correlation in the second cancer after accounting for the first cancer. Similarly, τ_1 is the spatial precision parameter for the first cancer, while τ_2 is the residual precision for the second cancer after accounting for the first.

Model implementation

Let y_{ij} be our outcome of interest corresponding to cancer i

in region j. We will assume that y_{ij} is a continuous variable, e.g., incidence rates, that is related to a set of explanatory variables through the regression model:

$$y_{ij} = x_{ij}^T \beta_i + w_{ij} + \varepsilon_{ij} \quad [5]$$

where x_{ij} is a $p_i \times 1$ vector of explanatory variables specific to cancer i within region j, β_i is the slopes corresponding to cancer i, w_{ij} is the spatial effects that collectively follow the BDAGAR distribution described in section "A BDAGAR model", $\varepsilon_{ij} \sim^{ind} N\left(0, 1/\sigma_i^2\right)$ capture additional heterogeneity and variability independent of spatial variation, where σ_i^2 is the residual variance for cancer i. The regression model is extended to the following specific Bayesian hierarchical framework with the posterior distribution $p(\beta, w, \eta, \rho, \tau, \sigma \mid y)$ proportional to

$$\begin{aligned} & p(\rho) \times p(\eta) \times \prod_{i=1}^{2} \left\{ IG\left(1/\tau_i \mid a_{\tau_i}, b_{\tau_i}\right) \times IG\left(\sigma_i^2 \mid a_\sigma, b_\sigma\right) \right. \\ & \left. \times N\left(\beta_i \mid \mu_\beta, \ V_\beta^{-1}\right) \right\} \times N\left(w \mid 0, Q_w\right) \\ & \times \prod_{i=1}^{2} \prod_{j=1}^{k} N\left(y_{ij} \mid x_{ij}^T \beta_i + w_{ij}, \sigma_i^2\right) \end{aligned} \quad [6]$$

where $\beta = \{\beta_1, \beta_2\}$, $\tau = \{\tau_1, \tau_2\}$, $\sigma = \{\sigma_1, \sigma_2\}$ and $\eta = \{\eta_0, \eta_1\}$, and $IG(\cdot \mid a, b)$ is the inverse-gamma distribution with shape and rate parameters a and b, respectively.

We sample the parameters from the posterior distribution in Eq. [6] using Markov chain Monte Carlo (MCMC) with Gibbs sampling and random walk metropolis (29) as implemented in the rjags package within the R statistical computing environment. To compare and assess models, we use the Widely Applicable Information Criterion (WAIC) (30,31), which is computed as

$$\text{WAIC} = -2\widehat{elppd} = -2\left(\widehat{lppd} - \hat{p}_{WAIC}\right) \quad [7]$$

where $\widehat{elppd}$ is the expected log point-wise predictive density for a new dataset and $\hat{p}_{WAIC}$ is the estimated effective number of parameters, which is sum of posterior variance of the log predictive density for each data point. WAIC is easy to compute using posterior samples.

Results

We analyze a data set extracted from the SEER*Stat database using the SEER*Stat statistical software (32). We consider 2 cancers, lung (ICD-O-3: C340-C349) and esophagus (ICD-O-3: C150-C159), where the outcome is the 5-year average crude incidence rates per 100,000 population in the years from 2012 to 2016 across 58 counties in California, USA, calculated from the software directly. County-level explanatory variables for each cancer, that possibly affect the incidence rates, are available and include adult cigarette smoking rates in percentage (smoke_{ij}), percentages of residents younger than 18 years old (young_{ij}), older than 65 years old (old_{ij}), with education level below high school (edu_{ij}), percentages of unemployed residents (unemp_{ij}), black residents (black_{ij}), male residents (male_{ij}), uninsured residents (uninsure_{ij}) and percentages of families below the poverty threshold (poverty_{ij}). All covariates, except adult cigarette smoking rates, are county attributes extracted from the SEER*Stat database (33) for the years 2012–2016. As a potential common risk factor for both lung and esophageal cancer, adult cigarette smoking rates for 2014–2016 were obtained from the California Tobacco Control Program (34).

