**Introduction**

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.

**Project Overview**

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, COVID-19 and other socioeconomic factors were investigated to determine any kinds of associations. All cancer data was from the Texas Cancer Registry and processed via the SEER\*Stat software.

**Explanation of Histologic Types:**

There are multiple classifications of lung cancer called 'histologic types' based on the appearance of the cancerous cell under a microscope. Only pathologists have the expertise to identify the specific visual differences between histologic cell types. 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.

As shown above, carcinomas (cancer forms in the skin/ tissue cells lining internal organs) are far more common than sarcomas (forms in connective tissue cells like fat and blood vessels). Most carcinomas are non-small cell; adenocarcinomas and squamous cell carcinomas are the two most common histologic types classified under this category.

**Using the Dashboard**

The sidebar on the left can be used to navigate through the various sections of the dashboard. The top of each page contains a blurb detailing its contents. There are multiple maps being constructed in addition to thorough sampling procedures, so please be patient with the page as it may need 1-2 minutes to load the appropriate results once a selection is made.

**Citations**:

(1) Li J, Guo W, Ran J, et al. Five-year lung cancer mortality risk analysis and topography in Xuan Wei: a spatiotemporal correlation analysis. BMC Public Health 19, 173 (2019). <https://doi.org/10.1186/s12889-019-6490-1>

(2) Siegel RL, Miller, KD, Jemal A. Cancer statistics, 2020. CA: A Cancer Journal for Clinicians. Vol 70; 1. January 8, 2020. <https://doi.org/10.3322/caac.21590>

(3) 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.

**General Trends Explanation:**

On this “General Trends” page of the dashboard, the first set of plots show the spatial relationships of lung cancer in order to visualize where there may be counties/ regions with abnormally high rates of specific lung cancers. The second set of plots show the temporal trends of different lung cancer types between 1995 and 2015. In addition, through the selection of inputs, the trends regarding specific age and/or gender groups can be investigated.

Disclaimer: Starting in 2001, a new category called “non-small cell carcinoma” was added under this classification, which may be the cause of the increase in cases during this time.

**SIR Plots & INLA Modeling Page**:

This **SIR Plots & INLA Modeling** page of the dashboard visualizes the spatiotemporal relationships of both the **true and modeled** data for each histologic type of lung cancer at both the **state and county** levels for the following three metrics:

(1) Observed (true) number of lung cancer cases

(2) Expected number of lung cancer cases: this number was calculated based on 5-year population census data. Essentially, state-wide

(3) Standardized Incidence Ratios (SIR’s). SIR = Expected/ Observed

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”.

**What is the importance of modeling the data?** Independence between each possible pair of observations (in this case, an observation is the relative risk or rate of lung cancer in a county for some year) is an assumption made when analyzing data. However, this is an unfair assumption because there exist **three types of correlations** that must be considered in order to accurately understand lung cancer trends:

(1) Spatial: Counties that are close to one another often share socioeconomic traits and topographies.

(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 and contribute to additional variation

For this project’s analysis, a combination of the **Bernardinelli Model** (2) and **Leroux Model** (3) was used to model the relative risk of lung cancer. It was implemented through the R-INLA (4) software by the methods outlined in both Moraga (5) and Rubio-Gomez V (6).

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.

(2) Bernardinelli L, Clayton DG, Pascutto C, Montomoli C, Ghislandi M, Songini M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* 14: 2433–43.

<https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780142112>

(3) Leroux B, Lei X, Breslow N. (1999). Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. *In Statistical Models in Epidemiology, the Environment and Clinical Trials*, edited by M Halloran and D Berry, 135–78. New York: Springer-Verlag.

<https://link.springer.com/chapter/10.1007/978-1-4612-1284-3_4>

(4) Rue H, Martino S, Chopin N. (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B, 71(2):319{392}.

<http://www.r-inla.org/>

(5) Moraga P. (2020). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. CRC Press. ISBN: 978-0367357955

<https://www.paulamoraga.com/book-geospatial/index.html>

(6) Rubio-Gomez V. (2020). *Bayesian Inference with INLA.* CRC Press. ISBN: 978-1138039872

<https://becarioprecario.bitbucket.io/inla-gitbook/index.html>

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: July 1, 1990 to July 1, 1999; April 1, 2000 to July 1, 2010; April 1, 2010 to July 1, 2019 Washington, DC.

