**Introduction**

Of all cancer sites, lung cancer has the second highest incidence and is the leading cause of mortality in the world1. In the United States alone, approximately 228,280 people are projected to be diagnosed with it in 2020; 14,380 of which will be Texans2. The 5-year relative survival rate for lung cancer is 19%, which is one of the lowest, particularly when compared to other high-incidence cancers such as prostate (98%) and breast (90%)2. As a result, it has become increasingly important to determine sub-populations and area-types that have an increased risk for the disease3. There are multiple practical options for both early detection (low-dose spiral CT screening, chest X-rays (CXR), sputum cytology, etc.) and treatment (resection, chemotherapy, etc.) for these individuals most at risk 3-5.

In December of 2019, the world was introduced to the novel COVID-19 virus, a pandemic that has taken thousands of lives and infected millions worldwide; in the United States, medical institutions and resources have become overwhelmed, leading to a state of emergency across the nation6. The CDC has officially stated that cancer increases one’s risk of severe illness from COVID-197. Previous work has suggested that cancer patients have a higher death rate from coronavirus8, particularly lung cancer patients who are typically older in addition to having cardiovascular/ respiratory comorbidities, smoking-related lung damage, abnormalities in airway/ pulmonary tissue, and are on immunosuppressing treatment9.

Previous studies have indicated evidence of spatiotemporal relationships among lung cancer in Maine, South Carolina and Kentucky10-13. The goal of this analysis was to determine (1) spatiotemporal relationships and trends; (2) county-level socioeconomic risk factors for lung cancer; (3) correlations with and effects of the COVID-19 pandemic in Texas. Finally, an easy-to-use interactive dashboard app was created to visualize the results of the study, which can be accessed by the following link: . Such information will be vital for policy and health leaders to best understand lung cancer and efficiently utilize early detection/ treatment resources.

**Data Collection & Preparation**

Lung cancer data for every year between 1995 and 2017 was collected from the Texas Cancer Registry (TCR) and processed via SEER\*Stat, a software that ensures strict data quality measures14. State law requires that health care professionals and institutions report information regarding cancer diagnoses and treatments to the TCR, meaning the data is as exhaustive as possible within Texas15. The analysis was conducted by histologic type as each is hypothesized to have a unique etiology, onset, and effect on the patient. The gray boxes in **Fig 1** detail the four histologic types that were used to divide the lung cancer dataset for this analysis.

**Figure 1:** Relative Distribution of Lung Cancer Diagnoses in the U.S. from 2013-2017 by Histologic Type16

A close up of a map

Description automatically generated

The following were the histologic codes, as determined by the third edition of the *International Classification of Diseases for Oncology* (ICD-0-3), used to define each histologic type above. Only microscopically confirmed, invasive/ metastatic lung cancer cases were considered (“/3”). Lung cancer was defined by the ICD-0-3/WHO 2008 site classification “Lung and Bronchus” within the SEER\*Stat software. Small cell carcinoma was classified under 8002 and 8041-5. Adenocarcinoma was classified under 8050, 8140-1, 8143-4, 8145, 8147, 8201, 8250-5, 8260, 8262, 8290, 8310, 8320, 8323, 8333, 8401, 8441, 8470, 8480-1, 8490, 8507, 8550-1 and 8570-6. Squamous cell carcinoma was classified under 8051-2, 8070-6, 8083-4, 8094, 8120, 8123 and 8130. Adenocarcinoma and squamous cell carcinoma are the most common non-small cell histologic types. Other non-small cell carcinomas were classified under 8003-4, 8012-4, 8021-2, 8030-3, 8046, 8082, 8200, 8240-6, 8249, 8430, 8560 and 8562. Sarcomas, non-specified carcinomas, and other specified types were excluded due to their limited prevalence among the population and lack of relation to the four categories defined in **Fig 1**.

Population count data in 1995, 2000, 2005, 2010, and 2015 was collected from the US Census18. These were the years used for the socioeconomic associations’ analysis. The county-level socioeconomic factors included were poverty and rurality. Poverty rate (%) data was collected from the US Decennial Census and, for 2015, the 5-year American Community Survey (ACS)19. County-level poverty rates for the years of 1995 and 2005 were extrapolated by taking the average of the 1990 and 2000 data/ 2000 and 2010 data, respectively. Quantifying the rurality of each county in Texas was conducted by utilizing their “Rural Urban Continuum Codes”, which are calculated by population and commuter data as well as adjacency to nearby metropolitan areas21-22. Because the scores are only available for the years 1993, 2003, and 2013, the most recent year before the desired year of analysis was used. Data regarding the frequency of COVID-19 cases and fatalities by county in Texas was collected from the TX DSHS, which has been updated daily since March 4th, 202022. All datasets were cleaned and processed (accounting for inconsistences/ missing data and reformatting to make the process of analysis more efficient) via both Excel and R.