We analyzed this data set using the Bayesian hierarchical model [6]. The county-level maps of the raw incidence rates per 100,000 population for the two cancers are shown in *Figure 1*. The maps exhibit the evidence of correlation across space and between cancers. Cutoffs for the different levels of incidence rates are quantiles for each cancer. For both lung and esophageal cancer, in general, incidence rates are higher in counties located in the northern areas than those in southern part. The four counties in the center including Amador, Calaveras, Tuolumne and Mariposa have relatively high incidence rates compared to the neighboring counties. Overall, counties with similar levels of incidence rates tend to depict some spatial clustering.

For our analysis, we specified the following prior distribution

$$\begin{aligned} & p(\eta, \rho, \tau, \sigma, w) = \\ & \prod_{i=1}^{2} Unif\left(\rho_i \mid 0,1\right) \prod_{i=0}^{1} N\left(\eta_i \mid 0, 10^2\right) \times \prod_{i=1}^{2} N\left(\beta \mid 0, 10^3\right) \\ & \times \prod_{i=1}^{2} IG\left(1/\tau_i \mid 2, 0.1\right) \times \prod_{i=1}^{2} IG\left(\sigma_i^2 \mid 2,1\right) \times N\left(w \mid 0, Q_w(\tau, \ \rho)\right) \end{aligned} \quad [8]$$

where $Unif(\cdot \mid a, b)$ denotes the Uniform density over (0, 1) and $Q_w(\tau, \rho)$ is the BDAGAR precision matrix of w given in Eq. [3].

We fit the BDAGAR model using the two different cancer orders, i.e., [esophagus] × [lung | esophagus] and the reverse ordering [lung] × [esophagus | lung]. We will refer to these orderings simply as [lung | esophagus] and [esophagus | lung], respectively. *Table 1* presents measures for model fit using the WAIC. We also compare BDAGAR with the "Generalized Multivariate Conditional

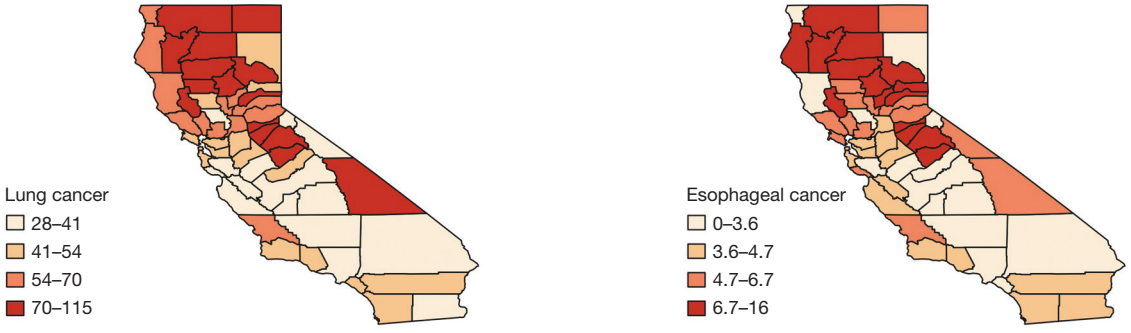

Figure 1 Maps of 5-year average incidence rates per 100,000 population for lung and esophageal cancer in California, 2012–2016.

Table 1 Model comparison using WAIC statistics for cancer data analysis

Model	lppd	*pWAIC*	WAIC
BDAGAR (esophagus | lung)	–261.31	45.32	613.27
BDAGAR (lung | esophagus)	–155.12	51.72	413.68
GMCAR (esophagus | lung)	–264.51	46.09	621.19
GMCAR (lung | esophagus)	–156.51	52.05	417.12

BDAGAR, bivariate directed acyclic graphical autoregressive; GMCAR, generalized multivariate conditional autoregression; WAIC, Widely Applicable Information Criterion.

