Analysis on Lung Cancer Death Distributions and the Contributions of Tobacco Smoke and Cumulative Radiation for Hanford Site Workers

Kuan-Hung Yeh[1], Yen-Lin Lai[1], Bryan Lei[1], Muchuan Niu[2]

1. Department of Biostatistics, Fielding School of Public Health, UCLA, Los Angeles, California 90095, USA.
2. Department of Environmental Health Sciences, Fielding School of Public Health, UCLA, Los Angeles, California 90095, USA.

I. INTRODUCTION

Lung cancer is one of the most diagnosed cancer types and the leading contributor to cancer mortalities, causing an estimated 1.8 million deaths in 2020 worldwide [1]. Although a declining trend in annual lung cancer incidence and deaths has been observed in developed countries [2], an incidence of more than 230,000 cases and about 130,000 deaths were still estimated for 2021 in the United States [3]. Thus, reducing lung cancer cases and deaths remains a critical aspect for better health.

In the past decades, extensive research effort has been directed to delineate the causes of lung cancer to protect population health. According to the U.S. Centers for Disease Control and Prevention (CDC), cigarette smoking is the top risk factor for lung cancer development, which has been linked to 80% ~ 90% of lung cancer deaths [4]. Other risk factors such as inherited genetics, poor diet, air pollution, and occupational exposures have also been shown to act either independently or collectively in determining the lung cancer epidemiology [5]. Specifically, one of the earlier studies on lung cancer risks by Peterson et al. [6] in the early 90s investigated the association between occupational ionizing radiation and lung cancer deaths, adjusted for tobacco smoke using case – cohort design in the Hanford Site, WA. They found no association between cumulative radiation dose (in millisievert, mSv) and lung cancer deaths and the adjustment for tobacco smoke did not change this finding. The estimated excess relative risk (ERR) was - 0.7% per 10 mSv increment in radiation exposure; however, the upper limit of the 95% confidence interval (CI) was 9.9%, indicating a potential positive relationship between radiation risk and lung cancer deaths. Nevertheless, no data presentation on lung cancer deaths in relation to ages and different work types was provided, which is possibly problematic since a distinct cumulative amount of exposure could be applied to different age groups and kinds of work. Further, this study did not associate lung cancer deaths directly with tobacco smoke but rather treated it as a mediator due to their research focusing on radiation. Since cigarette smoking is the top risk factor for causing lung cancer, a separate analysis focusing on the association between tobacco smoke and lung cancer would be of great value.

Here, we applied the same dataset that Peterson et al. [6] was previously working with at the Hanford Site to conduct a complementary analysis on lung cancer death distributions across different age groups and work types, as well as comparing the contribution between tobacco smoke and radiation exposure to provide better data visualization. This report would be helpful for researchers to learn about identifying hidden information from a large dataset.

II. METHODS & MATERIALS

A. Data Set

The dataset *HFLCAA01_d1 Lung Cancer File* conducted in this study was downloaded from the Comprehensive Epidemiologic Data Resource (CEDR) of the U.S. Department of Energy [7]. This dataset contains 7077 observations on 514 white males worked at the Hanford Site from 1965 to 1985. A total of 35 variables were included in this dataset covering birth and death years, tobacco use information, radiation information, lung cancer deaths identification, etc.

B. Data Cleaning and Variable Selection

In this study, we used *SAS® Studio* from *SAS® OnDemand for Academics* to conduct all data management and analysis. We are aiming at displaying the lung cancer death distributions toward four major factors: age, work type, amount of tobacco smoke, and cumulative radiation exposure. To determine the total number of workers in this dataset, medical records with follow – ups under the same subject ID were deleted and only the last observation for each ID was kept for counts. After selecting the last medical record for each subject ID, we finally obtained 514 subjects from the dataset.

