AmeriCancer

Ben Shea, Sarah Tsay, and Kate Wright

12/14/2020

Contents

1 Abstract

Much has been learned about colorectal cancer risk over the past two decades, and many analyses have focused on both individual level factors that influence cancer survival, including tumor grade at diagnosis, tumor size, and treatment options. The scientific community has also explored the influence of environmental level factors on colorectal cancer risk, including socioeconomic inequalities and racial disparities. Factors such as healthcare access, obesity, physical activity, diet, and chronic conditions (e.g., diabetes mellitus) have been associated with the incidence of colorectal cancer. In this analysis, we attempt to answer the following question: what are the key risk factors for colorectal cancer deaths in the 100 largest US cities?

We used data from the City Health Dashboard (https://www.cityhealthdashboard.com/metrics), maintained by NYU Langone Health, to analyze count of deaths due to colorectal cancer outcomes from the 100 most populous US cities. Initially, poisson regression models were constructed to understand key risk factors given the count nature of the outcome but, due to overdispersion in the poisson model, we ultimately used the negative binomial regression model. Environmental factors, socioeconomic factors, and clinical care factors were significantly associated with count of colorectal cancer deaths, but individual clinical outcome metrics were not as strongly associated with colorectal cancer deaths. This showed that factors in the literature that were related to incidences of colorectal cancer did not necessarily align with factors that had a significant effect on colorectal cancer deaths at a macro (ie. large cities) level.

2 Introduction

Colorectal cancer accounts for 8.8% of all cancer deaths[1] and is the third most common form of cancer deaths in men and women in the United States[2]. Much work has been done to examine etiology of the disease, often in specific populations. The incidence and mortality rate is highest for African American males[3].

Individual risk factors for colorectal cancer include obesity or overweight, physical inactivity, poor diet and nutrition, smoking, and diabetes[4], which many studies have explored these risks in detail. Chronic health conditions, such as diabetes and hypertension, have been associated with colorectal cancer incidence and mortality[5,6]. Other environmental and health system factors have been explored, including links between socioeconomic status (SES) and colorectal cancer deaths. There is also evidence from US longitudinal mortality studies[8] that individuals in more deprived areas or lower education and income groups had higher mortality and incidence rates than their more affluent counterparts. Patients with low SES have been found to receive less adjuvant therapy and have worse survival and mortality rates than those with high SES.[7,8]

The City Health Dashboard breaks the data down into five categories: Health Outcomes, Social and Economic Factors, Health Behavior, Physical Environment, and Clinical Care. Variables from each of these categories were selected for analysis because of their links to cancer as either predictors or risk factors for cancer incidence. For a detailed list of the variables included, refer to the detailed data dictionary in the appendix. Note that all predictors are continuous variables.

More than 1/2 of all cases and deaths are attributable to modifiable risk factors, such as smoking, an unhealthy diet, high alcohol consumption, physical inactivity, and excess body weight, and thus potentially preventable.[10] While many studies have focused on specific risk factors or populations, this study uniquely examines a variety of risk factors across urban populations at a macro-level. Furthermore, some of the variables collected at the city level are reflective of individual level risk factors for colorectal cancer. For instance, diabetes, high blood pressure, smoking, physical inactivity are all measured on a macro level in this dataset. The key question of this analysis is: what are the key city-level risk factors for deaths due to colorectal cancer in the 100 largest cities in the United States?

3 Review of Literature and Domain Expertise

Chronic health conditions, such as diabetes and hypertension, have been associated with colorectal cancer incidence and mortality.[5,6] Type II diabetes and colorectal cancer share common risk factors, including high body mass index and central adiposity, low physical activity, cigarette smoking, and diet.[5] Hypertension is also strongly associated with colorectal cancer. Adrenergic receptor stimulation, involved in the development of hypertension, has also been implicated in metastasis development in colon cancer.[6]

A study in New Zealand[7] on the impact of changing socioeconomic inequalities in cancer incidence and mortality posits that cancer may be increasingly responsible for the mortality gap between high and low socioeconomic position groups in high-income countries. There is also evidence from US longitudinal mortality studies[8] that individuals in lower education and income groups had higher mortality and incidence rates than their more affluent counterparts. Patients with colorectal cancer who were uninsured or insured by Medicaid or commercial HMOs had higher mortality rates than patients with commercial fee-for-service insurance.[11]

Built environment, including access to parks, and environmental health are also factors in colorectal cancer mortality. Aspects of the neighborhood environment have been shown to contribute independently to overall mortality. In addition, research has demonstrated that social and built neighborhood characteristics shape opportunities for and barriers to health promotion.[10]

Colorectal cancer incidence and mortality are highest in African Americans (52 per 100,000) and lowest in American Hispanics (37 per 100,000).[3] Comparative studies with Native Africans (<5 per 100,000) suggest environmental influence, rather than genetic susceptibility.[3] Studies suggest that risk is high because of excessive intake of animal meat and fat products and differences in colonic bacterial metabolism.[3,12] Alcohol intake is another major risk factor for colorectal cancer; moderate drinking increases colorectal cancer risk[13]. Acetaldehyde from alcohol metabolism leads to activation of cancer promoting cascades (e.g., oxidative stress).[13]

While other studies have provided insight on colorectal cancer epidemiology, including proposed mechanisms of risk factor actions, this study will uniquely be examining macro-level observations of some important individual risk factors for colorectal cancer mortality across America's largest cities.

4 Research and Analysis Methods

The NYU Langone Health dataset has over 750 American cities, with this study's outcome of interest being number of deaths due to colorectal cancer per 1,000,000 people in each city. The sample size has been limited to the 100 most populous cities since cities with fewer than 1,000,000 people would have little to no cases on a per 1,000,000 basis. Furthermore, it was important to link stage-at-diagnosis and colorectal cancer incidence rates to this data set, as it is one of the best predictors for survival from any type of cancer. Information on the colorectal cancer incidence rates and percent of colorectal cancer cases that were considered late-stage at diagnosis was gathered at the county level from the SEER cancer database[1], then linked to the cities based on the corresponding county. For cities that span multiple counties, a population-weighted average was calculated.

We first performed exploratory data analysis to have a better understanding of the data, particularly potential multicollinearity among the covariates. Assessing the correlation matrices in Figure 1, frequent physical distress, life expectancy, income inequality, and preventive services have the strongest linear association with count of colorectal cancer deaths. Furthermore, within the Health Outcomes metrics, many of the covariates are highly correlated with each other, such as, the prevalence of diabetes and prevalence of high blood pressure and diabetes and frequency of physical distress. This may cause multicollinearity and confounding in the Health Outcomes model, which we assess when fitting the poisson regression models.

Given that the original data had decimals (e.g. 13.1 colorectal cancer deaths per 100,000 people), we multiplied the outcome by 10 to become the number of colorectal cancer deaths per 1 million people for each city so we would have a discrete count value; we also confirmed that none of the covariates have to be scaled up. We fitted various poisson and negative binomial models to determine the best model and assess which metrics have a statistically significant association with the count of colorectal cancer deaths. Our models included an offset of 1 million to account for the counts per 1 million people. We will also test goodness of fit with the best poisson regression model and, if overdispersion occurs, we will use the negative binomial model instead.

5 Findings and Analysis

First, we fit the poison models and then assessed goodness of fit. We calculated deviance statistic/df and, as this metric is much greater than 1 across all poisson models, overdispersion occurs and the actual variability in the count of colorectal cancer deaths is much greater than the variability predicted by the model. Furthermore, the distribution of the number of deaths due to colorectal cancer (Figure 2) does not follow the poisson distribution whose mean is estimated from the outcome variable (Figure 3), therefore the poisson regression assumptions do not hold. For this reason, we will compare negative binomial models as this model has looser assumptions for the mean-variance relationship.

