Spatiotemporal & Socioeconomic Lung Cancer Relationships in Texas Between 1995 and 2015

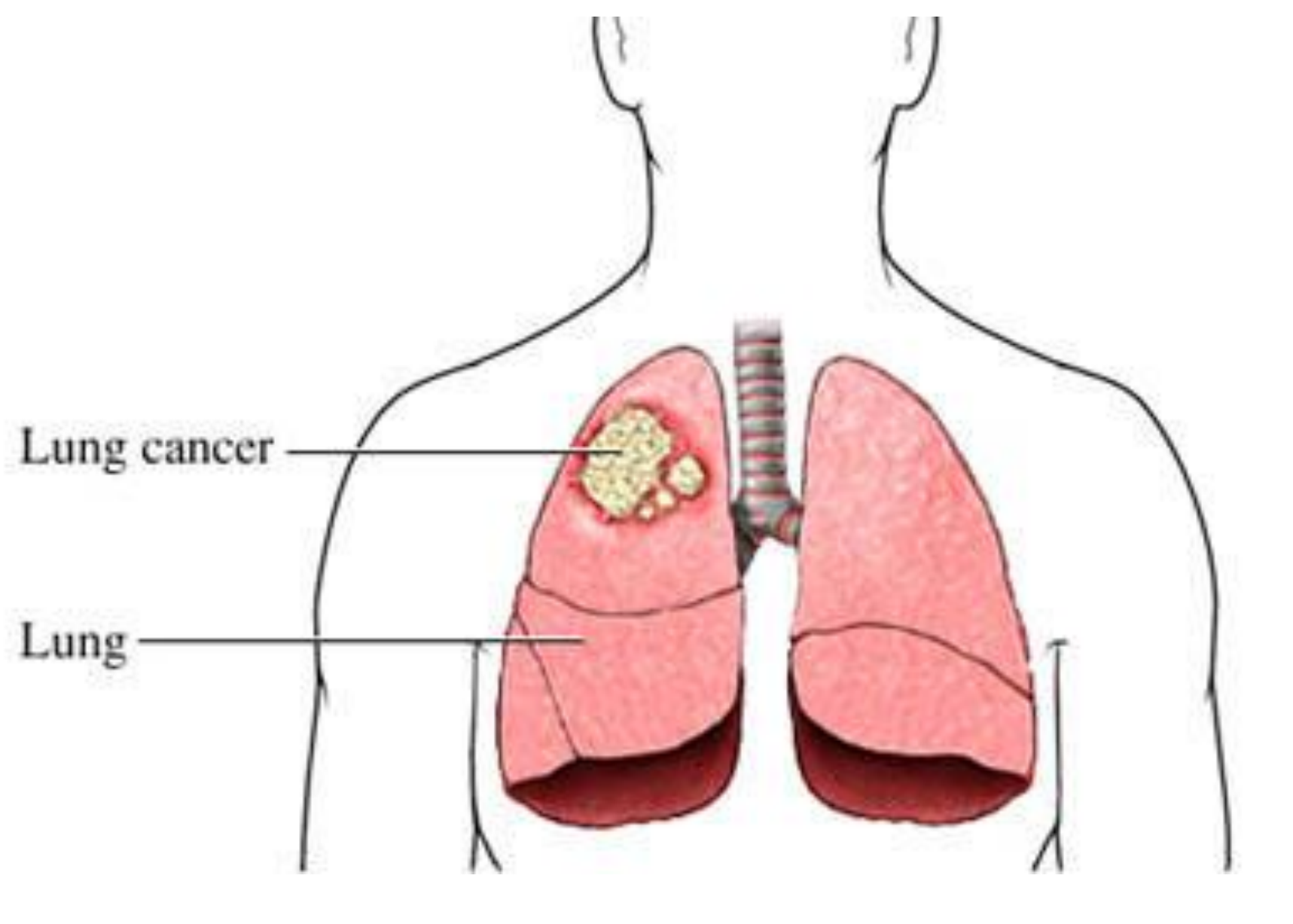
Dileka Gunawardana with Mentorship from Dr. Cici Bauer

CPRIT Summer Internship Program; UT Houston School of Public Health

Abstract

In the United States, lung cancer has the second highest incidence rate and the highest mortality rate of any cancer. This study investigated county-level spatial and temporal trends of four lung cancer histologic types in the state of Texas between 1995 and 2015. The results were then published on an online interactive dashboard for the purpose of assisting public health officials in the allocation of state funds. A combination of the Bernardinelli and Leroux models was used to find the relative risk (smoothed version of SIR) for each county in each year of the study. Implementation was by R'S INLA software, which is a faster alternative to MCMC sampling for Bayesian models, conducted via a Laplace approximation of the marginal posterior distribution. Although most lung cancer types have been trending downward in recent years due to decreased smoking use, adenocarcinoma has seen a rise in the last 20 years, driven mostly by females. The areas of the highest modeled relative risk for all years of the study tended to be in the eastern region of the state. While there were no significant associations between relative risk and county-level poverty rate, there were some between relative risk and rurality. The more metropolitan that a county was, the higher its risk, relative to the rest of the state, to lung cancer.