<https://www2.census.gov/programs-surveys/popest/tables/1990-2000/counties/asrh/casrh48.txt>

<https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-counties.html>

<https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-detail.html>

Moraga P. (2020). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. CRC Press. ISBN: 978-0367357955

**Socioeconomic Patterns**

This **Socioeconomic Associations** page of the dashboard investigates the existence of potential associations between county-level characteristics (i.e. poverty and rurality) and different lung cancer histologic types."

1: Metro area (Population > 1 million)

2: Metro area (1 million > Population > 250,000)

3: Metro area (250,000 > Population)

4: Urban area adjacent to metro area (Population > 20,000)

5: Urban area NOT adjacent to metro area (Population > 20,000)

6: Urban area adjacent to metro area (19,999 > Population > 2,500)

7: Urban area NOT adjacent to metro area (19,999 > Population > 2,500)

8: Rural area OR (2,500 > Population) adjacent to metro area

9: Rural area OR (2,500 > Population) NOT adjacent to metro area

It would be of interest to incorporate the county traits above into a model so that it has as much information as possible to most accurately predict relative risk:

More information about how this model was constructed can be found under the **SIR Plots & INLA Modeling** page. The map of Texas below shows the new relative risk for the selected year and selected histologic type based on this model.

If there is no relationship between a county characteristic and a certain histologic type of lung cancer, this will be reflected by the distribution of that trait’s parameter. For example, if small cell carcinoma is selected, and the distribution for beta\_1 indicates that it has a high probability of being 0, then it’s likely that there’s no association between a county’s rurality and its rate of small cell carcinoma in Texas. These distributions and their respective summary statistics (based on the model defined above) are below.

United States Department of Agriculture Economic Research Service. ERS Rural-Urban Continuum Codes (RUCC). 1993, 2003, 2013.

<https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx#.UYJuVEpZRvY>

Method of Using RUCC for Rurality Data: 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

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884169/>

U.S. Census Bureau. Historical County Level Poverty Estimates Tool. 1960-2010.

<https://www.census.gov/library/visualizations/time-series/demo/census-poverty-tool.html>

**COVID-19 and Lung Cancer**

This “COVID-19 and Lung Cancer” page of the dashboard contains side-by-side plots comparing the relative rates of lung cancer and the recent COVID-19 epidemic that has been significantly straining Texas hospitals. It is hypothesized that COVID-19 can have adverse effects on peoples’ lungs, so when assigning health funding, it’s essential to consider these two diseases in relation to one another.

A subject of interest would be to determine whether there is a relationship between how counties are affected by COVID-19 and how they are affected by lung cancer. In order to investigate the potential associations, two spatial models were constructed for, respectively, the two lung cancer metrics above. Both models are based on whichever COVID-19 metric is selected. The model descriptions as well as the distributions of the parameter are below. If there's a high likelihood that the parameter is 0, then there's likely no association between the COVID-19 and lung cancer indicators in that model.

Texas Department of State Health Services. Texas COVID-19 Data: Accessible Dashboard Data.

<https://dshs.texas.gov/coronavirus/additionaldata.aspx>

The Lancet Oncology. COVID-19: global consequences for oncology. *Lancet Oncol* 2020; **21**: 467

<https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(20)30175-3/fulltext>

Yang K, Sheng Y, Huang C, Xiong N, Jieng K, Lu H. Clinical Characteristics, outcomes, and risk factors for mortality in patients with cancer and COVID-19 in Hubei, China: a multicentre, retrospective, cohort study. *Lancet Oncol*; Published online May 29, 2020.

<https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(20)30310-7/fulltext>

Doesn’t:

Lee L, Cazier JB, et al. COVID-19 mortality in patients with cancer on chemotherapy or other anticancer treatments: a prospective cohort study. *Lancet*; published online May 28, 2020. [DOI:](https://doi.org/10.1016/S0140-6736(20)31173-9) https://doi.org/10.1016/S0140-6736(20)31173-9

<https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31173-9/fulltext>