REANALYSIS OF THE NATIONAL CANCER INSTITUTE'S ACRYLONITRILE COHORT STUDY BY IMPUTATION OF MISSING SMOKING INFORMATION

by

Michael A. Cunningham

B.S., University of Pittsburgh, 1992

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Michael A. Cunningham

It was defended on

April 25, 2006

and approved by

Gary M. Marsh, Ph.D.
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Ada O. Youk, Ph.D.
Research Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Evelyn O. Talbott, Dr.P.H.
Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

REANALYSIS OF THE NATIONAL CANCER INSTITUTE'S ACRYLONITRILE COHORT STUDY BY IMPUTATION OF MISSING SMOKING INFORMATION

Michael A. Cunningham, M.S.

University of Pittsburgh, 2006

A cohort study of workers exposed to the chemical acrylonitrile (AN) was carried-out in the late 1980s by the National Cancer Institute (NCI) to determine if there were any excess cancer risks associated with workplace exposures to AN. The results of the study did not show any overwhelming evidence that AN exposure was related to increased cancer risk, but did yield several results worth noting. Firstly, the authors reported an overall lung cancer risk of 3.6 for ever-smokers versus never-smokers, which appeared to be much too low. Secondly, there was a slight increase in the lung cancer relative risk due to exposure in the upper quintile of cumulative AN exposure. Lastly, there was a large proportion of missing smoking information for the employees selected in the sample.

Because results of occupational cohort studies such as the NCI's are used as the basis for determining health risks associated with workplace exposures and because acrylonitrile is widely used in the manufacturing of plastics, it is very important from a public health perspective to eliminate any possible sources of confounding or bias. The goal of this reanalysis is to address the issues of missing smoking information and the low overall lung cancer relative risk in ever-smokers to determine if the slight excess in the highest AN exposure category appears to be valid. This was accomplished using imputation, a procedure that predicts a smoking status for

the missings based on complete observations. The NCI analyses were then repeated with the imputed data to see if there were any differences in the overall smoking lung cancer RR or the lung cancer RR in the upper quintile of AN exposure.

The overall lung cancer RR due smoking could not be increased dramatically using the weighting schemes in this paper. Also, the lung cancer RRs in the upper quintile of AN exposure were not much lower than those in the original NCI study, so their analysis with the missing smoking information does not appear to have been biased. However, the smoking adjusted lung cancer RRs for cumulative AN exposure using the imputed data have a much flatter exposure-response trend than the NCI analysis, which, when combined with the only slightly elevated RR in the upper exposure group, could be used as evidence against an increased lung cancer risk due to high AN exposure.

TABLE OF CONTENTS

1. INTRODUCTION 1
   1.1. Original NCI Study Cohort 3
   1.2. Case-Cohort Design Description 4
   1.3. NCI Case-Cohort Design 6
   1.4. NCI Results 7
2. STATEMENT OF THE PROBLEM 10
3. METHODS 10
   3.1. Reproduction of NCI Results 10
      3.1.1. NCI sub-cohort data files 10
      3.1.2. NCI sub-cohort demographic data 11
      3.1.3. NCI sub-cohort distribution of lung cancer cases 12
      3.1.4. Proportional hazards models to obtain relative risks 14
      3.1.5. Reanalysis comparison to original NCI results 15
   3.2. Sub-cohort Smoking Status Variables 16
      3.2.1. Original smoking status variables 16
      3.2.1. Agreement between smoking status variables 17
      3.2.2. Smoking status variables created by NCI 19
      3.2.3. Analysis of discordant smoking status variables 21
   3.3. Smoking Information for Sub-Cohort 22
      3.3.1. Demographic information 22
      3.3.2. Smoking and AN exposure 24
      3.3.3. Smoking and lung cancer status 26
      3.3.4. Smoking status by plant 29
      3.3.5. Smoking by race and gender 30

3.3.6. Smoking and year of birth....31
4. REANALYSIS....33
4.1. Objectives....33
4.2. Reassignment of Smoking Status Variable....34
4.2.1. Employees whose smoking histories did not agree....34
4.2.2. Reallocation of unknowns....37
4.2.3. Logistic regression model to predict smoking status for unknowns....39
4.2.4. Proposed reallocation of unknowns....41
4.3. Imputations....42
4.3.1. Description....42
4.3.2. Unweighted imputation description....44
4.3.3. Stata imputation procedures....45
4.3.4. Unweighted imputation for 3 allocations....46
4.3.5. Weighted imputation description....49
4.3.6. Weighted imputation for 3 allocations....51
5. RESULTS....54
5.1. Overall Lung Cancer RR Due to Smoking....54
5.1.1. Analysis of 3 reallocations – including missing data....54
5.1.2. Analysis of 3 reallocations – imputations....55
5.2. Lung Cancer RR in Upper Quintile of AN Exposure....60
5.2.1. Average cell counts for imputed data by reallocation and weight....60
5.2.2. Proportional hazards models for imputed datasets....62
5.2.3. Comparison of RRs from NCI study to imputed datasets....69
5.2.4. Comparison of full cohort adjusted RRs from NCI study to imputed datasets....73
6. DISCUSSION....75
7. CONCLUSIONS....79
APPENDIX A: Example SAS program used in analysis....81
BIBLIOGRAPHY....89

LIST OF TABLES

Table 1: Lung cancer relative risk by quintile of AN exposure – NCI study results.....8
Table 2: Frequency distribution of sub-cohort employees by demographic and exposure variables.....11
Table 3: Frequency distribution of sub-cohort lung cancer cases by demographic and exposure variables.....13
Table 4: Exposed Lung Cancer Cases by Quintile of Cumulative AN Exposure.....13
Table 5: Lung cancer relative risk by quintile of AN exposure – NCI study results and reanalysis.....15
Table 6: Frequency distribution of NCI smoking status variables.....17
Table 7: Cross tabulation of QSMOKER and MSMOKER variables.....18
Table 8: Description of ever-, never-smoker smoking status variables.....20
Table 9: NCI assignment of a smoking status (SMKSTAT) based on QSMOKER and MSMOKER variables.....21
Table 10: Distribution of sub-cohort employees with discordant smoking histories by AN exposure.....22
Table 11: Frequency distribution of sub-cohort employees with missing smoking information by demographic and exposure variables.....23
Table 12: Prevalence of smoking for sub-cohort employees.....24
Table 13: Prevalence of smoking by AN exposure status.....24
Table 14: Distribution of employees with missing smoking information by AN exposure status.....25
Table 15: Smoking history by quintile of AN exposure for employees with known smoking history.....25

Table 16: Smoking history by quintile of AN exposure for all employees in sub-cohort 26
Table 17: Smoking history of cases and controls 26
Table 18: Smoking history of cases and controls, complete smoking histories only 27
Table 19: Smoking history by AN exposure status - lung cancer cases 28
Table 20: Smoking history by AN exposure group - lung cancer cases 28
Table 21: Smoking history by plant for entire sub-cohort 29
Table 22: Smoking history by plant for lung cancer cases 30
Table 23: Smoking history by race for entire sub-cohort 30
Table 24: Smoking history by gender for entire sub-cohort 31
Table 25: Smoking history by year of birth for entire sub-cohort 32
Table 26: Smoking history by year of birth for sub-cohort employees with known smoking histories 32
Table 27: Reallocation of smoking status for employees whose smoking status differs between the questionnaire and medical records 36
Table 28: Frequency of smoking status for each of the 3 new reallocation schemes 37
Table 29: Crude odds ratios obtained for the original data, and the minimum, maximum, and proportional allocation schemes 38
Table 30: Results of logistic regression to determine significant predictors of ever-smoker 40
Table 31: Proposed weighting for cases and controls for the 3 reallocation schemes 50
Table 32: Example of frequency weighting 51
Table 33: Overall lung cancer RR due to smoking for the analyses without imputation of missing data 55
Table 34: Descriptive statistics for the 100 imputed RRs by each reallocation 59
Table 35: Average cell counts for the unweighted imputations by AN exposure group 61
Table 36: Average cell counts for the weighted imputations by AN exposure group 62
Table 37: Descriptive statistics for the 100 imputed RRs by quintile of AN exposure and allocation 67

Table 38: Smoking adjusted lung cancer RRs by quintile of AN exposure and allocation.....70
Table 39: Final adjustments of the full cohort lung cancer RRs in the upper quintile of AN exposure.....74

LIST OF FIGURES

Figure 1: Case-cohort schematic.....5
Figure 2: Case-cohort schematic for NCI study 6
Figure 3: Reallocation flowchart 34
Figure 4: Flowchart of required analyses for the 3 reallocations.....41
Figure 5: Distribution of 100 imputed RRs for unweighted and weighted q_smoke variable 56
Figure 6: Boxplots of 100 imputed RRs for unweighted and weighted q_smoke variable 56
Figure 7: Distribution of 100 imputed RRs for unweighted and weighted m_smoke variable 57
Figure 8: Boxplots of 100 imputed RRs for unweighted and weighted m_smoke variable 57
Figure 9: Distribution of 100 imputed RRs for unweighted and weighted unk_smoke variable 58
Figure 10: Boxplots of 100 imputed RRs for unweighted and weighted unk_smoke variable 58
Figure 11: Distribution of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted q_smoke variable.....63
Figure 12: Distribution of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted m_smoke variable.....63
Figure 13: Distribution of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted unk_smoke variable.....64

Figure 14: Boxplots of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted q_smoke variable 65
Figure 15: Boxplots of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted m_smoke variable 65
Figure 16: Boxplots of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted unk_smoke variable 66
Figure 17: Boxplots of 100 imputed smoking adjusted RRs by quintile of AN exposure for unweighted and weighted q_smoke variable 68
Figure 18: Boxplots of 100 imputed smoking adjusted RRs by quintile of AN exposure for unweighted and weighted m_smoke variable 68
Figure 19: Boxplots of 100 imputed smoking adjusted RRs by quintile of AN exposure for unweighted and weighted unk_smoke variable 69
Figure 20: Smoking adjusted lung cancer RR by quintile of AN exposure for analysis models that included unknown smoking histories 71
Figure 21: Smoking adjusted lung cancer RR by quintile of AN exposure for q_smoke imputations 71
Figure 22: Smoking adjusted lung cancer RR by quintile of AN exposure for m_smoke imputations 72
Figure 23: Smoking adjusted lung cancer RR by quintile of AN exposure for unk_smoke imputations 72

1. INTRODUCTION

Acrylonitrile (AN) is a colorless liquid chemical that is most commonly found in organic solvents such as acetone, benzene, carbon tetrachloride, ethyl acetate, and toluene. The largest AN users are companies that make acrylic or modacrylic fibers, high impact ABS (acrylonitrile-butadiene-styrene) plastics and SAN (styrene-acrylonitrile), which is used in automotive products, household goods, and packaging materials (EPA). Acyrlonotrile had also been used as a pesticide and tobacco fumigant, but this was discontinued in the late 1970's (Blair, et al. 1998).

The primary routes of acrylonitrile exposure are inhalation and contact with the skin (EPA). Because AN does not occur naturally and is rarely found in air, most exposures are occupational. Because of any possible workplace exposure, OSHA, the Occupational Safety and Health Administration, which is responsible for the safety and health of the country's workforce, has set allowable exposure limits to acrylonitrile. The permissible exposure limit (PEL) set by OSHA is 2 ppm (parts per million) as an 8 hour time weighted average with no skin or eye contact (OSHA). The short term exposure ceiling (15 minutes) was set at 10 ppm.

The IARC, the International Agency for Research on Cancer, initiated a program in 1969 to evaluate the carcinogenic risk of certain chemicals to humans. These results, published as monographs, are prepared under the direction of international groups of experts who evaluate and review evidence on the carcinogenicity of certain substances. Types of information included in these monographs are exposure data, review of cancer studies in humans, studies of cancer in laboratory animals, evaluation of carcinogenicity and its mechanisms, and a summary and evaluation. The monographs are updated regularly as new information or studies are made available.

The IARC reports that past acrylonitrile studies have been inconclusive as to the relationship between AN exposure and cancer. Studies from the 1970s and 1980s did indicate a possible increase in lung cancer among exposed workers, but had potential design problems such as insufficient follow-up time, small sample sizes, lack of exposure data, and possible confounding due to smoking. Conversely, AN exposure tests in laboratory animals produced statistically significant results for increases in certain types of malignant and benign tumors. Based on reviews of these studies, the IARC evaluation states there is 'sufficient evidence' of carcinogenicity in laboratory animals. For humans, there is 'inadequate evidence' of carcinogenicity. The 'inadequate evidence' classification is based on the insufficient power, statistical significance, or quality of the available studies. The overall evaluation of AN as group 2B by the IARC means there is insufficient evidence of increased cancer risk in humans, but there is sufficient evidence of increased cancer risk in laboratory animals (IARC 1999). Similarly, the United States Environmental Protection Agency (EPA) has classified AN as a group B1 carcinogen, a probable human carcinogen (EPA).

Because of the shortcomings if these past studies, the National Cancer Institute (NCI) and the National Institute for Occupational Health and Safety (NIOSH) conducted a comprehensive cohort study to determine if there were any potential cancer risks due to occupational exposures to acrylonitrile. The results of the study were published by Blair et al. in 1998. The exposure assessment portion of the study was published at the same time by Stewart et al. (1998).

1.1. Original NCI Study Cohort

Because the greatest exposures could be expected at plants where acrylonitrile is produced, plants or facilities of these types were sought out for inclusion in the study. The final cohort was assembled from 8 different AN plants. The basic types of AN production at each facility were as follows: 4 involved in AN monomer production, 3 produced AN fibers, and 1 produced AN resin. All workers employed at any of the 8 study plants were included in the cohort. The workers had to be employed prior to 1984 and after AN production had begun. All demographic and work history information was obtained from company records. Demographic data includes information such as date of birth, date of hire, race, gender, and vital status. Work history data was collected to determine an employee's work area and job title history, which are used in determining an individual's AN exposure. Study investigators performed personal monitoring of exposures at all 8 plants. These exposures were linked to an employee's job title and area to determine AN exposure over an entire work history. The total number of employees in the final cohort was 25,460 (Blair et al. 1998).

Because smoking is known to be a risk factor for lung cancer, it is important to adjust for smoking if lung cancer risk due to an occupational exposure is going to be evaluated. To address any possible confounding due to smoking, it is first necessary to determine an employee's smoking history. Because the costs and time associated with determining the smoking status of every individual in the cohort would be prohibitive, the authors chose to collect smoking information for a sub-sample of the entire cohort. Smoking histories are only gathered for employees in this sample, saving time and money, and the lung cancer risk due to occupational

exposures can then be adjusted for smoking. An analysis of this type is referred to as a case-cohort design.

1.2. Case-Cohort Design Description

Prentice proposed the case-cohort design in 1986. In this design, a sample of the entire cohort is selected and evaluated for a certain measure, smoking status, for example. Any additional cases that were not selected in this sample would also be evaluated for smoking status. The sampled members of the cohort and the additional cases are combined to form a sub-cohort. In addition to the demographic variable information that was available for all individuals in the full cohort, the individuals in the sub-cohort have the additional smoking status covariate. The sub-cohort is analyzed using a standard Cox proportional hazards model, but with slight modification (Prentice 1986; Barlow 1994; Barlow and Ichikawa 1999). The cases that were included in the original sample are used as controls until their failure time. The cases that were not in the original sample are only analyzed as cases at their failure time and never appear as controls. Figure 1 represents a case-cohort design for a cohort of size **N** with **c** cases. The sub-cohort, **n**, is selected from the entire cohort, **N**, with a sampling fraction **f = n/N**. The cases included in the sub-cohort sample are represented by $\mathbf{c_{sc}}$. The lower figure shows the final size of the sub-cohort, which is the sum of the sampled employees, **n**, and the cases not included in the original sample, $\mathbf{(c-c_{sc})}$.

