An Examination of Factors Affecting Incidence and Survival in Respiratory Cancers

Katie Frank Roberto Perez
Mentor: Dr. Kate Cowles

ISIB 2015
University of Iowa
College of Public Health

July 16th, 2015

Introduction

Research Goals

- Predict which demographic variables distinguish between types of respiratory cancers.
- Determine which variables affect survival of patients with various types of respiratory cancers.
- Identify what characteristics of county populations affect respiratory cancer incidence rates.

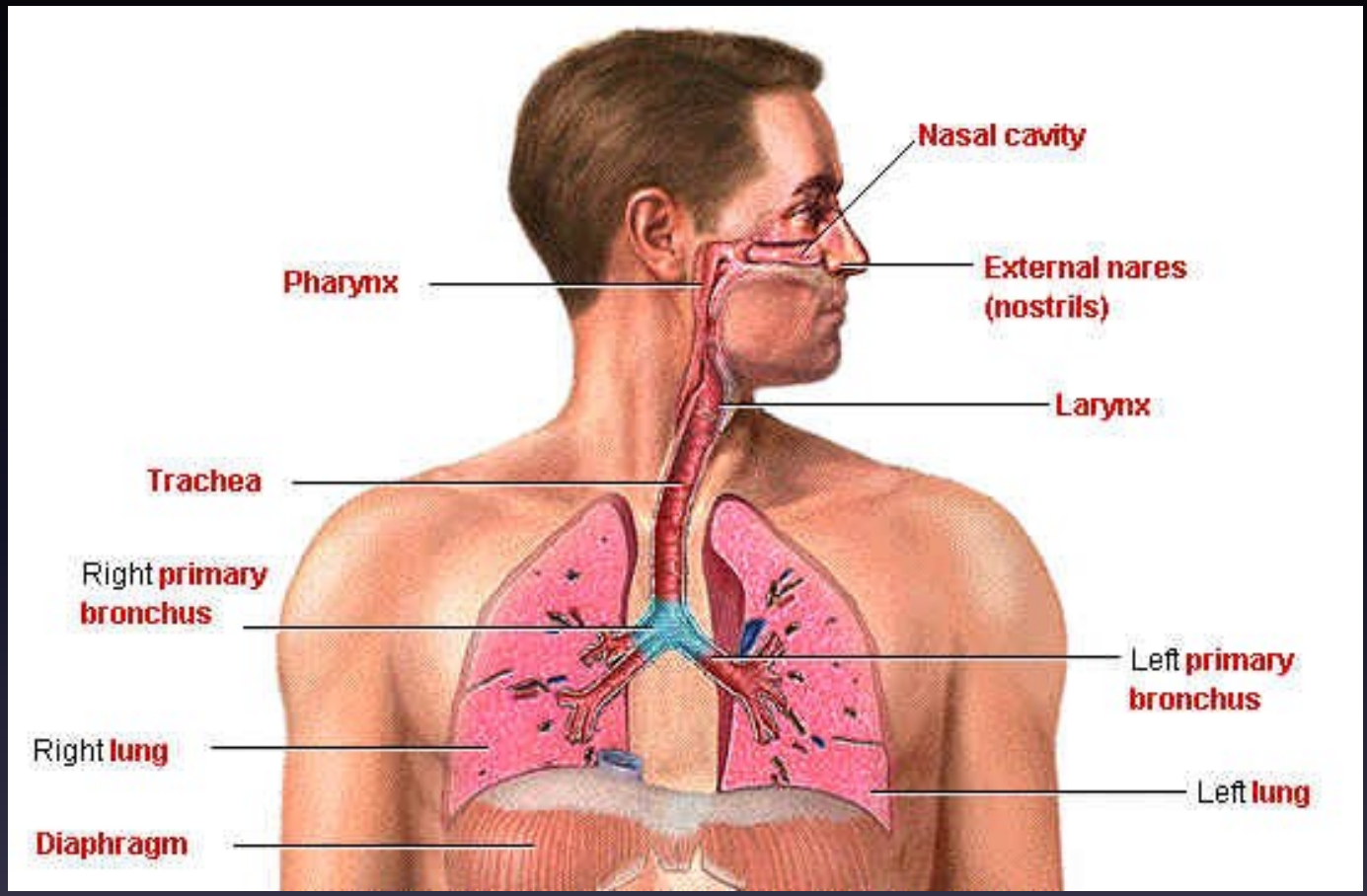
What Are Respiratory Cancers?

1. Cancers that affect any part of the respiratory system.
2. In the United States and worldwide, respiratory cancers are a leading cause of cancer-related deaths for both men and women.
3. According to the American Cancer Society, lung cancer in particular accounts for approximately 27% of all cancer deaths.
    - More deaths than prostate, breast, and colorectal cancers combined.
4. It is commonly known that smoking affects the risk of respiratory cancers, such as lung cancer.