Next, given that diabetes and frequent physical distress were highly correlated with each other, we assessed whether diabetes was a confounder on the effect of frequent physical distress on the count of colorectal cancer deaths. As the frequent physical distress coefficient changed by more than 10% (from 0.073 to 0.055, see Table 2), diabetes is a meaningful confounder on the association between frequent physical distress and count of colorectal cancer deaths.

We compared five negative binomial models across five areas: Health Behaviors, Social and Economic Factors, Physical Environment, Health Outcomes, and Clinical Care. Of the five models, health outcomes has the lowest AIC. However, the full model is still a better model, as it has a lower AIC. Using the elastic net method and comparing to the full model, the best negative binomial model was the one that used elastic net for feature selection.

Using the same covariates as the Poisson model with elastic net, the model is written out as follows:

log(Count of Colorectal Cancer Deaths per 1 Million people) = -999,987.10-0.008(Air Pollution)
-0.004(Binge Drinking)+0.0004(Frequent Physical Distress)-0.01(Housing with Potential Lead Risk)
-0.006(Income Inequality)-0.083(Life_Expectancy)-0.003(Limited_Access_to_Healthy_Foods)
-0.015(Obesity)-0.011(Preventive Services)+0.001(Racial/Ethnic Diversity)-0.003(Uninsured
+0.0002(Average Annual Colorectal Cancer Case Count)
-0.117(% Average Annual Colorectal Cancer Case Count in Late Stage)

Table 1: Poisson (1) vs Negative Binomial Model (2); feature selection via elastic net

	Dependent variable:	
	Deaths due to Colorectal Cancer per 1 million People	
	Poisson	*negative binomial*
	(1)	(2)
Constant	−999,987.30 (0.83)	−999,987.10 (1.82)
	p = 0.00***	p = 0.00***
Air_pollution____particulate_matter	−0.01 (0.01)	−0.01 (0.01)
	p = 0.13	p = 0.52
Binge_drinking	−0.003 (0.004)	−0.004 (0.01)
	p = 0.46	p = 0.63
Frequent_physical_distress	0.0004 (0.01)	0.0004 (0.02)
	p = 0.97	p = 0.99
Housing_with_potential_lead_risk	−0.01 (0.001)	−0.01 (0.003)
	p = 0.00***	p = 0.001***
Income_Inequality	−0.01 (0.001)	−0.01 (0.003)
	p = 0.0001***	p = 0.06*
Life_expectancy	−0.08 (0.01)	−0.08 (0.02)
	p = 0.00***	p = 0.0001***
Limited_access_to_healthy_foods	−0.003 (0.001)	−0.003 (0.002)
	p = 0.0002***	p = 0.09*
Obesity	−0.02 (0.003)	−0.02 (0.01)
	p = 0.0000***	p = 0.04**
Preventive_services	−0.01 (0.002)	−0.01 (0.01)
	p = 0.0000***	p = 0.03**
Racial_ethnic_diversity	0.002 (0.001)	0.001 (0.002)
	p = 0.03**	p = 0.46
Uninsured	−0.002 (0.002)	−0.003 (0.005)
	p = 0.28	p = 0.60
Average_Annual_Case_Count	0.0002 (0.0001)	0.0002 (0.0001)
	p = 0.0005***	p = 0.08*
Per_Average_Case_Count_Late_Stage	−0.12 (0.08)	−0.12 (0.18)
	p = 0.15	p = 0.52
Observations	100	100
Log Likelihood	−599.57	−475.26
θ		45.10*** (7.95)
Akaike Inf. Crit.	1,227.14	978.53
Note:		*p<0.1; **p<0.05; ***p<0.01

We can see that this model has the lowest AIC at 978.527, and that housing with potential lead risk, life expectancy, obesity, and preventive services all have statistically significant (p-value <0.05) associations with colorectal cancer deaths. As an example for one of the main effects, the variable “obesity” has a coefficient of -0.017 that is statistically significant. This means that for every 1% unit increase in obesity in the population, the count of colorectal cancer deaths decreases by $e^{0.02} = 1.017 \sim 1$ death per 1 million people on average, holding all other covariates constant. Furthermore, with 95% confidence, the incidence rate ratio for the association between count of colorectal cancer deaths and 1% of obesity versus 0% obesity is between 0.97 and 1.00 (see appendix for rcode calculation of the confidence interval).

6 Discussion

A key finding was the difference between known risk factors for colorectal cancer incidence in literature compared to the variables associated with colorectal cancer mortality in our model. Main factors such as health behaviors, SES, and built environment were associated with mortality, while other hypothesized variables, such as diabetes and hypertension, were not. While the five year survival rate of localized (stage 1) cancer at diagnosis is 90.2%, late stage diagnosis is significantly lower, at 14.3%.[1] Because colorectal cancer mortality is highly associated with stage at diagnosis, our findings reflect issues around environment and health-seeking behaviors, which have a large impact on mortality, rather than incidence. Although there is increasing evidence that diabetes mellitus and hypertension are associated with increased cancer incidence[5,6], these chronic conditions may lead to more health-seeking behavior, which may decrease mortality. Preventative and therapeutic management of colon cancer is compromised by SES, such as disparities in educational and insurance status, screening and treatment patterns, social support, and access to healthcare.[13] With large disparities in many of the variables included in this analysis continuing to persist, inequalities in colorectal cancer mortality in major cities are likely to continue to evolve. These findings could provide rich evidence for nationwide policy targets, as well as geographic locations for enhanced screening and health education programs in the future.

7 Limitations

While city level data provides interesting and insightful information, there are more granular levels of data that may provide more information and nuance to what is happening in these urban communities. Census tract data, for instance, allows for more information about social and economic variables, which would be important, especially in a larger city like New York. Unfortunately the national mortality data is not feasible at the neighborhood level.[8] Similarly, this dataset looks at population-wide metrics and lacks individual level measures, which we know are important for risk of cancer incidence and mortality[12].

Some of these measures were approximated. For example, the analysis used county-level incidences and % of incidence cases diagnosed as late stage as an approximation city-level colorectal cancer metrics. Furthermore, where one city spans across multiple counties, the weighted average across counties was used as a proxy for these metrics at the city level.

It is also important to note that the years of data are not uniform across all covariates. The data on colorectal cancer deaths was collected between 2015 and 2017. While most covariate measures were taken before 2017, there were a couple that used a 2018 five-year estimate, so that may cause some noise in the findings. We made the assumption that many of these covariate trends do not change substantially across a three to four year timeframe. Similarly, not all covariates were age-, race-, and gender-adjusted. Many covariates may be associated with one another (e.g, late stage diagnosis is more prevalent among those in lower SES groups, which may also result in less favorable medical care[8]) and thus higher mortality rates have more complex associations beyond these findings.

Although death from colorectal cancer is an important outcome, there are other outcome measures that may add further context. For instance, incidence rates or prevalence rates may show greater connection to variables that are known to be highly associated with colorectal cancer. Using only death by colorectal cancer may miss some individuals who, while diagnosed with colorectal cancer, have died from other causes, and therefore are missed in our model.

8 Future Scope

As discussed previously, this study lacks individual level outcomes, which are important to inform the risk of death from cancer. Although county-level factors, such as tumor stage at diagnosis, were investigated, this study is intended to look at population-level data. However, an important area of future study is to better understand how individual level risk factors can be effectively tracked and monitored at more macro levels, and whether the results correlate to individual level risk.

Other future research questions could include narrowing to similar measures at the census tract level and seeing if that geographic measure is a better predictor of individual risk factors. Similarly, these questions could be applied to other types of cancer, and expanded to include other covariates associated with cancer (e.g. diet, other comorbidities like COPD or cardiovascular disease). Other data sets, such as the American Community Survey, may also have other variables of interest.

This study investigated colorectal cancer mortality between 2015-2017. Analysis of long-term trends concurrent with the evolving nature of SES and urbanization within cities would add further utility to this study. Furthermore, widening the dataset beyond 100 cities could be assessed in future studies to explore associations between different covariates, particularly smaller cities (or rural areas).