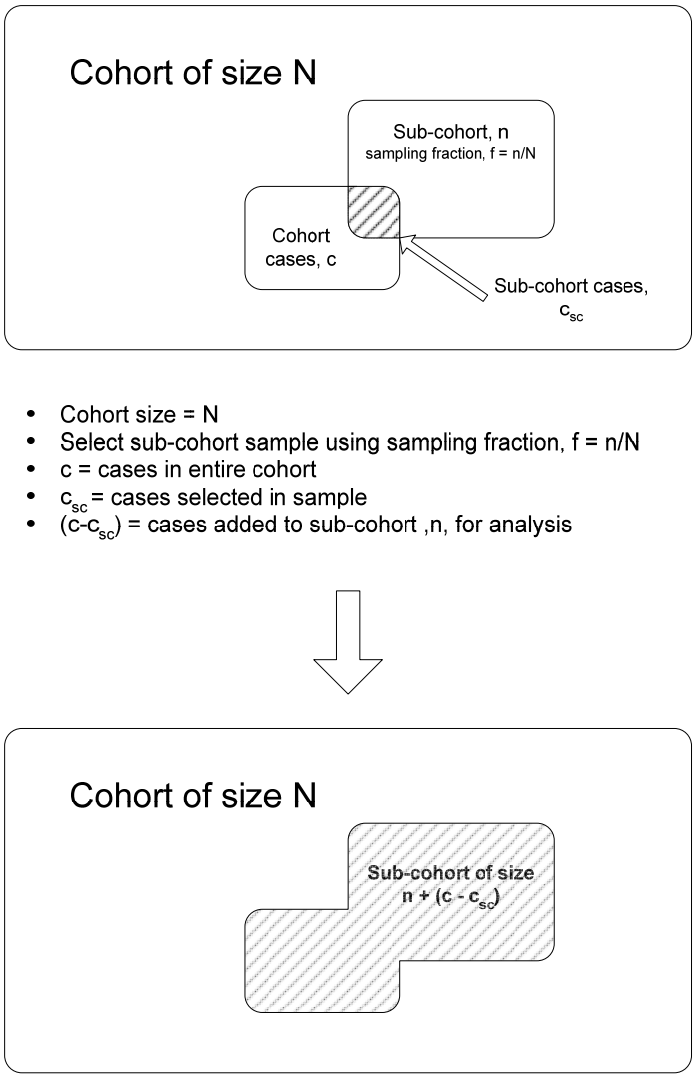

Figure 1: Case-cohort schematic

1.3. NCI Case-Cohort Design

In the NCI study, smoking information was gathered for a 10% sample of the entire cohort. The sample was drawn systematically by selecting employment records for every 10th individual after a random starting point was determined. Employment records for the cases (not in the original systematic sample) were also collected. The lung cancer relative risks (RR) for AN exposure, adjusted for smoking, were calculated using EPICURE's PEANUTS module, which analyzes proportional hazards models, specifically case-cohort designs (Blair et al. 1998). The schematic for the NCI sub-cohort is shown in Figure 2.

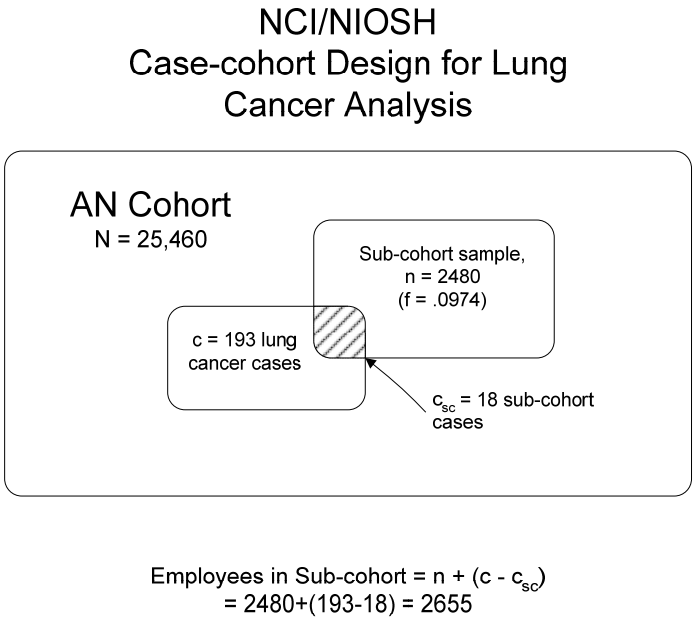

Figure 2: Case-cohort schematic for NCI study

The sub-cohort size is 2655 employees, with 193 lung cancer cases, 18 of which were selected in the systematic sample and are therefore eligible to be controls in the risk sets.

1.4. NCI Results

There are several results published in the NCI paper that are of interest. The reported rate ratio for lung cancer in ever- versus never-smokers was 3.6. This number appears to be very low and is atypical of the lung cancer relative risk of 8.0-10.0 observed in ever- versus never-smokers. In the U.S. Department of Health and Human Services *The Health Consequences of Smoking* (2004), the relative risk due to smoking in current versus never-smokers is reported as approximately 23 in males and 13 in females. Although this only involves current versus never smokers and not the ever- versus never-smoker discussed in the NCI paper, this gives an indication of the magnitude of the lung cancer relative risk due to smoking.

Another of the findings of the NCI paper was an elevated relative risk (RR) of lung cancer in the upper quintile of AN exposure. Although this result was not significant and there was no statistically significant relative risk trend through the increasing cumulative AN exposure groups, the paper concluded that there may be evidence of "carcinogenic activity at the highest levels of exposure." Because of the lack of statistical significance, the recommendation was additional follow-up to determine if there is an increase in lung cancer relative risk in the upper cumulative AN exposure groups.

Table 1: Lung cancer relative risk by quintile of AN exposure – NCI study results

	Quintile of exposure										
	1 (lowest)		2		3		4		5 (highest)		
	O	RR	O	RR	O	RR	O	RR	O	RR	p for trend
Cumulative exposure for full cohort	27	1.1	26	1.3	28	1.2	27	1.0	26	1.5	0.65
Cumulative exposure for full smoking subcohort (not adjusted for smoking)	27	0.8	26	1.1	28	1.0	27	0.9	26	1.5	0.70
Cumulative exposure for smoking subcohort with information on cigarette use (not adjusted for smoking)	5	0.3	6	0.9	7	1.0	13	1.0	9	1.7	0.80
Cumulative exposure for smoking subcohort adjusted for ever cigarette use	5	0.3	6	0.8	7	1.0	13	0.9	9	1.6	0.99
Cumulative exposure for smoking subcohort adjusted for number of cigarettes per day	5	0.3	8	0.7	7	1.1	13	1.0	9	1.7	0.96
Full cohort with RR values adjusted for smoking	-	1.1	-	1.0	-	1.1	-	0.9	-	1.4	-

O - Observed
RR - Relative risk

Table 1 is reproduced from the NCI study (Blair et al. 1998). The first row of the table shows the elevated lung cancer RR of 1.5 in the upper quintile of AN exposure. The p-value for trend is a non-significant 0.65. The last row of the table shows the exposure RRs for the full cohort adjusted for smoking. To get this 'adjusted' value, the authors calculated the proportional change in RR for the sub-cohort (for which smoking information was available) adjusted for and not adjusted for smoking. This proportional change was applied to the values in the first row of the table to obtain the values in the last row. For example, in the 5th quintile of AN exposure, the RR for the smoking sub-cohort (not adjusted for smoking) is 1.7. The RR for the smoking sub-

cohort adjusted for ever cigarette use is 1.6. The full cohort with RR adjusted for smoking (last row of table) is calculated in Equation 1.4.1.

$$RR_{adj}^{full} = RR_{unadj}^{full} \times \frac{RR_{adj}^{sub}}{RR_{unadj}^{sub}} = 1.5 \times \frac{1.6}{1.7} = 1.4 \qquad \text{(Equation 1.4.1.)},$$

where:

RR_{unadj}^{full} is the RR for the full cohort unadjusted for smoking (row 1)

RR_{adj}^{sub} is the RR for the sub-cohort adjusted for smoking (row 4)

RR_{unadj}^{sub} is the RR for the sub-cohort unadjusted for smoking (row 3)

The basis of the paper's conclusion, the relative risk of 1.5 in the upper quintile of exposure, was unadjusted for smoking. As shown, the relative risk only dropped to 1.4 after adjustment. The authors admit that this result is surprising because the smoking prevalence increased by exposure quintile. Confounding due to smoking, which could be possible, as the authors point out, because much of the smoking information was missing, was ruled out for two reasons. Firstly, the adjustment for smoking did not dramatically change the relative risks in the AN exposure groups (1.5 to 1.4 in the upper quintile). The smoking histories, which were missing for almost two-thirds of the cases, were dismissed as a possible reason for this slight change after adjustment. The authors noted that because the available information did not indicate confounding, there should be no reason to believe that the missing information would be any different and show confounding. Secondly, there were no excess risks noted in other diseases related to smoking in the AN exposed individuals, and if there was confounding, it would effect the other smoking-related conditions and not just lung cancer.

2. STATEMENT OF THE PROBLEM

Two of the issues from the paper will be the focus of this thesis:

- the low overall lung cancer RR of 3.6 for ever- versus never-smokers
- the minimal change in the lung cancer RR in the upper quintile of AN exposure after adjustment for smoking (1.5 to 1.4)

By analyzing the smoking histories and systematically allocating those with missing smoking information as ever- or never-smokers, it is hoped that the overall lung cancer RR due to smoking can be raised to a more realistic level, while at the same time observing an adjusted lung cancer RR in the upper AN exposure group less than the reported 1.4.

3. METHODS

3.1. Reproduction of NCI Results

3.1.1. NCI sub-cohort data files

Before beginning an analysis of the missing smoking information and reallocating the unknowns as ever- or never-smokers, it is necessary to reproduce the results of the NCI paper. In 2001, Marsh et al. performed a reevaluation of the lung cancer RRs in the NCI cohort study using external rates as the basis for comparison. To perform the reevaluation, a copy of the study data was obtained from the original study authors. The sub-cohort data was obtained as a SAS (SAS Institute Inc. 1999) dataset with 2655 observations (individuals) and 492 variables. Information in the dataset includes employee id, vital status, plant, race, sex, smoking status, and AN exposure data. An MS Excel spreadsheet with much of the same information was also

obtained. The SAS dataset will subsequently be referred to as SMK_NCI and the Excel file as SMOKECUM.

3.1.2. NCI sub-cohort demographic data

Table 2: Frequency distribution of sub-cohort employees by demographic and exposure variables

Demographic Variable	Plant 1 n (%)		Plant 2 n (%)		Plant 3 n (%)		Plant 4 n (%)		Plant 5 n (%)		Plant 6 n (%)		Plant 7 n (%)		Plant 8 n (%)		Total n (%)	
Race																		
White	183	(92.9)	183	(88.4)	160	(89.9)	219	(75.0)	715	(92.0)	248	(82.4)	202	(89.8)	416	(87.0)	2326	(87.6)
Nonwhite	14	(07.1)	23	(11.1)	14	(07.9)	71	(24.3)	61	(07.9)	53	(17.6)	23	(10.2)	62	(13.0)	321	(12.1)
Unknown	0	(00.0)	1	(00.5)	4	(02.2)	2	(00.7)	1	(00.1)	0	(00.0)	0	(00.0)	0	(00.0)	8	(00.3)
Gender																		
Male	169	(85.8)	157	(75.8)	118	(66.3)	201	(68.8)	634	(81.6)	263	(87.4)	206	(91.6)	379	(79.3)	2127	(80.1)
Female	28	(14.2)	50	(24.2)	60	(33.7)	91	(31.2)	143	(18.4)	38	(12.6)	19	(08.4)	99	(20.7)	528	(19.9)
Year of Birth																		
<1925	30	(15.2)	8	(03.9)	24	(13.5)	29	(09.9)	94	(12.1)	50	(16.6)	17	(07.6)	161	(33.7)	413	(15.6)
1925-1934	43	(21.8)	35	(16.9)	38	(21.3)	47	(16.1)	161	(20.7)	68	(22.6)	37	(16.4)	142	(29.7)	571	(21.5)
1935-1944	74	(37.6)	78	(37.7)	37	(20.8)	71	(24.3)	258	(33.2)	60	(19.9)	75	(33.3)	87	(18.2)	740	(27.9)
1945-1954	41	(20.8)	57	(27.5)	59	(33.1)	100	(34.2)	215	(27.7)	76	(25.2)	77	(34.2)	66	(13.8)	691	(26.0)
>1955	9	(04.6)	29	(14.0)	20	(11.2)	45	(15.4)	48	(06.2)	47	(15.6)	19	(08.4)	22	(04.6)	239	(09.0)
Unknown	0	(00.0)	0	(00.0)	0	(00.0)	0	(00.0)	1	(00.1)	0	(00.0)	0	(00.0)	0	(00.0)	1	(00.0)
AN exposure																		
Unexposed	40	(20.3)	108	(52.2)	128	(71.9)	33	(11.3)	232	(29.9)	110	(36.5)	37	(16.4)	198	(41.4)	886	(33.4)
Exposed	157	(79.7)	99	(47.8)	50	(28.1)	259	(88.7)	545	(70.1)	191	(63.5)	188	(83.6)	280	(58.6)	1769	(66.6)
<0.13 ppm-yrs	45	(28.7)	33	(33.3)	4	(08.0)	148	(57.1)	184	(33.8)	87	(45.5)	54	(28.7)	100	(35.7)	655	(37.0)
0.13 to 0.57 ppm-yrs	32	(20.4)	32	(32.3)	7	(14.0)	22	(08.5)	89	(16.3)	44	(23.0)	40	(21.3)	62	(22.1)	328	(18.5)
0.57 to 1.5 ppm-yrs	37	(23.6)	13	(13.1)	6	(12.0)	16	(06.2)	67	(12.3)	34	(17.8)	29	(15.4)	51	(18.2)	253	(14.3)
1.5 to 8.0 ppm-yrs	26	(16.6)	12	(12.1)	13	(26.0)	36	(13.9)	103	(18.9)	20	(10.5)	51	(27.1)	55	(19.6)	316	(17.9)
>8.0 ppm-yrs	17	(10.8)	9	(09.1)	20	(40.0)	37	(14.3)	102	(18.7)	6	(03.1)	14	(07.4)	12	(04.3)	217	(12.3)
Total	197	(07.4)	207	(07.8)	178	(06.7)	292	(11.0)	777	(29.3)	301	(11.3)	225	(08.5)	478	(18.0)	2655	(100.0)

Table 2 shows how the sub-cohort employees are distributed among the eight plants for several demographic variables and exposure histories. The table shows that the sub-cohort employees are predominately white males. Plant 5 is the largest, accounting for 777 of the 2655 employees (29.3%). The other seven plants range from 18% to 7% of the employees in the sub-cohort. Workers ever-exposed to AN make-up 67% of the sub-cohort. Among the exposed, 37% are from the lowest AN cumulative exposure group, 18.5% from group 2, 14.3% from group 3, 18% from group 4, and 12% are from the upper quintile of cumulative AN exposure. The cumulative AN exposure cut points are set at 0.13 ppm-years, 0.57 ppm-years, 1.50 ppm-yrs, and 8.00 ppm-years to match the NCI paper. Neither the NCI_SMK nor SMOKECUM

datasets had variables for the AN cumulative exposure groups, so they had to be determined from the exposure data. The cumulative AN was included in the SMOKECUM file, but had to be converted to ppm-years. Once the cumulative AN was converted to ppm-years, the quintile of exposure was assigned to the appropriate group according to the cut points. Equation 3.1.2.1. demonstrates this for an employee whose lifetime cumulative AN was given in the SMOKECUM spreadsheet as 85.2 ppm-days.

$$85.2\,ppm-days \times \frac{1\_year}{365\_days} \times \frac{5}{7} = 0.1667\,ppm-years \quad \text{(Equation 3.1.2.1.)}$$

This employee would fall into the 2nd quintile of cumulative AN exposure (0.13 to 0.57 ppm-years). The $\frac{5}{7}$ multiplier assumes an employee works 5 out of 7 days a week. The cumulative AN exposure calculations are relatively straightforward at this point because only the total AN exposure is required, but become more complex when the individual risk-sets must be developed at the case failure times because an employee's cumulative AN exposure group changes with time.

3.1.3. NCI sub-cohort distribution of lung cancer cases

There are 193 lung cancer cases in the entire cohort. Table 3 shows the distribution of the lung cancer cases by race, gender, year of birth, and cumulative AN exposure group.