Lung cancer deaths were counted based on the *caseflag* variable from the dataset. *Caseflag* is defined as any worker in the cohort who died from lung cancer and had been monitored for radiation exposure for at least 3 years. Consequently, there is a possibility that a worker might have died from lung cancer but was not flagged as a case. Since *caseflag* was the only variable that provided information about lung cancer deaths in the cohort, we were not able to identify other lung cancer deaths if the worker did not meet the caseflag critiria.

In this project, age is generally defined as:

$$year\ of\ death\ (yrbirth)\ -\ year\ of\ birth\ (yrdeath)$$

Since lung cancer deaths identification in the original study ended by the year of 1980, we used the year of 1980 to determine the patients' final age for this study if the patients remained alive after the study. For *yrdeath*, "99" assumes the worker was still alive at the year of 1980. Workers with no medical record were logged using "98". Ages were stratified into 5 groups according to the data density (i.e., < 60 years, 60 – 65 years, 65 – 70 years, 70 – 75 years and 75 – 80 years) for a better visualization of lung cancer deaths at different ages.

Work types were identified from the *title for job held longest (titlejob)* from the dataset, defined by the job that the worker was assigned to and worked the longest at the Hanford Site. We extracted the first single digit code from this variable to located work types. A total of four work types were identified, including white collar, craftsman, nuclear, and service.

Tobacco use information was recorded upon the medical exams taken at 1 – or 2 – year intervals. Data with values of "97" and "98" suggested that the worker did smoke rarely but the amount could not be determined, and no medical exam had been taken, respectively. We classified the workers recorded in "99" as non – smokers, as "99" suggested that the worker either did not use tobacco, or the year of beginning tobacco use could not be determined. The amount of tobacco use was determined by total pack years [8] smoked till 1980. The amount of tobacco use in pack years for the workers was calculated by:

$$number\ of\ packs\ smoked\ per\ day\ (amtobuse) \\ \times\ number\ of\ smoking\ years$$

The *number of smoking years* is defined by the year that the worker quitted smoking subtracted by the year he began smoking. If the years of both quit – and begin – smoking were 0, we regarded them as non – smokers. If the year of quit – smoking was missing, we treated these workers as chronic smokers; whereas, if only the year of begin – smoking was missing, we treated them as smokers by assuming that they started tobacco use at the age of 18. The end year for tobacco smoke was also confined to 1980 to match the end year used for age calculation. Pack years smoked were stratified into 5 strata, with 0 pack years smoked (non – smoker), 0 – 20 pack years, 20 – 40 pack years, and 40 – 60 pack years smoked.

Occupational radiation data from the files for the Hanford Site workers were used to retrieve radiation related information. *Cumulative radiation exposure (cumraexp)* provided the annual whole – body dose equivalent to external radiation the subject received in the unit of millisieverts (mSv). Cumulative radiation exposure was divided into 4 strata: 0 – 19.9 mSv, 20 – 49.9 mSv. 50 – 150 mSv, and > 150 mSv [6].

C. Statistical Methods

Lung cancer death rate is defined as the proportion of lung cancer deaths within each stratum, calculated by dividing the respective counts of lung cancer deaths by the total counts for each stratum and was expressed using percentages. We calculated the relative risks (RR) of tobacco smoke and cumulative radiation exposure, using non – smoker (pack year = 0) and the radiation amount of 0 – 19.9 mSv as reference levels, respectively. We calculated relative risks (RR) using the ratio of the reference's risk to each stratum's risk through two-by-two tables, showing below.

	Lung Cancer Death	Not Lung Cancer Death
Unexposed	A	B
Exposed	C	D

$$Relative\ Rsik\ (RR) = \ (A/A + B)/(C/C + D)$$

Lastly, we used the Wald confidence limits and applied the 95% confidence interval (CI) to the RR results by SAS default.