Autoregression (GMCAR)" models (*26*). In both BDAGAR and GMCAR models, the conditional order [esophagus] × [lung | esophagus] has a smaller WAIC (hence better fit to the data) than the reverse ordering. Meanwhile, within each order, BDAGAR seems to excel over the GMCAR with lower scores in both model fit and effective number of parameters, as seen in the values of $\widehat{elppd}$ and $\hat{p}_{WAIC}$, respectively. The preference of WAIC for [lung | esophagus] is also corroborated by the posterior distribution of η_0 and η_1 from BDAGAR shown in *Figure 2*. In [esophagus | lung], the parameter η_1 has posterior mean of –1.94 and a 95% credible interval (–3.94, –0.58). This shows significant negative values that offset part of the significant positive effect of η_0 with a mean of 7.58 and a 95% credible interval of (2.82, 13.94). For [lung | esophagus], η_0 is significantly positive with a mean of 17.58 and 95% credible interval of (11.62, 27.84), while η_1 tends to be positive with a mean of 1.1 but with a 95% credible interval (–0.77, 2.73) that includes 0. Consequently, we present the following results and analysis for [lung | esophagus] which seems to be the preferred model.

Table 2 summarizes the parameter estimates from the BDAGAR model corresponding to [lung | esophagus]. For fixed effects, the increasing percentage of residents younger than 18 years old significantly reduces the incidence rate for esophageal cancer, while the percentage of residents older than 65 years old has a significantly opposite effect for lung cancer. Unsurprisingly, higher adult cigarette smoking rates significantly increase the incidence rates for both lung and esophageal cancer. After accounting for these explanatory variables, the residual random effects still exhibit spatial association patterns for both cancers. Turning to spatial correlations, ρ_1 measures the residual spatial correlation (posterior mean 0.08) for esophageal cancer after accounting for the explanatory variables and ρ_2 measures the spatial correlation (posterior mean 0.5) for lung cancer after accounting for the explanatory variables and also the effect of esophageal cancer. The small point estimates and narrower credible interval for ρ_1 indicate greater confidence in weaker spatial correlation for esophageal cancer; the moderate value of ρ_2 and a wider credible interval suggest higher spatial correlation for lung cancer. Turning to the spatial precision of random effects for each cancer, the estimates of $\{\tau_1, \tau_2\}$ are indicative of esophageal cancer having larger variability, although we must keep in mind that τ_2 is the conditional marginal precision for lung cancer after accounting for esophageal cancer and, therefore, may not be directly comparable to τ_1.