Introduction



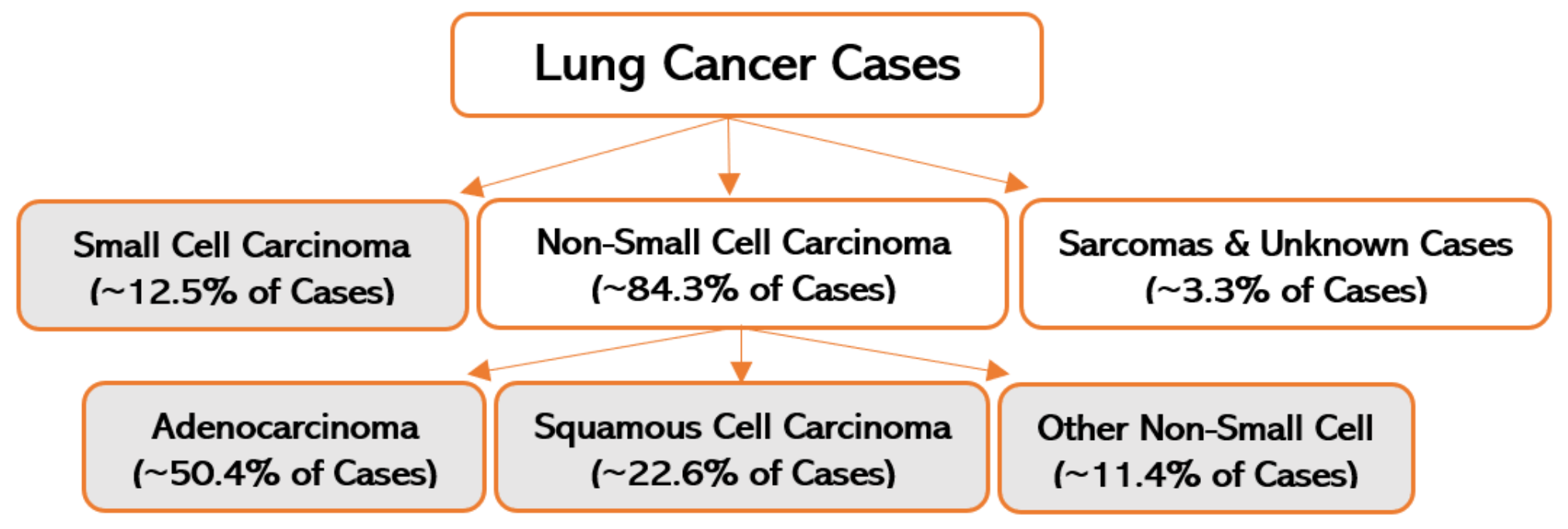
As shown by the diagram on the left, a lung cancer (LC) tumor can have devastating effects on how a person breathes and lives. Lung cancer is the leading cause of cancer mortality in the world, with many people dying due to late-stage diagnoses. As a result, it has become increasingly important to determine sub-populations and area-types that have an increased risk for the disease. In the United States alone, approximately 228,280 people are projected to be diagnosed with lung cancer in 2020. **The goal of this study** was to model the spatial (across different counties), temporal (over time), and spatiotemporal relationships of lung cancer across the state of Texas. In addition, other socioeconomic factors including rurality and poverty rates were investigated to determine any kinds of potential associations. In addition, an interactive online dashboard app was created via *R Shiny* software to describe the models for and display the results of the study.

Contact Information

Dileka Gunawardana: Incoming junior studying statistics at Rice University
Email: sdg2@rice.edu / dilekag21@gmail.com
Phone: (203) – 644- 2968
Dr. Cici Bauer: Assistant professor at UT Health Science Center in Houston
Email: cici.x.bauer@uth.tmc.edu
Website: https://cicibauer.netlify.com

Methods of Analysis

Lung cancer data was collected from the Texas Cancer Registry and processed via the SEER*Stat software. There are multiple classifications of lung cancer, called **histologic types**, based on the appearance of the cancerous cell under a microscope. Each has a unique etiology, therefore affecting people differently, meaning each should be studied individually. The grey boxes in the following diagram show the divisions of and relative prevalence of the four histologic types included in this analysis.


Lung Cancer Cases

Small Cell Carcinoma (~12.5% of Cases)

Non-Small Cell Carcinoma (~84.3% of Cases)

Sarcomas & Unknown Cases (~3.3% of Cases)

Adenocarcinoma (~50.4% of Cases)

Squamous Cell Carcinoma (~22.6% of Cases)

Other Non-Small Cell (~11.4% of Cases)

5-year population data was collected from the U.S. Census Bureau and divided into **30 demographic groups** determined by race, age, and gender. It was assumed that each of these groups had a unique lung cancer rate (this is because, for example, the elderly tend to have higher rates than the young). This data, along with lung cancer frequencies, was used to calculate expected cases and SIR's (for **demographic group k**, **county i** & **year j**):

$$LC\ Rate_{kj} = \frac{n\ of\ Texans\ Diagnosed\ with\ LC_{kj}}{n\ of\ Texans_{kj}} \tag{1}$$

$$Expected\ Cases_{ij} = \sum_{k=1}^{30} n\ of\ Texans_{kij} \cdot LC\ Rate_{kj} \tag{2}$$

$$SIR_{ij} = \frac{Observed\ Cases_{ij}}{Expected\ Cases_{ij}} \tag{3}$$

SIRs are useful because they have a straightforward interpretation. A value greater than 1 indicates a potential high incidence county or "hot spot" that may be at particular risk for lung cancer. A value less than 1, on the other hand, indicates a "cold spot". **Independence** between each possible pair of observations (in this case, an observation is the lung cancer SIR in a county for a given year) is an assumption made when analyzing data. However, this is unfair to make because there exist **three types of correlation** that must be considered through a **model** to accurately understand lung cancer trends:

1. Spatial: Counties close to each other share topographic and other traits
2. Temporal: Various events/ anomalies happen in certain years that are unrelated to lung cancer but still influence the counties' rates
3. Spatiotemporal: The two effects above may interact with one another

By getting rid of uncertainty and unnecessary noise, **model-based relative risk (RR),** provides a smoothed version of SIR, is less vulnerable to abnormalities, and is generally considered more accurate. A combination of the **Bernardinelli Model and Leroux Model** (below) was used to model each county's relative risk of lung cancer. It was implemented through the R-INLA software by the methods outlined in both Moraga and Rubio-Gomez V.

$$Observed\ LC\ Cases_{ij} \sim Poisson(Expected\ LC\ Cases_{ij} \cdot RR_{ij}) \tag{4}$$

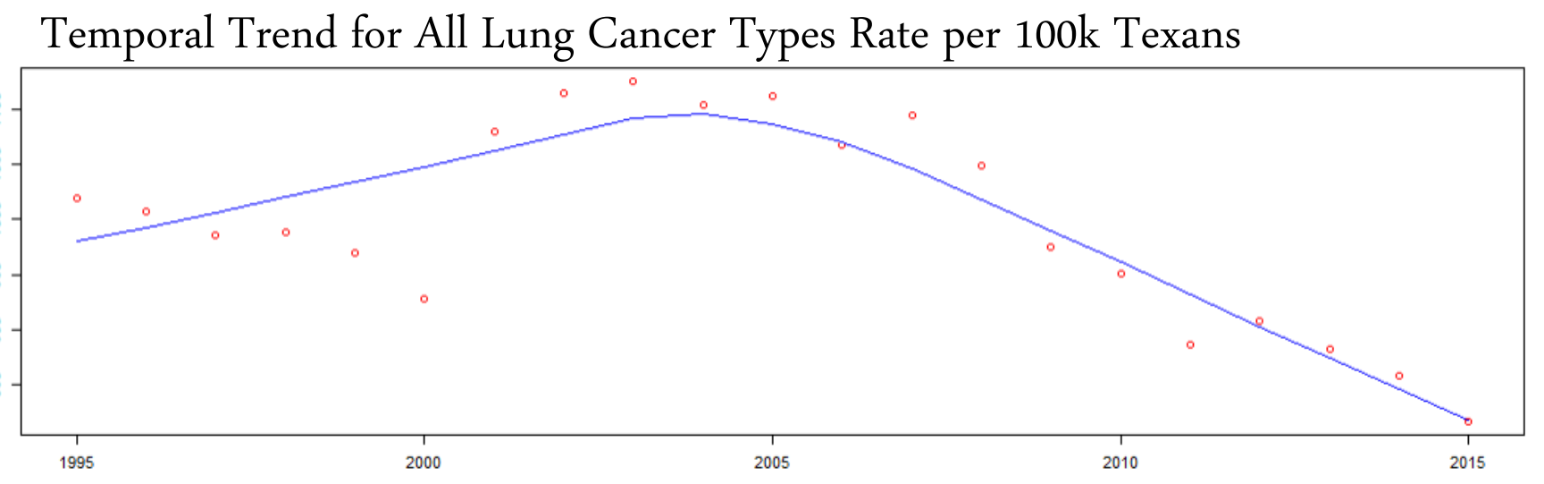
$$log(RR_{ij}) = \alpha + s_i + t_j + \delta_{ij} \tag{5}$$

This model estimates relative risk (RR) for **county i** in **year j** where:

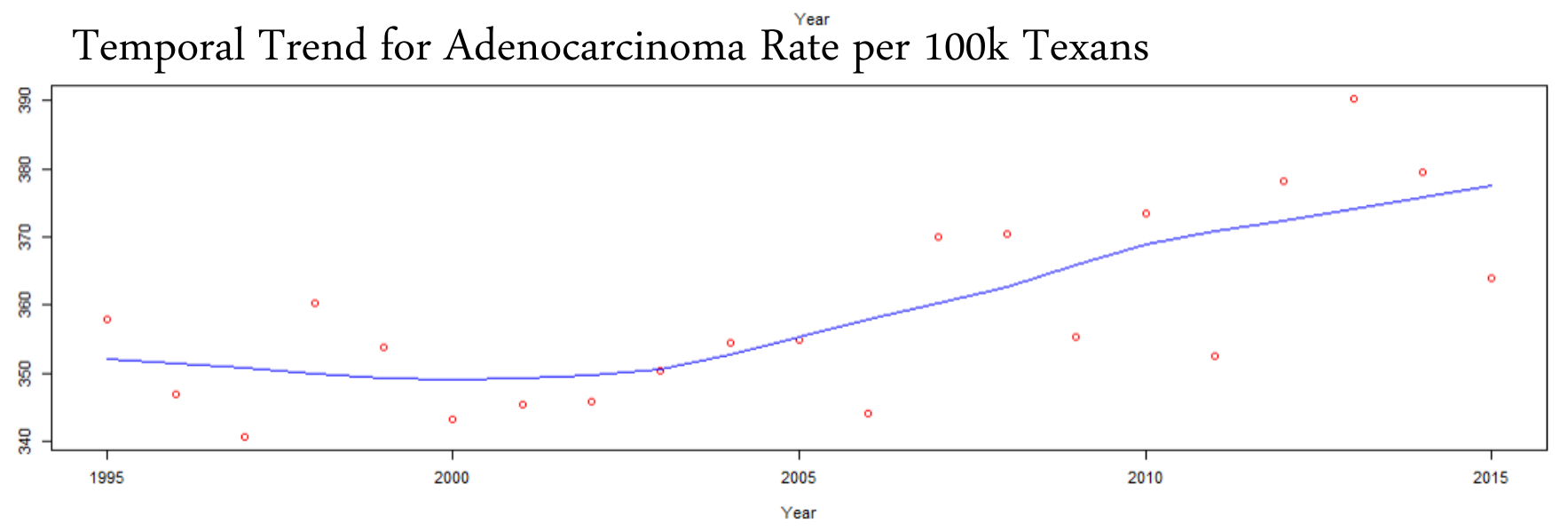
- **α** is the intercept
- **s** represents the spatial effects via a neighborhood matrix and Leroux parameter to determine the spatial dependency of the data
- **t** represents the temporal effects via a Markovian random-walk model of order two
- δ represents the spatiotemporal effects via a completely random, independent and identically distributed model.

