# Estimating colorectal cancer risk in Wisconsin counties

Joule Voelz

February 2023

**Abstract**

In this project, I estimate the relative risk of colorectal cancer over the period of 2015-2019 across Wisconsin counties by fitting a hierarchical Poisson model with structured and unstructured spatial effects and three covariates: smoking, poverty, and rate of preventative endoscopies. While the signs of the coefficients in the estimated model make intuitive sense, the model's fitted values do not fully capture the empirical relative risk and autocorrelation among counties. There seems to be a greater relative risk of colorectal cancer incidence and mortality in the northeast region of the state – most notably in Forest, Langlade, and Oneida counties. Further research is required to determine the covariates associated with this variation.

## 1 Introduction

Colorectal cancer is a type of cancer that develops in the large intestine when a benign tumor (called a polyp) becomes malignant. While colorectal cancer typically affects older adults, recent studies in the United States have shown the risk for colorectal cancer has more than doubled for people born in 1990 compared to people born in 1950. Apart from a family history of the disease, risk factors for colorectal cancer include sedentary lifestyle, being overweight or obese, smoking, heavy alcohol use, low-fiber-high-fat diets or diets high in processed meats.[2]

This project seeks to estimate the relative risk for colorectal cancer across counties in the state of Wisconsin using data from the National Cancer Institute State Cancer Profiles. Relative risk is modeled as a linear function of three covariates: smoking, poverty, and preventative endoscopies. Bringing together this data, I estimate the average annual incidence of and mortality from colorectal cancer between the years 2015-2019 as a hierarchical Poisson model using INLA (integrated nested Laplace approximation) in R.

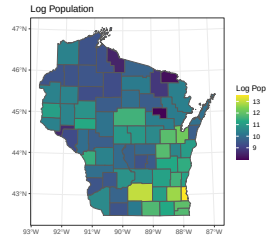
## 2 Data

Data for this project was obtained from the US National Cancer Institute's State Cancer Profiles, a website that brings together and visualizes data from State Cancer Registries, the US Census Bureau, and the American Community Survey.[1] Incidence data includes the average annual count of colorectal cancer cases and deaths in each of 71 Wisconsin counties for the years 2015-2019 from the National Program of Cancer Registries. Because colorectal cancer occurs more often in older adults, I obtained data for adults split into two strata: adults ages 18-64, and adults 65 and over. When the average incidence was 3 or fewer cases per year, the number was not reported. For this reason, 3 counties did not report average annual cases and 12 counties did not report average annual deaths. These counties are excluded from the sample when fitting the model.

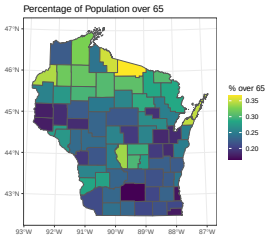
In order to calculate expected disease risk, I obtained county population data (2015-2019) by age group, sourced from the US Census Bureau and American Community Survey. I obtained data on three possible covariates: smoking, poverty, and rate of colorectal endoscopies (preventative screenings). Data from the Behavioral Risk Factor Surveillance System and the National Health Interview Survey provide estimates of the percentage of adults over 18 who currently smoke and/or have ever been a smoker. Data from the US Census Bureau provides estimates on the percentage of people in each county who live below the poverty line. Data from the Behavioral Risk Factor Surveillance System and the National Health Interview Survey provide estimates on the percentage of adults over 50 who have ever received a colorectal endoscopy screening.

## 3 Methods

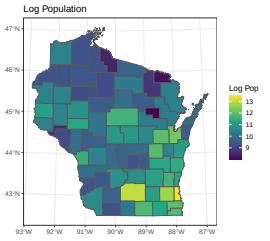
Following the methodology outlined in Paula Moraga's *Geospatial Health Data* (2019), I model the average annual number of cases $Y_i$ in county $i$ as a Poisson distribution whose expected value depends on the relative risk of colorectal cancer $\theta_i$ in county $i$.[3] $E_i$ is the expected number of cases in county $i$ if county $i$ were to behave as the state of Wisconsin does as a whole.