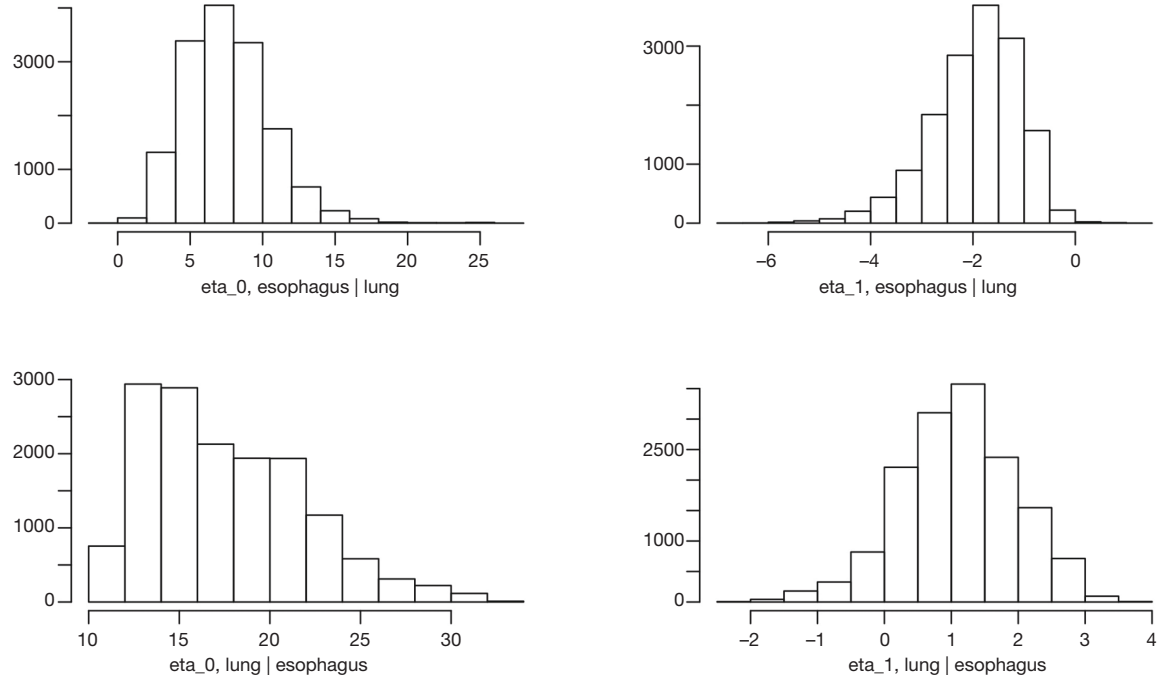

Figure 2 Posterior samples of linking parameters η_0 and η_1 from BDAGAR models. BDAGAR, bivariate directed acyclic graphical autoregressive.

Table 2 Parameter estimates (posterior means) for the California cancer incidence rate data from BDAGAR model. The numbers inside braces indicates the lower and upper bounds for the 95% credible intervals

Parameters	Esophagus	Lung
Intercept	18.75 (4.55, 32.72)	7.19 (–47.07, 61.87)
Smoke	0.27 (0.12, 0.41)	1.27 (0.28, 2.3)
Young	–0.23 (–0.45, –0.01)	–0.75 (–1.94, 0.44)
Old	0.14 (–0.03, 0.31)	2.61 (1.62, 3.61)
Edu	0.02 (–0.1, 0.14)	–0.25 (–1.04, 0.54)
Unemp	–0.07 (–0.26, 0.12)	0.52 (–0.79, 1.84)
Black	0.16 (–0.08, 0.39)	0.8 (–0.82, 2.41)
Male	–0.04 (–0.19, 0.12)	0.14 (–0.95, 1.26)
Uninsure	–0.31 (–0.53, –0.09)	–0.08 (–1.11, 0.94)
Poverty	0.32 (–0.33, 0.96)	0.23 (–3.96, 4.48)
ρ_i	0.08 (0, 0.25)	0.5 (0.03, 0.97)
τ_i	2.72 (0.96, 6.69)	19.41 (2.47, 54.36)
σ_i^2	2.05 (1.39, 3.05)	0.93 (0.18, 3.87)

BDAGAR, bivariate directed acyclic graphical autoregressive.

Figure 3 shows the estimated correlation between lung and esophageal cancer in each of 58 counties. This map also seems to be consistent with the estimates of η. Correlations between lung and esophageal cancers in all counties are significantly positive with large means at around 0.97–1 which are due to the highly positive values in η_0. This indicates that esophageal cancer is highly correlated with lung cancer. However, in general, the correlation between the two cancers increases slightly from the center to marginal areas, especially for those with fewer counties in the neighborhood.

Finally, *Figure 4* provides further visual corroboration of the goodness of fit for the BDAGAR mode corresponding to [lung | esophagus]. Here, we see that the posterior mean of the incidence rates for lung and esophageal cancer are very consistent with the raw incidence rates shown in *Figure 1*. Given the significant effect of adult cigarette smoking rates on incidence rates for both cancers, the higher fitted incidence rates in the northern areas are in accordance with higher smoking rates in same counties as shown in *Figure 5*. Though the smoking rates are also high in the middle part, the relatively lower fitted incidence rates may be due to the offset of negative spatial random effects for these counties.