We further applied the Logistic Regression Model to tobacco smoke and cumulative radiation exposure to test their respective associations with lung cancer death probabilities. In our model, the outcome Y represents the probability for lung cancer death (0 = alive or not died of lung cancer, and 1 = died of lung cancer). The coefficient β_1 is the expected change in log odds of lung cancer death per unit change in *pack_year*; and β_2 is the expected change in log odds of lung cancer death per unit change in *cumraexp*. An error term ε_i was incorporated into our model as a residual variable that accounts for a lack of perfect goodness of fit.

The equations for the Logistic Regression models in this project are shown as below:

$$Y = \beta_0 + \beta_1 * pack\_year + \varepsilon_i$$
$$Y = \beta_0 + \beta_2 * cumraexp + \varepsilon_i$$

where:

- Y = the probability for the lung cancer death
- β_0= intercept of the model
- β_1 = coefficient of *pack_year*
- β_2 = coefficient of *cumraexp*
- ε_i = error term

The log probabilities of the total amount of tobacco smoke and the cumulative radiation exposure are reported as log – average percent changes with 95% CI according to the model results.

III. RESULTS

A. *Lung cancer death distributions across different ages and work types*

Deaths specifically related to lung cancer were identified and the respective lung cancer death counts were shown for each age group and work type (Table 1). “Others” in Table 1 indicates that the subject was either alive at the end – year of 1980 or had died due to other causes. Among all age groups, 65 – 70 years had the most lung cancer deaths with a number count of 27, accounting for 16.1% of total subjects in this group. Surprisingly, workers less than 60 years of age had 26 lung cancer deaths (second to 65 – 70 years), as well as the highest lung cancer death rate of 32.9%. The group of 75 – 80 years had the least number of lung cancer deaths of 3, with the lowest death rate of 4.2%. Overall, a decreasing pattern in lung cancer death rates was observed with increasing years of age.

Table 1. Lung cancer deaths and proportions for different age groups and work types.

	Lung cancer deaths (%)	Others[a] (%)	Total
Number of subjects	92	422	514
Age groups (yrs*)			
< 60	26 (32.9)	53 (67.1)	95
60 - 65	21 (22.1)	74 (77.9)	168
65 - 70	27 (16.1)	141 (83.9)	101
70 - 75	15 (14.9)	86 (85.1)	71
75 - 80	3 (4.2)	68 (95.8)	79
Work types			
White collar	26 (15.1)	146 (84.9)	172
Craftsmen	37 (20.0)	148 (80.0)	185
Nuclear	6 (11.8)	45 (88.2)	51
Service	23 (21.7)	83 (78.3)	106

*: Years of age. %: percent proportion of the stratum total was displayed, defined as lung cancer death rate. [a]: Others provide the number of patients that were either alive in 1980 or had died due to other causes.

The number of Lung cancer deaths for craftsmen was the most across all work types, with 37 deaths, accounting for 20% of the total workers for this work type (Table 1). Service work showed the highest lung cancer death rate of 21.7% with 23 deaths. Interestingly, the work type of nuclear, which is commonly considered the work that involves more radiation exposures than other work types, had the least number of lung cancer deaths, as well as the least rate of death due to lung cancer.

B. *Lung cancer deaths for tobacco smoke and radiation*

Using a similar method, we further investigated lung cancer deaths and the respective death rates to the stratified amount of tobacco smoke and cumulative radiation exposures (Table 2). Results showed that lung cancer deaths occurred the most for the group of non-smokers, with a total of 63 workers. Nevertheless, the death rate for non-smokers was the lowest, with only 15.8% of the total compared to other exposure strata. In comparison, workers with more than 60 pack years of smoking had the highest lung cancer death rate of 40% (Table 2)

For cumulative radiation exposures, the most lung cancer deaths were found in the group of 0 – 19.9 mSv, with 31 workers died and a death rate of 17.9%. Exposure level of 50.0 – 149.9 mSv had the highest lung cancer death rate of 22.1%; while exposure more than 150 mSv had the lowest death rate of 14.7%.

Table 2. Lung cancer deaths and proportions for tobacco smoke and radiation.

