Annals of Epidemiology 27 (2017) 42–51

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

# Space-time variation of respiratory cancers in South Carolina: a flexible multivariate mixture modeling approach to risk estimation

Rachel Carroll PhD [a,*], Andrew B. Lawson PhD [a], Russell S. Kirby PhD [b], Christel Faes PhD [c], Mehreteab Aregay PhD [a], Kevin Watjou MS [c]

[a] Department of Public Health, Medical University of South Carolina, Charleston
[b] Department of Community and Family Health, University of South Florida, Tampa
[c] Interuniversity Institute for Statistics and Statistical Bioinformatics, Hasselt University, Agoralaan 1, Diepenbeek, Belgium

ARTICLE INFO

*Article history:*
Received 9 April 2016
Accepted 23 August 2016
Available online 31 August 2016

*Keywords:*
Spatiotemporal
Bayesian
Multivariate
Respiratory cancers
Mixture modeling
MCMC
Lung and bronchus cancer

ABSTRACT

*Purpose:* Many types of cancer have an underlying spatiotemporal distribution. Spatiotemporal mixture modeling can offer a flexible approach to risk estimation via the inclusion of latent variables.
*Methods:* In this article, we examine the application and benefits of using four different spatiotemporal mixture modeling methods in the modeling of cancer of the lung and bronchus as well as "other" respiratory cancer incidences in the state of South Carolina.
*Results:* Of the methods tested, no single method outperforms the other methods; which method is best depends on the cancer under consideration. The lung and bronchus cancer incidence outcome is best described by the univariate modeling formulation, whereas the "other" respiratory cancer incidence outcome is best described by the multivariate modeling formulation.
*Conclusions:* Spatiotemporal multivariate mixture methods can aid in the modeling of cancers with small and sparse incidences when including information from a related, more common type of cancer.


## Introduction

Respiratory cancers as a whole (ICD-9-CM codes: 160–163 and 165) are among the most common types of cancer worldwide, and lung and bronchus cancers (ICD-9-CM codes: 162.2–162.5, 162.8, and 162.9) are among the highest annual incidence cancers in the United States. Although incidence of lung and bronchus cancers as well as all respiratory cancers has been decreasing nationally [1], within the state of South Carolina incidence has increased in recent years. Cancers of the respiratory system display spatiotemporal (ST) patterns [2–8] and affect people of all ages, races, and genders, though not equally [1,9]. Thus, it is of interest to explore this occurrence via the use of complex ST statistical modeling.

Lung and bronchus cancers (LBCs) differ from "other" respiratory cancers (ORCs) in their risk factors as well as their rarity. The vast majority of LBCs are attributed to first- and second-hand tobacco smoke exposure and, to a lesser extent, occupational exposures [10,11]. Conversely, ORCs, which includes laryngeal, nasopharyngeal, mediastinal, pleural, and related cancers, are quite rare. ORCs also have associations with alcohol consumption [12], ethnicity, and co-occurrence of other diseases, viruses, or neoplasms [13–15].

Because these types of cancers affect related areas of the body, it is of interest to explore the usefulness of multivariate modeling for LBC with ORC. Furthermore, it is hypothesized that the multivariate modeling of these two diseases will improve the fit and prediction of the less common disease (ORC) by way of the additional information provided by the more common disease (LBC). This multivariate extension of a mixture model can easily allow predetermined information to be shared between the diseases of interest and ultimately lead to stronger, more robust results with the potential for selection to occur. In this exploration, information is shared via the use of common random effects between diseases, which represent unobserved effects, correlation, or both between the diseases. We believe this multivariate modeling could aid in explaining the etiology of the outcomes of interest. Most cancer epidemiologic studies focus on specific cancers, yet when studied ecologically cancer incidence often is elevated across several types of cancer. By developing multivariate mixture models, we are able to examine the joint effects of covariates on several distinct types of cancer simultaneously, within a spatial framework. Insights from these models may provide clues to environmental and social factors

* Corresponding author. Department of Public Health, Medical University of South Carolina, Charleston, SC 29425, USA. Tel.: 843-876-1100; fax: 843-792-6000.
*E-mail address:* mosra@musc.edu (R. Carroll).

http://dx.doi.org/10.1016/j.annepidem.2016.08.014
1047-2797/

hitherto unexplored in more traditional approaches to cancer epidemiology.

This article is developed as follows. First, we describe the available data. Second, we examine the different models to be applied. Next, we display the results of using these methods to the outcomes of interest. Finally, we discuss the results and draw conclusions.

## Materials and methods

Data for this study included measures of incidence of respiratory cancers for each of the 46 counties in the state of South Carolina, United States, annually for the years (1996–2009), and predictors from a variety of data sets. Because our outcome of interest was the incidence of either LBC or ORC, a conditionally independent Poisson distribution was a reasonable model for these data. This is a commonly assumed model for small area counts in disease mapping [16] and is appropriate because the Poisson distribution is a discrete frequency distribution that provides the probability of events occurring in a given area. It is described as follows:

$$y_{ij}\big|\mu_{ij} \sim Pois\left(\mu_{ij}\right)$$

$$\mu_{ij} = e_{ij}\theta_{ij}$$

with $y_{ij}$ the incidence of cancers in county $i$ at time $j$ and $\mu_{ij}$ as the mean of the Poisson distribution such that it is the product of the expected rate of disease, $e_{ij}$, and the relative risk, $\theta_{ij}$, for county $i$ at time $j$. The ICD-9-CM codes are listed and described in supplementary Table A.1, and a labeled county map is included in the supplementary materials as Figure A.1.