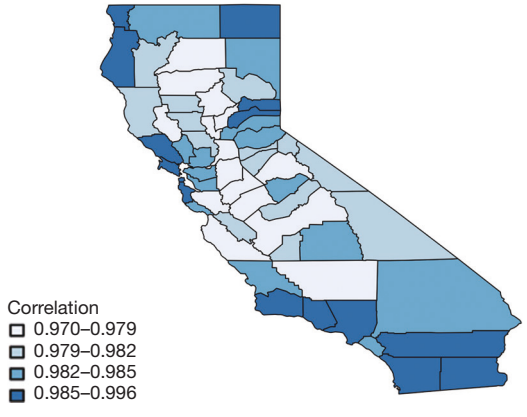

Figure 3 Estimated correlation between lung and esophageal cancer in each of 58 counties of California.

Discussion

We have extended a recently proposed class of DAGAR models (15) for univariate disease mapping to bivariate "BDAGAR" models that can be applied to estimate spatial correlations for two correlated cancers. The BDAGAR model retains the interpretation of DAGAR models clearly separating the spatial correlation for each cancer from any inherent or endemic association between the two cancers. The BDAGAR model can still be efficiently computed using MCMC algorithms. Our analysis of incidence rates from lung and esophagus cancer demonstrates the efficiency of BDAGAR and its improved performance, as measured by WAIC, over existing alternatives such as the GMCAR models. In fact, it has been reported that DAGAR tended to outperform CAR in univariate models (15). It is, therefore, not unexpected that BDAGAR will outperform the bivariate CAR models. We are currently exploring these issues in greater detail and even extending our analysis to more than two cancers. We expect to report our findings in future manuscripts.

While we have restricted our attention only to cancer incidence rates, BDAGAR models can also be used with time-to-event data to investigate geographical patterns in the hazard function. For example, each patient in a study may provide multiple survival times from the onset of each of two cancers along with his or her county of residence. The BDAGAR model can become an excellent alternative to CAR and different MCAR models in spatial survival analysis (18,25,35-37).

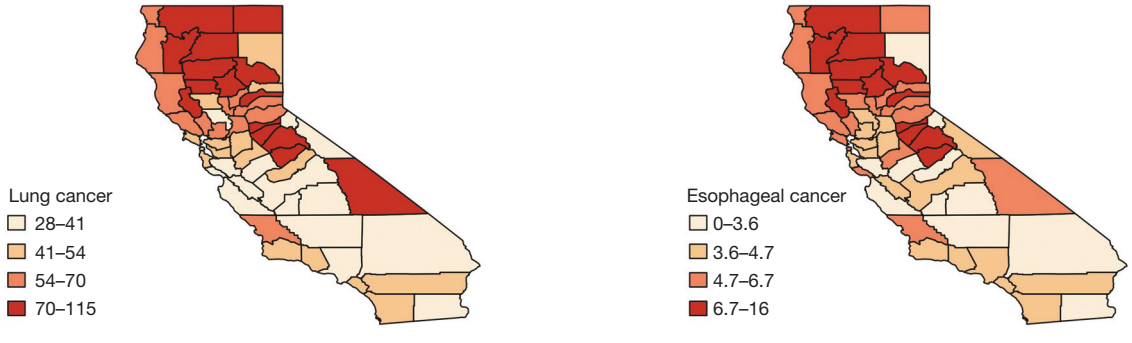

Figure 4 Maps of posterior mean incidence rates per 100,000 population for lung and esophageal cancer in California.

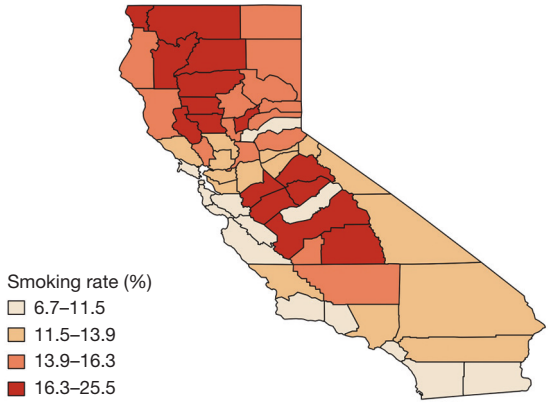

Figure 5 Maps of adult cigarette smoking rates in percentage in California, 2014–2016.