	Lung cancer deaths (%)	Others[a] (%)	Total	RR
Number of subjects	92	422	514	
Tobacco smoke (pack years)				
Non – smokers	63 (15.8)	335 (84.2)	398	-
0 < - 20	1 (25.0)	3 (75.0)	4	1.8 (0.2, 17)
20 < - 40	7 (23.3)	23 (76.7)	30	1.6 (0.7, 3.5)
40 < - 60	13 (21.0)	49 (79.0)	62	1.3 (0.8, 2.3)
> 60	8 (40.0)	12 (60.0)	20	3.3* (1.4, 7.7)
Cumulative radiation (mSv)				
0 - 19.9	31 (17.9)	142 (82.1)	173	-
20.0 - 49.9	26 (17.2)	125 (82.8)	151	1.0 (0.7, 1.3)
50.0 - 149.9	21 (22.1)	74 (77.9)	95	1.2 (0.8, 1.7)
≥ 150	14 (14.7)	81 (85.3)	95	0.9 (0.6, 1.4)

%: percent proportion of the stratum total was displayed, defined as lung cancer death rate. [a]: Others provide the number of patients that were either alive in 1980 or had died due to other causes. RR: relative risk, calculated using non – smoker and 0 - 19.9 mSv as references. *: statistical significance.

Relative risks (RR) for tobacco smoke and cumulative radiation exposure were calculated using non – smokers and 0 – 19.9 mSv as reference levels. For tobacco smoke, all RRs were larger than 1.0, suggesting a higher lung cancer death risk compared to non – smokers. Significant increase in RR was observed for > 60 pack years of smoke. In comparison, cumulative radiation exposure showed no consistent pattern of increased risks for lung cancer and no significant RR was observed.

C. *Lung cancer death probabilities associated with tobacco smoke and radiation*

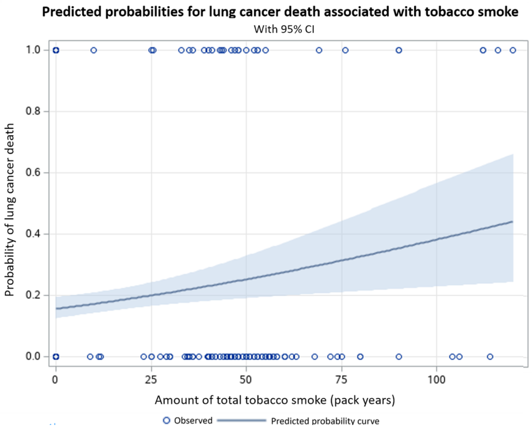

Figure 1. Predicted probabilities for lung cancer death associated with tobacco use. Observed shows either the worker was identified as a lung cancer death “1” or no “0”. Blue line gives the predicted log – probability for the corresponding pack years of tobacco smoke. Shaded region shows the upper and lower boundary for 95% CI.

Logistic Regression Model was applied to predict lung cancer death probabilities in relation to tobacco smoke (in pack years) and cumulative radiation exposure (mSv), with the shaded area representing 95% CI for the log – probability of lung cancer death. As shown in Figure 1, a consistently increasing log – probability of lung cancer death was displayed with increasing pack years smoked. The estimate for pack years of tobacco smoke was 0.012, with p = 0.005 (Table 3), suggesting a significantly elevated probability for lung cancer death with more exposure to tobacco smoke.

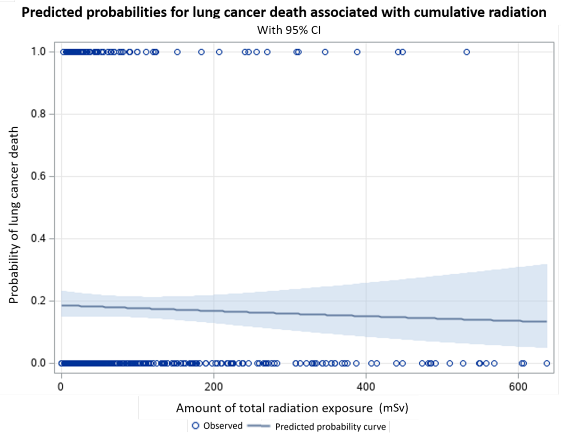

Figure 2. Predicted probabilities for lung cancer death associated with cumulative radiation. Observed shows either the worker was identified as a lung cancer death “1” or no “0”. Blue line gives the predicted log – probability for the corresponding pack years of tobacco smoke. Shaded region shows the upper and lower boundary for 95% CI.