Data Sets

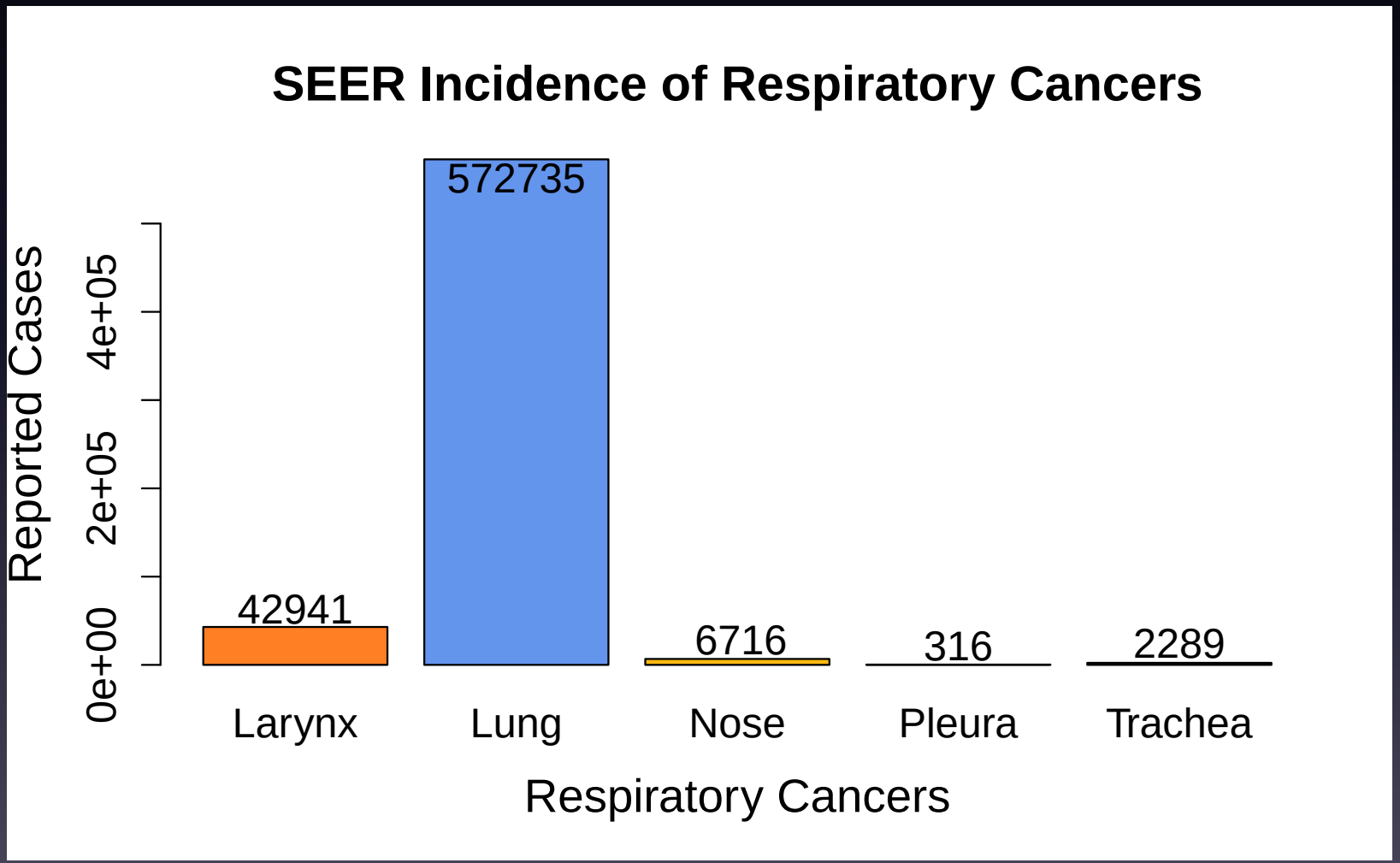
1. The Surveillance, Epidemiology, and End Results (SEER) data set is a compilation of cancer data on all incident cancer cases from 1973 to 2012 in certain regions of the United States.
    - Includes data on approximately 625,000 respiratory cancer cases.
    - Contains 131 variables.
    - Iowa is a SEER State.
2. The National Cancer Institute's Small Area Estimates for Cancer Risk Factors and Screening Behavior data set contains estimates on smoking prevalence by county for the years 1997 to 1999 and 2000 to 2003.
    - Includes data on current smokers and ever smokers for both sexes.

