Lung Cancer Survival (SEER 2004–2015): Data-Driven Insights on Stage and Outcomes

Code

Markuss Tomass Saule

2025-10-06

1 . Overview

This project uses data from the SEER cancer registry to explore patterns in lung and bronchus cancer survival.
The goal is to see how clinical and demographic factors—like **stage at diagnosis**, **tumor size**, and **lymph node involvement** relate to the chance of surviving two years after diagnosis.

I built a complete workflow: loading and cleaning data, defining survival outcomes, running basic statistical tests, and using a logistic regression model to identify the strongest predictors of survival.

Key Questions - How does stage at diagnosis relate to survival? - Are there survival differences by sex or race? - Which variables are most predictive of two-year survival?

Note: Two-year survival is used as the main outcome (because follow-up data beyond five years is limited in this subset).

2 . Renaming long column names into shorter ones:

Show

age_grp	sex	race	site_code	stage_sum	dx_year	surv_months	vital_status	tumor_size	nodes_pos	mets_dx	rad_recode	chemo_recode	no_surg_rsn	seq_r
55-59 years	Male	Black	674	Blank(s)	1981	0136	Dead	Blank(s)	Blank(s)	Blank(s)	None/Unknown	No/Unknown	Surgery performed	1st of or mo prima
60-64 years	Male	Black	341	Blank(s)	1985	0084	Dead	Blank(s)	Blank(s)	Blank(s)	None/Unknown	No/Unknown	Surgery performed	2nd o or mo prima
75-79 years	Female	White	164	Blank(s)	1977	0113	Dead	Blank(s)	Blank(s)	Blank(s)	None/Unknown	No/Unknown	Surgery performed	1st of or mo prima
70-74 years	Female	White	447	Blank(s)	1984	0025	Dead	Blank(s)	Blank(s)	Blank(s)	None/Unknown	No/Unknown	Surgery performed	2nd o or mo prima
45-49 years	Male	Other (American Indian/AK Native, Asian/Pacific Islander)	343	Blank(s)	1992	0147	Dead	Blank(s)	00	Blank(s)	Beam radiation	No/Unknown	Surgery performed	One prima only
60-64 years	Female	White	509	Blank(s)	1975	0162	Dead	Blank(s)	Blank(s)	Blank(s)	None/Unknown	No/Unknown	Surgery performed	One prima only

3 . Defining the Lung Cancer Sample

(I keep Lung & Bronchus cancers (site codes 340–349), years 2004–2015, and first primary cancers only.)

Show

```
[1] 116902
```

4 . Create Survival Outcomes (Primary = 2-Year)

SEER sometimes stores months as zero-padded text (e.g., “0060”). Convert to numeric first, then define outcomes:

2-year survival: surv_months >= 24

5-year survival: surv_months >= 60 (kept for reference)