Table 3: Frequency distribution of sub-cohort lung cancer cases by demographic and exposure variables

Demographic Variable	Plant 1 n (%)	Plant 2 n (%)	Plant 3 n (%)	Plant 4 n (%)	Plant 5 n (%)	Plant 6 n (%)	Plant 7 n (%)	Plant 8 n (%)	Total n (%)
Race									
White	8 (88.9)	6 (100.0)	11 (100.0)	23 (92.0)	38 (97.4)	25 (86.2)	10 (100.0)	58 (90.6)	179 (92.7)
Nonwhite	1 (11.1)	0 (00.0)	0 (00.0)	2 (08.0)	1 (02.6)	4 (13.8)	0 (00.0)	6 (09.4)	14 (07.3)
Unknown	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)
Gender									
Male	9 (100.0)	3 (50.0)	10 (90.9)	21 (84.0)	36 (92.3)	29 (100.0)	10 (100.0)	58 (90.6)	176 (91.2)
Female	0 (00.0)	3 (50.0)	1 (09.1)	4 (16.0)	3 (07.7)	0 (00.0)	0 (00.0)	6 (09.4)	17 (08.8)
Year of Birth									
<1925	6 (66.7)	1 (16.7)	9 (81.8)	12 (48.0)	16 (41.0)	16 (55.2)	8 (80.0)	47 (73.4)	115 (59.6)
1925-1934	3 (33.3)	2 (33.3)	1 (09.1)	9 (36.0)	16 (41.0)	11 (37.9)	2 (20.0)	16 (25.0)	60 (31.1)
1935-1944	0 (00.0)	2 (33.3)	1 (09.1)	4 (16.0)	7 (17.9)	1 (03.4)	0 (00.0)	0 (00.0)	15 (07.8)
1945-1954	0 (00.0)	1 (16.7)	0 (00.0)	0 (00.0)	0 (00.0)	1 (03.4)	0 (00.0)	0 (00.0)	2 (01.0)
>1955	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	1 (01.6)	1 (00.5)
AN exposure									
Unexposed	1 (11.1)	5 (83.3)	7 (63.6)	2 (08.0)	15 (38.5)	9 (31.0)	2 (20.0)	18 (28.1)	59 (30.6)
Exposed	8 (88.9)	1 (16.7)	4 (36.4)	23 (92.0)	24 (61.5)	20 (69.0)	8 (80.0)	46 (71.9)	134 (69.4)
<0.13 ppm-yrs	0 (00.0)	0 (00.0)	0 (00.0)	9 (39.1)	2 (08.3)	4 (20.0)	0 (00.0)	13 (28.3)	28 (20.9)
0.13 to 0.57 ppm-yrs	1 (12.5)	0 (00.0)	0 (00.0)	1 (04.3)	6 (25.0)	6 (30.0)	3 (37.5)	8 (17.4)	25 (18.7)
0.57 to 1.5 ppm-yrs	5 (62.5)	1 (100.0)	0 (00.0)	2 (08.7)	4 (16.7)	7 (35.0)	1 (12.5)	7 (15.2)	27 (20.1)
1.5 to 8.0 ppm-yrs	1 (12.5)	0 (00.0)	0 (00.0)	1 (04.3)	6 (25.0)	2 (10.0)	3 (37.5)	15 (32.6)	28 (20.9)
>8.0 ppm-yrs	1 (12.5)	0 (00.0)	4 (100.0)	10 (43.5)	8 (33.3)	1 (05.0)	1 (12.5)	3 (06.5)	28 (20.9)
Total	9 (04.7)	6 (03.1)	11 (05.7)	25 (13.0)	39 (20.2)	29 (15.0)	10 (05.2)	64 (33.2)	193 (100.0)

As would be expected, due to the distribution of the sub-cohort employees, the majority of the lung cancer cases are white males. Over 90% (175/193) of the cases occur in employees born before 1934. It is also interesting to note that although Plant 8 accounts for only 18% of the sub-cohort employees, it accounts for one-third (64/193) of the cases. In the unexposed workers, there are 59 cases. Table 4 shows the distribution of the 134 exposed cases by cumulative AN exposure quintile.

Table 4: Exposed Lung Cancer Cases by Quintile of Cumulative AN Exposure

lung cancer	Quintile of AN Exposure 1	2	3	4	5	Total
case	28	25	27	28	26	134
(row %)	20.90	18.66	20.15	20.90	19.40	100.00

The lung cancer cases are very evenly distributed among the cumulative AN exposure quintiles. If the frequency of cases are compared to those in Table 1 (the NCI study's Table 9), it can be seen that there are 27, 26, 28, 27, and 26 cases in exposure quintiles 1 through 5, respectively. The observed deaths are off by one for quintiles 1, 2, 3, and 4. The SMOKECUM and

NCI_SMK data (the original data from NCI) were double-checked and always yielded the observed deaths as shown in Table 4. It appears that the discrepancy may just be with the NCI Table 9. Even slight manipulation of the AN cumulative exposure group cut points will not yield observed deaths that match the NCI paper. The values in Table 4 will be assumed to be correct.

3.1.4. Proportional hazards models to obtain relative risks

The next step is to run the proportional hazard models to get the overall lung cancer RR due to smoking and then to determine the lung cancer RRs by cumulative AN exposure quintile, adjusted and unadjusted for smoking. These values will be compared to the NCI Table 9 to ensure that the results are similar. Because of software availability and advances, SAS (SAS Institute Inc. 1999) will be used instead of EPICURE to run the proportional hazards regression models. Ichikawa and Barlow (1998) developed a SAS macro specifically designed to analyze case-cohort data that also allows for selection of different weighting schemes, handling of ties, stratification, and covariate selection. The smoking status variable used by the authors to determine ever- or never-smoker status will be discussed in the "Smoking Information" section. Example SAS code is given in Appendix A. The results are shown in Table 5.

Table 5: Lung cancer relative risk by quintile of AN exposure – NCI study results and reanalysis

		Quintile of exposure[b]									
		1 (lowest)		2		3		4		5 (highest)	
		O	RR[a]	O	RR	O	RR	O	RR	O	RR
Cumulative exposure for full smoking subcohort (not adjusted for smoking)	NCI	27	0.8	26	1.1	28	1.0	27	0.9	26	1.5
	Reanalysis	28	0.91	25	1.10	27	0.98	28	0.97	26	1.54
Cumulative exposure for smoking subcohort with information on cigarette use (not adjusted for smoking)	NCI	5	0.3	6	0.9	7	1.0	13	1.0	9	1.7
	Reanalysis	4	0.35	8	0.92	8	0.95	12	1.00	11	1.73
Cumulative exposure for smoking subcohort adjusted for ever cigarette use	NCI	5	0.3	6	0.8	7	1.0	13	0.9	9	1.6
	Reanalysis	4	0.31	8	0.85	8	0.92	12	0.84	11	1.56
Cumulative exposure for smoking subcohort adjusted for number of cigarettes per day	NCI	5	0.3	8	0.7	7	1.1	13	1.0	9	1.7
	Reanalysis	3	0.28	6	0.72	8	1.07	11	0.95	11	1.78

[a] RR values adjusted for gender and race

[b] Quintile cut points at 0.13 ppm-yrs, 0.57 ppm-yrs, 1.50 ppm-yrs, and 8.00 ppm-yrs

3.1.5. Reanalysis comparison to original NCI results

Overall, the relative risks obtained in the reanalysis agree with those from the NCI study. The RRs in the original study are only reported to one decimal place, so the degree of difference cannot be calculated exactly, but the results are close enough to rule out any significant problem with the modeling in SAS or possible programming errors. As pointed out in earlier, the observed deaths in the reanalysis do not match the NCI study exactly. The differences are more pronounced when the smoking status variable is introduced. For example, for the 'cumulative

exposure for smoking sub-cohort with information on cigarette use (not adjusted for smoking)' category in the table, there are 40 observed deaths in the NCI study, but 43 in the reanalysis. The reanalysis uses the original data, so it again appears as if the problem may be with the table itself and not in the actual data or with data manipulation.

3.2. Sub-cohort Smoking Status Variables

3.2.1. Original smoking status variables

The whole reason for collecting smoking information was to enable the researchers to adjust their models for smoking, which is known to be a risk factor for lung cancer, which also happens to be one of the outcomes of interest. The methods the authors used to gather and assign smoking histories to employees in the sub-cohort must be explored. Information on an employee's smoking history was obtained from one of three sources: medical records, interviews with the actual employees, or interviews with an employee's next-of-kin. The cigarette smoking information gathered in the interviews included age started smoking, ever- or never-smoker, number of years smoked, amount smoked, and cigar or other tobacco use. The results for each employee were coded and stored in the smoking analysis dataset (SMK_NCI). The smoking status of interest here is the binary smoking variable, ever- or never-smoker. Four variables, QSMOKER, MSMOKER, SMKSTAT, and SMKSTAT2 are used to indicate ever- or never-smoker. QMSMOKER and MSMOKER are not directly labeled in the SMK_NCI dataset, but appear to indicate an employee's smoking status based on employee interview (QSMOKER) or employee medical records (MSMOKER). These variables were coded as follows:

0 – never-smoker
1 – ever-smoker
8,9 – missing smoking information

Table 6 shows the frequency distribution of the QSMOKER and MSMOKER variables for the entire sub-cohort.

Table 6: Frequency distribution of NCI smoking status variables

variable	description	code	description	Freq.	Cumul. Freq.	Cumul. %
QSMOKER	smoking status from interview	0	never	697	697	26.3%
		1	ever	1193	1890	71.2%
		8	unknown	765	2655	100.0%
MSMOKER	smoking status from medical records	0	never	384	384	14.5%
		1	ever	651	1035	39.0%
		8	unknown	1126	2161	81.4%
		9		494	2655	100.0%

3.2.2. Agreement between smoking status variables

The authors reported that 1890 (71%) of the 2655 individuals in the sub-cohort completed an interview. This is verified in the last column of the QSMOKER variable, where 71.2% of the 2655 sub-cohort employees were either coded as a never- or ever-smoker. Similarly, the NCI paper reported the smoking information based on medical records was available for 1035 employees, which can be seen as the cumulative percent for MSMOKER equal to 0 or 1 (39.0%). A value of 8 or 9 for an individual indicates an unknown smoking status based on either the questionnaire or medical records. With either variable, the percentage of available smoking information is still rather low (39.0% for medical records and 71.2% for interviews).

In order to maximize the available information, the authors created two additional variables, SMKSTAT and SMKSTAT2, which combine the medical records and interviews to determine if an employee was or was not a smoker.

But before doing this, the QSMOKER and MSMOKER results were compared to determine how closely the medical records and interviews agreed. The NCI paper reports an agreement of 86% between the medical records and interviews using only those 1035 employees with available medical records. Table 7 compares the MSMOKER and QSMOKER values.

Table 7: Cross tabulation of QSMOKER and MSMOKER variables

			QSMOKER			
			never	ever	unk	
			0	1	9	total
MSMOKER	never	0	248	69	67	384
	ever	1	42	463	146	651
	unk	9	407	661	552	1620
		total	697	1194	774	2665

The reported agreement is calculated using only the concordant and discordant pairs for those individuals with a known smoking status from both sources. The agreement is calculated as (248+463)/(248+69+42+463) = 711/822 = 86.5%. When looking at the discordant pairs, it can be seen that there are 69 employees who were coded as never-smoker in their medical records but were coded as ever-smoker in their questionnaires. There were 42 employees coded as ever-smoker in the medical records but who were coded as never-smoker from the questionnaire.

This result can be verified in Stata (StataCorp. 2005), which also produces the kappa statistic, which is a measure of agreement often used when comparing results from two or more raters. The Stata output below shows the results:

```
           |        QSMOKER
   MSMOKER |     never       ever |     Total
-----------+----------------------+----------
     never |       248         69 |       317
      ever |        42        463 |       505
-----------+----------------------+----------
     Total |       290        532 |       822
```

```
             Expected
Agreement    Agreement      Kappa   Std. Err.          Z       Prob>Z
------------------------------------------------------------------------
  86.50%       53.37%     0.7104     0.0348       20.42       0.0000
```

The kappa statistic equals 0.71, indicating a strong agreement between the medical record and interview smoking information (Rosner 2000). The 'agreement' in the output equals the 86.5% calculated earlier. But if the entire 3x3 table as shown above was analyzed in STATA, the results are quite different.

```
           |              QSMOKER
   MSMOKER |     never        ever     unknown |     Total
-----------+---------------------------------+----------
     never |       248          69          67 |       384
      ever |        42         463         146 |       651
   unknown |       407         661         552 |     1,620
-----------+---------------------------------+----------
     Total |       697       1,193         765 |     2,655

             Expected
Agreement    Agreement      Kappa   Std. Err.          Z       Prob>Z
------------------------------------------------------------------------
  47.57%       32.40%     0.2245     0.0121       18.48       0.0000
```

In this case the agreement is only 47.6% with a kappa statistic of 0.225. The difference is largely due to the number of individuals with missing smoking information from their medical records but who had smoking information available from the interviews.

3.2.3. Smoking status variables created by NCI

Based on the 86.5% agreement between the two smoking history sources, Blair created the two additional smoking variables, SMKSTAT and SMKSTAT2 to maximize this smoking information. The SMKSTAT variable is a combination of smoking information based on both the interviews and medical records. SMKSTAT2 is the same as QSMOKER, but with all of the unknown values recoded to 9. The logic used by Blair to assign employees as ever- or never-

smokers was to use the interview (questionnaire) when the smoking history was known (0 – never, 1 – ever), but to use the medical records if the interview smoking history was unknown (9 – unknown). A summary of the ever- never-smoker variables is shown in Table 8.

Table 8: Description of ever-, never-smoker smoking status variables

Original smoking history variables

variable	smoking information based on:	levels	description
QSMOKER	interview (questionnaire)	0	never
		1	ever
		8	unknown
MSMOKER	medical records	0	never
		1	ever
		8	unknown
		9	

Additional variables created in NCI analysis

variable	smoking information based on:	levels	description
SMKSTAT	interview (questionnaire) and medical records	0	never
	use interview if not unknown, then medical records	1	ever
	if available	9	unknown
SMKSTAT2	interview (questionnaire)	0	never
	same as QSMOKE but unknowns recoded	1	ever
		9	unknown

The main ever-, never-smoker variable used in this paper for smoking descriptives and the NCI study is the SMKSTAT variable. This was the variable created in the NCI analysis that maximized an employees smoking history by looking at both their medical records and the employee (or family) interviews. The cross classification in Table 9 demonstrates the logic used for the creation of the ever-, never-smoker variable used in the NCI paper.

Table 9: NCI assignment of a smoking status (SMKSTAT) based on QSMOKER and MSMOKER variables

		SMKSTAT (interviews and medical records)			
QSMOKER	MSMOKER	0	1	9	Total
0	0	248	-	-	248
	1	42	-	-	42
	8	268	-	-	268
	9	139	-	-	139
1	0	-	69	-	69
	1	-	463	-	463
	8	-	455	-	455
	9	-	206	-	206
8	0	67	-	-	67
	1	-	146	-	146
	8	-	-	403	403
	9	-	-	149	149
	Total	764	1339	552	2655

The table shows that there are 552 (403 + 149) employees where the ever-, never-smoker information is missing from both interviews and medical records. There are two occurrences when the QMSMOKER variable disagreed with MSMOKER. In both instances the interview information was assumed to be correct and took precedence over the medical information. The SMKSTAT variable recoded 42 employees as never-smoker who were ever-smoker in their medical records, and recoded 69 employees as ever-smokers who were noted as never-smokers in their records. Of the 765 employees who did not have smoking information from interviews (QSMOKER = 8), 213 (67 + 146) did have information available in the medical records.

3.2.4. Analysis of discordant smoking status variables

Because the authors assumed that the questionnaire information was correct in the 111 individuals whose medical records disagreed with their questionnaires, it is important to check the distribution of these employees by AN exposure and lung cancer status. Table 10 shows how the 111 discordant employees are distributed.

Table 10: Distribution of sub-cohort employees with discordant smoking histories by AN exposure

NCI disposition	lung cancer status	Cumulative AN exposure group						
		unexposed	1	2	3	4	5	total
assigned as ever-smoker	control	22	11	9	7	9	10	68
	case	0	0	1	0	0	0	1
assigned as never-smoker	control	6	12	10	4	8	1	41
	case	0	0	0	0	0	1	1
	total	28	23	20	11	17	12	111

Only 2 lung cancer cases were among those employees with discrepancies between their medical record and questionnaire smoking statuses. One of these was in the 2nd quintile of AN exposure and was assigned as an ever-smoker based on the questionnaire. The other, and more important case, occurred in the upper quintile of AN exposure. This employee was assigned as a never-smoker by the authors based on the questionnaire, although his smoking history in the medical records indicated ever-smoker. As will be shown later, there are only 11 lung cancer cases in the upper quintile of AN exposure with known smoking histories, 8 ever- and 3 never-smokers. Because the three never-smoking cases include this one case described above and the never-smoking cases account for 27% of the total cases in the upper quintile, this reassignment by the study authors likely has an important impact on the reported lung cancer RR for the upper quintile of AN exposure after adjustment for smoking.

3.3. Smoking Information for Sub-Cohort

3.3.1. Demographic information

Using the same ever- and never-smoking status variable as the study authors, SMKSTAT, the 552 employees with missing smoking information can be analyzed to see how the missing

vary among the lung cancer cases, the exposed workers, and other demographic variables to determine if there is a pattern in those with missing histories. This also allows for comparison of the NCI study results to this reanalysis. Table 11 shows the distribution of these employees with missing smoking information by race, gender, year of birth, and cumulative AN exposure group.