Respiratory System

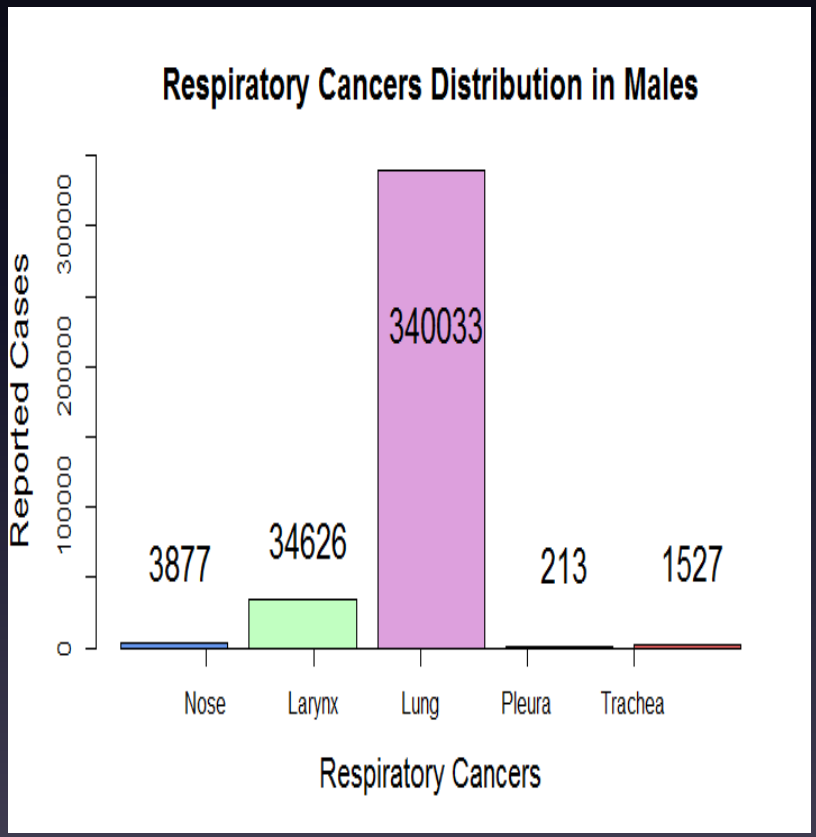


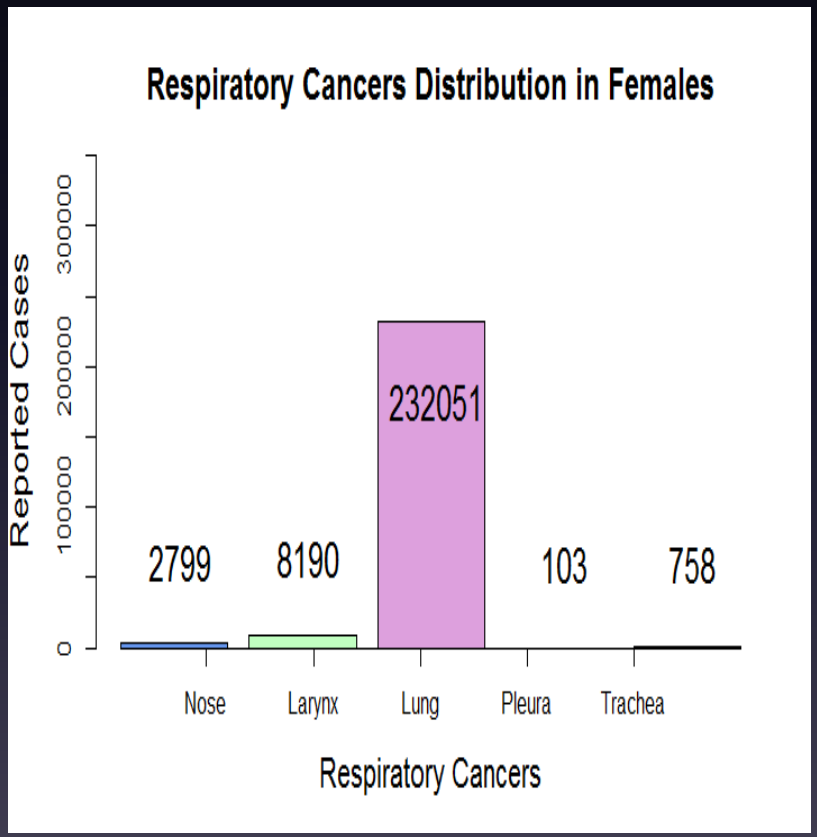
- Nose, Nasal Cavity, and Middle Ear
- Larynx
- Lung and Bronchus
- Trachea, Mediastinum, and Other Respiratory Organs
- Pleura