Results

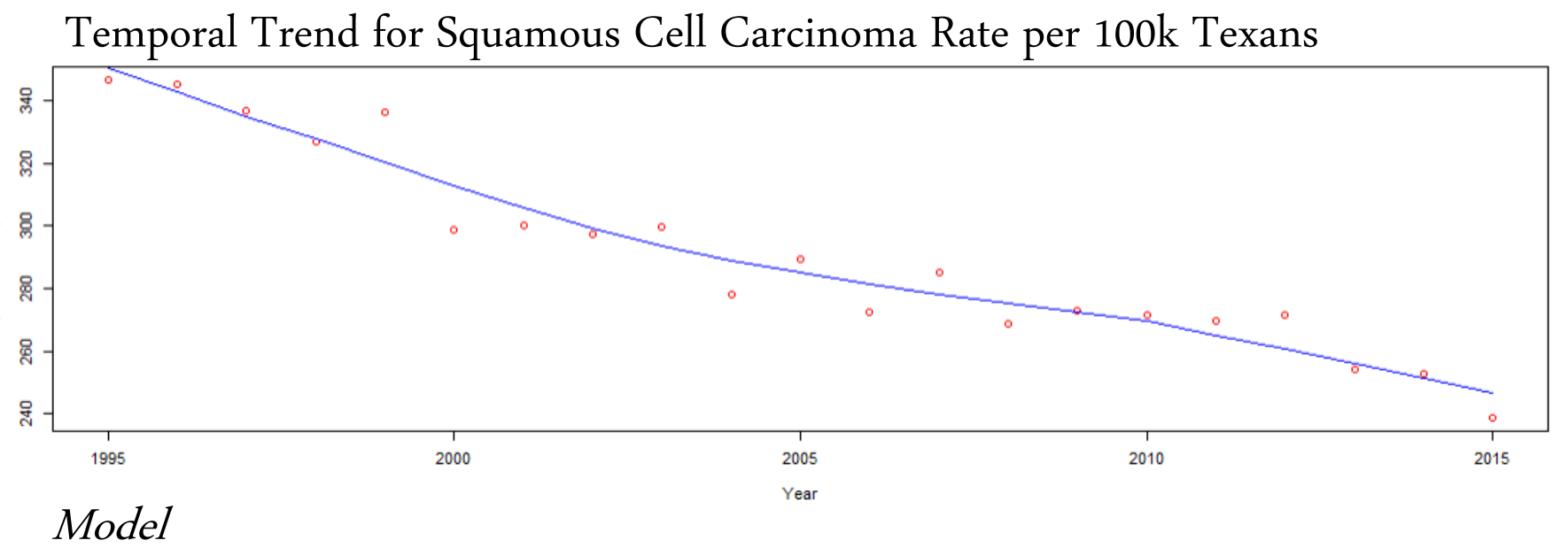
Temporal Trend for All Lung Cancer Types Rate per 100k Texans