Table 11: Frequency distribution of sub-cohort employees with missing smoking information by demographic and exposure variables

Demographic Variable	Plant 1 n (%)	Plant 2 n (%)	Plant 3 n (%)	Plant 4 n (%)	Plant 5 n (%)	Plant 6 n (%)	Plant 7 n (%)	Plant 8 n (%)	Total n (%)
Race									
White	30 (90.9)	22 (78.6)	37 (86.0)	54 (79.4)	140 (89.2)	53 (86.9)	22 (81.5)	115 (85.2)	473 (85.7)
Nonwhite	3 (09.1)	5 (17.9)	4 (09.3)	14 (20.6)	17 (10.8)	8 (13.1)	5 (18.5)	20 (14.8)	76 (13.8)
Unknown	0 (00.0)	1 (03.6)	2 (04.7)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	3 (00.5)
Gender									
Male	28 (84.8)	16 (57.1)	25 (58.1)	46 (67.6)	120 (76.4)	49 (80.3)	27 (100.0)	111 (82.2)	422 (76.4)
Female	5 (15.2)	12 (42.9)	18 (41.9)	22 (32.4)	37 (23.6)	12 (19.7)	0 (00.0)	24 (17.8)	130 (23.6)
Year of Birth									
<1925	4 (12.1)	2 (07.1)	11 (25.6)	8 (11.8)	24 (15.3)	19 (31.1)	5 (18.5)	67 (49.6)	140 (25.4)
1925-1934	9 (27.3)	4 (14.3)	8 (18.6)	18 (26.5)	41 (26.1)	22 (36.1)	4 (14.8)	35 (25.9)	141 (25.5)
1935-1944	12 (36.4)	12 (42.9)	9 (20.9)	15 (22.1)	48 (30.6)	7 (11.5)	10 (37.0)	18 (13.3)	131 (23.7)
1945-1954	8 (24.2)	7 (25.0)	11 (25.6)	24 (35.3)	38 (24.2)	2 (03.3)	8 (29.6)	10 (07.4)	108 (19.6)
>1955	0 (00.0)	3 (10.7)	4 (09.3)	3 (04.4)	6 (03.8)	11 (18.0)	0 (00.0)	5 (03.7)	32 (05.8)
Unknown	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)	0 (00.0)
AN exposure									
Unexposed	8 (24.2)	19 (67.9)	34 (79.1)	15 (22.1)	66 (42.0)	25 (41.0)	7 (25.9)	62 (45.9)	236 (42.8)
Exposed	25 (75.8)	9 (32.1)	9 (20.9)	53 (77.9)	91 (58.0)	36 (59.0)	20 (74.1)	73 (54.1)	316 (57.2)
<0.13 ppm-yrs	11 (44.0)	5 (55.6)	1 (11.1)	33 (62.3)	33 (36.3)	19 (52.8)	7 (35.0)	31 (42.5)	140 (44.3)
0.13 to 0.57 ppm-yrs	5 (20.0)	1 (11.1)	0 (00.0)	5 (09.4)	16 (17.6)	8 (22.2)	8 (40.0)	12 (16.4)	55 (17.4)
0.57 to 1.5 ppm-yrs	5 (20.0)	2 (22.2)	0 (00.0)	4 (07.5)	11 (12.1)	7 (19.4)	0 (00.0)	16 (21.9)	45 (14.2)
1.5 to 8.0 ppm-yrs	3 (12.0)	1 (11.1)	2 (22.2)	5 (09.4)	16 (17.6)	2 (05.6)	5 (25.0)	12 (16.4)	46 (14.6)
>8.0 ppm-yrs	1 (04.0)	0 (00.0)	6 (66.7)	6 (11.3)	15 (16.5)	0 (00.0)	0 (00.0)	2 (02.7)	30 (09.5)
Total	33 (06.0)	28 (05.1)	43 (07.8)	68 (12.3)	157 (28.4)	61 (11.1)	27 (04.9)	135 (24.5)	552 (100.0)

Plants 5 and 8 account for 28.4% and 24.5% of the missing values, respectively. Plant 5 is the largest in the study and Plant 8 was the plant that had a large proportion of the cases. Almost 75% of the missing values occur for employees born prior to 1944 and 57% of the smoking histories are missing for employees ever-exposed to AN.

The NCI paper reported that employees who ever-smoked cigarettes made up 66% of the sample. A frequency tabulation of the SMKSTAT for the entire sub-cohort of employees (with available smoking information) results are shown in Table 12.

Table 12: Prevalence of smoking for sub-cohort employees

smoker	Freq.	Percent	Cum.
never	764	36.33	36.33
ever	1,339	63.67	100.00
Total	2,103	100.00	

The ever-smokers make up 1339 (64%) of the 2103 employees with a known smoking history. This is not exactly the same value reported in the original paper, but the 66% reported may only have included the employees in the 10% sample and not the entire sub-cohort as shown.

3.3.2. Smoking and AN exposure

The authors next looked at smoking status by exposure. For employees categorized as ever- or never-exposed to acrylonitrile (AN), NCI reported 56% ever-smokers among the never-exposed workers, while the prevalence of ever-smokers among the ever-exposed was 68%. Table 13 shows a cross-tabulation of the ever-, never-exposed to AN (ANSTATUS) against the smoking status variable, SMKSTAT.

Table 13: Prevalence of smoking by AN exposure status

	AN Exposure		
smoker	never	ever	Total
never	284	480	764
(row %)	37.17	62.83	100.00
(column %)	43.69	33.04	36.33
ever	366	973	1,339
(row %)	27.33	72.67	100.00
(column %)	56.31	66.96	63.67
Total	650	1,453	2,103
(row %)	30.91	69.09	100.00
(column %)	100.00	100.00	100.00

From Table 13, 56% of those never-exposed to AN were ever-smokers and 67% of those ever-exposed were ever-smokers. The results are close to those in the NCI study. In Table 14, the

AN status in the 552 employees with missing smoking information shows that 57% of those were ever-exposed to AN.

Table 14: Distribution of employees with missing smoking information by AN exposure status

AN Exposure	Freq.	Percent	Cum.
never	236	42.75	42.75
ever	316	57.25	100.00
Total	552	100.00	

Smoking information can also be tabulated by quintile of cumulative AN exposure, as in Table 15. For the 1453 employees with complete smoking information and ever-exposed to AN, the ever-smokers made up 63%, 63%, 68%, 72%, and 74% of the employees in the exposure quintiles, 1, 2, 3, 4, and 5, respectively. The smoking prevalence increases as the cumulative exposure increases.

Table 15: Smoking history by quintile of AN exposure for employees with known smoking history

	Quintile of AN Exposure					
smoker	1	2	3	4	5	Total
never	191	100	66	75	48	480
(row %)	39.79	20.83	13.75	15.63	10.00	100.00
(column %)	37.09	36.63	31.73	27.78	25.67	33.04
ever	324	173	142	195	139	973
(row %)	33.30	17.78	14.59	20.04	14.29	100.00
(column %)	62.91	63.37	68.27	72.22	74.33	66.96
Total	515	273	208	270	187	1,453
(row %)	35.44	18.79	14.32	18.58	12.87	100.00
(column %)	100.00	100.00	100.00	100.00	100.00	100.00

With the addition of the exposed employees with missing smoking information, it can be seen in Table 16 that of the 316 employees missing smoking information, 44% were in quintile 1, 17% in quintile 2, 14% from quintile 3, 15% from quintile 4, and 14% from the upper quintile of AN

exposure. Similarly, the proportion of missing smoking histories by cumulative AN exposure quintile is 0.21, 0.17, 0.18, 0.15, and 0.14 for quintiles one through five.

Table 16: Smoking history by quintile of AN exposure for all employees in sub-cohort

	Quintile of AN Exposure					
smoker	1	2	3	4	5	Total
never	191	100	66	75	48	480
(row %)	39.79	20.83	13.75	15.63	10.00	100.00
(column %)	29.16	30.49	26.09	23.73	22.12	27.13
ever	324	173	142	195	139	973
(row %)	33.30	17.78	14.59	20.04	14.29	100.00
(column %)	49.47	52.74	56.13	61.71	64.06	55.00
missing	140	55	45	46	30	316
(row %)	44.30	17.41	14.24	14.56	9.49	100.00
(column %)	21.37	16.77	17.79	14.56	13.82	17.86
Total	655	328	253	316	217	1,769
(row %)	37.03	18.54	14.30	17.86	12.27	100.00
(column %)	100.00	100.00	100.00	100.00	100.00	100.00

3.3.3. Smoking and lung cancer status

It is also important to analyze the missing smoking history pattern in the lung cancer cases and controls. Table 17 shows the distribution of smoking status for the cases and controls.

Table 17: Smoking history of cases and controls

lung cancer	smoker			
	never	ever	missing	Total
control	757	1,282	423	2,462
(row %)	30.75	52.07	17.18	100.00
(column %)	99.08	95.74	76.63	92.73
case	7	57	129	193
(row %)	3.63	29.53	66.84	100.00
(column %)	0.92	4.26	23.37	7.27
Total	764	1,339	552	2,655
(row %)	28.78	50.43	20.79	100.00
(column %)	100.00	100.00	100.00	100.00

From Table 17, only 17% (423/2462) of the controls are missing a smoking status, but approximately 67% (129/193) of the cases are missing the ever-, never-smoker value. Of the 552 missings, 423 (77%) are controls and 129 (23%) are lung cancer cases.

Table 18: Smoking history of cases and controls, complete smoking histories only

lung cancer	smoker: never	ever	Total
control	757	1,282	2,039
(row %)	37.13	62.87	100.00
(column %)	99.08	95.74	96.96
case	7	57	64
(row %)	10.94	89.06	100.00
(column %)	0.92	4.26	3.04
Total	764	1,339	2,103
(row %)	36.33	63.67	100.00
(column %)	100.00	100.00	100.00

Table 18 summarizes the cases and controls by smoking exposure for the non-missings only. Of the controls, 63% were ever-smokers, while the smoking prevalence in the lung cancer cases was 89%. Smoking causes about 90% of lung cancer deaths in men and about 80% in women, so the prevalence in this cohort, where 57 of the 64 cases were ever-smokers, appears to be in agreement with the general population.

The missing smoking information can also be analyzed for the 193 lung cancer cases by AN exposure status as in Table 19. The percentage of missing smoking information is 64% in the unexposed cases and 68% in the exposed cases.

Table 19: Smoking history by AN exposure status - lung cancer cases

smoker	AN Exposure never	ever	Total
never	3	4	7
(row %)	42.86	57.14	100.00
(column %)	5.08	2.99	3.63
ever	18	39	57
(row %)	31.58	68.42	100.00
(column %)	30.51	29.10	29.53
missing	38	91	129
(row %)	29.46	70.54	100.00
(column %)	64.41	67.91	66.84
Total	59	134	193
(row %)	30.57	69.43	100.00
(column %)	100.00	100.00	100.00

Table 20 shows the cases by cumulative AN exposure group. For quintiles 1 through 5, the missing percentages are 86%, 68%, 70%, 57%, and 58% respectively. The table also shows that are no never-smoking cases in exposure quintiles 1, 2, and 3 and only 1 and 3 never-smoking cases in quintiles 4 and 5. The ever-smoking cases are distributed relatively evenly across the exposure quintiles (4, 8, 8, 11, and 8), but over 30% (18/57) of the ever-smoking cases were never-exposed to AN.

Table 20: Smoking history by AN exposure group - lung cancer cases

smoker	Quintile of AN Exposure unexposed	1	2	3	4	5	Total
never	3	0	0	0	1	3	7
(column %)	5.08	0.00	0.00	0.00	3.57	11.54	3.63
ever	18	4	8	8	11	8	57
(column %)	30.51	14.29	32.00	29.63	39.29	30.77	29.53
missing	38	24	17	19	16	15	129
(column %)	64.41	85.71	68.00	70.37	57.14	57.69	66.84
Total	59	28	25	27	28	26	193
(column %)	100.00	100.00	100.00	100.00	100.00	100.00	100.00

The missingness of the smoking status variable for ever- or never-smoker can also be tabulated by other covariates in the original NCI dataset such as: plant, race, and gender.

3.3.4. Smoking status by plant

For all employees in the sub-cohort and all 8 plants in the study, the percentage of employees with a missing smoking status ranges from a maximum of 28% in plant 8 to 12% in plant 7, as shown in Table 21. The prevalence of ever-smokers ranges from a minimum of 50% (89/179) in plant 2 to a maximum of 73% (164/224) in plant 4.

Table 21: Smoking history by plant for entire sub-cohort

	plant								
smoker	1	2	3	4	5	6	7	8	total
never	55	89	57	60	207	88	75	133	764
(row %)	7.2%	11.6%	7.5%	7.9%	27.1%	11.5%	9.8%	17.4%	100.0%
(column %)	27.9%	43.0%	32.0%	20.5%	26.6%	29.2%	33.3%	27.8%	28.8%
ever	109	90	78	164	413	152	123	210	1339
(row %)	8.1%	6.7%	5.8%	12.2%	30.8%	11.4%	9.2%	15.7%	100.0%
(column %)	55.3%	43.5%	43.8%	56.2%	53.2%	50.5%	54.7%	43.9%	50.4%
missing	33	28	43	68	157	61	27	135	552
(row %)	6.0%	5.1%	7.8%	12.3%	28.4%	11.1%	4.9%	24.5%	100.0%
(column %)	16.8%	13.5%	24.2%	23.3%	20.2%	20.3%	12.0%	28.2%	20.8%
total	197	207	178	292	777	301	225	478	2655
(row %)	7.4%	7.8%	6.7%	11.0%	29.3%	11.3%	8.5%	18.0%	100.0%
(column %)	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

For only the 193 lung cancer cases by plant and smoking status, the missing rate is relatively constant across plants, ranging from 62% in plant 5 to 72% in plant 6 (Table 22). The distribution of the 129 cases with missing smoking information is not so even across the eight plants. Plant 2 accounts for only 3.1% of the employees with missing smoking information, while plant 8 accounts for 33.3% of these 129 lung cancer cases.

Table 22: Smoking history by plant for lung cancer cases

	plant								
smoker	1	2	3	4	5	6	7	8	Total
never	0	0	1	0	1	2	1	2	7
(row %)	0.0%	0.0%	14.3%	0.0%	14.3%	28.6%	14.3%	28.6%	100.0%
(column %)	0.0%	0.0%	9.1%	0.0%	2.6%	6.9%	10.0%	3.1%	3.6%
ever	3	2	3	8	14	6	2	19	57
(row %)	5.3%	3.5%	5.3%	14.0%	24.6%	10.5%	3.5%	33.3%	100.0%
(column %)	33.3%	33.3%	27.3%	32.0%	35.9%	20.7%	20.0%	29.7%	29.5%
missing	6	4	7	17	24	21	7	43	129
(row %)	4.7%	3.1%	5.4%	13.2%	18.6%	16.3%	5.4%	33.3%	100.0%
(column %)	66.7%	66.7%	63.6%	68.0%	61.5%	72.4%	70.0%	67.2%	66.8%
total	9	6	11	25	39	29	10	64	193
(row %)	4.7%	3.1%	5.7%	13.0%	20.2%	15.0%	5.2%	33.2%	100.0%
(column %)	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

3.3.5. Smoking by race and gender

The ever-, never-, and missing smoking information by race and gender are shown in Tables 23 and 24. 85% of the 552 missing smoking values are white and 14% are nonwhite. The ever-smoking prevalence is 65% (1196/1853) in whites and 57% (139/245) for nonwhites.

Table 23: Smoking history by race for entire sub-cohort

	race			
smoker	white	nonwhite	other	Total
never	657	106	1	764
(row %)	85.99	13.87	0.13	100.00
(column %)	28.25	33.02	12.50	28.78
ever	1,196	139	4	1,339
(row %)	89.32	10.38	0.30	100.00
(column %)	51.42	43.30	50.00	50.43
missing	473	76	3	552
(row %)	85.69	13.77	0.54	100.00
(column %)	20.34	23.68	37.50	20.79
Total	2,326	321	8	2,655
(row %)	87.61	12.09	0.30	100.00
(column %)	100.00	100.00	100.00	100.00

Table 24: Smoking history by gender for entire sub-cohort

smoker	gender male	female	Total
never	572	192	764
(row %)	74.87	25.13	100.00
(column %)	26.89	36.36	28.78
ever	1,133	206	1,339
(row %)	84.62	15.38	100.00
(column %)	53.27	39.02	50.43
missing	422	130	552
(row %)	76.45	23.55	100.00
(column %)	19.84	24.62	20.79
Total	2,127	528	2,655
(row %)	80.11	19.89	100.00
(column %)	100.00	100.00	100.00

85% of the 552 missing smoking values are males and 15% are female. The ever-smoking prevalence is 67% (1133/1705) in males and 52% (206/398) for females.

3.3.6. Smoking and year of birth

The year of birth of an employee is another variable that may have a relationship to an employees smoking status. Four dummy variables were created for the five year of birth groups. The cutoff points for the year of birth groups were set at: <1925, 1925-1934, 1935-1944, 1945-1954, and >=1955 to match the NCI study. The baseline was the >= 1955 birth year group. The percentages of missing smoking information for groups <1925, 1925-1934, 1935-1944, 1945-1954, and >=1955 are 34%, 25%, 18%, 16%, and 13%, respectively (Table 25). This decrease would be expected, as older employees may not have adequate records or would not be available for interview regarding his or her smoking history.