Geographic variations in cancer incidence, as well as exploring differences in urban/rural communities may also yield important findings. Previous research has suggested geographical patterns of colorectal cancer incidence and mortality rates, with survival after colorectal cancer diagnosis significantly worse among men residing in hotspot counties in the southern portion of the United States.[16] Future work could generate a larger dataset, group according to US census regions, and explore differences in colorectal cancer deaths, as well as differences in covariates that might be statistically significant for certain regions but not for others. Similar analyses could be conducted looking at urban areas versus rural areas.

9 Tables and Figures

Figure 1

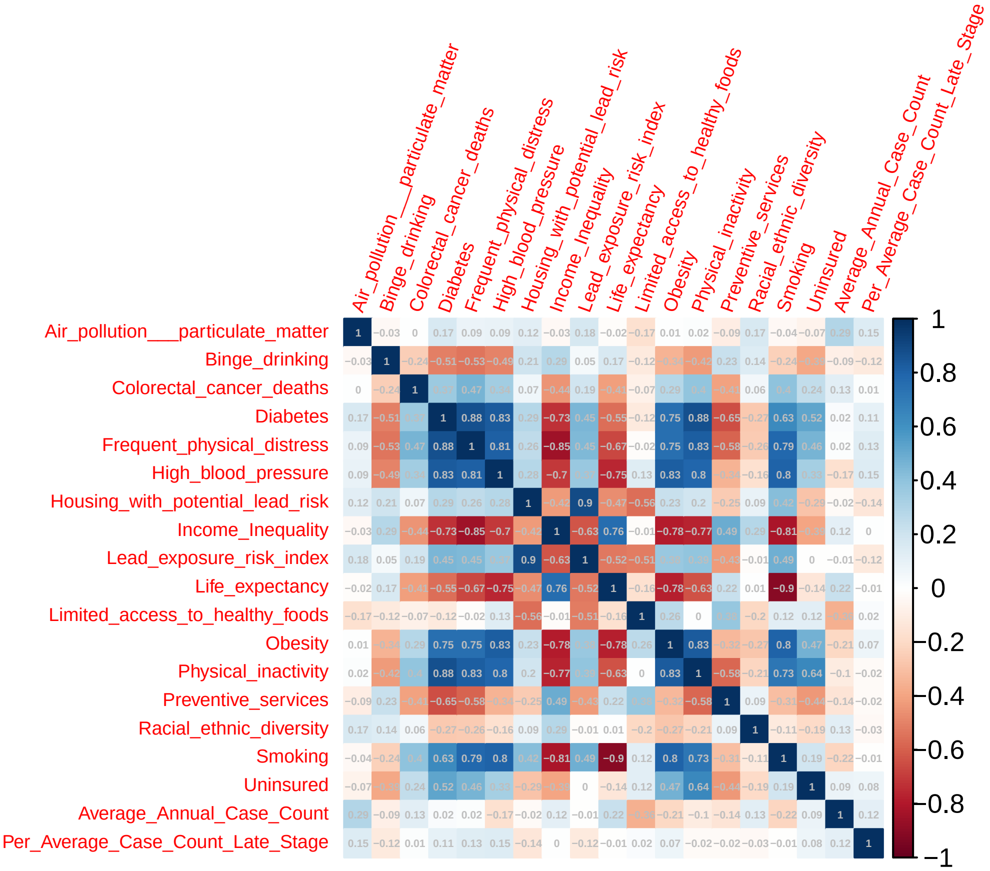


Figure 2:
Histogram of number of Colorectal
Cancer Deaths

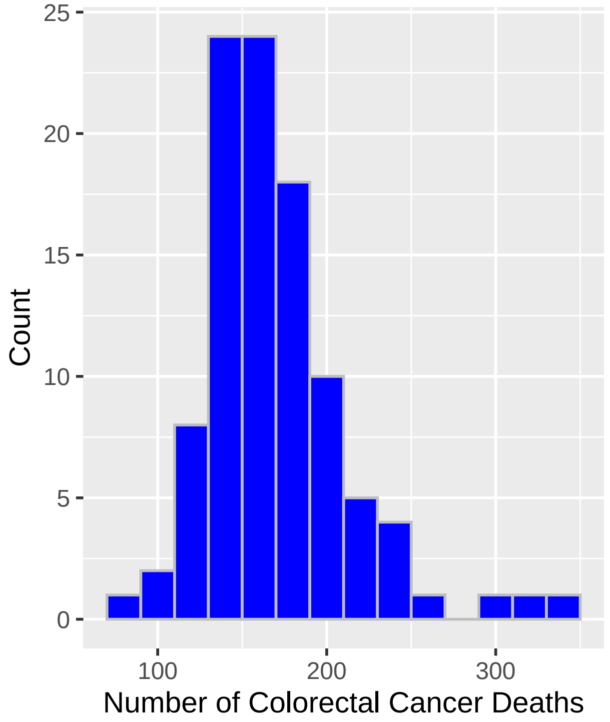


Figure 3:
Poisson Distribution w/
mean estimated from outcome

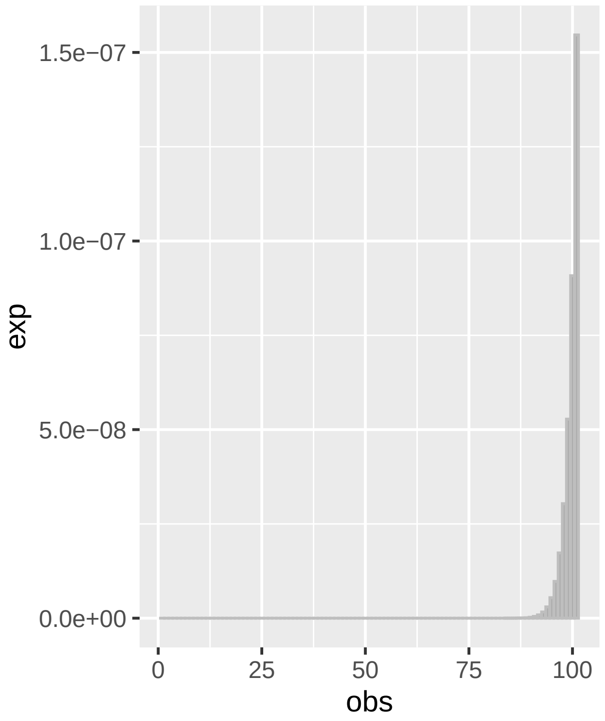


Table 2: Frequent Physical Distress (1) vs Frequent Physical Distress+Diabetes (2)

	Dependent variable:	
	Deaths due to Colorectal Cancer per 1 million People	
	(1)	(2)
Constant	−999,995.600 (0.117)	−999,995.600 (0.119)
	p = 0.000***	p = 0.000***
Diabetes		−0.020 (0.017)
		p = 0.255
Frequent_physical_distress	0.055 (0.009)	0.073 (0.019)
	p = 0.000***	p = 0.0002***
Observations	100	100
Log Likelihood	−494.256	−493.610
θ	28.711*** (4.686)	29.150*** (4.770)
Akaike Inf. Crit.	992.512	993.221
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3: Health Outcomes (1), Model w/ Elastic Net for Feature Selection (2), and Full Model (3)