SEER Incidence of Respiratory Cancers

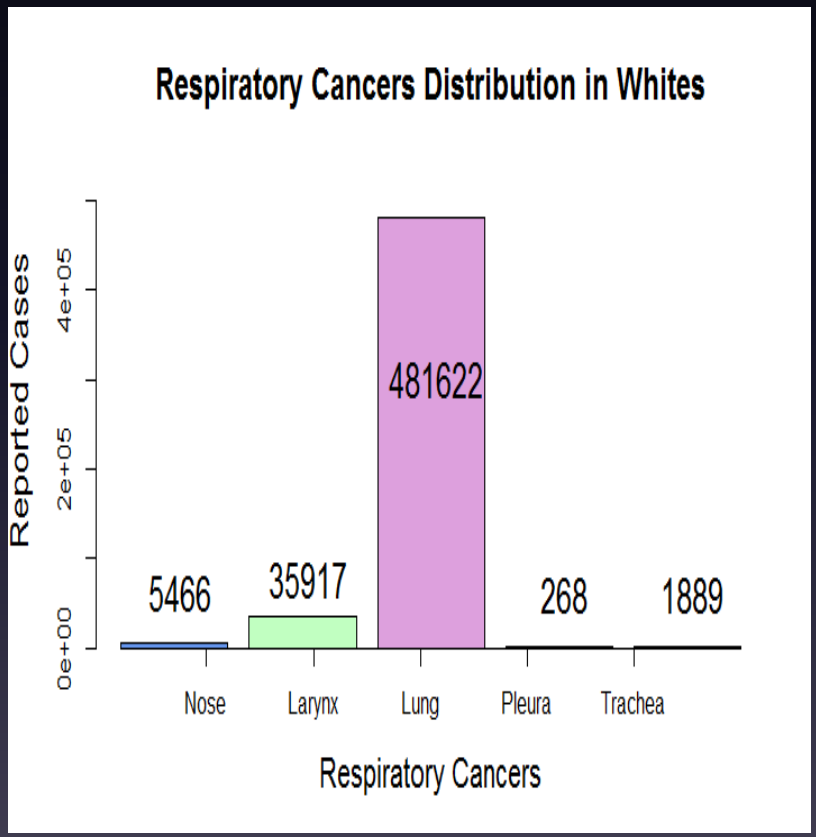


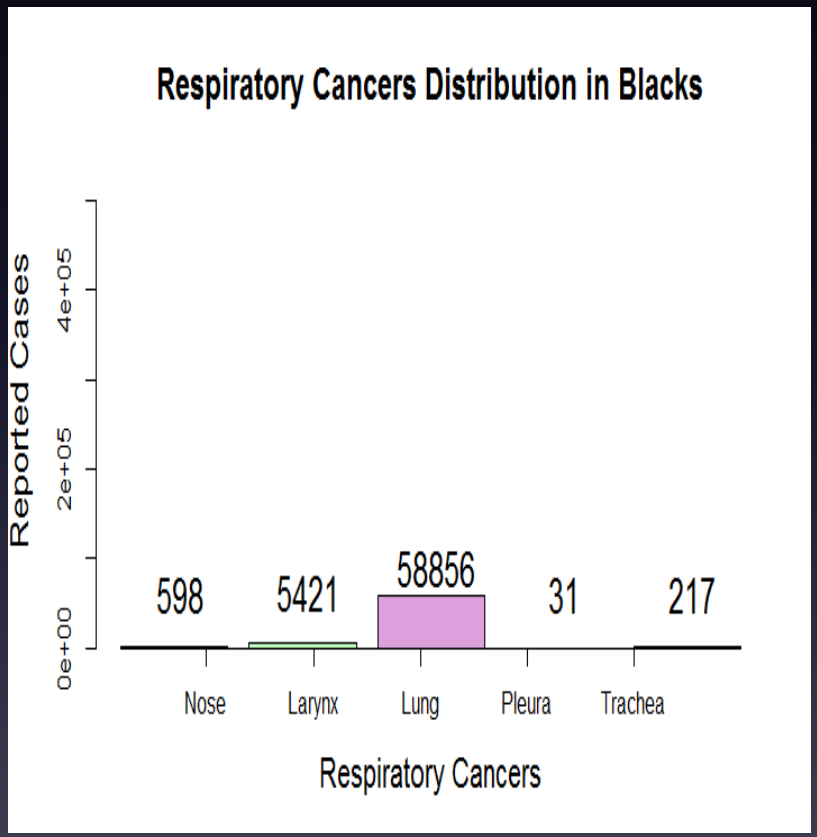
SEER Incidence of Respiratory Cancers by Sex





SEER Incidence of Respiratory Cancers by Race





Machine Learning

- Machine learning is a subfield of computer science that explores the construction and study of algorithms that can learn from and make predictions on data.
- We used classification models to predict cancer type given a variety of demographic variables.

Machine Learning

- We used the Naive Bayes, Neural Networks, Multinomial Logistic Regression and Support Vector Machines machine learning algorithms to predict cancer type given certain demographic variables.
- The variables used were age, race, hispanic origin, and region.
- To test the reliability of our models we divided our data in a training subset (80% of the data) and a testing subset (20% of the data).

Machine Learning (With Lung Cancer)

- Naive Bayes

	Nose	Larynx	Lung	Pleura	Trachea
Nose	1	0	0	0	1
Larynx	11	1	7	0	11
Lung	1311	7659	115475	71	372
Pleura	0	0	0	0	0
Trachea	12	1	12	0	55

Percent Correct: 92.4256%

- Neural Networks

	Nose	Larynx	Lung	Pleura	Trachea
Lung	1335	7661	115494	71	439

Percent Correct: 92.3952%

- Multinomial Logistic Regression

	Nose	Larynx	Lung	Pleura	Trachea
Nose	7	0	0	0	8
Larynx	0	0	0	0	0
Lung	1319	7661	115490	71	388
Pleura	0	0	0	0	0
Trachea	9	0	4	0	43

Percent Correct: 92.432%

- Support Vector Machines

	Nose	Larynx	Lung	Pleura	Trachea
Nose	0	0	0	0	0
Larynx	0	0	0	0	0
Lung	1335	7661	115494	71	439
Pleura	0	0	0	0	0
Trachea	0	0	0	0	0

Percent Correct: 92.3952%

Machine Learning (Without Lung Cancer)

- Naive Bayes

	Nose	Larynx	Pleura	Trachea
Nose	8	9	0	9
Larynx	1349	8364	79	337
Pleura	0	0	0	0
Trachea	80	38	6	140

Percent Correct: 81.6969%

- Neural Networks

	Nose	Larynx	Pleura	Trachea
Nose	41	32	4	31
Larynx	1322	8357	79	317
Trachea	74	22	2	138

Percent Correct: 81.92725%

- Multinomial Logistic Regression

	Nose	Larynx	Pleura	Trachea
Nose	14	2	0	15
Larynx	1360	8387	83	346
Pleura	0	0	0	0
Trachea	63	22	2	125

Percent Correct: 81.83127%

- Support Vector Machines

	Nose	Larynx	Pleura	Trachea
Nose	11	1	0	13
Larynx	1367	8389	83	352
Pleura	0	0	0	0
Trachea	59	21	2	121

Percent Correct: 81.78328%

Ecological Analysis with Poisson Regression

Poisson regression is a form of regression analysis that can be used to model count data.
Let x $\in \mathbb{R}^n$ is a vector of independent variables then:

$$log(E(Y|x)) \ = \ \alpha + \beta^{'} x$$

In our case Y represents the incidence of a certain cancer.

Ecological Tables

Table: Lung and Bronchus Cancer (Current Smoking 1997-1999)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4024	0.0978	-45.00	0.0000
40<	5.0801	0.0858	59.19	0.0000
Black	0.0429	0.0270	1.59	0.1121
Other	-0.7491	0.0476	-15.73	0.0000
Female	-0.0130	0.0657	-0.20	0.8426
Current1997_1999	0.0317	0.0020	15.77	0.0000
Female:Current1997_1999	-0.0112	0.0029	-3.80	0.0001

Ecological Tables

Table: Lung and Bronchus Cancer (Lifetime Smoking 1997-1999)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.1354	0.1193	-43.04	0.0000
40<	5.0676	0.0858	59.04	0.0000
Black	0.0854	0.0267	3.20	0.0014
Other	-0.7075	0.0477	-14.83	0.0000
Female	0.1650	0.1080	1.53	0.1267
Lifetime1997_1999	0.0286	0.0016	17.91	0.0000
Female:Lifetime1997_1999	-0.0048	0.0023	-2.10	0.0361