The BDAGAR models developed here proceeds from conditional specifications. Concerns may arise over the ordering of the variables in the hierarchical approach. While in the case of a few cancers, such as 2 in our case, one can evaluate models arising from the different orders, this strategy will become cumbersome with several cancers. For instance, even with 4 cancers, we will have 24 different models that will need to be evaluated and compared. This becomes impractical. A joint modeling approach, analogous to order-free MCAR models (22), can build rich spatial structures from linear transformations of simpler latent variables. For instance, we can develop alternate multivariate DAGAR, or MDAGAR models, using $w = \Lambda f$, where Λ is a suitably specified square matrix and f is a latent vector whose components follow independent univariate DAGAR distributions. Note that by modeling the joint distribution, the incompatibility of conditional model building (i.e., different joint distributions for different orderings) is avoided. However, the issue of the identifiability of Λ is raised, and careful specification of its structure is needed. These approaches will be further investigated elsewhere.

Finally, we caution against using the BDAGAR models developed here for causal inference because of the potential limitations of DAGs in causal analysis. While DAGs assume that relationships are directed and acyclic, truly cyclical or bidirectional relationships may exist as exceptions in causal processes (38). Moreover, the discrete observations DAGs hardly fully capture the underlying time-continuous causal processes and as a result, the conditional independence in DAGs can rarely be identified with causal structure (39). Instead, these models should be used for evincing relationships between the cancers and the strengths of spatial association for each cancer to posit new hypotheses and generate further research. For example, one such finding from these models is that the spatial association exhibited by esophageal cancer seems to be considerably less pronounced than for lung cancer (after controlling for esophageal cancer). This finding is similar to those from earlier spatial analysis of mortality rates for these two cancers (21). Whether this is an artefact of the model itself or of the specific dataset analyzed here or a result of further explanatory variables not accounted for here will need to be further investigated elsewhere.