	Dependent variable:		
	Deaths due to Colorectal Cancer per 1 million People		
	(1)	(2)	(3)
Constant	−999,990.400 (1.524)	−999,987.100 (1.821)	−999,986.700 (2.094)
	p = 0.000***	p = 0.000***	p = 0.000***
Diabetes	0.008 (0.020)		−0.010 (0.028)
	p = 0.701		p = 0.717
High_blood_pressure	−0.009 (0.010)		0.007 (0.012)
	p = 0.393		p = 0.567
Air_pollution____particulate_matter		−0.008 (0.012)	−0.009 (0.013)
		p = 0.518	p = 0.459
Binge_drinking		−0.004 (0.009)	−0.002 (0.010)
		p = 0.627	p = 0.869
Frequent_physical_distress	0.052 (0.020)	0.0004 (0.021)	0.009 (0.029)
	p = 0.010***	p = 0.986	p = 0.763
Housing_with_potential_lead_risk		−0.010 (0.003)	−0.010 (0.003)
		p = 0.001***	p = 0.002***
Income_Inequality		−0.006 (0.003)	−0.006 (0.003)
		p = 0.057*	p = 0.058*
Life_expectancy	−0.058 (0.017)	−0.083 (0.019)	−0.088 (0.023)
	p = 0.001***	p = 0.00002***	p = 0.0002***
Limited_access_to_healthy_foods		−0.003 (0.002)	−0.003 (0.002)
		p = 0.090*	p = 0.091*
Obesity	−0.014 (0.007)	−0.015 (0.007)	−0.015 (0.009)
	p = 0.053*	p = 0.031**	p = 0.092*
Physical_inactivity			0.00000 (0.010)
			p = 1.000
Preventive_services		−0.011 (0.005)	−0.012 (0.006)
		p = 0.028**	p = 0.038**
Smoking			−0.011 (0.016)
			p = 0.494
Racial_ethnic_diversity		0.001 (0.002)	0.001 (0.002)
		p = 0.459	p = 0.477
Uninsured		−0.003 (0.005)	−0.003 (0.006)
		p = 0.592	p = 0.600
Average_Annual_Case_Count	0.0002 (0.0001)	0.0002 (0.0001)	0.0002 (0.0001)
	p = 0.081*	p = 0.073*	p = 0.079*
Per_Average_Case_Count_Late_Stage	−0.085 (0.195)	−0.117 (0.179)	−0.156 (0.195)
	p = 0.664	p = 0.513	p = 0.424
Observations	100	100	100
Log Likelihood	−486.504	−475.263	−474.947
θ	34.199*** (5.708)	45.098*** (7.947)	45.447*** (8.020)
Akaike Inf. Crit.	989.007	978.527	985.893
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: Physical Environment (1), Clinical Care (2), and Full Model (3)

	Dependent variable:		
	Deaths due to Colorectal Cancer		
	negative binomial		*Poisson*
	(1)	(2)	(3)
Constant	−999,995.20 (0.23) p = 0.00***	−999,994.80 (0.19) p = 0.00***	−999,987.10 (0.96) p = 0.00***
Binge_drinking	−0.01 (0.01) p = 0.20		−0.0003 (0.005) p = 0.95
Diabetes			−0.01 (0.01) p = 0.36
Frequent_physical_distress			0.01 (0.01) p = 0.59
Life_expectancy			−0.08 (0.01) p = 0.00***
High_blood_pressure			0.01 (0.01) p = 0.14
Income_Inequality			−0.01 (0.002) p = 0.0001***
Racial_ethnic_diversity			0.002 (0.001) p = 0.04**
Physical_inactivity	0.01 (0.01) p = 0.21		0.001 (0.005) p = 0.85
Preventive_services			−0.01 (0.003) p = 0.0000***
Smoking	0.02 (0.01) p = 0.03**		−0.01 (0.01) p = 0.16
Uninsured			−0.003 (0.003) p = 0.26
Average_Annual_Case_Count			0.0002 (0.0001) p = 0.0004***
Per_Average_Case_Count_Late_Stage			−0.16 (0.09) p = 0.09*
Air_pollution___particulate_matter		−0.002 (0.02) p = 0.89	−0.01 (0.01) p = 0.09*
Housing_with_potential_lead_risk		0.001 (0.002) p = 0.74	−0.01 (0.001) p = 0.00***
Obesity			−0.02 (0.004) p = 0.0001***
Limited_access_to_healthy_foods		−0.001 (0.002) p = 0.64	−0.003 (0.001) p = 0.0003***
Observations	100	100	100
Log Likelihood	−496.10	−508.64	−597.94
θ	27.55*** (4.47)	20.86*** (3.28)	
Akaike Inf. Crit.	1,000.19	1,025.27	1,231.88
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: Socio-economic Factors (1), Clinical Care (2), and Full Model (3)

	Dependent variable:		
	Deaths due to Colorectal Cancer		
	(1)	(2)	(3)
Constant	−999,995.20 (0.11) p = 0.00***	−999,994.30 (0.18) p = 0.00***	−999,986.70 (2.09) p = 0.00***
Air_pollution_____particulate_matter			−0.01 (0.01) p = 0.46
Binge_drinking			−0.002 (0.01) p = 0.87
Diabetes			−0.01 (0.03) p = 0.72
Frequent_physical_distress			0.01 (0.03) p = 0.77
Life_expectancy			−0.09 (0.02) p = 0.0002***
High_blood_pressure			0.01 (0.01) p = 0.57
Income_Inequality	−0.01 (0.001) p = 0.00***		−0.01 (0.003) p = 0.06*
Racial_ethnic_diversity	0.004 (0.002) p = 0.02**		0.001 (0.002) p = 0.48
Housing_with_potential_lead_risk			−0.01 (0.003) p = 0.002***
Obesity			−0.01 (0.01) p = 0.10*
Physical_inactivity			0.0000 (0.01) p = 1.00
Preventive_services		−0.02 (0.005) p = 0.0001***	−0.01 (0.01) p = 0.04**
Smoking			−0.01 (0.02) p = 0.50
Uninsured		0.004 (0.004) p = 0.40	−0.003 (0.01) p = 0.60
Average_Annual_Case_Count			0.0002 (0.0001) p = 0.08*
Per_Average_Case_Count_Late_Stage			−0.16 (0.19) p = 0.43
Limited_access_to_healthy_foods			−0.003 (0.002) p = 0.10*
Observations	100	100	100
Log Likelihood	−492.99	−498.40	−474.95
θ	29.59*** (4.86)	26.21*** (4.24)	45.45*** (8.02)
Akaike Inf. Crit.	991.98	1,002.80	985.89
Note:			*p<0.1; **p<0.05; ***p<0.01

Models	Overdispersion.Test..Deviance.
Health Outcomes	7.006306
Socioeconomic Factors	7.365442
Health Behavior	7.913696
Physical Environment	9.943345
Clinical Care	8.091684
Full Model	6.109073
Model with Feature Selection via Elastic Net	5.862815

10 References

1) NIH National Cancer Institute Surveillance, Epidemiology, and End Results Program, "Cancer Stat Facts: Colorectal Cancer". https://seer.cancer.gov/statfacts/html/colorect.html

2) Center for Disease Control and Prevention, "Colorectal (Colon) Cancer" updated June 8, 2020. https://www.cdc.gov/cancer/colorectal/statistics/index.htm

3) Sharma, Sumit, & O'Keefe, Stephen J D. (2007). Environmental influences on the high mortality from colorectal cancer in African Americans. Postgraduate Medical Journal, 83(983), 583-589.

4) American Cancer Society. Colorectal Cancer Facts & Figures 2017-2019. Atlanta: American Cancer Society; 2017. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2017-2019.pdf

5) De Bruijn, K. M. J, Arends, L. R, Hansen, B. E, Leeflang, S, Ruiter, R, & Van Eijck, C. H. J. (2013). Systematic review and meta-analysis of the association between diabetes mellitus and incidence and mortality in breast and colorectal cancer. British Journal of Surgery, 100(11), 1421-1429.

6) Sud, S., O'Callaghan, C., Jonker, C, et al. (2018). Hypertension as a predictor of advanced colorectal cancer outcome and cetuximab treatment response. Curr Oncol, 25(6): e516-e526.

7) Teng, Andrea M, Atkinson, June, Disney, George, Wilson, Nick, & Blakely, Tony. (2017). Changing socioeconomic inequalities in cancer incidence and mortality: Cohort study with 54 million person-years follow-up 1981-2011. International Journal of Cancer, 140(6), 1306-1316.