### *Case study data*

The outcomes of interest were acquired from the South Carolina Community Assessment Network data sets [17]. The respiratory cancer data was collected in two steps. First, a table of all county-level respiratory cancer incidences in the state of South Carolina annually for years 1996–2009 was collected. Similar data for lung and bronchus cancers were also collected. Then, the ORC incidences outcome was calculated by subtracting the counts of LBC from that of the all respiratory cancer data set. Note that all county boundaries remained constant over the study time period.

After these calculations, the resulting statistics were available for each disease classification to be used for analysis. Summary statistical figures are available as Figures A.2 and A.3 in the supplementary materials; these figures confirm that both incidence and rate of disease were increasing over the study time period for both LBC and ORC. For the 44,668 diagnosed LBC across South Carolina during these study years, there was approximately a mean incidence of 69 cases per county per year ranging from 5 to 321 cases per county leading to the following rate of disease: 0.0008. For the 4077 diagnosed ORC across South Carolina during these study years, there was a mean incidence of six cases per county per year with a range of 0–30 cases per county leading to the following rate of disease: 0.00007. Although these data had no missing values at the county level, there was some censoring where counts of 1–4 were given the value 5 and counts of 5–10 were given the value of 10. This occurs in about 5.3% ($n = 34$) of the county-level measures across the study time for the all respiratory cancer incidence table and 7.5% ($n = 48$) of the county-level measures across the study time for the lung and bronchus cancer incidence table. In these cases, the censored values were assumed to be the true value, and because this occurs in such a small portion of the data, we believe that the impact of this assumption was negligible. However, when considering stratification by age, the amount of censored data does become more of a hindrance. Furthermore, the distribution of incidences by gender was nearly identical; thus, we performed this analysis using the county-level population without stratification.

The indirect standardized incidence rate (SIR) per county over time for each disease classification are displayed in Figure 1 (LBC) and Figure 2 (ORC). The SIR was calculated as the ratio of the observed cancer incidences to the expected rates of disease for each of the 46 counties and can be useful as a first step in data analysis [18]. Therefore, an SIR of one indicates that the observed incidence is equal to that of the expected count for a particular county at a particular time. This expected rate of disease is the product of the above overall rate of disease (calculated as the total statewide incidences across the study time divided by total statewide population over the study time, this rate is also the horizontal line present in supplemental Figure A.3) and the county-level population per each study year, and it is the same as that which was used in the Poisson modeling. The expected rate was not age group specific because, at the population level, age-specific variation was very small across counties as well as over time. Furthermore, if there were age effects present, their impact would be very marginal on the results. These plots are discussed further in the results section. The SIR is an estimator of the relative risk, thus estimated values for $\theta_{ij}$ under our models are comparable, albeit smoothed, to the SIR estimates.

The predictors for this analysis were obtained from a variety of resources as they were both demographic and environmental measures as well as spatial, temporal, or both in structure. The demographic predictors came from the Area Health Resource Files [19] data set and consisted of proportion of persons with health insurance (pHI), proportion of African American population (pAA), unemployment rate of those 16 years or older (UER), and proportion of persons in poverty (pppov). The "proportion" forms of the predictors were calculated by taking the ratio of a "number of persons" measure to the county-level populations acquired from the South Carolina Community Assessment Network data sets. The three environmental variables were average daily sunlight (sun) acquired from the North America Land Data Assimilation System [20], average in-home radon concentrations (pCi/L) based on in-home test kit results analyzed by the South Carolina Department of Health and Environmental Control laboratory [21] and statewide average annual rainfall from the National Oceanic and Atmospheric Administration [22]. The seven selected predictors were chosen based on availability and reasonable amount of collinearity. Additionally, variables selected as predictors have been included in previous research examining associations of environmental exposures and socioeconomic status with cancer incidence both overall and more specifically with cancer of the lung and bronchus [1,23–25]. However, these predictors were selected primarily for their ability to perform as representations of the behavior of spatial, temporal, and ST variables. These predictors were standardized (zero mean and variance of one) for the model fitting.

Among the predictor variables, radon, pHI, and pAA were set to be spatially varying predictors. They were chosen as such because of either availability or lack of temporal variation in the distribution of the predictor determined via qualitative procedures. Their distributions are displayed in supplemental Figure A.4, and this display illustrates that spatial variation. pHI and pAA were recorded in the year 2000, whereas radon is an average of test results that is considered current through October 31, 2014 according to South Carolina Department of Health and Environmental Control.

The predictor rainfall was chosen as an example of a temporal varying predictor because its measures of average annual rainfall were collected per county for each of the study years, and this is the