**Analysis**

A close up of a map

Description automatically generatedA close up of a map

Description automatically generatedThe risk of lung cancer was assumed to be different for various demographic groups. This is because lung cancer rates tend to vary by age, sex, and race. In order to take this into consideration, every year, Texas’ population data was divided into 30 groups that were formed assuming 3 age-groups (<55 Years, 55-74 Years, 75+ Years), 2 genders (Male, Female), and 5 race groups (White, Black, American Indian, Asian and Pacific Islander, and Hispanic). For example, one demographic group was “black females between 55 and 74 years of age”. Based on state-wide lung cancer data, a general lung cancer rate was calculated for each demographic group. These rates were then used to calculate the expected number of lung cancer diagnoses in each county based on the number of people in each demographic group living in that county. Finally, the true number of lung cancer cases in each county were used to find its standardized incidence ratio (SIR). 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”. Assuming demographic group *k* and county *i* in year *j*, SIR was calculated by the following formulas23:

**Figure 4:** Model-based RR & Observed SIR from 1995-2015

*Dallas County (Pop. = 2,600,000)*

*Andrews County (Pop. = 18,000)*

A picture containing table

Description automatically generated

**Model Selection & Implementation**

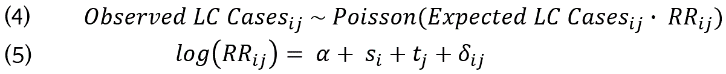
Independence between each possible pair of observations is an unfair assumption in this analysis because there exist three types of correlations in county-level lung cancer data that must be taken into consideration in a model:

(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 influence 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 Model24 and Leroux Model25 was used to model the relative risk of lung cancer 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; and **𝛿** represents the spatiotemporal effects via a completely random, independent and identically distributed model. This model was selected after calculating the information loss (sum of WAIC, DIC, and -∑ log (CPO)) of seven variations of the model for the lung cancer data of each histologic type26. Interestingly, the model with the least information loss of the original data varied by histologic type. The model options and their information loss can be found in the Appendix (**Figs 2-3**). The models were implemented through the R-INLA27 software by the methods outlined in both Moraga28 and Rubio-Gomez V29. 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. **Fig 4** helps visualize the effects of the model – the significant yearly variation in the SIR for smaller counties tends to be due to small sample size, which is why mod eled relative risk fluctuates less in the less populated Andrews County.

**Results**

Detailed results by demographic group & county can be found by the interactive dashboard (link in introduction).

A close up of a map

Description automatically generated

*General* – For all histologic cancer types, **Fig 5** shows that lung cancer rates in Texas increased and then decreased between 1995 and 2015, all trends being mostly driven by the elderly. Males and females had very similar patterns for small cell carcinoma and “other” non-small cell carcinomas. The gap between genders has decreased in recent years, a result that has been seen throughout the US21. However, although squamous cell carcinoma rates have decreased significantly among men, the rate has remained relatively stagnant among women. Adenocarcinoma rates, on the other hand, have remained relatively stagnant among men but increased significantly among females. **Fig 7** shows this concerning trend, which is mainly driven by women above the age of 75. Rates for different histologic types and demographic groups can be found on the dashboard. **Fig 8** visualizes the spatiotemporal trends across counties in Texas by relative risk (modeled version of SIR). Redder counties are at particularly high risk for lung cancer. **Figs 9-12** in the Appendix show these trends by histologic type.

**Figure 8**: Modeled (Eq. 4 & 5) Relative Risk (RR) by County in 1995, 2000, 2005, 2010 & 2015 for All Histologic Types

A close up of a map

Description automatically generated

**Figure 13**: Rurality Score by County in TX in 2015

A close up of a map

Description automatically generated

*Socioeconomic* – Rurality (**Fig 13**) and poverty (**Fig 14**) were two county-level characteristics hypothesized to be related to lung cancer. A modification of the model (Eq 5) was made to incorporate these two datasets as components of α:



**Figure 14**: Poverty Rate (%) by County in TX in 2015

A close up of a map

Description automatically generated **Figure 16**: 95% Confidence Intervals for the Poverty and Rurality Parameter Distributions by Histologic Type

A screenshot of a cell phone

Description automatically generated

**Fig 15** (in the Appendix) displays the relative risk (RR) by county in Texas assuming this modified model (Eq 6). Non-informative priors were assumed for both parameters in the analysis. As outlined in **Fig 16**, the parameter distributions yielded from the model (Eq 6) included 0 in their 95% CI’s for most histologic types. Overall, this provides strong evidence that county-level poverty rate has an insignificant effect on lung cancer risk. However, there were significant negative relationships between poverty and all lung cancer types, small cell carcinoma, and adenocarcinoma. This implies that adenocarcinoma and small cell carcinoma rates tend to be higher in Texas’ less rural and more metropolitan counties.