8) Singh, Gopal K, & Jemal, Ahmedin. (2017). Socioeconomic and Racial/Ethnic Disparities in Cancer Mortality, Incidence, and Survival in the United States, 1950–2014: Over Six Decades of Changing

Patterns and Widening Inequalities. Journal of Environmental and Public Health, 2017, 1-19.

9) Aarts, M.J.,Lemmens, V., et al. (2010). Socioeconomic status and changing inequalities in colorectal cancer: A review of the associations with risk, treatment, and outcome. European Journal of Cancer, October 2010, 46(15):2681-2695.

10) Islami F, Goding Sauer A, Miller KD, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA Cancer J Clin. 2018;68:31-54.

11) Roetzheim RG, Pal N, Gonzalez EC, Ferrante JM, Van Durme DJ, Krischer JP. Effects of health insurance and race on colorectal cancer treatments and outcomes. Am J Public Health.

12) Agrawal, S, et al. Colorectal cancer in African Americans. American Journal of Gastroenterology. 100(3):515-523, March 2005.

13) Rossi, M.; Jahanzaib Anwar, M.; Usman, A.; Keshavarzian, A.; Bishehsari, F. Colorectal Cancer and Alcohol Consumption—Populations to Molecules. Cancers 2018, 10, 38.

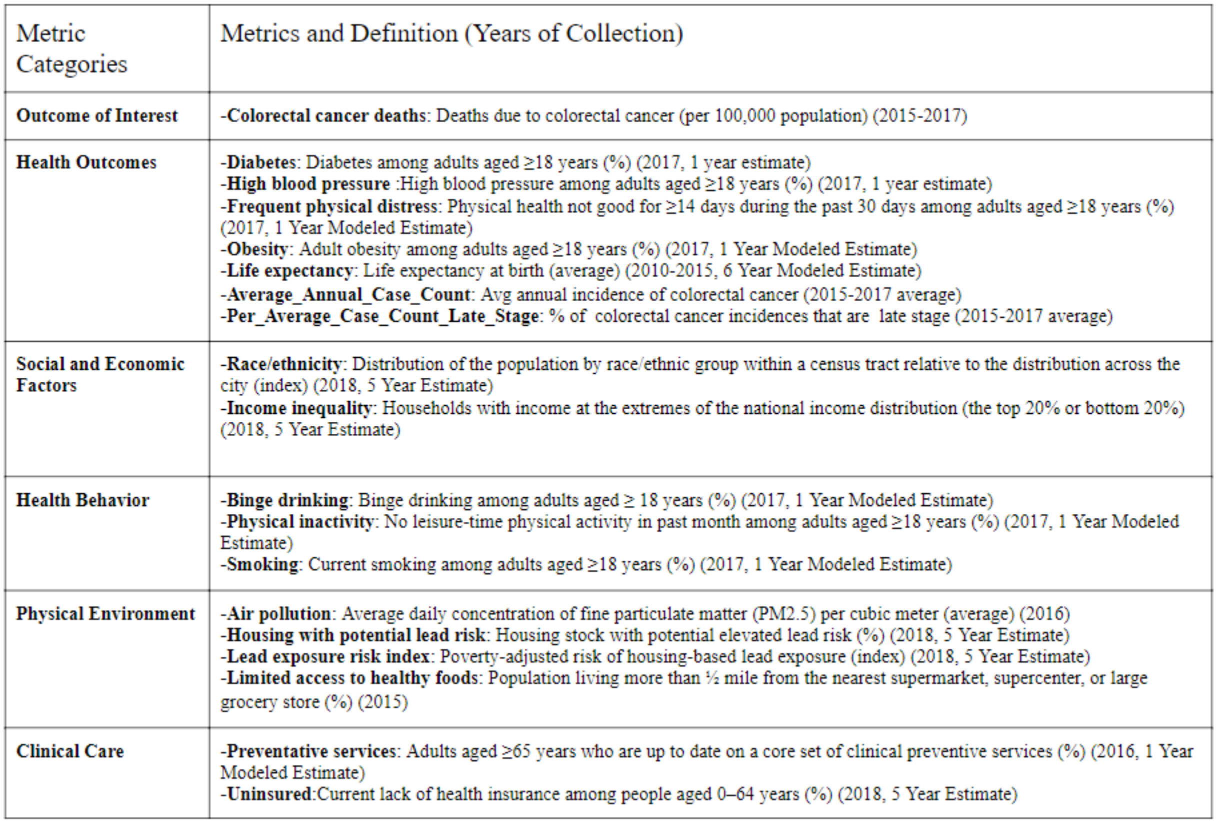
14) 2000;90(11):1746-1754. doi:10.2105/ajph.90.11.1746

15) Gomez, SL, Shariff-Marco, S, DeRouen, M, et al. THe Impact of neighborhood social and built environment factors across the cancer continuum: Current research, methodological considerations, and future directions. Cancer. 2015; 121(14): 2314-2330.

16) Rogers CR, Moore JX, Qeadan F, Gu LY, Huntington MS, Holowatyj AN. Examining factors underlying geographic disparities in early-onset colorectal cancer survival among men in the United States. Am J Cancer Res. 2020;10(5):1592-1607. Published 2020 May.

11 Appendix

11.1 Data Dictionary



Metric Categories	Metrics and Definition (Years of Collection)
Outcome of Interest	-**Colorectal cancer deaths**: Deaths due to colorectal cancer (per 100,000 population) (2015-2017)
Health Outcomes	-**Diabetes**: Diabetes among adults aged ≥18 years (%) (2017, 1 year estimate) -**High blood pressure** :High blood pressure among adults aged ≥18 years (%) (2017, 1 year estimate) -**Frequent physical distress**: Physical health not good for ≥14 days during the past 30 days among adults aged ≥18 years (%) (2017, 1 Year Modeled Estimate) -**Obesity**: Adult obesity among adults aged ≥18 years (%) (2017, 1 Year Modeled Estimate) -**Life expectancy**: Life expectancy at birth (average) (2010-2015, 6 Year Modeled Estimate) -**Average_Annual_Case_Count**: Avg annual incidence of colorectal cancer (2015-2017 average) -**Per_Average_Case_Count_Late_Stage**: % of colorectal cancer incidences that are late stage (2015-2017 average)
Social and Economic Factors	-**Race/ethnicity**: Distribution of the population by race/ethnic group within a census tract relative to the distribution across the city (index) (2018, 5 Year Estimate) -**Income inequality**: Households with income at the extremes of the national income distribution (the top 20% or bottom 20%) (2018, 5 Year Estimate)
Health Behavior	-**Binge drinking**: Binge drinking among adults aged ≥ 18 years (%) (2017, 1 Year Modeled Estimate) -**Physical inactivity**: No leisure-time physical activity in past month among adults aged ≥18 years (%) (2017, 1 Year Modeled Estimate) -**Smoking**: Current smoking among adults aged ≥18 years (%) (2017, 1 Year Modeled Estimate)
Physical Environment	-**Air pollution**: Average daily concentration of fine particulate matter (PM2.5) per cubic meter (average) (2016) -**Housing with potential lead risk**: Housing stock with potential elevated lead risk (%) (2018, 5 Year Estimate) -**Lead exposure risk index**: Poverty-adjusted risk of housing-based lead exposure (index) (2018, 5 Year Estimate) -**Limited access to healthy foods**: Population living more than ½ mile from the nearest supermarket, supercenter, or large grocery store (%) (2015)
Clinical Care	-**Preventative services**: Adults aged ≥65 years who are up to date on a core set of clinical preventive services (%) (2016, 1 Year Modeled Estimate) -**Uninsured**:Current lack of health insurance among people aged 0–64 years (%) (2018, 5 Year Estimate)