Ecological Tables

Table: Larynx Cancer (Current Smoking 1997-1999)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.0562	0.2969	-13.66	0.0000
40<	4.4987	0.2522	17.84	0.0000
Black	0.4022	0.0947	4.25	0.0000
Other	-0.8298	0.2034	-4.08	0.0000
Female	-0.7595	0.2927	-2.60	0.0095
Current1997_1999	0.0394	0.0067	5.85	0.0000
Female:Current1997_1999	-0.0207	0.0136	-1.52	0.1277

Ecological Tables

Table: Larynx Cancer (Lifetime Smoking 1997-1999)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.0940	0.3773	-13.50	0.0000
40<	4.4846	0.2522	17.78	0.0000
Black	0.4640	0.0927	5.01	0.0000
Other	-0.7682	0.2038	-3.77	0.0002
Female	-0.2454	0.4698	-0.52	0.6014
Lifetime1997_1999	0.0378	0.0054	7.03	0.0000
Female:Lifetime1997_1999	-0.0157	0.0104	-1.51	0.1316

Ecological Analysis Summary

- Blacks are at a higher risk than whites for most respiratory cancers while Others (American Indians/AK Natives and Asian/Pacific Islanders) are at lower risk when other variables are fixed. (Both Cases)
- Females are at a lower risk than males for most respiratory cancers when other variables are fixed.
- Smokers, both Current and Lifetime, are at a higher risk for respiratory cancers.

Survival Analysis

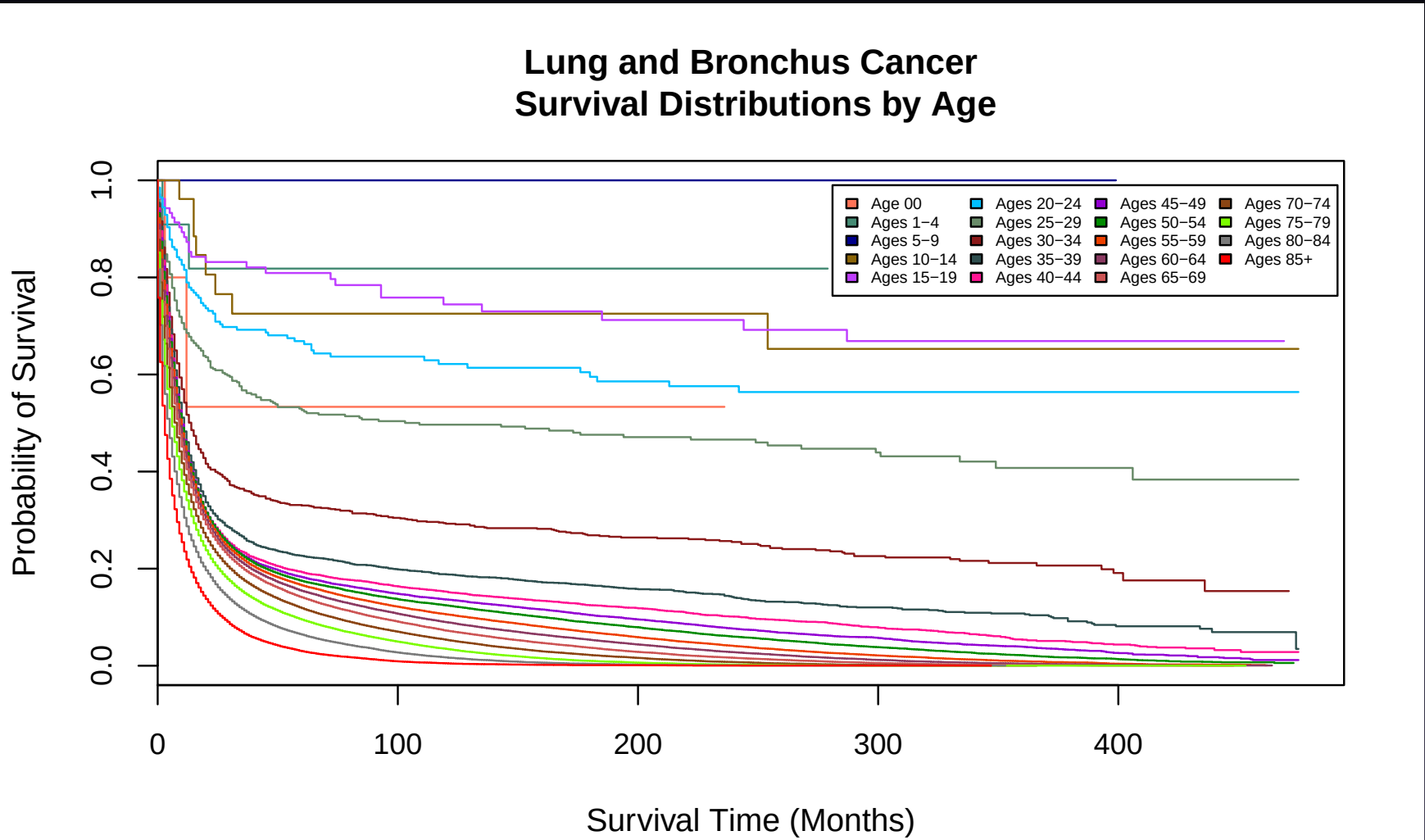
1. Performed survival analysis on SEER respiratory cancer data by using the Kaplan-Meier method.
2. Time-to-event: The time from the beginning of the observation period to death or withdrawal from the study.
3. Variables
   - Age at Diagnosis
   - Sex
   - Race
   - Tumor Grade