In Figure 2, the log – probability of lung cancer deaths generally decreased with more cumulative radiation exposures experienced. The estimate for cumulative radiation was -0.0006, with p = 0.53 (Table 3). However, the upper boundary of the 95% CI for log lung cancer death probability started to rise when cumulative radiation exposure exceeded 100 mSv. Thus, we could not rule out the possibility that increased radiation exposure might still be associated with elevated lung cancer death probability.

Table 3: Estimates and test statistics for tobacco smoke and cumulative radiation[a].

Variables	Estimate (SE[b])	P value
Tobacco smoke (pack years)	0.012 (0.004)	0.005*
Cumulative radiation (mSv)	-0.0006 (0.001)	0.53

[a]: Logistic Regression Model was applied to generate estimates for the variables. [b]: Standard error.
*: Statistical significance.

IV. DICUSSION

In this project, we conducted several complementary analyses on different variables, including age, work type, tobacco smoke, and cumulative radiation to investigate their contributions to lung cancer deaths. We also tried to utilize the Logistic Regression Model to present the lung cancer death probabilities associated with tobacco smoke and cumulative radiation exposure. We found that lung cancer death rate decreased with increasing ages, and the nuclear workers had the lowest rate of death. Tobacco smoke showed consistently increased RRs with one significance found for workers who had more than 60 pack years smoked. Significantly increased log – probability (p = 0.005) of lung cancer death was observed in relation to increased pack years of tobacco smoke. In contrast, cumulative radiation exposure generally showed no increased risks for lung cancer deaths and the regression analysis did not provide an identifiable trend in lung cancer death probability either.

The result of this project on lung cancer deaths and ages was not in correspondence with current lung cancer statistics that suggested higher incidence and lower survival for elderlies [9]. The reason behind this dichotomy between our study and existing findings relies on the fact that the data input for the Hanford Site workers logged lung cancer deaths only when the patient died from lung cancer and had been monitored for radiation exposure for at least 3 years during the study. In addition, elder populations might be suffering from many other diseases that could lead to mortalities as they grow older. Therefore, we postulate that less lung cancer deaths for older age groups could be attributed to them being allocated to missing data and other causes of deaths. As our sample size is relatively small by nature, the result here might not be generalized for larger population.

Logistic Regression Model analysis suggested that the log – probability of lung cancer deaths is significantly increased with more tobacco smoke (in pack years) (Figure 1). This finding is in accordance with previous knowledge that tobacco smoke is the key factor that increases lung cancer risks and subsequent deaths [4]. However, in our model, pack years of tobacco smoke was treated as the only variable associated with lung cancer death probability. Nonetheless, variables such as age and other types of diseases factors might work either independently or collectively to affect lung cancer deaths [4]. As a result, future studies may consider adding more components to the model and accounting for a wider variety of confounding variables to better consolidate this result.

The data quality of the Hanford Site workers file was quite poor, with many incomplete data due to loss of follow up or incorrect recordings. The data structure was also difficult for us to select a worker's most up to date record. Therefore, we selected the last available data for each worker for consistency issue, since each worker had at least one medical record according to the author of the original study [6]. Thus, the data we selected may not accurately represent the most up to date data for each participant. In addition, key variables in our study such as smoking information were not recorded in a consistent way during the research period due to response errors, which is typical among self – reported data.