11.2 R Code

```
knitr::opts_chunk$set(echo = TRUE,tidy.opts=list(width.cutoff=50),tidy=TRUE)


rm(list = ls())


gc(reset = TRUE)


# SETUP -----------------------------------------------------------------
library(rstudioapi)
library(readxl)
library(glmnet)
library(rvest)
library(lubridate)
library(readxl)
library(stargazer)
library(MASS)
library(corrplot)
library(gridExtra)
library(tidyverse)

getActiveDocumentContext()$path
working_path <- dirname(getActiveDocumentContext()$path)
setwd(working_path)

# READ IN DATA --------------------------------------------------------------
#read in Data from https://www.cityhealthdashboard.com/metrics
data<-read_csv("data/CHDB_data_city_all v9_0.csv")

#read in incidence of colorectal cancer and % late stage colorectal cancer from SEERS website
#average incidence of colerectal cancer and % of cancer that are late stage from 2013-2017
cancer_stage <- read_csv("data/city_county_LateStage.csv")

#read in county population sizes from census.gov: https://www.census.gov/data/datasets/time-series/demo/
county_pop <- read_excel("data/co-est2019-annres.xlsx",skip=3)
county_pop$...1 <- str_remove(county_pop$...1,".")

#read in state name to abbreviation crosswalk
state_abbrev <- read_xlsx("data/state name to abbrev crosswalk.xlsx")

#read in state region crosswalk
state_region <- read_xlsx("data/STATE REGION CROSSWALK.xlsx")

#Read in top 200 cities and clean up to match to subset_data
# taken from https://worldpopulationreview.com/us-cities
```

```r
top_200_cities <- read_csv("data/top_200_cities.csv")
```

```r
# CLEAN AND FILTER CITY HEALTH DASHBOARD DATA ----------------------------------------------------

#look at the variables that could be associated with Colorectal cancer deaths
variable_list <- c("Diabetes","High blood pressure",
                   "Frequent physical distress","Obesity","Life expectancy",
                   "Racial/ethnic diversity",
                   "Income Inequality","Binge drinking",
                   "Physical inactivity","Smoking",
                   "Air pollution - particulate matter",
                   "Housing with potential lead risk",
                   "Lead exposure risk index",
                   "Limited access to healthy foods",
                   "Colorectal cancer deaths",
                   "Preventive services", "Uninsured")

#subset to only city name, metric name, year, and estimate (value),
#look at total_population metrics, and subset to only metrics of interest
subset_data <- data[which((data$group_name=="total population") &
                            (data$metric_name %in%
                               variable_list)),][c("city_name","state_abbr",
                                                   "metric_name","data_yr_type","est")]

#NOTE: San Juan, Paradise, and Arlington aren't in the
#Langone Dataset, so include the next 3 largest cities
top_100_cities <- top_200_cities %>% slice(1:103)

names(top_100_cities)[which(names(top_100_cities)
                            %in% c("name","usps"))] <- c("city_name","state_abbr")

#change names for "Boise" and "New York City" to "Boise City" and "New York"
top_100_cities$city_name[which(top_100_cities$city_name %in%
                                 c("Boise", "New York City"))] <- c("New York","Boise City")

#filter for just top 100 cities in subset_data
city_state <- paste0(top_100_cities$city_name,", ",top_100_cities$state_abbr)
data_city_state <- paste0(subset_data$city_name,", ",subset_data$state_abbr)

subset_top_100_cities <- subset_data[which(data_city_state %in% city_state),]

#data checks
cities_transformed <- subset_top_100_cities %>%
  spread(metric_name, est)
subset_top_100_cities$metric_year <- paste0(subset_top_100_cities$metric_name,
                                           "_",subset_top_100_cities$data_yr_type)
cities_years_check <- subset_top_100_cities[c("city_name","state_abbr","metric_year","est")] %>%
  spread(metric_year, est)

#since there's only 1 value for each metric for each city, remove data_yr_type
final_data_long <- subset_top_100_cities[c("city_name","state_abbr","metric_name","est")]

#convert from long to wide format; have each metric name as a column
cleaned_df <- final_data_long %>% spread(metric_name, est)

# ADD AIR QUALITY DATA FOR ANCHORAGE AND HONOLULU ----------------------------------------------------

#Data represent modeled estimates produced by Community Multiscale Air Quality (CMAQ) model and
#do not include estimates for Alaska and Hawaii.
#add values for air pollution for honolulu and anchorage from other sources
honolulu_2016_air_pollution <- 13.8
#https://health.hawaii.gov/cab/files/2019/07/aqbook_2016.pdf pg 18
anchorage_2016_air_pollution <- 6.5 #https://www.muni.org/Departments/OCPD/Planning/AMATS/MTP/2040/Chpa

cleaned_df$`Air pollution - particulate matter`[which(cleaned_df$city_name=="Anchorage")] <-
```

```r
  anchorage_2016_air_pollution
cleaned_df$`Air pollution - particulate matter`[which(cleaned_df$city_name=="Honolulu")] <-
  honolulu_2016_air_pollution

#replace spaces with _ in column names
names(cleaned_df)<-str_replace_all(names(cleaned_df), c(" " = "_" , "," = "" ))
names(cleaned_df)<-str_replace_all(names(cleaned_df), c("-" = "_" , "," = "" ))
names(cleaned_df)<-str_replace_all(names(cleaned_df), c("/" = "_" , "," = "" ))

# ADD IN COUNTY % LATE STAGE DATA ---------------------------------------------------------------------
county_df <- county_pop %>% separate(...1, c("County","State"), sep = ", ") %>%
      slice(2:3143) %>% dplyr::select(County, State,"2013","2014","2015","2016","2017")

county_df$County <-sub("\\ County", "", county_df$County)


#change county for San Diego from Jim Wells to San Diego
cancer_stage$County[which((cancer_stage$County=="Jim Wells") & (cancer_stage$City=="San Diego"))] <- "Sa

#adjust a couple county names in county dataset
county_df$County[which(county_df$County=="Anchorage Municipality")] <- "Anchorage"
county_df$County[which(county_df$County=="Chesapeake city")] <- "Chesapeake"
county_df$County[which(county_df$County=="Orleans Parish")] <- "Orleans"
county_df$County[which(county_df$County=="Norfolk city")] <- "Norfolk"
county_df$County[which(county_df$County=="Virginia Beach city")] <- "Virginia Beach"

cancer_stage$`Average Annual Count` <- as.numeric(cancer_stage$`Average Annual Count`)
cancer_stage_pop <- cancer_stage %>%
  left_join(state_abbrev, by =c("state_abbr"="Abbreviation")) %>%
  left_join(county_df, by =c("County","US State"="State")) %>%
  mutate(pop_13_17=`2013`+`2014`+`2015`+`2016`+`2017`,
         `Average Cases with Late Stage Count` = `Average Annual Count`*
```

```
           `Percent of Cases with Late Stage`/100) %>%
  dplyr::select(City, state_abbr,County,`Average Annual Count`,
                `Average Cases with Late Stage Count`,`pop_13_17`)

#take weighted average for average annual count and cases with late stage count
city_cancer_stage <- cancer_stage_pop %>%
  mutate(total_case_count =`Average Annual Count`*`pop_13_17`,
         total_late_stage_count =
           `Average Cases with Late Stage Count`*`pop_13_17`) %>%
  group_by(City,state_abbr) %>%
  summarise(Average_Annual_Case_Count =
              sum(total_case_count, na.rm=TRUE)/sum(pop_13_17),
            Per_Average_Case_Count_Late_Stage =
              sum(total_late_stage_count, na.rm=TRUE)/sum(pop_13_17)
            /Average_Annual_Case_Count, .groups = 'drop')

cleaned_df <- cleaned_df %>% left_join(city_cancer_stage, by=c("city_name"="City","state_abbr"))

# SCALE UP OUTCOME (COLORECTAL CANCER DEATHS) ----------------------------------------------------------
#multiply Colorectal_cancer_deaths and Breat_cancer_deaths by 10 to make it discrete
cleaned_df$Colorectal_cancer_deaths <- cleaned_df$Colorectal_cancer_deaths*10

#placeholder NaN as 0; Hennepin County (for Minneapolis), Decatur (for Wichita),
#and Shelby (for St. Paul) all have 3 or fewer avg annual cases of Colorectal Cancer; set to 0