Kaplan-Meier Survival Analysis

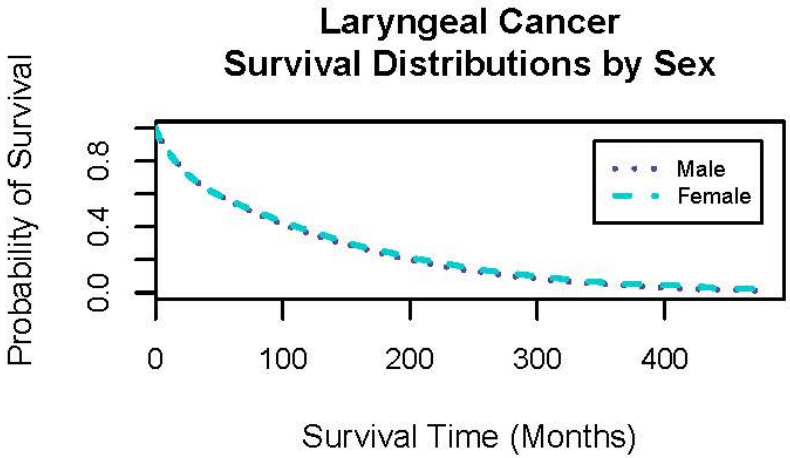
- Used to estimate a population survival curve from a sample.
- Allows computation of survival over time with censored data.
- Let S(t) be the survival probability at any given time for a member from a given population.

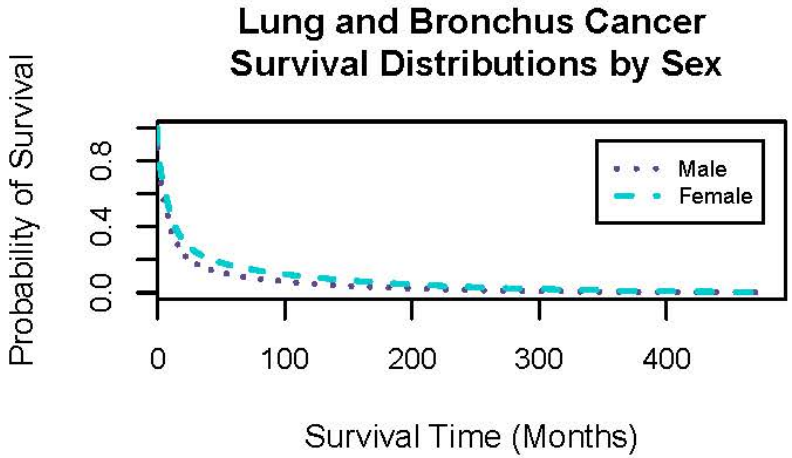
$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