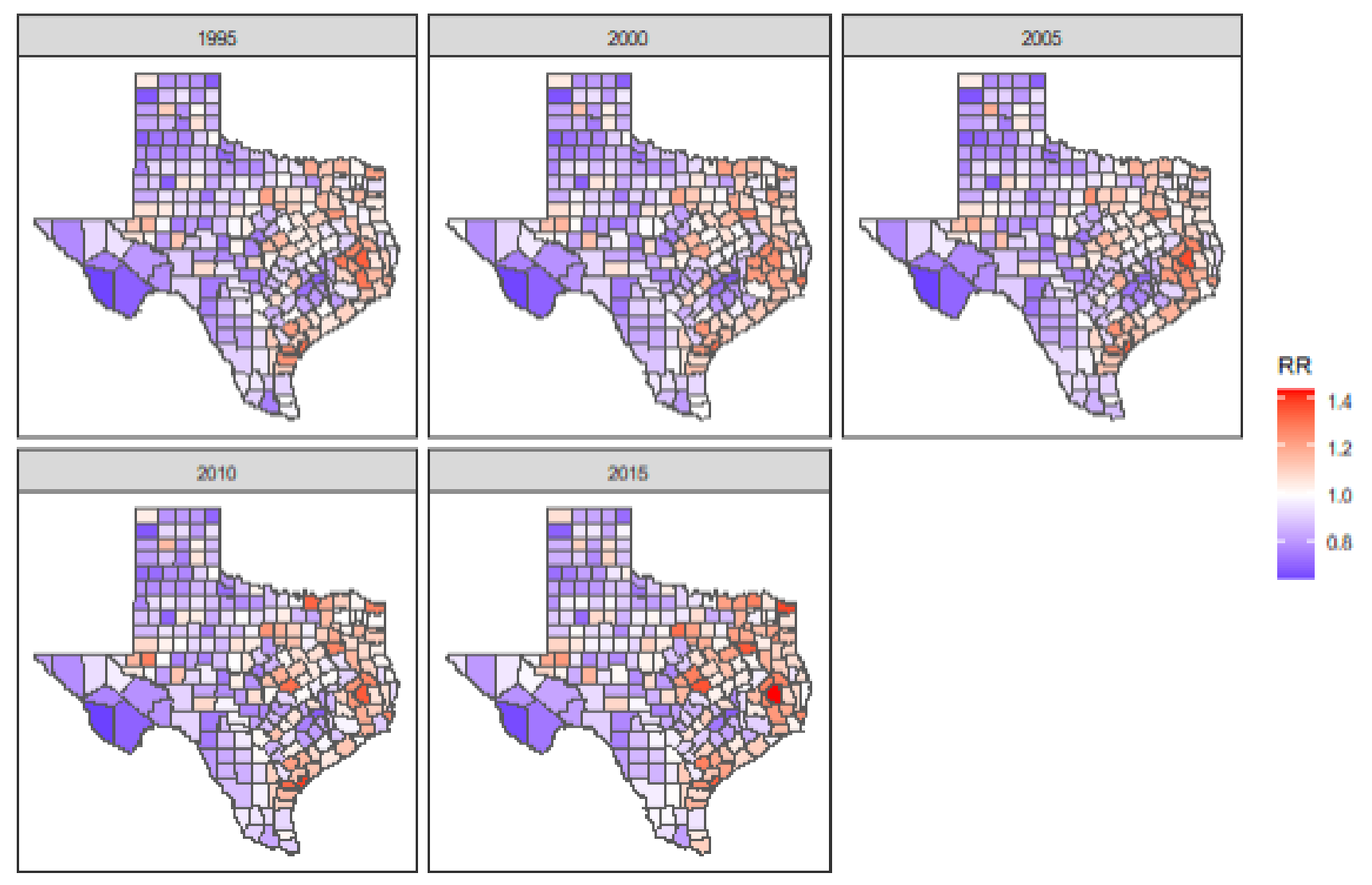
Temporal Trend for Adenocarcinoma Rate per 100k Texans



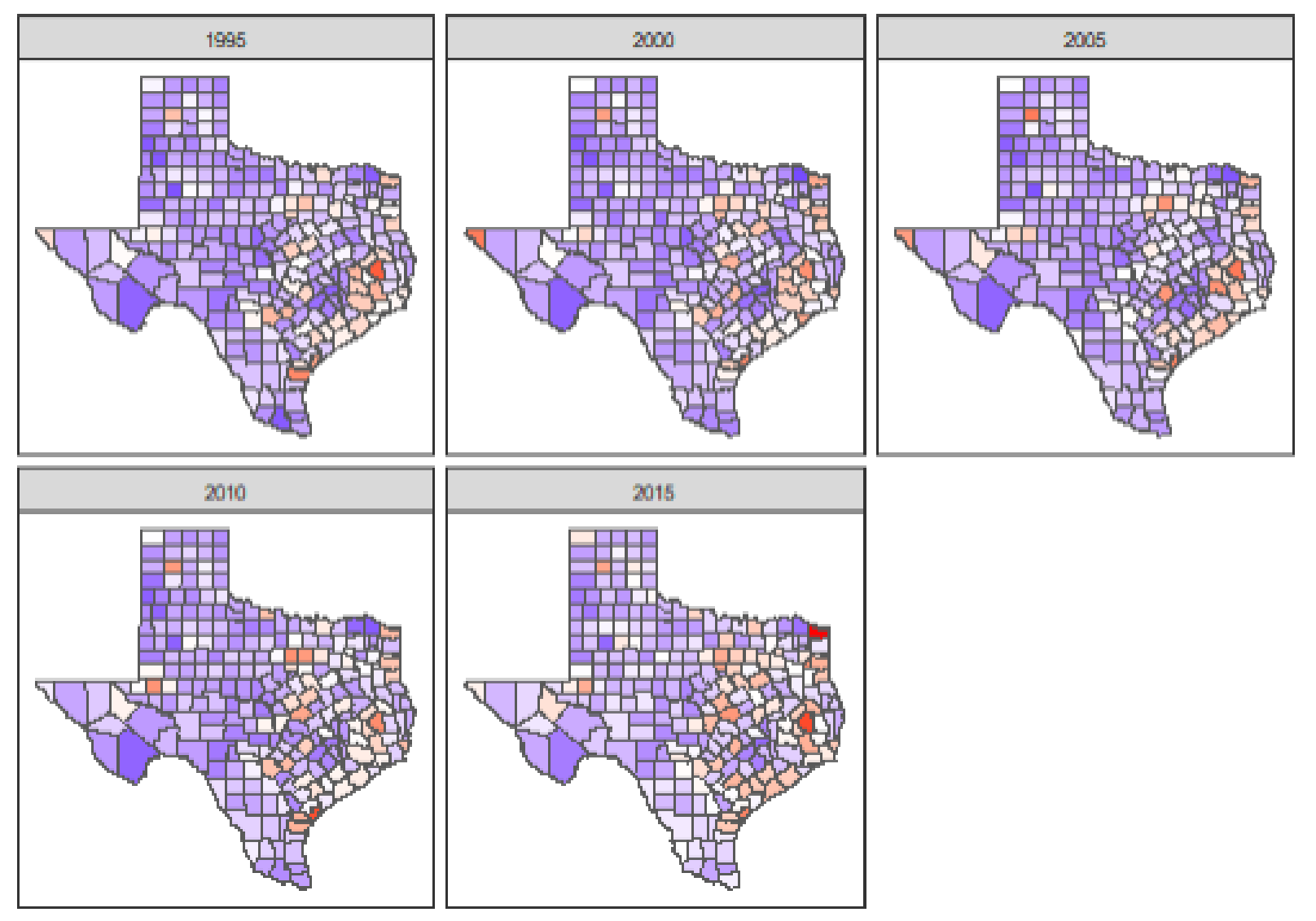
Temporal Trend for Squamous Cell Carcinoma Rate per 100k Texans