cleaned_df$Per_Average_Case_Count_Late_Stage[which(
  is.na(cleaned_df$Per_Average_Case_Count_Late_Stage))] <- 0
```

```
#break down metrics into groups
outcome <- "Colorectal_cancer_deaths"
health_outcomes <- c("Diabetes","High_blood_pressure","Frequent_physical_distress",
                     "Life_expectancy", "Obesity","Average_Annual_Case_Count",
                     "Per_Average_Case_Count_Late_Stage")
social_economic_factors <- c("Income_Inequality","Racial_ethnic_diversity")
```

```r
health_behavior <- c("Binge_drinking","Physical_inactivity","Smoking")
physical_environment <- c("Air_pollution___particulate_matter",
                          "Housing_with_potential_lead_risk",
                          "Limited_access_to_healthy_foods")
clinical_care <- c("Preventive_services","Uninsured")

health_outcomes_df <- cleaned_df[c(outcome, health_outcomes)]
social_economic_factors_df <- cleaned_df[c(outcome, social_economic_factors)]
health_behavior_df <- cleaned_df[c(outcome, health_behavior)]
physical_environment_df <- cleaned_df[c(outcome, physical_environment)]
clinical_care_df <- cleaned_df[c(outcome, clinical_care)]
```

```r
#colorectal cancer deaths will be per 1 million instead of per 100K.
#no other values are on a per 100k people rate
offset_value <- rep(1000000, dim(cleaned_df)[1])

colorectal_poisson_model_function<- function(df){
  poisson_model <- glm(Colorectal_cancer_deaths ~.,family="poisson", offset = offset_value,
                       data=df)
  return(poisson_model)
}

#Testing if Diabetes is a confounder for Frequent Physical distress
frequent_physical_distress_model <- glm(Colorectal_cancer_deaths ~
                                          Frequent_physical_distress, family="poisson",
                                        offset=offset_value, data=cleaned_df)
diabetes_frequent_physical_distress_model <- glm(Colorectal_cancer_deaths ~
                                                   Diabetes+
                                                   Frequent_physical_distress,
                                                 family="poisson",
                                                 offset=offset_value, data=cleaned_df)

#Health Outcomes as the predictor
```

```
health_outcomes_model <- colorectal_poisson_model_function(health_outcomes_df)

#Poisson Model with SES as predictors
social_economic_factors_model<-colorectal_poisson_model_function(social_economic_factors_df)

#Poisson Model with health behavior as predictors
health_behavior_model <- colorectal_poisson_model_function(health_behavior_df)

#Poisson Model with physical environment as predictors
physical_environment_model <- colorectal_poisson_model_function(physical_environment_df)

#Poisson Model with clinical care as predictors
clinical_care_model<- colorectal_poisson_model_function(clinical_care_df)

#All covariates
full_model <- glm(Colorectal_cancer_deaths ~ Air_pollution___particulate_matter+
                    Binge_drinking+Diabetes+
                    Frequent_physical_distress+Life_expectancy+
              High_blood_pressure+Income_Inequality+Racial_ethnic_diversity+
                Housing_with_potential_lead_risk+
                Obesity+Physical_inactivity+Preventive_services+Smoking+Uninsured+
                Average_Annual_Case_Count+Per_Average_Case_Count_Late_Stage+
                Limited_access_to_healthy_foods,
              family="poisson", offset = offset_value,  data=cleaned_df)

# Poisson elastic net--------------------------------------------------
x <- cleaned_df %>% dplyr::select(-c(Colorectal_cancer_deaths,city_name, state_abbr)) %>%
  data.matrix()
y <- cleaned_df %>% dplyr::select(Colorectal_cancer_deaths) %>%
  data.matrix()
poisson_elasticnet_col_cancer_model <- cv.glmnet(x, y, family= "poisson")
```

```r
#lambda.min is the value of lambda that gives minimum mean cross-validated error
#poisson_elasticnet_col_cancer_model$lambda.min

#covariates with a slope estimate are included in best model
poisson_elasticnet_coeff <- coef(poisson_elasticnet_col_cancer_model, s = "lambda.min")

#fit new model
poisson_model_enet <- glm(Colorectal_cancer_deaths ~ Air_pollution___particulate_matter+
                            Binge_drinking+Frequent_physical_distress+
                            Housing_with_potential_lead_risk+Income_Inequality+
                            Life_expectancy+Limited_access_to_healthy_foods+
                            Obesity+Preventive_services+Racial_ethnic_diversity+Uninsured+
                            Average_Annual_Case_Count+
                            Per_Average_Case_Count_Late_Stage, family="poisson",
                          offset = offset_value,
                          data=cleaned_df)

# Negative Binomial Regression ------------------------------------------------------------
deaths.negbinom <- glm.nb(Colorectal_cancer_deaths ~ Air_pollution___particulate_matter+
                            Binge_drinking+Frequent_physical_distress+
                            Housing_with_potential_lead_risk+Income_Inequality+
                            Life_expectancy+Limited_access_to_healthy_foods+
                            Obesity+Preventive_services+Racial_ethnic_diversity+Uninsured+
                            Average_Annual_Case_Count+
                            Per_Average_Case_Count_Late_Stage+offset(offset_value),
                          data=cleaned_df, link=log)
```

```r
#CI for obesity in negative binomial model
#95% CI for Obesity Slope
coef(deaths.negbinom)[9]
```

```
##     Obesity
## -0.01547581
```

```r
exp(coef(deaths.negbinom)[9])
```

```
##   Obesity
## 0.9846433
```

```r
#sqrt(vcov(deaths.negbinom)) #obesity is the 9th variable
exp(coef(deaths.negbinom)[9]*1 + c(-1, 1)*
      1.96*1*sqrt(vcov(deaths.negbinom)[9,9]))
```

```
## [1] 0.9708991 0.9985821
```

```r
#covariates to include in poisson model to have minimum mean cross-validated error:
rownames(poisson_elasticnet_coeff)[which(poisson_elasticnet_coeff != 0)]
```

```
##  [1] "(Intercept)"                        "Air_pollution___particulate_matter"
##  [3] "Binge_drinking"                     "Frequent_physical_distress"
##  [5] "Housing_with_potential_lead_risk"   "Income_Inequality"
##  [7] "Life_expectancy"                    "Limited_access_to_healthy_foods"
##  [9] "Obesity"                            "Preventive_services"
## [11] "Racial_ethnic_diversity"            "Average_Annual_Case_Count"
## [13] "Per_Average_Case_Count_Late_Stage"
```

```r
#Elastic Net Coefficients
poisson_elasticnet_coeff
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                                 1
## (Intercept)                        11.4804806467
## Air_pollution___particulate_matter -0.0058612802
## Binge_drinking                     -0.0023063295
## Diabetes                            .
## Frequent_physical_distress          0.0024657074
## High_blood_pressure                 .
## Housing_with_potential_lead_risk   -0.0075309680
## Income_Inequality                  -0.0046799885
## Lead_exposure_risk_index            .
## Life_expectancy                    -0.0685798126
## Limited_access_to_healthy_foods    -0.0019766409
## Obesity                            -0.0131771495
## Physical_inactivity                 .
## Preventive_services                -0.0120720188
## Racial_ethnic_diversity             0.0014779871
## Smoking                             .
## Uninsured                           .
## Average_Annual_Case_Count           0.0001749795
## Per_Average_Case_Count_Late_Stage  -0.0643112125
```

```
#testing other models
summary(glm(Colorectal_cancer_deaths ~ Binge_drinking,
            family="poisson", offset = offset_value,  data=cleaned_df))$coeff
summary(glm(Colorectal_cancer_deaths ~ Diabetes,
            family="poisson", offset = offset_value,  data=cleaned_df))$coeff
summary(glm(Colorectal_cancer_deaths ~ Frequent_physical_distress,