Table 25: Smoking history by year of birth for entire sub-cohort

	Year of Birth					
smoker	<1925	1925-1934	1935-1944	1945-1954	>=1955	Total
never	68	115	213	249	119	764
(row %)	8.90	15.05	27.88	32.59	15.58	100.00
(column %)	16.46	20.10	28.78	36.03	49.79	28.78
ever	205	316	396	334	88	1,339
(row %)	15.31	23.60	29.57	24.94	6.57	100.00
(column %)	49.64	55.24	53.51	48.34	36.82	50.43
missing	140	141	131	108	32	552
(row %)	25.36	25.54	23.73	19.57	5.80	100.00
(column %)	33.90	24.65	17.70	15.63	13.39	20.79
Total	413	572	740	691	239	2,655
(row %)	15.56	21.54	27.87	26.03	9.00	100.00
(column %)	100.00	100.00	100.00	100.00	100.00	100.00

For only the 2103 employees with known smoking histories, the tabulation of the year of birth by never-, ever-smoker shows that the smoking prevalence decreases with the employee's age is shown in Table 26. The employees born before 1925 had a smoking prevalence of 75%, while the prevalence drops to 57% and 43% in the last two year of birth groups.

Table 26: Smoking history by year of birth for sub-cohort employees with known smoking histories

	Year of Birth					
smoker	<1925	1925-1934	1935-1944	1945-1954	>=1955	Total
never	68	115	213	249	119	764
(row %)	8.90	15.05	27.88	32.59	15.58	100.00
(column %)	24.91	26.68	34.98	42.71	57.49	36.33
ever	205	316	396	334	88	1,339
(row %)	15.31	23.60	29.57	24.94	6.57	100.00
(column %)	75.09	73.32	65.02	57.29	42.51	63.67
Total	273	431	609	583	207	2,103
(row %)	12.98	20.49	28.96	27.72	9.84	100.00
(column %)	100.00	100.00	100.00	100.00	100.00	100.00

To summarize the smoking information, 20.8% (552/2655) of the 2655 employees in the smoking sub-cohort were missing a smoking status. The percentage of missings in the lung cancer cases and controls varies greatly. Smoking information is missing for 66.8% (129/193) of the cases, in contrast to a missingness rate of 17.2% (423/2462) in the controls.

4. REANALYSIS

4.1. Objectives

In order to determine if the missing smoking information affected the results of the NCI paper, in particular the overall relative risk of lung cancer in ever- versus never-smokers and the smoking adjusted lung cancer relative risk of 1.6 in the upper quintile of AN exposure, a systematic method for assigning a smoking status to these unknowns is proposed. Specifically, reassigning a smoking status (ever- or never-smoker) to the employees without a smoking history will be performed to achieve the following:

- Increase the overall relative risk for lung cancer in ever- versus never-smokers from 3.6 to 8.0-10.0
- Increase the smoking status rate of response from 82.8% in the controls and 33.2% in the cases to 100% (all of the unknowns will be assigned as ever- or never-smokers)
- Rerun the proportional hazards models to find the smoking adjusted lung cancer RR due to AN exposure in the upper quintile and compare to the NCI result

One advantage of reassigning all of the unknowns as ever- or never-smokers is that all of the employees in the sub-cohort will be used in the proportional hazard regression model. If the

model includes the SMKSTAT variable (ever- or never-smoker) used in the original NCI study and the sub-cohort is analyzed with this missing smoking history covariate, the risk set will only consist of 64 cases (strata). On the other hand, if the remaining 129 lung cancer cases are assigned as ever- or never-smokers, all 193 lung cancer cases would be modeled, resulting in 193 strata in the risk set.

4.2. Reassignment of Smoking Status Variable

4.2.1. Employees whose smoking histories did not agree

This reassignment of unknown smoking histories will first focus on the employees whose smoking histories disagreed in the medical records and questionnaires. Figure 3 is a flowchart describing this process.

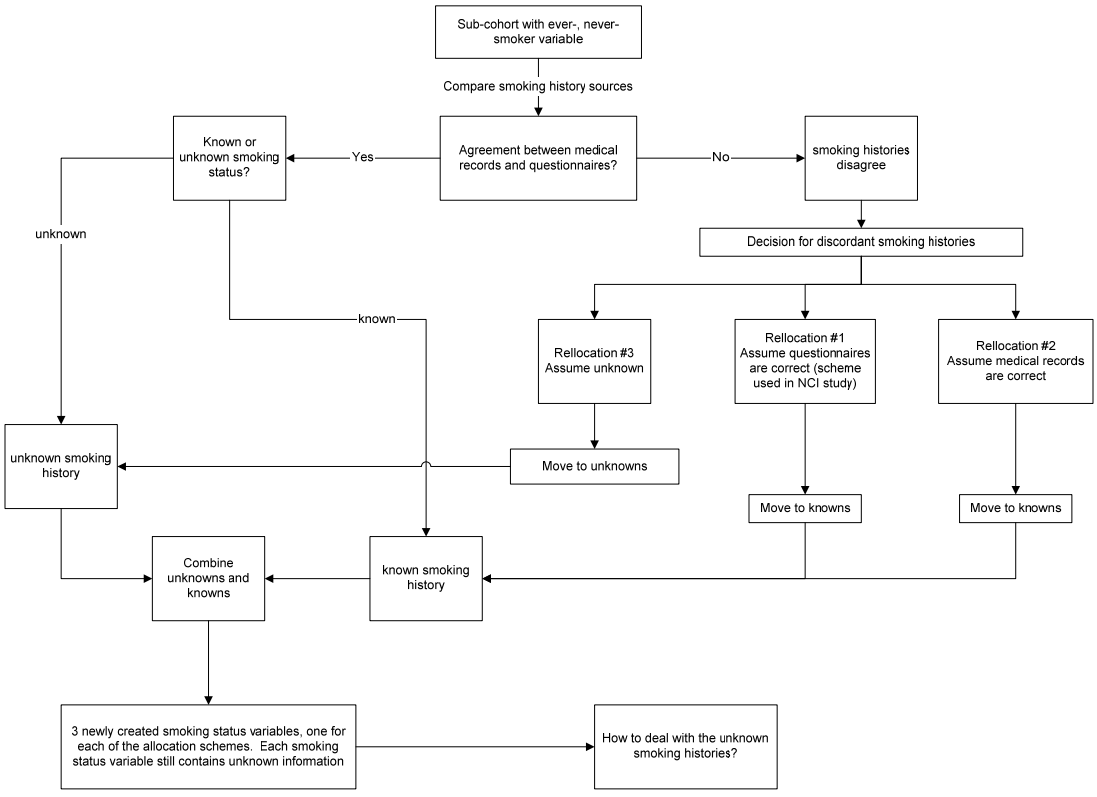

Figure 3: Reallocation flowchart

If the medical records and questionnaires were in agreement as to whether an employee was or was not a smoker, the smoking status will be assumed to be correct, known or unknown. If there is disagreement in the medical records and questionnaire, three different reallocation schemes are proposed. Reallocation #1 will assume that the questionnaire information is correct. Therefore, the 111 discordant smoking histories will be assigned a smoking status based on the interviews. The other employees are assigned as ever- or never-smokers according to the available information (medical or interview). This was the method used in the NCI study. Reallocation #2 assumes that the medical records are correct for these 111 individuals. Reallocation #3 will reassign the 111 as unknown, assuming that the smoking histories are unknown because there is disagreement in the two smoking history sources. Table 27 summarizes these reallocation schemes.

Table 27: Reallocation of smoking status for employees whose smoking status differs between the questionnaire and medical records

Reallocation #1 - Assume questionnaire smoking status is correct

		original data medical records				
		never	ever	unknown		
questionnaire (interview)	never	248	42	407	→	111 employees remain in their original cells and are assigned based on questionnaire (NCI's method)
	ever	69	463	661		
	unknown	67	146	552		

Reallocation #2 - Assume medical record smoking status is correct

		original data medical records				**reallocated data** medical records		
		never	ever	unknown		never	ever	unknown
questionnaire (interview)	never	248	42	407	→	317	0	407
	ever	69	463	661		0	505	661
	unknown	67	146	552		67	146	552

Reallocation #3 - Assume neither are correct, assign as unknown

		original data medical records				**reallocated data** medical records		
		never	ever	unknown		never	ever	unknown
questionnaire (interview)	never	248	42	407	→	248	0	407
	ever	69	463	661		0	463	661
	unknown	67	146	552		67	146	663

Reallocations #1 and #2 result in the same number of missings, 552, while reallocation #3 results in 663 unknowns. Three new smoking status variables were created to replace the NCI's ever-, never-smoking status indicator. They are: q_smoke, m_smoke, and unk_smoke, and have been named to indicate the allocation of the 111 discordant observations. Table 28 details this allocation.

Table 28: Frequency of smoking status for each of the 3 new reallocation schemes

allocation	description of reallocation[a]	new smoking variable	smoking status	Freq.	%
1	Assume questionnaire smoking status is correct	q_smoke[b]	never	764	28.8
			ever	1339	50.4
			unknown	552	20.8
			total	2655	100.0
2	Assume medical record smoking status is correct	m_smoke	never	791	29.8
			ever	1312	49.4
			unknown	552	20.8
			total	2655	100.0
3	Assume both are unknown	unk_smoke	never	722	27.2
			ever	1270	47.8
			unknown	663	25.0
			total	2655	100.0

[a] - reallocating only the 111 observations where the smoking status differs between the medical records and the questionnaires
[b] - this is the same as the NCI study variable 'smkstat' used to describe ever- or never-smoker

The prevalence of smoking in reallocations #1 and #2 are similar, 50.4% and 49.4% respectively. Reallocation #3, which assumes both are unknown, has a smoking prevalence of 47.8%. The percent of unknowns after allocation is 20.8% in schemes 1 and 2 but increases to 25% in scheme 3. Once the three new variables have been created, there are still employees missing a smoking status.

4.2.2. Reallocation of unknowns

The question still remains as to how to assign a smoking history to the remaining employees with a missing smoking status. If the lung cancer cases are stratified by cumulative AN exposure group and smoking status, crude odds ratios for smoking and lung cancer can be calculated within each AN exposure group, ignoring the unknowns. If the goal is to raise the overall lung cancer risk due to smoking, the unknowns could be assigned in such a way as to maximize this odds ratio within each stratum. If all of the unknown cases are assigned as ever-

smokers and all of the unknown controls are assigned as never-smokers, the diagonal cells of each stratum would increase. This would maximize the odds ratio within each stratum and for the total. Conversely, if all of the unknown cases were assigned as never-smokers and all of the controls were assigned as ever-smokers, the crude odds ratio would be minimized. The most logical, and believable, assignment of the unknowns would be to assign the unknown controls and cases in the same proportions as the known controls and cases. Table 29 shows the crude odds ratios obtained for the original data, the minimum, the maximum, and the proportional allocation as described previously.

Table 29: Crude odds ratios obtained for the original data, and the minimum, maximum, and proportional allocation schemes

			AN cumulative exposure group																			Total		
			Unexposed			1			2			3			4			5						
			S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U
Original sub-cohort	lung cancer	case	18	3	38	4	0	24	8	0	17	8	0	19	11	1	16	8	3	15	57	7	129
		control	348	281	198	320	191	116	165	100	38	134	66	26	184	74	30	131	45	15	1282	757	423
	smoking prevalence (%)	case	85.7			100.0			100.0			100.0			91.7			72.7			89.1		
		control	55.3			62.6			62.3			67.0			71.3			74.4			62.9		
	lung cancer OR		4.84			-			-			-			4.42			0.92			4.81		
Minimum	lung cancer	case	18	41	-	4	24	-	8	17	-	8	19	-	11	17	-	8	18	-	57	136	-
		control	546	281	-	436	191	-	203	100	-	160	66	-	214	74	-	146	45	-	1705	757	-
	smoking prevalence (%)	case	30.5			14.3			32.0			29.6			39.3			30.8			29.5		
		control	66.0			69.5			67.0			70.8			74.3			76.4			69.3		
	lung cancer OR		0.23			0.07			0.23			0.17			0.22			0.14			0.19		
Maximum	lung cancer	case	56	3	-	28	0	-	25	0	-	27	0	-	27	1	-	23	3	-	186	7	-
		control	348	479	-	320	307	-	165	138	-	134	92	-	184	104	-	131	60	-	1282	1180	-
	smoking prevalence (%)	case	94.9			100.0			100.0			100.0			96.4			88.5			96.4		
		control	42.1			51.0			54.5			59.3			63.9			68.6			52.1		
	lung cancer OR		25.69			-			-			-			15.26			3.51			24.46		
Proportional	lung cancer	case	51	8	-	28	0	-	25	0	-	27	0	-	26	2	-	19	7	-	176	17	-
		control	458	369	-	393	234	-	189	114	-	151	75	-	205	83	-	142	49	-	1538	924	-
	smoking prevalence (%)	case	86.4			100.0			100.0			100.0			92.9			73.1			91.2		
		control	55.4			62.7			62.4			66.8			71.2			74.3			62.5		
	lung cancer OR		5.14			-			-			-			5.26			0.94			6.22		

S - ever-smoker
~S - never-smoker
U - smoking status unknown

The table uses the smoking status variable from the NCI paper, SMKSTAT. The crude lung cancer odds ratio (OR) using the original data is 4.81 without stratifying by AN exposure group. The minimum and maximum attainable ORs are 0.19 and 24.5, respectively. Even with this high OR of 24.5 overall for the maximum allocation, the OR in the uppermost AN exposure quintile is only 3.51. This is due to the 3 non-smoking cases in this stratum. The reallocation schemes #2 and #3 shown in Table 28 will lower this cell count to 2 non-smoking cases, which should allow a greater increase within this stratum. The proportionally allocated data yield similar results to the original, as expected, with slight differences due to rounding in the cells.

Although these tables are crude odds ratios based on the stratum specific frequencies observed in the data and were not analyzed with the proportional hazards models, they do show, in the case of the maximum allocation, it should be possible to increase the overall lung cancer relative due to smoking from 3.6 to between 8.0 and 10.0. It does also demonstrate that if the unknowns are simply allocated as ever- or never-smokers in the same proportion as the knowns, that the overall lung cancer RR would likely not be very different from the 3.6 in the original study.

4.2.3. Logistic regression model to predict smoking status for unknowns

Because the ever-, never- smoking status variable is binary (0 – never, 1 – ever), logistic regression can be used to determine which of the covariates are significant predictors of smoking status. Demographic variables such as the employee's year of birth, location of employment (plant), race, and gender will be included in the model. Lung cancer, which is likely to be a significant predictor of ever-smokers, will also be included. The baseline employee is a white male born in 1955 or later who worked in plant 8, does not have lung cancer, and was never

exposed to AN. The results of the logistic regression on the 2103 complete records, with the event ‘ever-smoker’ as the dependent variable, are shown in Table 30.

Table 30: Results of logistic regression to determine significant predictors of ever-smoker

Logistic regression — Number of obs = 2103; LR chi2(20) = 162.20; Prob > chi2 = 0.0000; Log likelihood = -1296.9712; Pseudo R2 = 0.0588

smoke	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lung	3.124805	1.29707	2.74	0.006	1.385157	7.04931
yob_lt_25	3.849399	.8348264	6.22	0.000	2.516467	5.888363
yob_25_35	3.908099	.7468675	7.13	0.000	2.687166	5.683772
yob_35_45	2.604454	.4547217	5.48	0.000	1.849703	3.667172
yob_45_55	1.871962	.318869	3.68	0.000	1.340614	2.613908
ancum1	1.286419	.1789655	1.81	0.070	.9794089	1.689666
ancum2	1.149289	.1847574	0.87	0.387	.8386717	1.57495
ancum3	1.277242	.2328886	1.34	0.180	.8934441	1.825907
ancum4	1.414466	.2435062	2.01	0.044	1.009379	1.982122
ancum5	1.354296	.2732655	1.50	0.133	.9119287	2.011251
female	.6406542	.082283	-3.47	0.001	.4980797	.8240404
nonwhite2	.8508262	.127886	-1.07	0.282	.6337216	1.142308
nonwhite3	3.245685	3.727275	1.03	0.305	.341826	30.81824
plant1	1.321757	.275449	1.34	0.181	.8785464	1.98856
plant2	.8864419	.1759991	-0.61	0.544	.6006867	1.308135
plant3	1.223061	.2712366	0.91	0.364	.791914	1.888939
plant4	2.501397	.5177097	4.43	0.000	1.667289	3.75279
plant5	1.453574	.2190033	2.48	0.013	1.081908	1.952916
plant6	1.399542	.2585697	1.82	0.069	.97437	2.010241
plant7	1.217748	.2399034	1.00	0.317	.8276861	1.791633

From Table 30, lung cancer, year of birth, and gender (female) are significant predictors of smoking status. Several of the plants and one of the AN exposure quintiles are also significant. Race is not. The odds ratio for ever-smoker decreases in the age groups, which suggests that the older employees were more likely to be smokers than the younger employees. This was also observed in the table of smoking status versus year of birth, which showed the same decrease in smoking prevalence as the employees got younger.