(a) Log Population

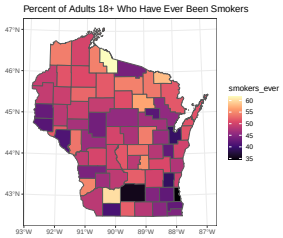



(b) % over 65

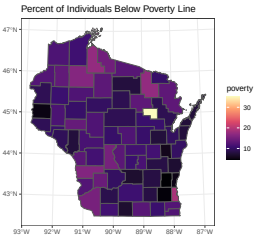



(c) Log Pop Density

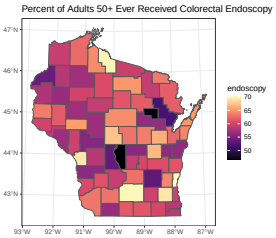
Figure 1: Mapping population




(a) Current Smokers




(b) Poverty




(c) Endoscopies

Figure 2: Mapping covariates

$$Y_i|\theta_i \sim Poisson(E_i \times \theta_i) \tag{1}$$

We assume that the relative risk $\theta_i$ can be modeled as a function of the covariates, plus two noise terms.

$$\log \theta_i = \beta_0 + \beta_1 Smoking_i + \beta_2 Poverty_i + \beta_3 Endoscopies_i + u_i + v_i \tag{2}$$

Here $Smoking_i$ is the percentage of adults in county $i$ who have ever been smokers, $Poverty_i$ is the percentage of people in that county who live below the poverty line, and $Endoscopies_i$ is the percentage of adults over 50 who have ever received a colorectal endoscopy. We expect the coefficient $\beta_1$ to be positive, as smoking is a risk factor for colorectal cancer. We expect $\beta_2$ to be positive as well, since poverty is associated with poor diet and other risk factors for colorectal cancer. We expect $\beta_3$ to be negative, since a higher rate of endoscopy screenings should correlate with more preventative care and fewer incidents and deaths from colorectal cancer.

$$u_i|u_{-i} \sim N(\bar{u}_{\delta_i}, \frac{1}{\tau_u n_{\delta_i}}) \qquad v_i \sim N(0, \frac{1}{\tau_v}) \tag{3}$$

$u_i$ is a structured spatial effect that captures the correlation between county $i$ and its neighboring counties. It is distributed normally around the average structured spatial effect of its neighboring counties ($\delta_i$), with a variance term that decreases with the number of neighbors $n_{\delta_i}$. $v_i$ is an unstructured spatial effect that is centered at zero and uncorrelated among counties.

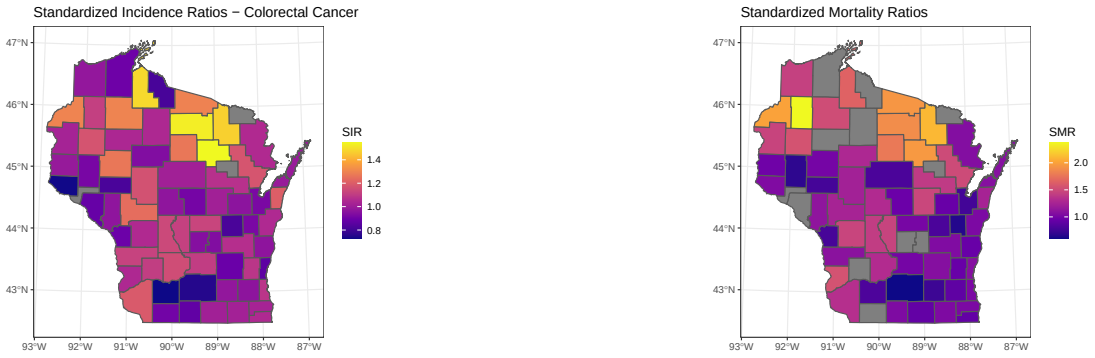
$$E_i = \sum_j^m r_j^{(s)} n_j^{(i)} \tag{4}$$

The expected risk $E_i$ of county $i$ is calculated by summing the product of $r_j^{(s)}$ (incident rate in stratum $j$ across the entire state) with $n_j^{(i)}$ (population of stratum $j$ in county $i$). In this project I work with only two strata: people 64 and younger, and people 65 and older.