            family="poisson", offset = offset_value,  data=cleaned_df))$coeff
summary(glm(Colorectal_cancer_deaths ~ High_blood_pressure,
            family="poisson", offset = offset_value,  data=cleaned_df))$coeff
```

```
stargazer(poisson_model_enet, deaths.negbinom,
          title="Poisson (1) vs Negative Binomial Model (2); feature selection via elastic net",
          dep.var.labels="Deaths due to Colorectal Cancer per 1 million People",
          report="vcsp*", intercept.bottom = FALSE,
          header=FALSE, single.row=TRUE, digits =2, digits.extra = 2, type = 'latex')
```

```
corrplot_data <-  cleaned_df %>% dplyr::select(everything(), -city_name, -state_abbr)


#correlation matrix
corrplot(cor(corrplot_data),
         method="color", addCoef.col="grey",number.cex= 0.3, tl.cex=0.6, tl.srt=70)
```

```
clc_dist <- ggplot(data = cleaned_df, aes(Colorectal_cancer_deaths)) +
  geom_histogram(binwidth=20, fill="blue", color="grey")+
  xlab("Number of Colorectal Cancer Deaths") + ylab("Count") +
  ggtitle("Figure 2: \nHistogram of number of Colorectal\n Cancer Deaths")


y <- cleaned_df$Colorectal_cancer_deaths
y_length <- length(cleaned_df$Colorectal_cancer_deaths)
lambda <- mean(cleaned_df$Colorectal_cancer_deaths)
tbl <- NULL
for (k in 0:100) {
  tbl <- rbind(tbl,c(obs=sum(y==k),
                     exp=y_length*exp(-lambda) * lambda^k / factorial(k)))
}


tbl <- as.data.frame(tbl)
poisson_dist_df <- as.data.frame(tbl)
poisson_dist <- ggplot(data=poisson_dist_df, aes(x=seq(nrow(tbl)),y=exp))+
  geom_bar(stat="identity",fill="darkgrey",color="grey")+
  ggtitle("Figure 3:\nPoisson Distribution w/ \nmean estimated from outcome")+
  xlab("obs")


grid.arrange(clc_dist, poisson_dist, nrow = 1)
```

```
#testing if diabetes is a confounder on the association of
#frequent physical distress on colorectal cancer deaths
stargazer(frequent_physical_distress_model, diabetes_frequent_physical_distress_model,
          title="Frequent Physical Distress (1) vs Frequent Physical Distress+Diabetes (2)",
          dep.var.labels="Deaths due to Colorectal Cancer per 1 million People",
          report="vcsp*",
          intercept.bottom = FALSE, header=FALSE, single.row=TRUE, digits =3,
          digits.extra = 2, type="latex")


#model summary for categories
stargazer(health_outcomes_model,poisson_model_enet,
          full_model,
          title="Health Outcomes (1), Model w/ Elastic Net for Feature Selection (2), and
          Full Model (3)",
          dep.var.labels="Deaths due to Colorectal Cancer per 1 million People",
          report="vcsp*",
          intercept.bottom = FALSE, header=FALSE, single.row=TRUE, digits =3,
          digits.extra = 2, type="latex")

stargazer(health_behavior_model,physical_environment_model,
          full_model, title="Physical Environment (1), Clinical Care (2), and Full Model (3)",
          dep.var.labels="Deaths due to Colorectal Cancer", report="vcsp*",
          intercept.bottom = FALSE,
          header=FALSE, single.row=TRUE, digits =2, digits.extra = 2)

stargazer(social_economic_factors_model,clinical_care_model,full_model,
title="Socio-economic Factors (1), Clinical Care (2), and Full Model (3)",
dep.var.labels="Deaths due
          to Colorectal Cancer", report="vcsp*",
intercept.bottom = FALSE,
          header=FALSE, single.row=TRUE, digits =2, digits.extra = 2)
```

```
# Negative Binomial Regression ----------------------------------------------------------
colorectal_nb_model_function<- function(df){
  nb_model <- glm.nb(Colorectal_cancer_deaths ~.+offset(offset_value), data=df, link=log)
  return(nb_model)
}


frequent_physical_distress_nb_model <- glm.nb(Colorectal_cancer_deaths ~
                                                Frequent_physical_distress+offset(offset_value),
                                              data=cleaned_df)
diabetes_frequent_physical_distress_nb_model <- glm.nb(Colorectal_cancer_deaths ~ Diabetes+
                                                         Frequent_physical_distress+
                                                         offset(offset_value), data=cleaned_df)


#Health Outcomes as the predictor
```

```r
health_outcomes_nb_model <- colorectal_nb_model_function(health_outcomes_df)

#SES as predictors
social_economic_factors_nb_model<-colorectal_nb_model_function(social_economic_factors_df)

#health behavior as predictors
health_behavior_nb_model <- colorectal_nb_model_function(health_behavior_df)

#physical environment as predictors
physical_environment_nb_model <- colorectal_nb_model_function(physical_environment_df)

#clinical care as predictors
clinical_care_nb_model<- colorectal_nb_model_function(clinical_care_df)

#All covariates
full_nb_model <- glm.nb(Colorectal_cancer_deaths ~
                          Air_pollution___particulate_matter+Binge_drinking+Diabetes+
                          Frequent_physical_distress+Life_expectancy+
                          High_blood_pressure+Income_Inequality+
                          Racial_ethnic_diversity+Housing_with_potential_lead_risk+
                          Obesity+Physical_inactivity+
                          Preventive_services+Smoking+Uninsured+
                          Average_Annual_Case_Count+Per_Average_Case_Count_Late_Stage+
                          Limited_access_to_healthy_foods+
                          offset(offset_value), data=cleaned_df, link=log)

#elastic net
deaths.negbinom <- glm.nb(Colorectal_cancer_deaths ~ Air_pollution___particulate_matter+
                            Binge_drinking+Frequent_physical_distress+
                            Housing_with_potential_lead_risk+Income_Inequality+
                            Life_expectancy+Limited_access_to_healthy_foods+
                            Obesity+Preventive_services+Racial_ethnic_diversity+
                            Uninsured+Average_Annual_Case_Count+
                            Per_Average_Case_Count_Late_Stage+offset(offset_value),
                          data=cleaned_df, link=log)
```

```
goodness_of_fit <- data.frame(Models = c("Health Outcomes","Socioeconomic Factors",
                      "Health Behavior","Physical Environment",
                      "Clinical Care","Full Model","Model with Feature Selection via Elastic Net"),
                      `Overdispersion Test (Deviance)`=
                        c(deviance(health_outcomes_model)/
                            health_outcomes_model$df.residual,
                          deviance(social_economic_factors_model)/
                            social_economic_factors_model$df.residual,
                          deviance(health_behavior_model)/
                            health_behavior_model$df.residual,
                          deviance(physical_environment_model)/
                            physical_environment_model$df.residual,
                          deviance(clinical_care_model)/
                            clinical_care_model$df.residual,
                          deviance(full_model)/
                            full_model$df.residual,
                          deviance(poisson_model_enet)/
                            poisson_model_enet$df.residual
                          )
                      )


goodness_of_fit %>% kableExtra::kable() %>% kableExtra::kable_styling()
```

Evaluation the assumptions of Poisson regression models:

1) Linearity: The log of the mean colorectal cancer deaths rate is a linear function of the covariates
2) Independence: Each city's count of colorectal cancer should not be affected by other cities
3) Stationarity: The incidence rate is time invariant, as we wouldn't see higher rates of colorectal cancer at different times of the year
4) Poisson Outcome: The outcome, number of colorectal cancer deaths per 1,000,000 people, is a count

per unit of time or space; however, when assessing goodness of fit, the deviance statistic is not close to 1, therefore this assumption does not hold