A logical next step would be to predict a smoking status for the 552 employees with missing smoking information based on this model. Because all of these covariates are known for the 552 employees, the model could be used to predict whether or not a specific employee was a smoker or non-smoker based on lung cancer status, year of birth, AN cumulative exposure group, gender, plant, and race. But using the predicted probabilities may not achieve the desired effect of increasing the overall lung cancer RR due to smoking.

4.2.4. Proposed reallocation of unknowns

Figure 4 summarizes the next proposed steps in the analysis.

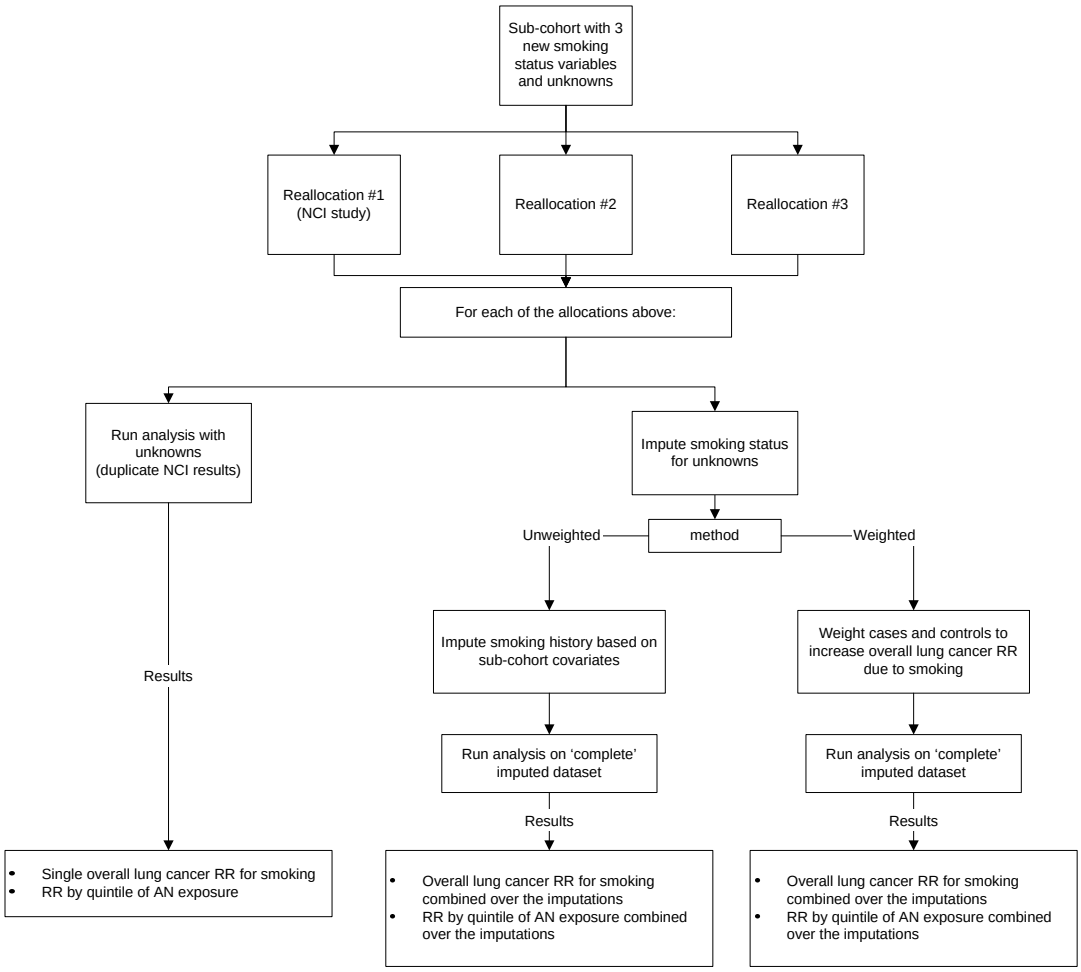

Figure 4: Flowchart of required analyses for the 3 reallocations

The first step will be to run the analysis of the case-cohort data using the three newly created smoking status variables. The q_smoke variable will reproduce the NCI study results. The other variables, m_smoke and unk_smoke will also be modeled. At this point, the analysis will include employees with an unknown smoking status, thus the goal of increasing the smoking status response rate to 100% will not be achieved. These individuals will drop out of the analysis because the smoking information is missing. The other approach will be to impute a smoking status for the unknowns and then analyze the sub-cohort as if the smoking status had been observed for all individuals. The dataset would be complete, but the overall lung cancer relative risk may not increase from the 3.6 reported in the NCI study. Therefore, 2 imputation schemes are proposed; an unweighted, which will use a method similar to the regression model discussed earlier and a weighted, which will apply more influence to certain individuals to force an increase to the overall lung cancer RR due to smoking. Details of the imputation process are discussed in the next section. The overall lung cancer RR due to smoking and the lung cancer RRs by AN cumulative exposure quintile will be obtained for each of the three allocation schemes using each of the three analyses: modeling with the unknowns, unweighted imputation, and weighted imputation.

4.3. Imputations

4.3.1. Description

Imputation of missing data is an approach that involves filling-in the missing data and then analyzing this imputed dataset as if there were no missing data (Rubin 1976; Schafer 1999; Little and Rubin 2002). In the NCI smoking sub-cohort, all of the employees with missing

smoking information were removed from any analysis using the smoking history covariate. This is called case-deletion (Little and Rubin 2002). If it can be assumed that the deleted observations are a representative sub-sample of the complete observations, then the case-deletion method may be reasonable. But in the NCI data, the proportions of employees with missing smoking histories are much different in the cases and controls, so the deleted cases are not a representative sub-sample of the complete data (the cases make up 23% (129/552) of the missing data but only 3% (64/2103) of the complete data).

For the NCI smoking sub-cohort, imputation would involve filling-in a smoking status for the 552 employees whose status was missing, and then running the proportional hazard regression on this imputed dataset to obtain the parameter estimates. The smoking status would be imputed using a model (e.g. logistic regression) of the complete case information with selected covariates (lung cancer status, gender, year of birth, etc.).

Multiple imputation is an extension of imputation where the imputation procedure is repeated m times. For each of the m independent imputations, an entire imputed dataset is created (Schafer 1999). For the NCI data, if $m = 5$, there would be 5 newly created datasets, where the employees with a missing smoking status would have an imputed smoking status, s_{m1}, s_{m2}, s_{m3}, s_{m4}, and s_{m5}. , where s_{mi} is the imputed smoking status for the i^{th} imputation, $i = 1, 2, 3, 4, 5$. The smoking status imputed for each of the missings will be generated independently for each imputation, based on the results of the modeling. The complete observations from the original data would be duplicated for each of the m datasets and would retain their original information. Each of the m datasets are analyzed separately and the results are combined using the methods discussed by Little and Rubin (2002) to obtain final estimates. For the estimates, the mean is used as the summary for the m imputations.

4.3.2. Unweighted imputation description

Several procedures have been developed in Stata by Royston (Royston 2004; Royston 2005a; Royston 2005b) to perform imputation of missing values. One of the procedures, **`ice`**, is designed for multivariate imputation, where there may be missing values for more than one covariate as well as the dependent variable (Van Buuren and Oudshoorn 1999). The other macro, **`uvis`**, is of interest here. It is a univariate imputation procedure designed to impute missing values for a single dependent variable given fully observed covariates (independent variables). The NCI data can be imputed using this program because only the smoking status (dependent variable) is missing and the covariates are fully observed (except for one employee in the controls whose year of birth was missing). This univariate imputation procedure uses a regression command, such as logistic regression, to predict the value of the outcome (dependent variable) based on the covariates in the regression model. For the NCI data, the outcome variable is smoking status, which is a binary outcome, so logistic regression is appropriate. Predicted values of the smoking status variable are calculated for those individuals with missing smoking information based on the regression of the complete data. The univariate imputation algorithm defined by Van Buuren et al. (1999) and incorporated by Royston (2004; 2005a) in Stata is as follows:

Assumptions:

- Let Y_{obs} be observations where the outcome is fully observed and Y_{mis} indicates the observations where the outcome is missing
- Missing values occur only in the outcome variable, Y_{mis}
- Dependent variables X are completely observed
- Let X_{obs} denote the predictors for the complete observations, Y_{obs}

- Let X_{mis} denote the predictors for the observations with a missing outcome, Y_{mis}

Steps:

- Calculate $W = \left(X'_{obs} X_{obs}\right)^{-1}$, $\hat{\beta} = WX'_{obs} Y_{obs}$, and $\hat{Y}_{obs} = X_{obs}\hat{\beta}$
- Bootstrap option:
    - Draw bootstrap sample of nonmissing observations
    - Regress the dependent variable Y_{obs} on X_{obs}
    - Estimate $\hat{\beta}_*$
- Calculate predicted values $\hat{Y}_{mis} = X_{mis}\hat{\beta}_*$
- For missing values $i = 1,..., n_{mis}$ find the observation where $\hat{Y}_{obs}$ is closest to $\hat{Y}_{mis,i}$
- Take Y_{obs} of that observation as the imputed value of i.
- Repeat steps m times to get $Y_{mis}^{(1)}, Y_{mis}^{(2)},..., Y_{mis}^{(m)}$, the multiple imputations of the outcome variable

4.3.3. Stata imputation procedures

The **uvis** procedure in Stata follows this algorithm (with the bootstrap option) and follows the syntax below (Royston 2005a):

```
uvis regression_cmd yvar xvarlist [if] [in] [weight],
     gen(newvarname)[boot seed(#)]
```

The relevant options in **uvis** are:

- *regression_cmd* – type of regression used to impute the missing Y values (logistic, conditional logistic, etc.)

- *yvar* – dependent variable, Y, for which values will be imputed (non-missing values will not be imputed)
- *xvarlist* – prediction covariates used in regression model
- **gen**(*newvarname*) – imputed Y values can be stored under a user defined *newvarname.* Contains the original (nonmissing) and imputed (originally missing) values of *yvar*.
- `boot` – specifies bootstrap sampling be used to obtain the predicted parameter estimate $\hat{\beta}_*$
- `seed(#)` – enables user to set seed so imputations can be reproduced

The logistic regression output showed which of the covariates in the dataset significantly predicted smoking status. They were: lung cancer status, year of birth, gender, and one of the AN exposure quintiles. These covariates will be used in the univariate imputations (Van Buuren et al. 1999). Although some of the plants were significant, plant will not be used in the prediction equation because it is not one of the covariates used for adjustment in the proportional hazards models run in the original NCI study.

4.3.4. Unweighted imputation for 3 allocations

The unweighted imputations were run in Stata for each of the three reallocations discussed previously and were performed 100 times, creating 300 additional smoking status variables. The **`uvis`** command was embedded in a do loop to iterate from 1 to 100 imputations.

The Stata code used is as follows:

Reallocation #1, using smoking status variable "q_smoke":

```
. forvalues i = 1(1)100 {
  2. uvis logistic q_smoke lung female ancum1 ancum2 ancum3 ancum4 ancum5  yob_lt_25
yob_25_35 yob_35_45 yob_45_55, gen(imp_qv2_`i') boot seed(9432`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

552 missing observations on q_smoke imputed from 2103 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

The **uvis** procedure is called 100 times and will impute values of "q_smoke" based on the lung cancer, gender, AN exposure, and year of birth covariates. The new variable "imp_qv2_i" will contain the imputed values of smoking status for the 552 missing observations and the 2103 existing smoking statuses. The bootstrap option was specified and the seed was set at an initial value of 9432`i' and will be incremented from i = 1 to 100. The 2nd to last line from the output describes the process for the first imputation and states that "552 missing observations on q_smoke imputed from 2103 complete observations". This line is reproduced for each of the 100 imputations (but is not shown to conserve space). The other reallocations are similar except for the dependent variable and the names of the newly imputed smoking variables.

Reallocation #2, using smoking status variable "m_smoke":

```
. forvalues i = 1(1)100 {
  2. uvis logistic  m_smoke lung female ancum1 ancum2 ancum3 ancum4 ancum5  yob_lt_25
yob_25_35 yob_35_45 yob_45_55, gen(imp_mv2_`i') boot seed(49932`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

552 missing observations on m_smoke imputed from 2103 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

In this case, "m_smoke" is imputed using the same predictors.

Reallocation #3, using smoking status variable "unk_smoke":

```
. forvalues i = 1(1)100 {
  2. uvis logistic   unk_smoke lung female ancum1 ancum2 ancum3 ancum4 ancum5
yob_lt_25 yob_25_35 yob_35_45 yob_45_55, gen(imp_unkv2_`i') boot seed(3888`i')
```

```
  3. }
[imputing by drawing from conditional distribution with bootstrap]

663 missing observations on unk_smoke imputed from 1992 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

"unk_smoke" is imputed using the same predictors, but notice the difference in the number of imputed observations, 663. This is because the 111 discordant smoking statuses were assumed to be unknown in Reallocation #3.

The 100 new smoking status variables that were imputed for each of the 3 reallocation scenarios were then exported to text files along with the employee ID so these new variables could be matched to the employee in the SMK_NCI SAS dataset. The proportional hazards regression models can then be run for each of the newly imputed smoking status variables to get the lung cancer RRs or other parameter estimates for each imputation. These estimates would then be combined using Rubin's method to obtain the final set of estimates. And because there are no missing smoking status variables, all 2655 employees in the sub-cohort will be included in the analysis.

But the same issue of the lung cancer relative risk still arises. Because the complete observations are used to impute the missing values, the unknowns will be assigned as ever- or never-smokers in a similar manner as the complete observations, which alone yielded the overall RR of 3.6. Analysis of the imputed datasets would likely result in similar RRs. To overcome this, the smoking status for the unknown cases and controls will be imputed using separate regression models to force the lung cancer relative risk due to smoking into the 8.0 to 10.0 range.

4.3.5. Weighted imputation description

From analysis of the 2x2 tables shown earlier, the relative risk can theoretically be maximized by coding the unknown lung cancer cases as ever-smokers and assigning the unknown controls as never-smokers. This would be the most extreme assignment of the unknowns and would not be very believable, because the smoking prevalence in the complete information is not 100% in the lung cancer cases and 0% in the controls. A method of weighting is proposed that will incorporate both the AN cumulative exposure group and the case/control status into the weighting. Imputation will be carried out, but this time using a weighted logistic regression model as the prediction equation.

The weighted regressions will again use the q_smoke, m_smoke, and unk_smoke smoking status variables created for each of the three reallocation plans. Because the prevalence of smoking increases with the cumulative AN exposure and smoking is related to lung cancer, the cases in the upper quintile of AN exposure will be weighted so that an unknown case in that group is more likely to be imputed as an ever-smoker. Likewise, the controls in the upper quintile will be weighted in such a way that the individual will be more likely reassigned as a never-smoker than an ever-smoker.

Table 31 shows the proposed weighting scheme. The ever-smoking cases are applied weights of 1, 2, 4, 6, 8, and 10 for cumulative AN exposure quintiles 0 (unexposed), 1, 2, 3, 4, and 5, respectively. The never-smoking cases are assigned a weight of 1. In the controls, the ever-smokers are unweighted. The never-smoking controls are weighted in a similar manner as the ever-smoking cases, increasing in magnitude from 1 (unweighted) in the unexposed, to 10 in the upper quintile of AN exposure. The table shows the frequency of the complete observations within each of the AN exposure groups. The case weighted imputations will be based on 64, 64,

and 62 complete lung cancer cases for the variables q_smoke, m_smoke, and unk_smoke, respectively. The control weighted imputations will be based on 2039 complete observations for q_smoke and m_smoke, but only 1930 observations for unk_smoke.

Table 31: Proposed weighting for cases and controls for the 3 reallocation schemes

Case Weighting Applied to Complete Observations

AN	q_smoke				m_smoke				unk_smoke			
	ever-smoker		never-smoker		ever-smoker		never-smoker		ever-smoker		never-smoker	
exposure group	weight	obs.	weight	obs.	weight	obs.	weight	obs.	weight	obs.	weight	obs.
unexposed	1	18	1	3	1	18	1	3	1	18	1	3
1	2	4	1	0	2	4	1	0	2	4	1	0
2	4	8	1	0	4	7	1	1	4	7	1	0
3	6	8	1	0	6	8	1	0	6	8	1	0
4	8	11	1	1	8	11	1	1	8	11	1	1
5	10	8	1	3	10	9	1	2	10	8	1	2
		57		7		57		7		56		6

Control Weighting Applied to Complete Observations

AN	q_smoke				m_smoke				unk_smoke			
	ever-smoker		never-smoker		ever-smoker		never-smoker		ever-smoker		never-smoker	
exposure group	weight	obs.	weight	obs.	weight	obs.	weight	obs.	weight	obs.	weight	obs.
unexposed	1	348	1	281	1	332	1	297	1	326	1	275
1	1	320	2	191	1	321	2	190	1	309	2	179
2	1	165	4	100	1	166	4	99	1	156	4	90
3	1	134	6	66	1	131	6	69	1	127	6	62
4	1	184	8	74	1	183	8	75	1	175	8	66
5	1	131	10	45	1	122	10	54	1	121	10	44
		1282		757		1255		784		1214		716

The type of weight used in the upcoming weighted imputations is called a frequency weight in Stata (StataCorp. 2005). A simple example that demonstrates frequency weighting is shown in Table 32. The 2x2 table on the left shows a hypothetical crude odds ratio calculation for lung cancer and cigarette smoking. With no frequency weighting, the OR = 9.33. If a frequency weight of 10 is applied only to the ever-smoking lung cancer cases, as shown in the 2x2 table on the right, the odds ratio increases by a factor of 10 from 9.33 to 93.33. The frequency weight of 10 treats each of the ever-smoking cases as 10 individuals, so the cell count is multiplied by the frequency weight.

Table 32: Example of frequency weighting

No weighting

		smoker		
		never	ever	total
lung cancer	control	40	10	50
	case	15	**35**	50
	total	55	45	100
	OR =	(40*35)/(15*10) =		9.33

Frequency weighting = 10

		smoker		
		never	ever	total
lung cancer	control	40	10	50
	case	15	**350***	365
	total	55	360	415
	OR =	(40*350)/(15*10) =		93.33

* Frequency weighting treats each lung cancer ever-smoker as 10 individuals

If the frequency weight is added to the logistic regression prediction equation in the **`uvis`** procedure, the weighted observations will have more influence in predicting the smoking status of the unknowns. The frequency weighting is not adding observations to the dataset, but merely adding more influence to certain covariate combinations, such as ever-smoking lung cancer cases, that are used in predicting a smoking status for the unknowns.

4.3.6. Weighted imputation for 3 allocations

Performing the weighted imputations requires a few more steps than the unweighted imputation shown earlier. The basic steps are:

- Create control and case weights for each of the variables q_smoke, m_smoke, and unk_smoke according to Table 31
- Impute a smoking status for the q_smoke, m_smoke, or unk_smoke variables for the cases only, using the appropriate case weights
- Impute a smoking status for the q_smoke, m_smoke, or unk_smoke variables for the controls only, using the appropriate control weights

- Combine the case imputed and control imputed variables into a single smoking status variable for each of the three reallocations, q_smoke, m_smoke, and unk_smoke

100 weighted imputations were run for each of the three reallocations. The **`uvis`** command was embedded in a loop to iterate 1 to 100 imputations for the cases and then for the controls, using the appropriate weights. The Stata code used is as follows:

Reallocation #1, using smoking status variable "q_smoke":

Case imputation