*COVID-19* – As described in the introduction, it was of interest to investigate potential relationships between COVID-19 and lung cancer, particularly how the pandemic has stressed Texas’ oncological resources. As shown in **Fig 17**, there do not appear to be notable similarities in how the two have affected the state. For example, lung cancer rates appear to be lowest in the southern tip of Texas, but the same area has some of the highest relative incidences of COVID-19. In addition, as evidenced by the parameter distributions in **Figs 18-21** (in the Appendix), there do not appear to be any statistically significant spatiotemporal relationships between how lung cancer and COVID-19 have affected Texas counties. Each of these models takes into consideration both death rates (i.e. how successfully healthcare institutions can care for their patients) and case rates (i.e. areas where people are most at risk).

**Figure 17**: COVID-19 Cases & Lung Cancer Diagnoses per 100k Texans by County

A close up of a map

Description automatically generated

**Conclusions**

Throughout the state of Texas, between 1995 and 2015, there appeared to be a decreasing trend in yearly diagnoses among most histologic types of lung cancer. However, there was an increasing trend among Adenocarcinoma lung cancer, which is the most common type. In addition, spatial patterns from the model suggest that there is a higher risk of lung cancer (including Adenocarcinoma) in the eastern region of the state, which also contains the state’s largest three cities. This further supports the results of the statistical inference conducted, which provided evidence of a significant negative association between “ruralness” and lung cancer risk. In addition, there did not appear to be a relationship between how COVID-19 and lung cancer have affected Texas counties. In other words, counties with higher lung cancer risk or death rates did not appear to be more susceptible to COVID-19).

**Discussion**

The results of this research will be a valuable resource for Texas public health officials, policymakers, and oncological institutions. Lung cancer diagnostic and treatment funds have significant potential to save many peoples’ lives; however, in order to be most effective, they must be distributed to the individuals and areas that are most at risk. Based on the current information available, this project is different from previous related analyses in that it is a histologic analysis focused specifically in Texas. The state is unique in that it is home to three of the largest cities in the country *as well as* several extremely rural farming communities. Also, Texas’ population has a rare demographic makeup, particularly in regard to its relatively large Hispanic population18. In addition, no one has investigated the spatiotemporal patterns in how COVID-19 has stressed lung cancer healthcare resources and patients. Finally, the Bayesian model used and sampling methods via R-INLA may be utilized/ a starting point for similar epidemiological spatiotemporal analyses.

Cigarette smoking is thought to be responsible for up to 90% of lung cancers in men and 65% in women30. Due to highly successful prevention and awareness programs, cigarette smoking among adults in the United States has decreased from ~42% of the population in 1965 to ~20% in 201131. Small cell carcinoma and squamous cell carcinoma have been strongly linked with smoking32 which, coupled with the decrease in cigarette smoking, further substantiates the decreasing rates of these histologic types seen in Texas by this study’s results. It has also been hypothesized that increasing adenocarcinoma rates may be a result of tobacco refinement and filter vents in cigarettes that allow for deeper inhalation32. However, this still does not explain the increasing adenocarcinoma diagnoses among women, who have been smoking less in recent years. It is essential that non-smokers at a high risk of developing lung cancer are not ignored, particularly because these individuals have begun to represent an increasing proportion of one of the most common cancers in the world.

Previous studies have pointed toward radon, asbestos, hormonal therapies, arsenic, infections, pollution, pesticides, genetic predisposition, mining practices, paint, welding fumes, reactive chemicals, and exposure to solvents as being just a few of the potential risks for lung cancer12, 21,32. Increasing trends could also be attributed to better diagnostic tools and the increasing use of medical resources in more rural areas. In this case, the increasing trends would simply be uncovering cancers that had always existed, although this is likely only a component of recent trends. This project will hopefully provide the groundwork as well as be a piece of evidence in continuing to identify and even mitigate the sources causing lung cancer.

There are several limitations of this study that must be taken into consideration before drawing conclusions from the results. First, only *microscopically confirmed* lung cancer cases at a Texas medical institution were included in the analysis. This meant that any Texans who did not meet these criteria were excluded. It is likely that there are individuals who haven’t sought out medical treatment/ utilized the service. However, this number is likely small, and it is impossible to include such data because it simply has not been collected. Secondly, due to the recency of the COVID-19 pandemic, all data regarding testing of the illness has a significant lag period. In addition, lag periods vary by county, so a county with a very “low” COVID-19 incidence may simply have a backlog of data that has not been updated. The interactive dashboard app linked in the introduction will be updated weekly, so hopefully the data and any conclusions drawn will become more accurate once testing sites are not as overwhelmed and results continue to be released.