Show

```
   No   Yes
87658 29244
```

Show

```
    No    Yes
101636  15266
```

Show

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0       2       8      25      25     227    3310
```

5 . Descriptive Statistics

We'll look at stage and simple survival proportions (2-years).

Show

Case Counts by Summary Stage

stage_sum	n
Blank(s)	138
Distant	69158
Localized	15826
Regional	23671
Unknown/unstaged	8109

Show

2-Year Survival Proportions by Stage

stage_sum	survived_24m	Percent
Blank(s)	No	0.1
Blank(s)	Yes	0.0
Distant	No	52.1
Distant	Yes	7.1
Localized	No	4.7
Localized	Yes	8.9
Regional	No	12.0
Regional	Yes	8.3
Unknown/unstaged	No	6.2
Unknown/unstaged	Yes	0.7

6 . Chi-Square Tests (2-Year Survival)

Show

Chi-Square Test: Stage vs 2-Year Survival

Test statistic	df	P value
24261	4	0 * * *

Show

Chi-Square Test: Sex vs 2-Year Survival

Test statistic	df	P value
847.2	1	2.926e-186 * * *

Show

Chi-Square Test: Race vs 2-Year Survival

Test statistic	df	P value
169	3	2.1e-36 * * *

7 . t-Test & ANOVA (Tumor Size)

SEER encodes some special values for tumor_size (e.g., 888, 990, 999 = missing/masked). Remove those and NAs. Then compare mean tumor size by 2-year survival (t-test) and across stages (ANOVA).

Show

Tumor Size by 2-Year Survival

survived_24m	n	mean_tumor	sd_tumor
No	58322	61	109
Yes	25736	58	149

Show

t-Test: Tumor Size by 2-Year Survival (continued below)

Test statistic	df	P value	Alternative hypothesis	mean in group No
2.452	38400	0.0142 *	two.sided	60.74

mean in group Yes
58.2

Show

ANOVA: Tumor Size Across Stages

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
stage_sum	4	46536918	11634229	798.7	0
Residuals	84053	1.224e+09	14567	NA	NA

8 . Logistic Regression (Primary Outcome = 2-Year Survival)

Predictors: stage, sex, race, tumor_size, nodes_pos. Clean nodes_pos and ensure all factors have ≥2 levels.

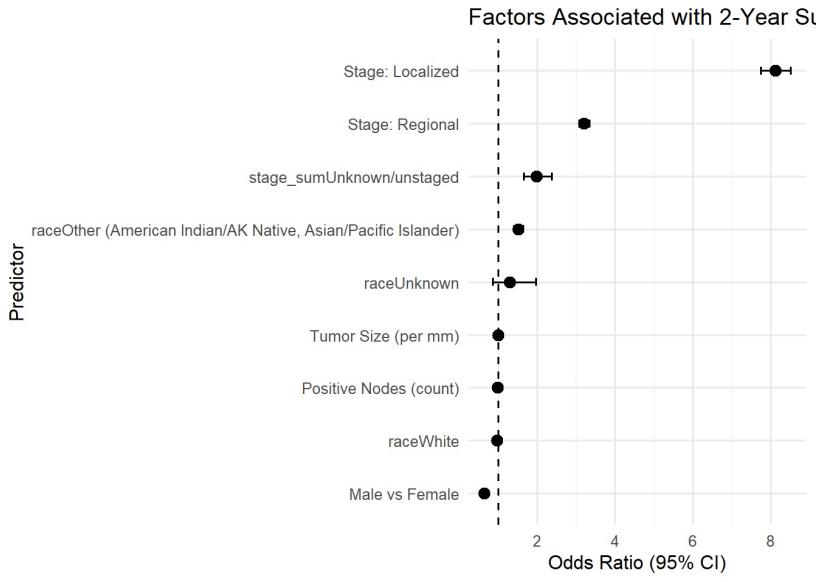
Show

Logistic Regression Results (2-Year Survival)

Predictor	Odds Ratio	Lower 95% CI	Upper 95% CI	p-value
(Intercept)	0.71	0.66	0.76	0.00
stage_sumLocalized	8.13	7.76	8.52	0.00
stage_sumRegional	3.20	3.07	3.34	0.00
stage_sumUnknown/unstaged	1.98	1.65	2.37	0.00
sexMale	0.64	0.62	0.67	0.00
raceOther (American Indian/AK Native, Asian/Pacific Islander)	1.52	1.40	1.64	0.00
raceUnknown	1.30	0.85	1.96	0.22
raceWhite	0.97	0.91	1.03	0.29
tumor_size	1.00	1.00	1.00	0.00
nodes_pos	0.98	0.98	0.98	0.00

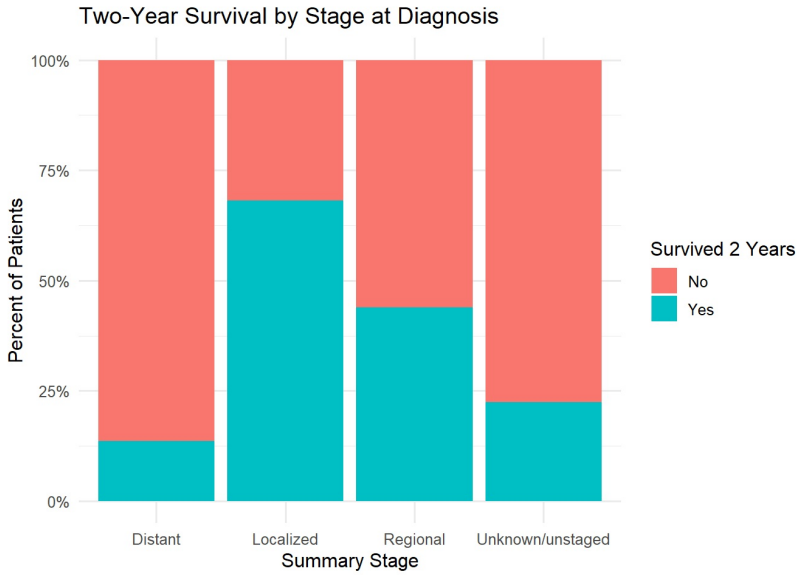
9 . Visualize Odds Ratios (Forest Plot)

Show



10 . Stage vs 2-Year Survival (Visual Summary)

Show



11 . Interpretation

The logistic regression shows that **stage at diagnosis** has the biggest influence on survival.
Patients diagnosed at a **localized** stage were roughly eight times more likely to survive two years than those with distant disease. Those diagnosed at the **regional** stage also showed better odds, though not as dramatically.

Tumor size and **number of positive lymph nodes** were both negatively associated with survival—larger tumors and more affected nodes meant worse outcomes.
There were smaller but noticeable differences by **sex** and **race**: men had slightly lower odds of survival, and patients identified as Asian, Pacific Islander, or American Indian tended to do somewhat better than White patients.

Overall, the findings reflect what's seen in real life: **early detection and limited spread are the strongest predictors of survival**.

12 . Limitations

- Some SEER fields include masked or missing data, especially for tumor size and lymph node counts.
- Treatment variables like chemotherapy and radiation don't specify timing or dosage, which limits how much we can infer about their impact.
- These analyses are descriptive, not causal—they identify associations, not direct effects.
- Two-year survival was used as the main measure because the five-year data in this subset had too few complete cases.

13 . Conclusion

This analysis shows how national registry data can reveal meaningful patterns in cancer outcomes.
Using SEER data from 2004–2015, I found that **earlier stage**, **smaller tumors**, and **fewer positive nodes** are consistently linked to better two-year survival in lung and bronchus cancer.

These results confirm the importance of early screening and detection. They also show how accessible statistical tools, such as R, can turn large public datasets into clear, actionable insights for healthcare research and decision-making.