```
. forvalues i = 1(1)100 {
  2. uvis logistic q_smoke lung  ancum1 ancum2 ancum3 ancum4 ancum5 [fw= wt_case_q] if
lung==1, gen(imp_q`i') boot seed(555`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

129 missing observations on q_smoke imputed from 64 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

Control imputation

```
. forvalues i = 1(1)100 {
  2.  uvis  logistic  q_smoke  cntr_ind   ancum1  ancum2  ancum3  ancum4  ancum5  [fw=
wt_ctrl_q] if cntr_ind==1, gen(imp_ctrl_q`i') boot seed(77789`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

423 missing observations on q_smoke imputed from 2039 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

For the case imputations, the logistic regression model is again used, but this time the only predictors in the model are lung cancer and the AN cumulative exposure quintiles. The **`[fw= wt_case_q]`** option specifies the variable that contains the case frequency weight, in this example the case weights for q_smoke would be assigned. The imputation is limited to the cases by the '`if lung==1`' command. As in the unweighted imputations, the bootstrap was used.

The controls were imputed using AN cumulative exposure groups but without the lung cancer variable as was done with the controls. The variable 'cntr_ind' was created and is just a binary indicator variable that equals 1 if the individual is a control and 0 if the individual is a case. This variable was created so the cases and controls could be weighted and imputed separately.

Reallocation #2, using smoking status variable "m_smoke":

Case imputation

```
. forvalues i = 1(1)100 {
  2. uvis logistic m_smoke lung  ancum1 ancum2 ancum3 ancum4 ancum5 [fw= wt_case_m] if
lung==1, gen(imp_m`i') boot seed(1212`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

129 missing observations on m_smoke imputed from 64 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

Control imputation

```
. forvalues i = 1(1)100 {
  2.  uvis  logistic  m_smoke  cntr_ind    ancum1  ancum2  ancum3  ancum4  ancum5  [fw=
wt_ctrl_m] if cntr_ind==1, gen(imp_ctrl_m`i') boot seed(8789`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

423 missing observations on m_smoke imputed from 2039 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

Reallocation #3, using smoking status variable "unk_smoke":

Case imputation

```
. forvalues i = 1(1)100 {
  2.  uvis  logistic    unk_smoke  lung    ancum1  ancum2  ancum3  ancum4  ancum5  [fw=
wt_case_unk] if lung==1, gen(imp_unk`i') boot seed(678`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

131 missing observations on unk_smoke imputed from 62 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

Control imputation

```
. forvalues i = 1(1)100 {
  2. uvis logistic   unk_smoke cntr_ind   ancum1 ancum2 ancum3 ancum4 ancum5 [fw=
wt_ctrl_unk] if cntr_ind==1, gen(imp_ctrl_unk`i') boot seed(789`i')
  3. }
[imputing by drawing from conditional distribution with bootstrap]

532 missing observations on unk_smoke imputed from 1930 complete observations.
[imputing by drawing from conditional distribution with bootstrap]
```

After combing the case imputations and the control imputations, the 100 new smoking status variables for each of the 3 reallocation scenarios were then exported to text files along with the employee ID so the new variables could be matched to the employee in the SMK_NCI SAS dataset.

5. RESULTS

5.1. Overall Lung Cancer RR Due to Smoking

5.1.1. Analysis of 3 reallocations – including missing data

The first goal of this analysis was to increase the overall relative risk for lung cancer due to smoking from the 3.6 reported in the NCI study to a range of 8.0 to 10.0. The results of running the proportional hazards models for the sub-cohort (without any imputation of the missing values) are shown in Table 33 along with the NCI study result. The dependent variable in the model is lung cancer with only one predictor, smoking status.

Table 33: Overall lung cancer RR due to smoking for the analyses without imputation of missing data

design	assignment of smoking status	outcome	predictor	lung cancer RR
NCI study	questionnaire	lung cancer	smkstat	3.6
NCI study reanalysis	questionnaire	lung cancer	smkstat	3.87
reallocation #1	questionnaire	lung cancer	q_smoke	3.87
reallocation #2	medical records	lung cancer	m_smoke	4.22
reallocation #3	unknown	lung cancer	unk_smoke	4.40

The NCI lung cancer RR of 3.6 could not be reproduced exactly; the reanalysis resulted in a slightly higher RR of 3.87. As should be expected, the lung cancer relative risks for the NCI study reanalysis with smkstat as the predictor and reallocation #1 with q_smoke as the predictor are identical. This is because the smkstat and q_smoke variables are the same. When the 111 discordant medical record and questionnaire observations are reallocated in #2 and #3, the RRs increase to 4.22 and 4.40, respectively. The increase is larger in reallocation #3, because those individuals with conflicting information were assumed to be unknown. Among the 111, there were 2 cases who, in q_smoke, were assigned a smoking history, but with the unk_smoke variable, the two cases would not have been modeled and assumed to have missing smoking histories.

5.1.2. Analysis of 3 reallocations – imputations

The overall RR can now be analyzed for the imputed data. Recall that for each of the three allocations, unweighted and weighted imputations were performed. This results in 600 new smoking status variables. For each variable, the proportional hazards model was run using SAS software to obtain the overall lung cancer relative risk due to smoking. For each of the six groups, the mean, median, percentiles, and other descriptive statistics can be calculated for the

100 relative risks obtained from the imputed datasets. Figures 5 – 10 show the distributions of the overall lung cancer relative risks for each of the three allocation schemes and weighting. The reference line in the box plots is the NCI study's 3.6 RR, the dashed line is the median.

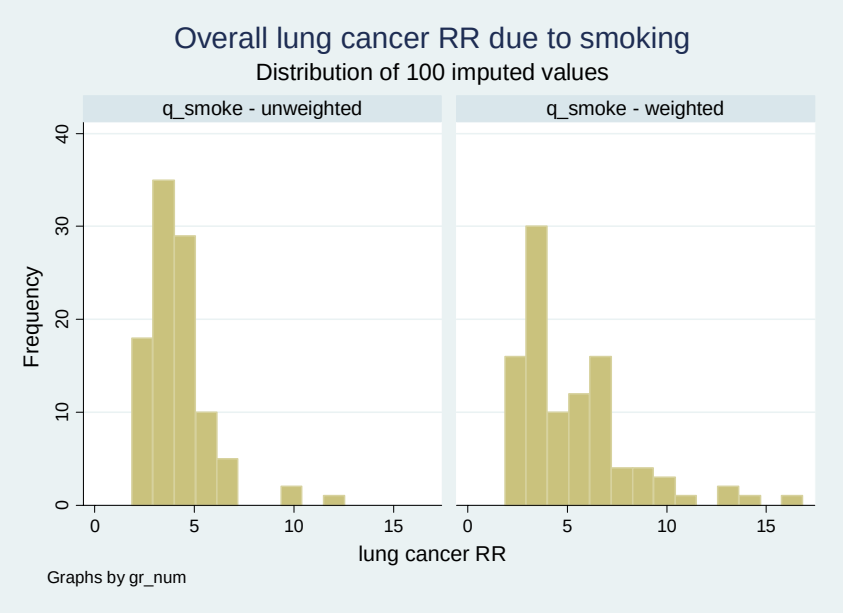

Figure 5: Distribution of 100 imputed RRs for unweighted and weighted q_smoke variable

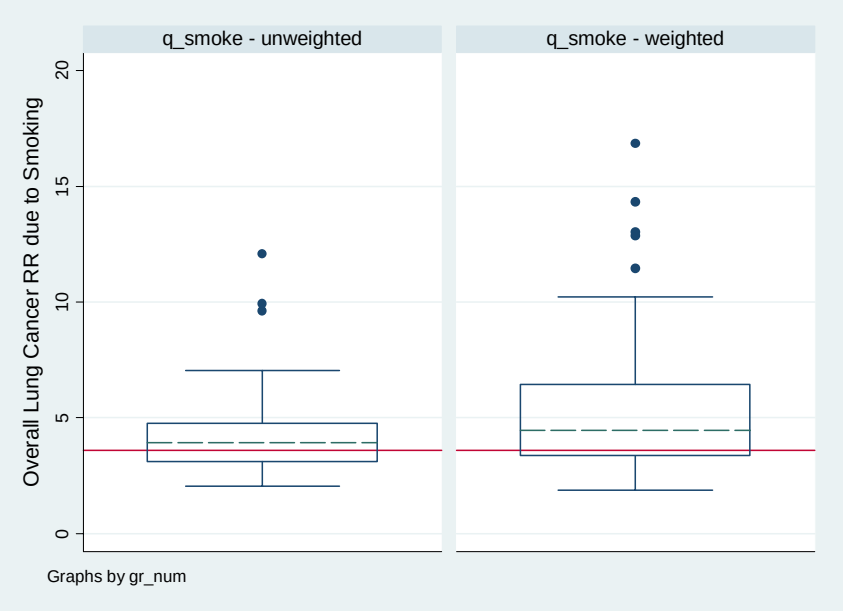

Figure 6: Boxplots of 100 imputed RRs for unweighted and weighted q_smoke variable

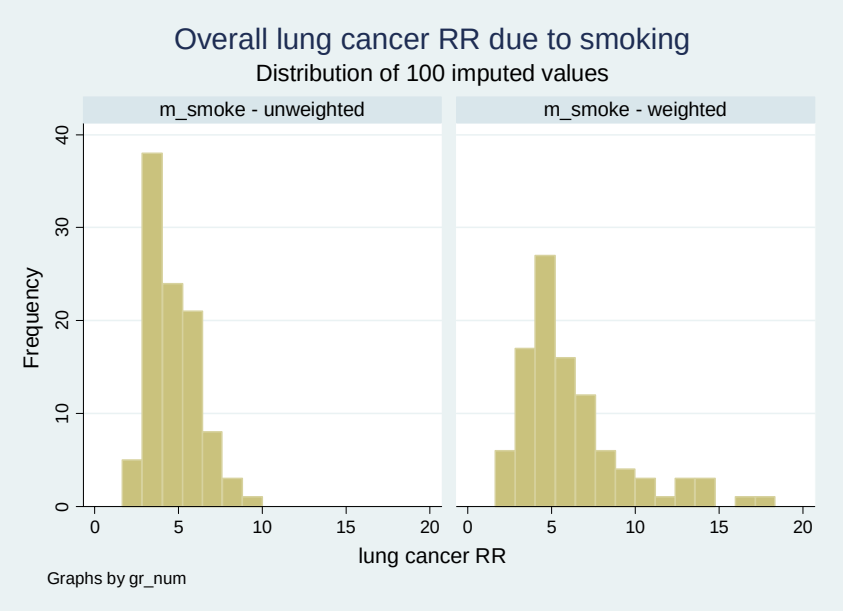

Figure 7: Distribution of 100 imputed RRs for unweighted and weighted m_smoke variable

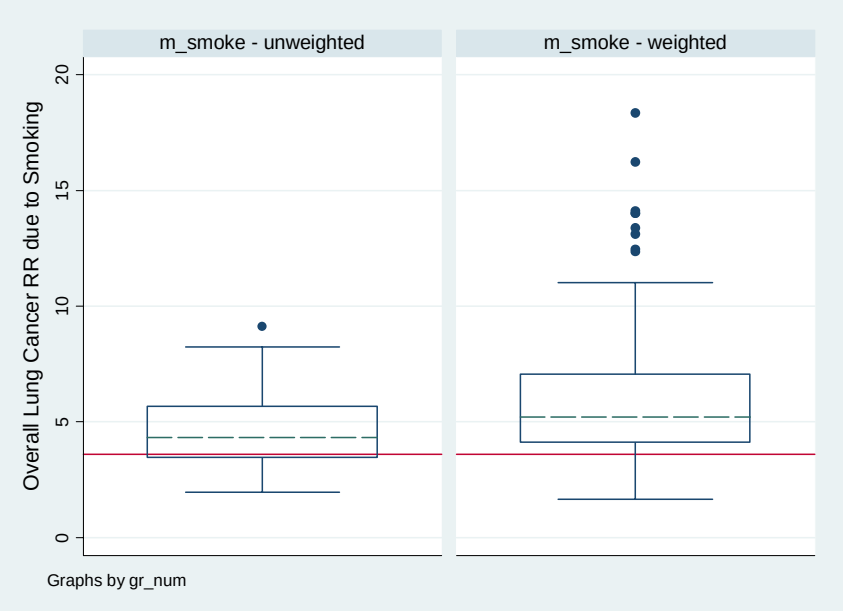

Figure 8: Boxplots of 100 imputed RRs for unweighted and weighted m_smoke variable

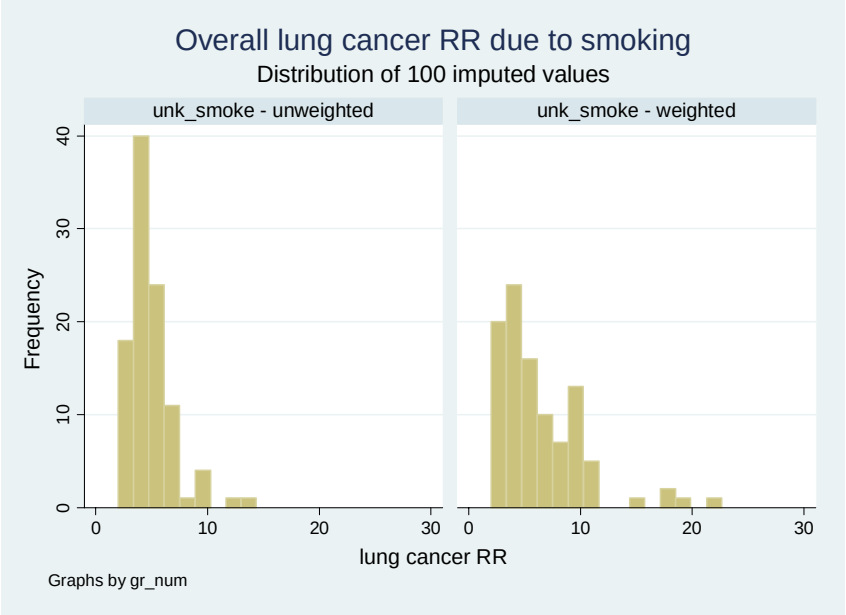

Figure 9: Distribution of 100 imputed RRs for unweighted and weighted unk_smoke variable