Modeled Relative Risk for All Lung Cancer Types by County

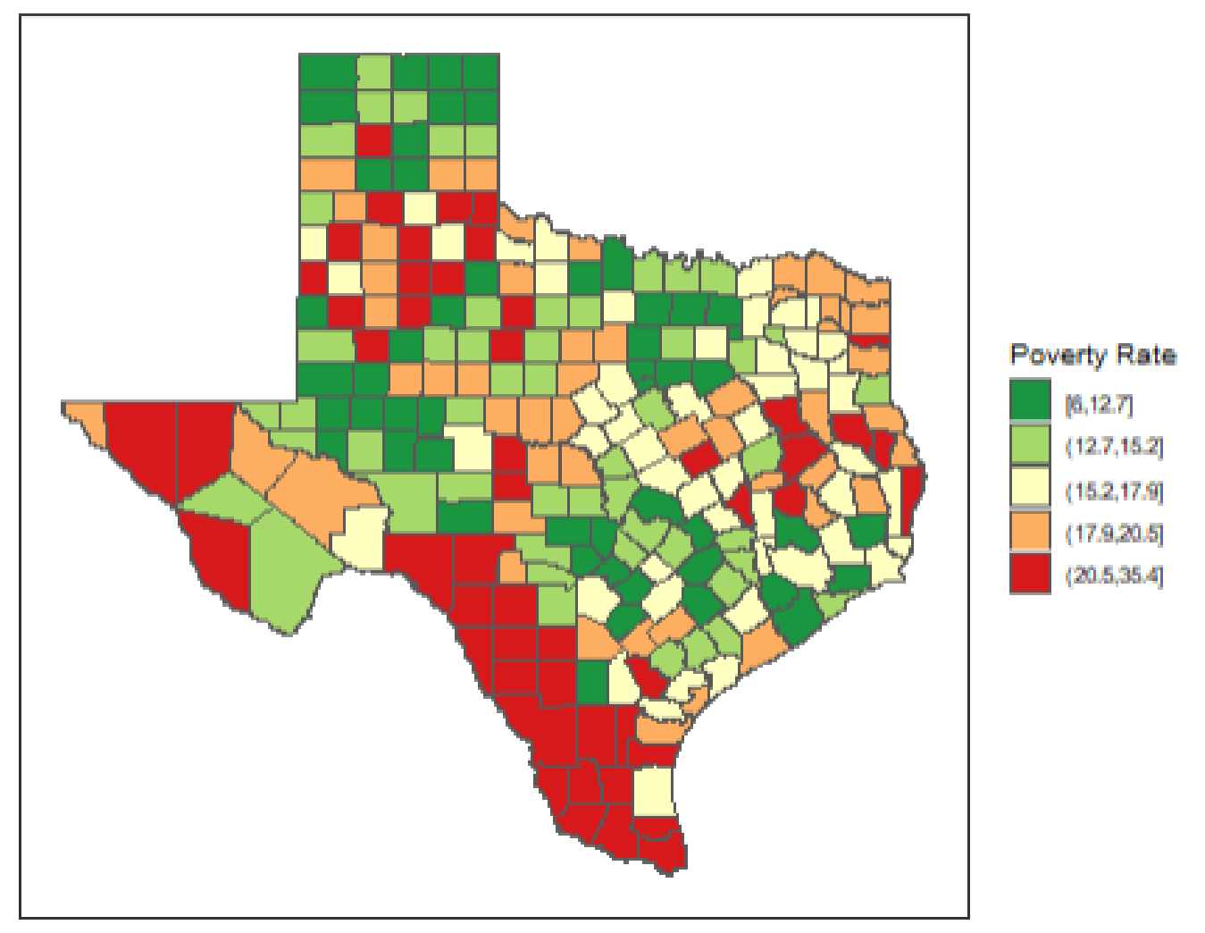


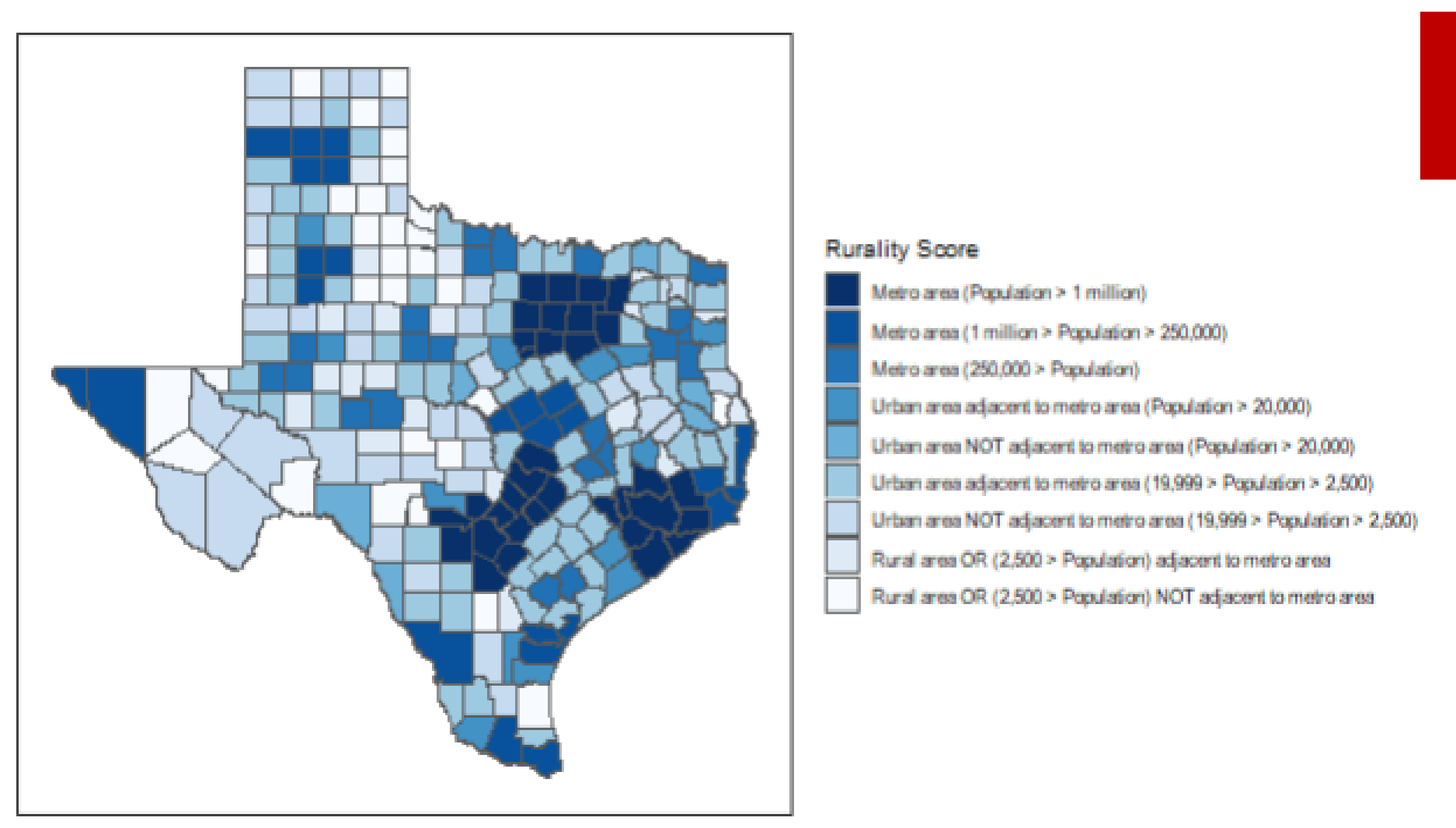
Modeled Relative Risk for Adenocarcinoma by County



Model Covariate Parameter Values for

Poverty Rate		Rurality Score	
Cancer Type	**95% C.I.**	**Cancer Type**	**95% C.I.**
All	(-0.0007, 0.0066)	All*	(-0.0214, -0.0024)
Small Cell	(-0.0019, 0.0122)	Small Cell*	(-0.0418, -0.0033)
Adenocarc.	(-0.0043, 0.0043)	Adenocarc.*	(-0.0318, -0.0088)
Squamous	(-0.0019, 0.0102)	Squamous	(-0.0219, 0.0103)
Other	(-0.0002, 0.0151)	Other	(-0.0367, 0.0030)