Besides, the labels of the missing data in this study were not defined well and the descriptions were ambiguous. For example, for the *amount of tobacco use* (*amtobuse*), there were three coding options for missing and undetermined data. The number of "97" means that "the worker did smoke rarely but the amount could not be determined"; "98" suggested "no medical exam taken yet"; and "99" reflected that "either the worker did not smoke, or the amount could not be determined". Therefore, it is extremely difficult to determine the amount of tobacco smoke when *amtobuse* equals "97" and "98". Thus, we denoted "97" and "98" as missing data and "99" as non-smokers. The same way of data selection was applied to analyze the cumulative radiation exposure.

V. CONCLUSIONS

In this project, we analyzed the lung cancer death counts and death rates for four variables, including age, work type, tobacco smoke, and cumulative radiation exposure. We found that lung cancer death rates decreased as the workers grew older. Results from Logistic Regression Model applied indicated that the log – probability of lung cancer deaths was significantly elevated with more pack years of smoking. Cumulative radiation exposure in different mSv ranges did not dramatically fluctuate the death rate of lung cancer and no trend in the log – probability of death was observed.

REFERENCES

1. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA: A Cancer Journal for Clinicians, 2021. **71**(3): p. 209-249.

2. Youlden, D.R., S.M. Cramb, and P.D. Baade, *The International Epidemiology of Lung Cancer: Geographical Distribution and Secular Trends.* Journal of Thoracic Oncology, 2008. **3**(8): p. 819-831.

3. American Cancer Association. *Key Statistics for Lung Cancer. 2021 [cited 2021 10/31];* Avilable from:https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

4. Centers for Disease Control and Prevention (CDC). *What Are the Risk Factors for Lung Cancer?* 2021 [cited 2021 10/31]; Available from: https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm.

5. Malhotra, J., et al., *Risk factors for lung cancer worldwide.* European Respiratory Journal, 2016. **48**(3): p. 889-902.

6. Petersen, G.R., et al., *A case-cohort study of lung cancer, ionizing radiation, and tobacco smoking among males at the Hanford Site.* Health Phys, 1990. **58**(1): p. 3-11.

7. CEDR, U.S.D.o.E. [cited 2021 11/7]; Available

from: https://oriseapps.orau.gov/cedr/DataFile.aspx?DataSet=HFLCAA01&DFile=HFLCAA01_1.

8. National Cancer Institute, NIH. [cited 2021 11/7]; Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pack-year.

9. UK, C.R. *Lung cancer mortality statistics*. [cited 2021 11/10]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/mortality#heading-One

Appendix

I. SAS Code

Table 1:

```
proc freq data = Age;
table agegroup*caseflag/nocum nopercent;
run;

proc freq data = HFLCAA_6 order = data;
table job*caseflag/nocum ;
run;
```

Table 2:

```
*/pack_year*smoking;
proc freq data = HFLCAA_6 order =
internal;
table pack_year_range*caseflag/nocum
nocol norow ;
run;

*/Relative Risk1;
data smoking_casecontrol;
  input Case $ Pack $ Count;
  datalines;
    Yes No 31
    No  No 142
    Yes one 14
    No  one 81
;
proc freq order=data
data=smoking_casecontrol;
  weight Count;
  tables Case*Pack / riskdiff relrisk;
run;

*/radiation;
proc freq data = HFLCAA_6 order = internal;
table cumuradiation*caseflag/nocum;
run;
```

Graph 1 & Table 3 first row:

```
proc logistic data=HFLCAA_7 plots=effect;
  model caseflag(event="the worker was selected
as a lung cancer case") = pack_year;
  output out=LogiOut_smoking
predicted=PredProb;
  effectplot fit /obs ;
run;
```

Graph 2 &Table 3 second row:

```
proc logistic data=HFLCAA_8 plots=effect;
  model caseflag(event="the worker was selected
as a lung cancer case") = cumraexp;
  output out=LogiOut_radiation
predicted=PredProb;
  effectplot fit /obs;
  run;
```