Acknowledgments

Funding: The work of the first and second authors have been supported in part by the Division of Mathematical Sciences (DMS) of the National Science Foundation (NSF) under grant 1916349 and by the National Institute of Environmental Health Sciences (NIEHS) under grants R01ES030210 and 5R01ES027027. The work of the third author was supported by the Division of Mathematical Sciences (DMS) of the National Science Foundation (NSF) under grant 1915803.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Peter Baade and Susanna Cramb) for the series "Spatial Patterns in Cancer Epidemiology" published in *Annals of Cancer Epidemiology*. The article was sent for external peer review organized by the Guest Editors and the editorial office.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/ace-19-41). The focused issue "Spatial Patterns in Cancer Epidemiology" was commissioned by the editorial office without any funding or sponsorship. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Koch T. Cartographies of disease: maps, mapping, and medicine. Redlands, CA: Esri Press, 2005.
2. Waller LA, Carlin BP, Xia H, et al. Hierarchical spatio-temporal mapping of disease rates. J Am Stat Assoc 1997;92:607-17.
3. Banerjee S. Spatial Data Analysis. Annu Rev Public Health 2016;37:47-60.
4. Schaible WL. Indirect estimators in US federal programs. vol. 108. Springer Science & Business Media, 2013.
5. Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data. Boca Raton, FL: CRC Press, 2014.
6. Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 2005;14:35-59.
7. Waller L, Carlin B. Handbook of Spatial Statistics. Gelfand AE, Diggle PJ, Fuentes M, et al. editors. Boca Raton, FL: Taylor and Franci. USA: Chapman and Hall CRC Press, 2010.
8. Waller LA, Gotway CA. Applied spatial statistics for public health data. vol. 368. John Wiley & Sons, 2004.
9. Lawson AB. Statistical methods in spatial epidemiology. John Wiley & Sons, 2013.
10. Rua H, Held L. Gaussian Markov Random Fields: Theory and Applications. Monographs on statistics and applied probability. Boca Raton, FL: Chapman and Hall/CRC Press, 2005.
11. Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological) 1974;36:192-236.
12. Besag J, York J, Mollié,A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 1991;43:1-20.
13. Anselin L. Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. Geographical Analysis 1988;20:1-17.
14. Kissling WD, Carl G. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecology and Biogeography 2008;17:59-71.
15. Datta,A, Banerjee S, Hodges JS, et al. Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. Bayesian Analysis 2019;14:1221-44.
16. Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2001;164:73-85.
17. Held L, Natário I, Fenton SE, et al. Towards joint disease mapping. Stat Methods Med Res 2005;14:61-82.
18. Diva U, Dey DK, Banerjee S. Parametric models for spatially correlated survival data for individuals with multiple cancers. Stat Med 2008;27:2127-44.
19. Martinez-Beneito MA. A general modelling framework for multivariate disease mapping. Biometrika 2013;100:539-53.
20. Marí-Dell'Olmo M, Martinez-Beneito MA, Gotsens M, et al. A smoothed ANOVA model for multivariate ecological regression. Stoch Environ Res Risk Assess 2014;28:695-706.
21. Kim H, Sun D, Tsutakawa RK. A Bivariate Bayes Method for Improving the Estimates of Mortality Rates With a Twofold Conditional Autoregressive Model. J Am Stat Assoc 2001;96:1506-21.
22. Jin X, Banerjee S, Carlin BP. Order-free co-regionalized areal data models with application to multiple-disease mapping. J R Stat Soc Series B Stat Methodol 2007;69:817-38.
23. Zhang Y, Hodges JS, Banerjee S. Smoothed ANOVA with spatial effects as a competitor to MCAR in multivariate spatial smoothing. Ann Appl Stat 2009;3:1805-30.
24. Gelfand AE, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 2003;4:11-25.
25. Carlin BP, Banerjee S. Hierarchical multivariate CAR models for spatio-temporally correlated survival data. Bayesian Statistics 2003;7:45-63.
26. Jin X, Carlin BP, Banerjee S. Generalized hierarchical multivariate CAR models for areal data. Biometrics 2005;61:950-61.
27. Agrawal K, Markert RJ, Agrawal S. Risk factors for adenocarcinoma and squamous cell carcinoma of the esophagus and lung. AME Med J 2018;3:35.
28. Shi WX, Chen SQ. Frequencies of poor metabolizers of cytochrome P450 2C19 in esophagus cancer, stomach cancer, lung cancer and bladder cancer in Chinese population. World J Gastroenterol 2004;10:1961-3.
29. Gamerman D, Lopes HF. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Chapman and Hall/CRC, 2006.

30. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable in- formation criterion in singular learning theory. Journal of Machine Learning Research 2010;11:3571-94.
31. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing 2014;24:997-1016.
32. Step 1: Calculating Age-adjusted Rates - SEER*Stat Tutorials. Available online: https://seer.cancer.gov/seerstat/tutorials/aarates/step1.html
33. Static County Attributes – SEER Datasets. Available online: https://seer.cancer.gov/seerstat/variables/countyattribs/static.html
34. California Department of Public Health, California Tobacco Control Program. California Tobacco Facts and Figures 2018. Sacramento, CA: California Department of Public Health, 2018.
35. Banerjee S, Wall MM, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. Biostatistics 2003;4:123-42.
36. Banerjee S, Dey DK. Semiparametric proportional odds models for spatially correlated survival data. Lifetime Data Anal 2005;11:175-91.
37. Cooner F, Banerjee S, McBean AM. Modelling geographically referenced survival data with a cure fraction. Stat Methods Med Res 2006;15:307-24.
38. Pearce N, Lawlor DA. Causal inference-so much more than statistics. Int J Epidemiol 2016;45:1895-903.
39. Aalen OO, Røysland K, Gran JM, et al. Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. Stat Methods Med Res 2016;25:2294-314.

doi: 10.21037/ace-19-41
Cite this article as: Gao L, Banerjee S, Datta A. Spatial modeling for correlated cancers using bivariate directed graphs. Ann Cancer Epidemiol 2020;4:8.