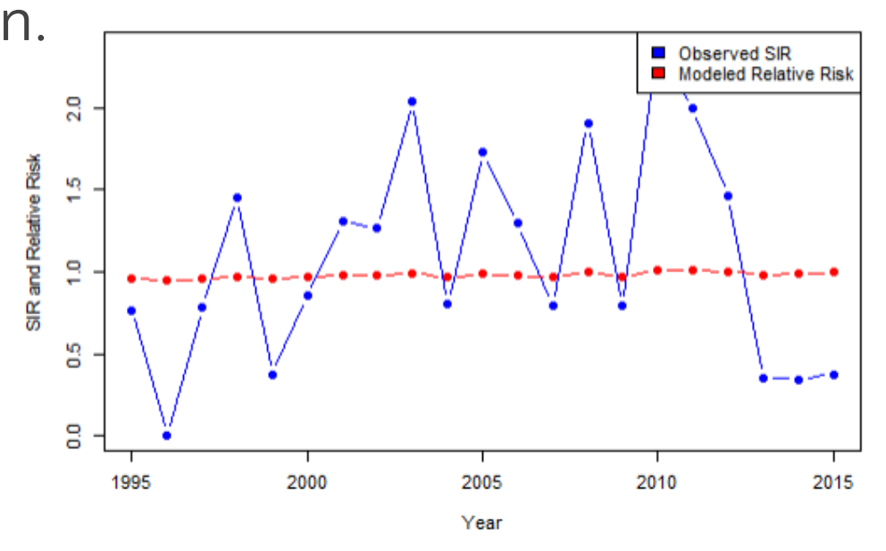


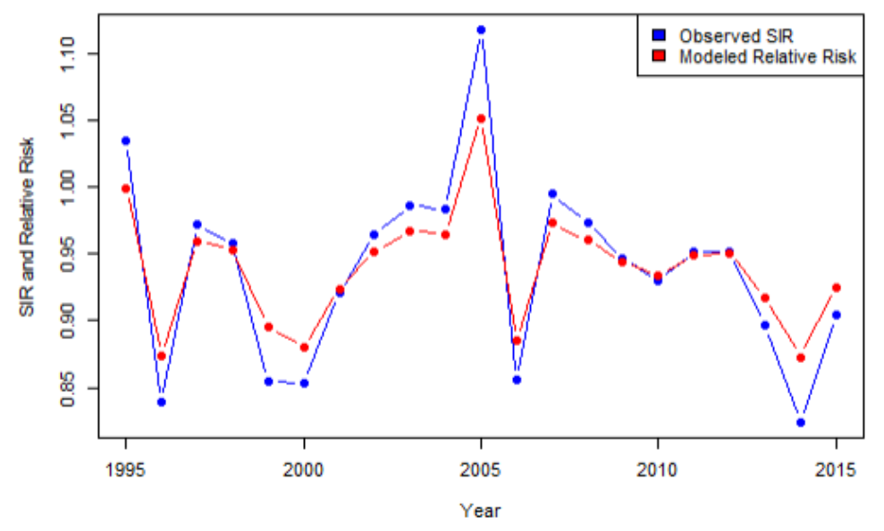

Rurality Score
Metro area (Population > 1 million)
Metro area (1 million > Population > 250,000)
Metro area (250,000 > Population)
Urban area adjacent to metro area (Population > 20,000)
Urban area NOT adjacent to metro area (Population > 20,000)
Urban area adjacent to metro area (19,999 > Population > 2,500)
Urban area NOT adjacent to metro area (19,999 > Population > 2,500)
Rural area OR (2,500 > Population) adjacent to metro area
Rural area OR (2,500 > Population) NOT adjacent to metro area

Conclusions

From the spatial trends map, it appears that there are several regions in Eastern Texas with abnormally high risks to lung cancer, notably Polk and Lamar counties. Although squamous cell cancer's prevalence (associated with smoking) in Texas has decreased, adenocarcinoma's has increased, with the rise being driven mainly by women.

Rurality was found to have significant negative correlations with multiple lung cancer types, suggesting that the more rural a county is, the lower that county's risk for lung cancer tends to be. The plots on the right show the effects of modeling. Andrews county (top) has a much smaller population, so even though its risk was more-or-less the same over the time period, its SIR fluctuated unnecessarily when compared to Dallas county (bottom), which had a generally decreasing trend.





Discussion & Future Directions

It's important to determine the counties in Texas that are most at risk for lung cancer for state public health officials to efficiently allocate funds and diagnostic resources to the appropriate demographic groups and counties. Currently, hospitals and health professionals are being inundated by the novel COVID-19 pandemic, meaning many oncology clinics aren't functioning at their normal levels. It would be of interest to investigate lung cancer and the virus alongside one another to determine how they not only affect patients but also health institutions at a macroscopic level.

Investigating the spatiotemporal patterns of lung cancer will also help set the groundwork for further analyses in high-incidence areas. In addition, this work may open the gates for establishing causal relationships. Previous literature has pointed to alternative potential causes of lung cancer – farming pesticides, mining, air pollution, radon, and asbestos – but none of these have been thoroughly investigated. This research project was limited in that it was focused on county-level characteristics. This is problematic because people with very different lifestyles can live in the same county. Future work should attempt a point-level analysis where specific lung cancer cases or neighborhoods can be studied alongside detailed lifestyle and air quality data.

Acknowledgements

I would like to acknowledge my mentor, Dr. Cici Bauer, for creating the original idea for the project as well as for providing me resources & guidance throughout the research process.

Dileka Gunawardana was supported by the UTHealth Innovation for Cancer Prevention Research Training Program Summer Undergraduate Fellowship (Cancer Prevention and Research Institute of Texas Grant #RP160015).

References

- Christian WJ, et al. Spatiotemporal Analysis of Lung Cancer Histological Types in Kentucky, 1995-2014. Cancer Control, Vol 26: 1-8. March 21, 2019. DOI: 10.1177/1073274819845873
- Houston KA, Mitchell KA, King J, White A, Ryan BM. Histologic Lung Cancer Incidence Rates and Trends Vary by Race/Ethnicity and Residential County. J Thorac Oncol. 2018; 13(4):497-509. doi:10.1016/j.jtho.2017.12.010
- Moraga P. (2020). Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny. CRC Press. ISBN: 978-0367357955
- Rubio-Gomez V. (2020). Bayesian Inference with INLA. CRC Press. ISBN: 978-1138039872

Main Data Sources

- Population Estimates Program, Population Division, U.S. Census Bureau. Intercensal Estimates of the Resident Population by Five-Year Age Groups, Sex, Race, and Hispanic Origin for Counties. Washington, DC.
- Texas Cancer Registry (www.dshs.state.tx.us/tcr) SEER*Stat Database, Limited_Use 1995-2017 Incidence, Texas statewide, Texas Department of State Health Services, created December 2019, based on NPCR-CSS Submission, cut-off 11/07/19.