Finally, this study was a county-level analysis. This is problematic because the same county can have several individuals with extremely different lifestyles and air quality, implying such individuals would also have very differing risks to lung cancer despite living in the same county. As a result, it’s very difficult to make conclusions that attempt such generalizations over such a large group of individuals. There are a limited number of characteristics that are unique to each county and applicable to every person within that county. In the future, it may be of interest to investigate more of these county-level characteristics such as the % of individuals in the county who have insurance, the number of oncology personnel providing services in the area, and even the landscape/ climate (i.e. whether the county is a desert or more tropical). Ideally, a point-level or zip-code analysis would be done throughout the state of Texas or even the US. Then, detailed individual-level and environmental data for each person developed could be gathered.

**Funding**

The author 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**

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

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

3 Knight S, Crosbie P, Balata H, Chudziak J, Hussell T, Dive C. (2017). Progress and prospects of early detection in lung cancer. *Open Biology*. 7(9):170070.

doi:10.1098/rsob.170070

4 Brawley OW, Flenaugh EL. (2014). Low-dose spiral CT screening and evaluation of the solitary pulmonary nodule. *Oncology (Williston Park).* 28(5):441-446.

5 Midthun DE. (2016). Early detection of lung cancer. *F100Res,* Vol 5: Faculty Rev-739,

doi:10.12688/f1000research.7313.1

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

7 Centers for Disease Control and Prevention (CDC). (2020). Coronavirus Disease: Certain Medical Conditions and Risk for Severe COVID-19 Illness. *CDC*.

8 Yang K, Sheng Y, Huang C, Xiong N, Jieng K, Lu H. (2020). 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*.

9 Passaro A. (2020) Testing for COVID-19 in lung cancer patients. *Ann Oncol*. doi:10.1016/j.annonc.2020.04.002

10 Christian WJ, et al. (2019). Spatiotemporal Analysis of Lung Cancer Histological Types in Kentucky, 1995-2014. *Cancer Control*, Vol 26: 1-8.

doi:10.1177/1073274819845873

11 Hosgood HD, Farah C, Black CC, Schwenn M, Hock JM. (2013). Spatial and temporal distributions of lung cancer histopathology in the state of Maine. *Lung Cancer*; 82(1):55‐62. doi:10.1016/j.lungcan.2013.06.018

12 Lewis DR, Pickle LW, Zhu L. Recent Spatiotemporal Patterns of US Lung Cancer by Histologic Type. (2017). *Front Public Health*; 5:82. doi:10.3389/fpubh.2017.00082

13 Carroll R, Lawson AB, et al. (2017). Space-time variation of respiratory cancers in South Carolina: A flexible multivariate mixture modeling approach to risk estimation. *Ann Epidemiol*; 27(1): 42-51.

doi:10.1015/j.annepidem.2016.08.014.

14 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.

15 Texas Cancer Registry. (2019). Cancer Reporting. (https://www.dshs.texas.gov/tcr/reporting.aspx)

16 Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review 1975-2017, National Cancer Institute. Table 15.28. Bethesda, MD, based on November 2019 SEER data submission, posted to the SEER web site, April 2020.

17 World Health Organization. (‎2013)‎. International classification of diseases for oncology (‎‎ICD-O)‎‎ – 3rd edition, 1st revision, 3rd ed. ISBN: 9789241548496. (https://apps.who.int/iris/handle/10665/96612)

18 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.

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

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

21 Houston KA, Mitchell KA, King J, White A, Ryan BM. (2018). Histologic Lung Cancer Incidence Rates and Trends Vary by Race/Ethnicity and Residential County. J Thorac Oncol. 13(4):497-509. doi:10.1016/j.jtho.2017.12.010

22 Texas Department of State Health Services. Texas COVID-19 Data: Accessible Dashboard Data. (https://dshs.texas.gov/coronavirus/additionaldata.aspx)

23 Moraga P. (2020). Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny. CRC Press. ISBN: 978-0367357955. (https://www.paulamoraga.com/bookgeospatial/index.html)

24 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.

25 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.

26 Cramb SM, Duncan EW, Baade PD, Mengersen KL. (2017). Investigation of Bayesian Spatial Models. Brisbane: Cancer Council Queensland and Queensland University of Technology (QUT).

27 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}.

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

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

30 Xie A. (2014). Smoking and lung cancer. *Ann Cardiothorac Surg*. 3(2):221.

doi: 10.3978/j.issn.2225319X.2014.03.09

31 National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta (GA): Centers for Disease Control and Prevention (US); 2014. 13, Patterns of Tobacco Use Among U.S. Youth, Young Adults, and Adults.

32 Furrukh M. (2013). Tobacco Smoking and Lung Cancer: Perception-changing facts. *Sultan Qaboos Univ* Med J. 13(3):345-358. doi:10.12816/0003255