I use the R package INLA to estimate Equation 1 using integrated nested Laplace approximation.

# 4 Results and Discussion

After cleaning and joining the data, I begin by calculating Standardized Incidence Ratios (SIR) and Standardized Mortality Ratios (SMR) for Wisconsin counties. SIR is the ratio of observed to expected cases of colorectal cancer observed in each county, while SMR is the ratio of observed to expected deaths from colorectal cancer. For colorectal incidences, the expected cases $E_i$ was calculated with two strata: adults under 65, and adults 65




(a) SIR

(b) SMR

Figure 3: Standardized Incidence and Mortality Ratios

and older. For mortality, expected mortality was calculated with only one stratum (adults over 18) due to a low incidence of death under 65 and hence missing data for many counties.

The results (Figure 3) seem to show a higher SIR and SMR in the northeast part of the state, concentrated around Oneida, Langlade, and Forest counties. These counties are in the less populated region of the state, where a higher percentage of the population is over 65. The maps of incidence and mortality ratios seem to show autocorrelation among neighboring counties.

Using the INLA package to estimate Equation 1 on colorectal cancer incidence yields the coefficients in Table 1. As expected, the coefficients on smokers and poverty are positive, but they are not significantly different from zero. On the other hand, the coefficient on endoscopies is negative, larger in magnitude, and significant. As we might hope, living in a county with a higher percentage of adults who have received preventative screening is associated with a lower relative risk of colorectal cancer. This may be due to the direct effect of the screenings, but could also be due to unmeasured confounder variables like access to medical facilities, higher rate of medical insurance, or income.

| | mean | sd | 0.025quant | 0.975quant |
|---|---|---|---|---|
| (Intercept) | 0.565 | 0.433 | -0.286 | 1.415 |
| smokers | 0.005 | 0.004 | -0.003 | 0.012 |
| poverty | 0.006 | 0.006 | -0.006 | 0.017 |
| endoscopy | -0.013 | 0.005 | -0.023 | -0.004 |

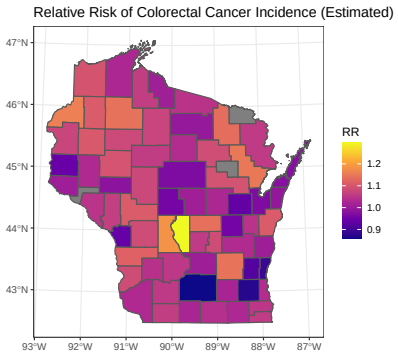
Table 1: Fitted coefficients (Incidence)

| | mean | sd | 0.025quant | 0.975quant |
|---|---|---|---|---|
| (Intercept) | 0.120 | 0.744 | -1.339 | 1.579 |
| smokers | 0.022 | 0.006 | 0.009 | 0.035 |
| poverty | 0.015 | 0.010 | -0.005 | 0.035 |
| endoscopy | -0.020 | 0.008 | -0.036 | -0.004 |

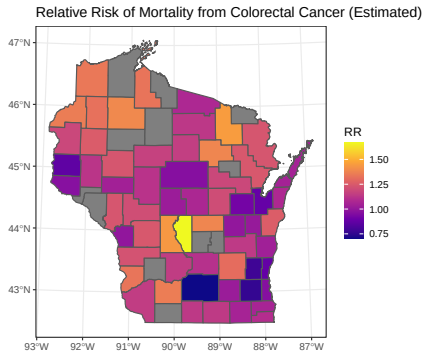
Table 2: Fitted coefficients (Mortality)