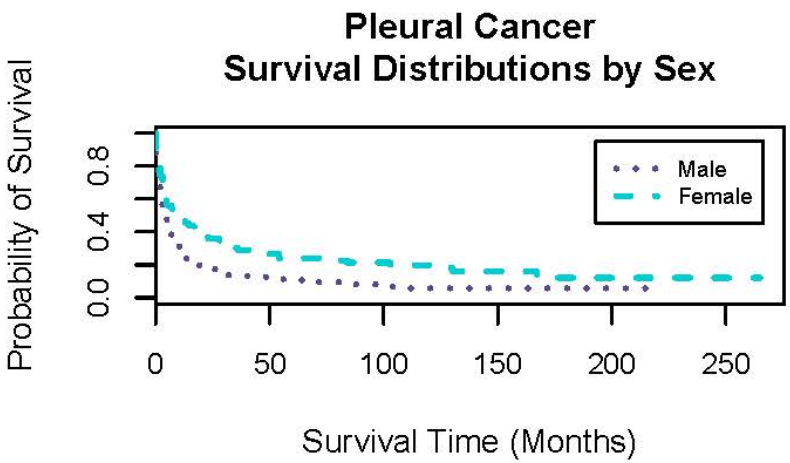
Age at Diagnosis Survival Analysis

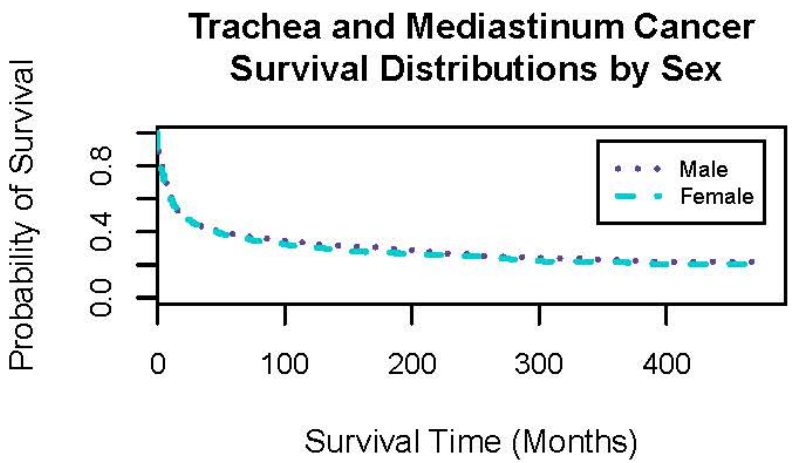


Sex Survival Analysis

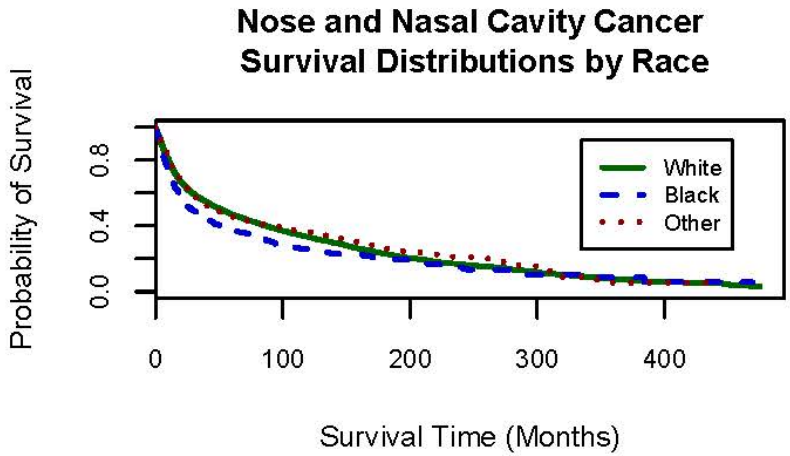


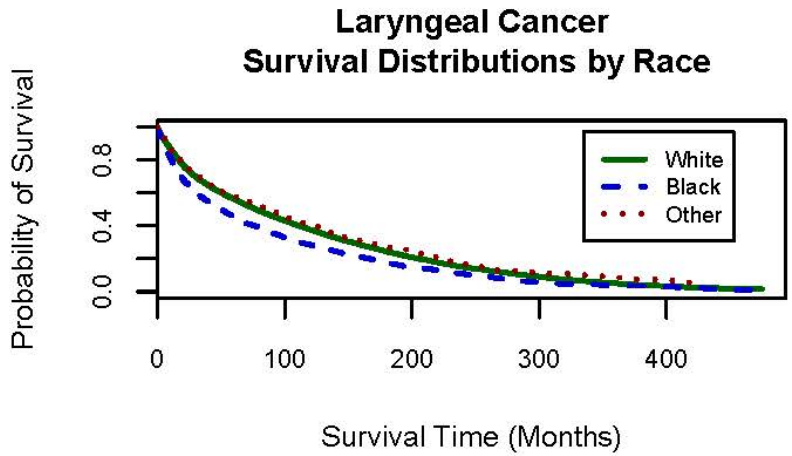


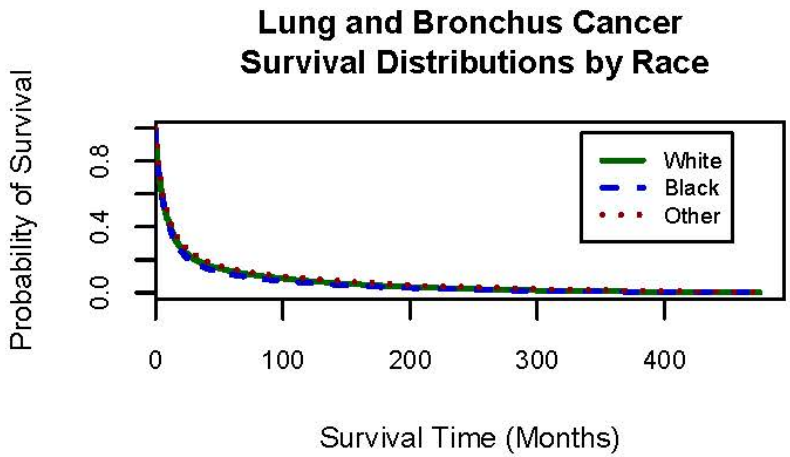


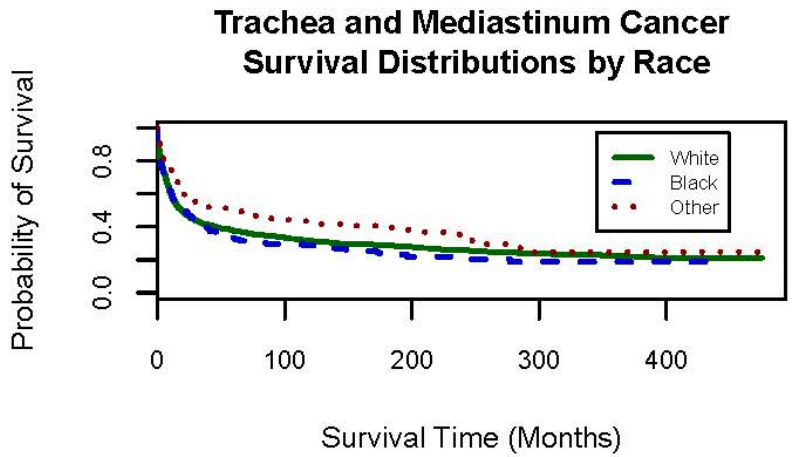


Race Survival Analysis

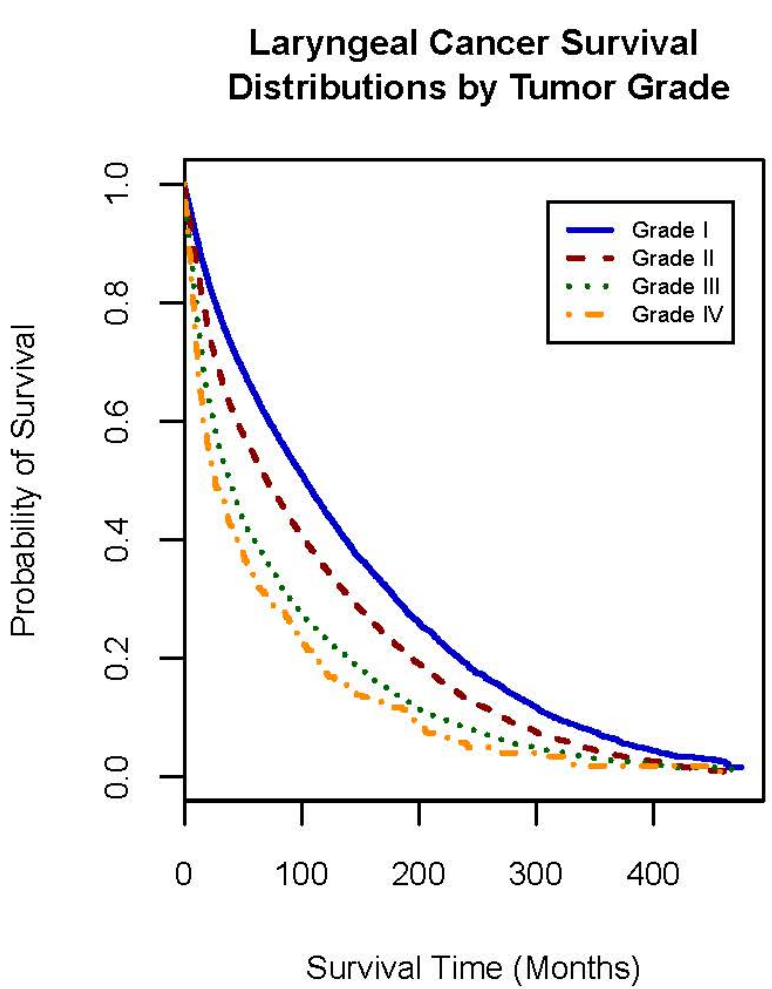


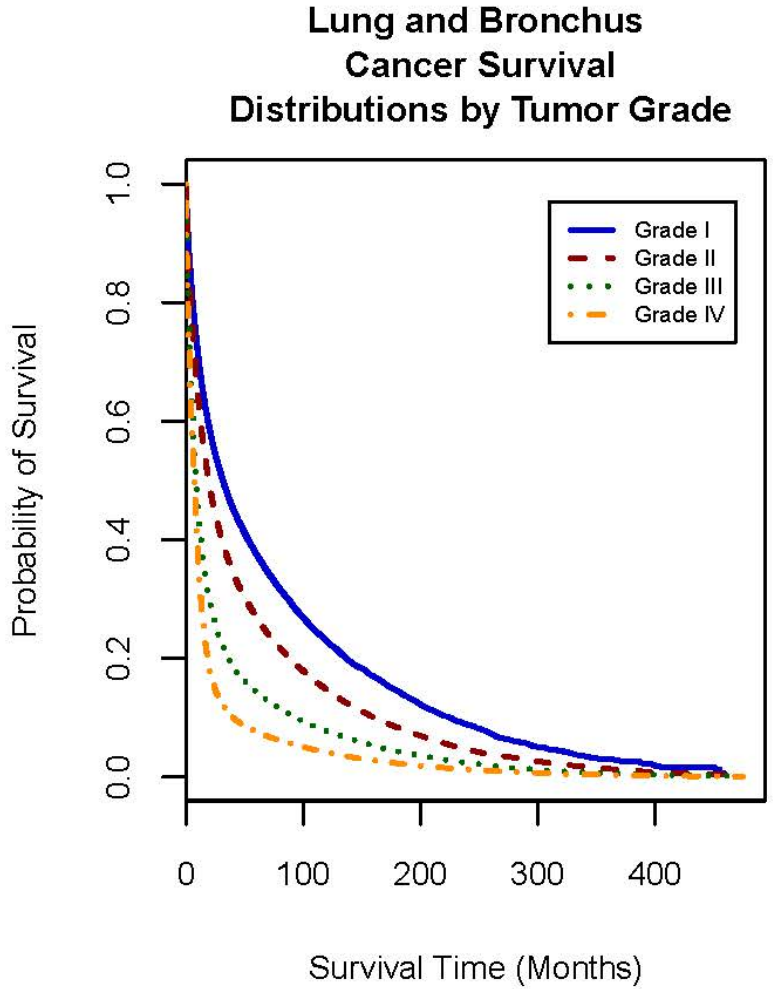






Tumor Grade Survival Analysis





Survival Analysis Summary

- Elderly people have the worst survival prognosis.
- Survival rates are lower for males than females for most respiratory cancers.
- For all respiratory cancers, American Indians/AK Natives and Asian/Pacific Islanders have the best survival prognosis, while Blacks have the worst survival prognosis.
- A higher tumor grade corresponds to a lower probability of survival.

Conclusions

- Current and Lifetime smoking are highly significant predictors for respiratory cancers.
- Age and Tumor Grade are good predictors for survival.
- Some respiratory cancer patients are able to live 30+ years with their cancer.
- Smoking data is a bit dated, would've been interesting to look at data from this decade.

References

- Respiratory System Organs [Photograph]. Retrieved from http://www.wickersham.us/anne/respiration.htm