Estimating the same model on colorectal cancer mortality yields coefficients with the same signs (Table 2). However, in this case the coefficients are all larger in magnitude. Now, the coefficient on smoking is positive and significant, and the coefficient on endoscopies is larger and remains significant as well. It is difficult to draw conclusions here, as we are missing more data on mortality (there are 12 counties with 3 or fewer deaths per year) than on incidence (there are 3 counties with 3 or fewer cases per year). Thus the sample suffers from self-selection bias and coefficients tend to reflect estimated effects for counties with a higher incidence of colorectal cancer. That being said, among these counties it seems that the rate of smoking and the rate of preventative screenings – or related confounding factors – are related to the relative risk of mortality from colorectal cancer, moreso than the risk of incidence. And this makes some sense. At least part of the incidence of colorectal cancer is probably due to genetic factors that may be uncorrelated with lifestyle choices like smoking or drinking, or with poverty. However, conditional on being genetically predisposed cancer, individuals with a healthier lifestyle, better access to healthcare, and a higher income probably have a better chance of surviving a fight with cancer.

While the estimated coefficients have sensible signs, the fitted values of relative risk do not seem fully capture

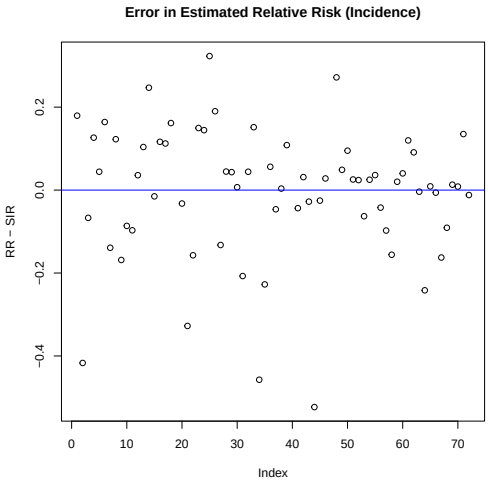



(a) Incidence

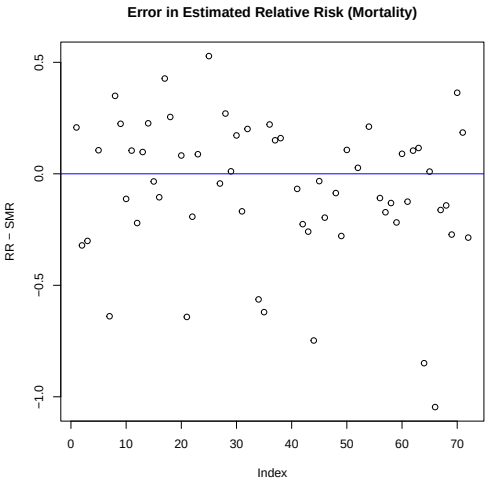



(b) Mortality

Figure 4: Estimated Relative Risk of Colorectal Cancer




(a) SIR Error




(b) SMR Error

Figure 5: Error in Estimated Relative Risk

the pattern of variation in the empirically calculated standardized incidence and mortality ratios (compare Figure 3 to Figure 4). A plot of the errors (Figure 5) shows that the fitted values are unable to explain some particularly high relative risk values.

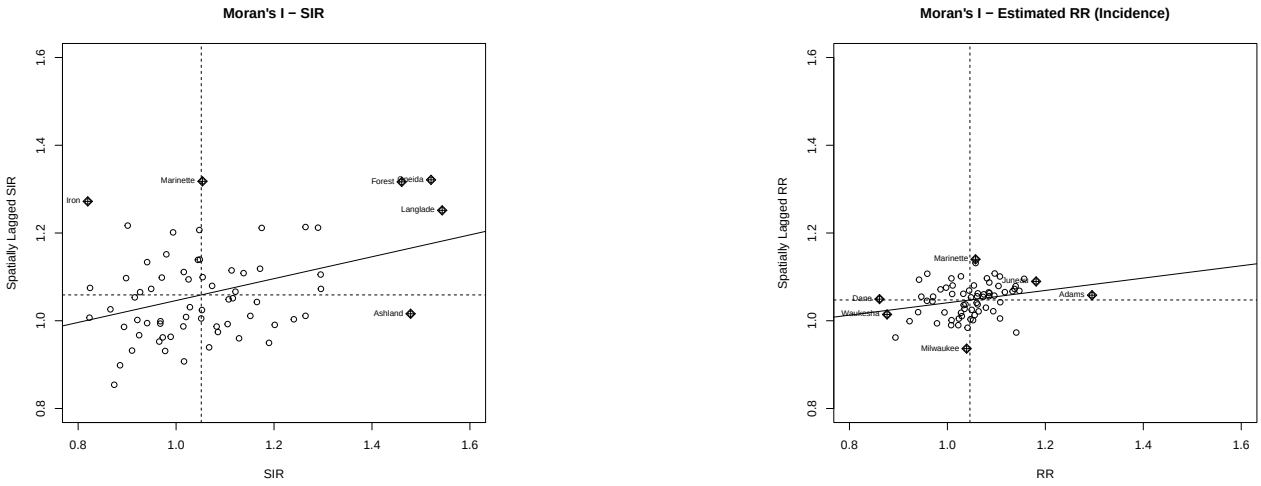
The predicted values also fail to capture the autocorrelation in the SIRs, as captured by Moran's $I$ statistic. Moran's $I$ measures the degree of global spatial autocorrelation in a dataset. It is calculated:

$$I = \frac{\sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i \neq j} w_{ij}(Y_i - \bar{Y})^2} \tag{5}$$

where $w_{ij}$ are weights that measure the distance between counties $i$ and $j$. $I$ is much larger than its expected value $E[I]$, when there is clustering in the data. The Moran's $I$ statistics for both the empirical SIR and the estimated RR indicate the presence of spatial autocorrelation with a 95% confidence level, but Moran's $I$ for SIR is 0.249, while Moran's $I$ for estimated RR is 0.141. In other words, the empirical risk of colorectal cancer is more spatially autocorrelated than the fitted estimates.

This can be seen clearly in plots of SIR and RR versus the weighted average of neighboring values, also known as spatially lagged values (Figure 6). The empirically calculated SIRs indicate higher spatial autocorrelation, especially in the northeast Forest, Langlade, and Oneida counties. In the plot of estimated RR's, there is less autocorrelation and less variation in general.

Taken together, this evidence points to the idea that there are important covariates missing from the model specification, or that the model may be misspecified in some way. There seems to be a higher relative risk of colorectal cancer incidence and mortality in the north and particularly northeast counties of Wisconsin, but so far this model specification cannot explain why.




(a) Empirical SIR

(b) Estimated RR

Figure 6: Moran's $I$ Scatterplot

# 5 Conclusion

In this project, I estimated the relative risk of colorectal cancer in Wisconsin counties by fitting a hierarchical Poisson model with structured and unstructured spatial effects. The model included three covariates: smoking, poverty, and rate of preventative endoscopies. While the signs of the coefficients on the estimated model made intuitive sense, the model's fitted values did not fully capture the variation observed in the empirically estimated SIR and SMRs. The fitted values also failed to capture the spatial autocorrelation observed the empirical data, as demonstrated by Moran's $I$ statistic. There seems to be a greater relative risk of colorectal cancer incidence and mortality in the northeast region of the state – most notably in Forest, Langlade, and Oneida counties – that is not explained by the model.

Given more time to work on this project, I would like to enrich the model with more relevant covariates, for example: alcohol risk, income, health insurance coverage, number of health facilities per county, percent rural population. I would also like to use a sample with more relevant strata to calculate a more accurate expected value of cases. For example, since colorectal cancer is more common in men, it would be helpful to have a sample stratified by both age and sex. Since colorectal cancer rates among young people are on the rise, it would also be informative to model relative risk over time.

# References

[1] National Cancer Institute. *State Cancer Profiles*. URL: `https://www.statecancerprofiles.cancer.gov/index.html`.

[2] Kathy Katella. *Colorectal Cancer: What Gen-Xers and Millennials Need to Know*. 2022. URL: `https://www.yalemedicine.org/news/colorectal-cancer-in-young-people`.

[3] Paula Moraga. *Geospatial Health Data*. Chapman Hall/CRC, 2019. URL: `https://www.paulamoraga.com/book-geospatial/`.