- Small Area Estimates for Cancer Risk Factors Screening Behaviors. National Cancer Institute, DCCPS, Statistical Methodology Applications Branch, released May 2010 (sae.cancer.gov). Underlying data provided by Behavioral Risk Factor Surveillance System (http://www.cdc.gov/brfss/) and National Health Interview Survey (http://www.cdc.gov/nchs/nhis.htm).
- Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2012), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2015, based on the November 2014 submission.
- The Respiratory System. (2012, July 7). Retrieved July 15, 2015, from http://www.nhlbi.nih.gov/health/health-topics/topics/hlw/system
- What are the key statistics about lung cancer? (2015, March 4). Retrieved July 15, 2015, from http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-key-statistics
- ISIB Program sponsored by the National Heart Lung and Blood Institute (NHLBI) T15 HL097622

Citation for R and R Packages

- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.
- David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2014). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4. http://CRAN.R-project.org/package=e1071
- Ian Marschner (2014). glm2: Fitting Generalized Linear Models. R package version 1.1.2. http://CRAN.R-project.org/package=glm2
- Ingo Feinerer and Kurt Hornik (2015). tm: Text Mining Package. R package version 0.6-2. http://CRAN.R-project.org/package=tm
- Therneau T (2014). A Package for Survival Analysis in S. R package version 2.37-7. http://CRAN.R-project.org/package=survival
- Venables, W. N. Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0