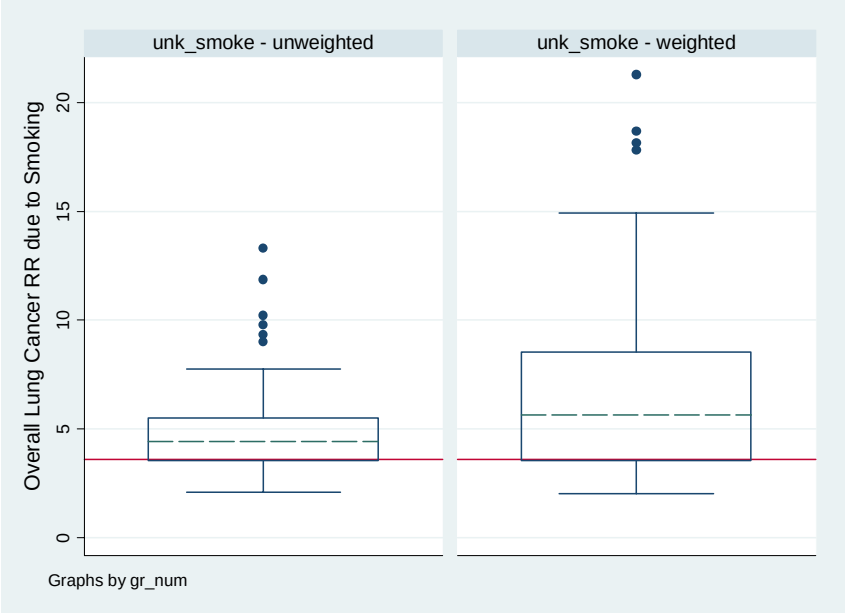

Figure 10: Boxplots of 100 imputed RRs for unweighted and weighted unk_smoke variable

Table 34 shows the descriptive statistics for the overall lung cancer RR for each of the allocation schemes and each of the imputation types (unweighted or weighted). The relative risks in the weighted imputations are higher than those of the unweighted imputations for each of the allocation schemes. This effect is most pronounced in reallocation 3, where the 111 discordant smoking histories were assumed to be unknown. The median RRs for the unweighted and weighted for unk_smoke are 4.41 and 5.62, respectively. None of the medians are up to the desired 8.0-10.0 value, but with the weighted m_smoke and weighted unk_smoke smoking variable, the 75th percentile value for the RRs are 7.06 and 8.53, respectively. This shows that in the weighted imputations for reallocations 2 and 3, an RR of 7.0-8.0 could occur with some regularity and not just happen due to chance alone.

Table 34: Descriptive statistics for the 100 imputed RRs by each reallocation

reallocation	imputed smoking variable	type of imputation	mean	min	max	percentiles				
						10	25	50	75	90
1	q_smoke	unweighted	4.17	2.04	12.07	2.50	3.10	3.93	4.76	5.95
		weighted	5.29	1.86	16.84	2.71	3.35	4.45	6.44	9.11
2	m_smoke	unweighted	4.64	1.95	9.12	3.00	3.44	4.33	5.66	6.92
		weighted	6.22	1.65	18.35	3.24	4.10	5.21	7.06	10.84
3	unk_smoke	unweighted	4.88	2.09	13.29	3.07	3.52	4.41	5.48	7.28
		weighted	6.37	2.01	21.29	2.80	3.52	5.62	8.53	10.26

The medians of the unweighted imputations are very comparable to the RRs from Table 33, which shows that the imputation of the smoking status based on the selected covariates yields similar results to the proportional hazard modeling of the complete observations. From Table 33, the RRs for q_smoke, m_smoke and unk_smoke are 3.87, 4.22, and 4.40, respectively. For the unweighted imputations in Table 34, the median RRs are 3.93, 4.33, and 4.41.

5.2. Lung Cancer RR in Upper Quintile of AN Exposure

5.2.1. Average cell counts for imputed data by reallocation and weight

The next part of the analysis is to rerun the proportional hazards models with the unweighted and weighted imputed values for each of the three reallocations schemes and see how the lung cancer RR due to AN exposure in the upper quintile is affected. Tables 35 and 36 show the average cell counts stratified by cumulative AN exposure group for the three allocations and each of the 100 unweighted and weighted imputations. The original sub-cohort (based on the NCI smoking status variable) values are also shown. Table 35 shows that the smoking prevalence increases by AN exposure quintile in a similar manner to the original data. The total crude odds ratios are also similar, only reaching a maximum of 5.59 for the unk_smoke allocation. The RR in the upper quintile of AN exposures ranges from 1.75 with the q_smoke variable to 2.98 using m_smoke. The assumption that the questionnaire information is correct never allows any of the imputed cell counts for the never-smoking cases to drop below the original sub-cohort value of 3, which prevents the RR in the upper stratum of AN exposure to increase significantly. When the discordant histories are assumed unknown (unk_smoke), the average count in this cell drops to 3.0 resulting in an RR of 2.74 due to smoking in the upper quintile.

The weighted imputation average cell counts are shown in Table 36. The two most noticeable differences are the smoking prevalence in the controls across the exposure groups and the lung cancer RR due to smoking in the upper quintile of AN exposure. In the m_smoke allocation, the smoking prevalence in the controls by exposure group is 53%, 60%, 59%, 61%, 66% and 65% in contrast to the original sub-cohort's 55%, 63%, 62%, 67%, 71%, and 74%.

This shows how the frequency weighting assigns more of the controls as never-smokers. The lung cancer RRs due to smoking in the upper quintile of AN exposure have now increased to 5.13 and 4.49 in the m_smoke and unk_smoke allocations, respectively. However, the total RRs have not significantly increased for any of the allocations in the weighted imputations.

Table 35: Average cell counts for the unweighted imputations by AN exposure group

			AN cumulative exposure group																			Total		
			Unexposed			1			2			3			4			5					
			S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U
Original sub-cohort	lung cancer	case	18	3	38	4	0	24	8	0	17	8	0	19	11	1	16	8	3	15	57	7	129
		control	348	281	198	320	191	116	165	100	38	134	66	26	184	74	30	131	45	15	1282	757	423
	smoking prevalence (%)	case	85.7			100.0			100.0			100.0			91.7			72.7			89.1		
		control	55.3			62.6			62.3			67.0			71.3			74.4			62.9		
	lung cancer OR		4.84			-			-			-			4.42			0.92			4.81		
q_smoke (average)	lung cancer	case	49	8.5	-	25	2.2	-	23	1.7	-	25	2	-	25	2.5	-	21	4	-	168	21	-
		control	450	360	-	387	227	-	186	111	-	149	72	-	201	81	-	139	48	-	1513	900	-
	smoking prevalence (%)	case	85.3			92.0			93.1			92.8			90.9			83.5			88.9		
		control	55.6			63.0			62.6			67.4			71.3			74.3			62.7		
	lung cancer OR		4.65			6.72			8.00			6.25			4.03			1.75			4.76		
m_smoke (average)	lung cancer	case	49	8.7	-	26	2	-	22	2.7	-	25	2	-	26	2	-	22	3	-	169	20.3	-
		control	429	382	-	390	225	-	187	110	-	146	75	-	200	83	-	130	58	-	1481	932	-
	smoking prevalence (%)	case	84.9			92.7			89.0			94.0			92.7			87.1			89.3		
		control	52.9			63.4			62.8			66.1			70.8			69.3			61.4		
	lung cancer OR		5.02			7.35			4.78			7.98			5.27			2.98			5.24		
unk_smoke (average)	lung cancer	case	50	8.2	-	26	1.6	-	23	1.7	-	25	2	-	26	2	-	23	3	-	171	18.1	-
		control	442	369	-	394	220	-	190	107	-	151	71	-	204	78	-	137	50	-	1517	896	-
	smoking prevalence (%)	case	85.8			94.2			93.1			94.0			92.7			88.2			90.4		
		control	54.5			64.1			63.9			68.0			72.3			73.2			62.9		
	lung cancer OR		5.05			9.05			7.62			7.33			4.89			2.74			5.59		

Table 36: Average cell counts for the weighted imputations by AN exposure group

			AN cumulative exposure group																			Total		
			Unexposed			1			2			3			4			5					
			S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U	S	~S	U
Original sub-cohort	lung cancer	case	18	3	38	4	0	24	8	0	17	8	0	19	11	1	16	8	3	15	57	7	129
		control	348	281	198	320	191	116	165	100	38	134	66	26	184	74	30	131	45	15	1282	757	423
	smoking prevalence (%)	case	85.7			100.0			100.0			100.0			91.7			72.7			89.1		
		control	55.3			62.6			62.3			67.0			71.3			74.4			62.9		
	lung cancer OR		4.84			-			-			-			4.42			0.92			4.81		
q_smoke (average)	lung cancer	case	49	8.7	-	24	3.2	-	22	2.5	-	24	3	-	26	1.9	-	22	4	-	167	22.6	-
		control	448	363	-	366	249	-	173	124	-	138	84	-	188	95	-	132	55	-	1443	970	-
	smoking prevalence (%)	case	84.9			88.4			89.8			89.8			93.1			85.9			88.1		
		control	55.3			59.5			58.1			62.1			66.4			70.5			59.8		
	lung cancer OR		4.57			5.17			6.33			5.37			6.79			2.55			4.95		
m_smoke (average)	lung cancer	case	50	8	-	24	3.1	-	22	2.6	-	24	3	-	25	2.1	-	23	2	-	168	20.8	-
		control	428	383	-	365	249	-	174	123	-	135	87	-	186	96	-	122	65	-	1410	1003	-
	smoking prevalence (%)	case	86.2			88.7			89.4			90.2			92.3			90.6			89.0		
		control	52.8			59.5			58.5			60.9			65.9			65.2			58.4		
	lung cancer OR		5.56			5.36			5.97			5.90			6.23			5.13			5.76		
unk_smoke (average)	lung cancer	case	50	8.4	-	24	3.4	-	22	2.2	-	24	3	-	25	2	-	23	3	-	168	21.1	-
		control	440	370	-	366	249	-	170	127	-	134	88	-	183	99	-	124	63	-	1417	996	-
	smoking prevalence (%)	case	85.5			87.6			91.1			90.6			92.7			89.8			88.9		
		control	54.3			59.5			57.2			60.3			64.9			66.2			58.7		
	lung cancer OR		4.96			4.82			7.63			6.32			6.87			4.49			5.61		

5.2.2. Proportional hazards models for imputed datasets

The imputed smoking status variables can now be modeled using SAS' proportional hazards regression procedure to obtain the RRs in each quintile AN exposure, adjusted for smoking (Appendix A). These results would be equivalent to the 1.6 reported in the NCI study (last column of row 4 of Table 1). Figures 11-13 show the distribution of the lung cancer RRs in the upper quintile of AN exposure for each of the allocations, both unweighted and weighted.

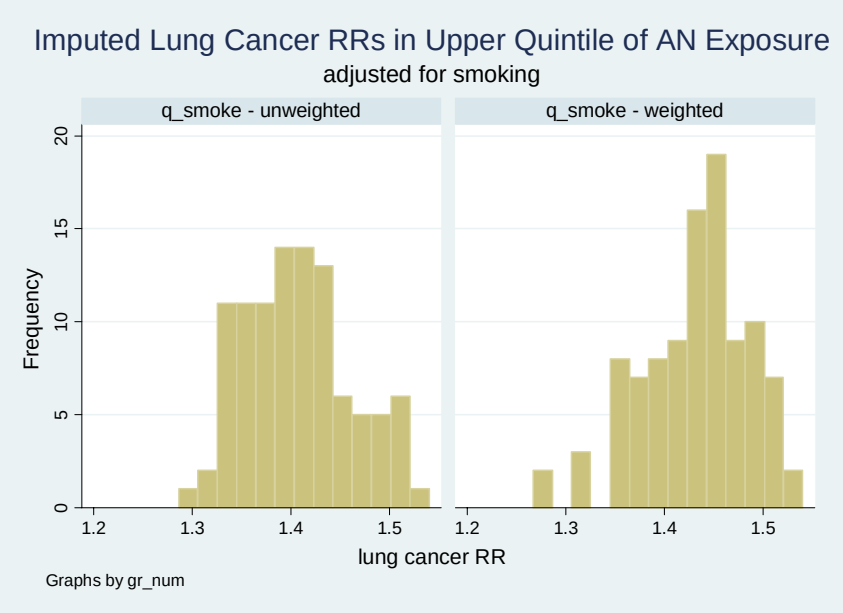

Figure 11: Distribution of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted q_smoke variable

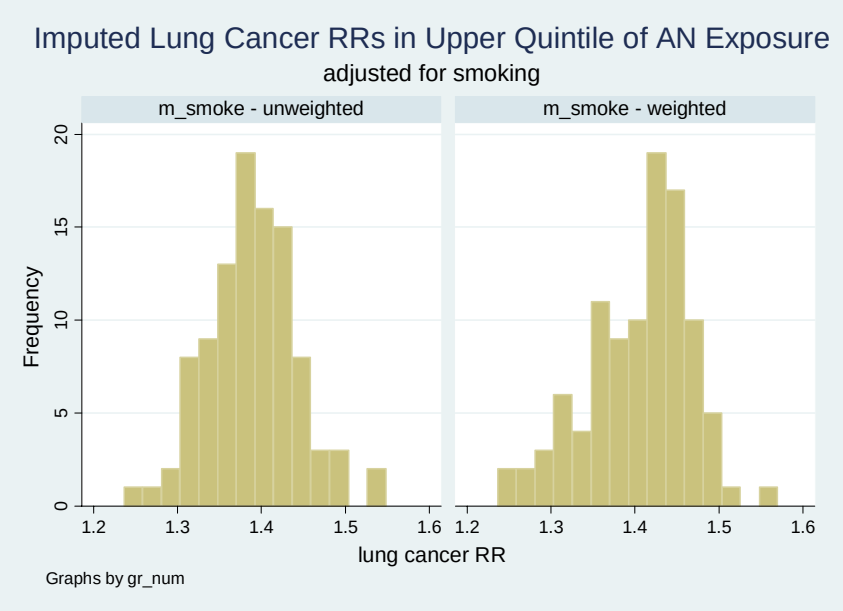

Figure 12: Distribution of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted m_smoke variable

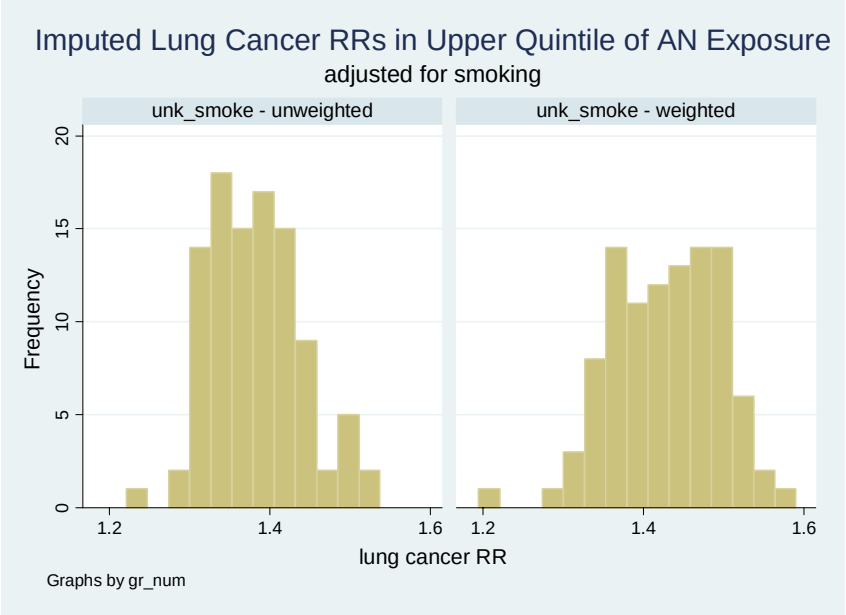

Figure 13: Distribution of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted unk_smoke variable

The histograms show that the distributions of these RRs appear to be more normally distributed than the overall RRs shown in Figures 5-10. These smoking adjusted RRs also appear to be lower in the unweighted imputations than the weighted. Because the weighting was designed to increase the overall lung cancer RR due to smoking, it seems that it has also slightly increased the lung cancer RR due to AN exposure in the upper quintile. Figures 14-16 are boxplots of the same RRs for the 100 imputations for each scenario. The reference line at 1.6 is the reported RR from the NCI study.

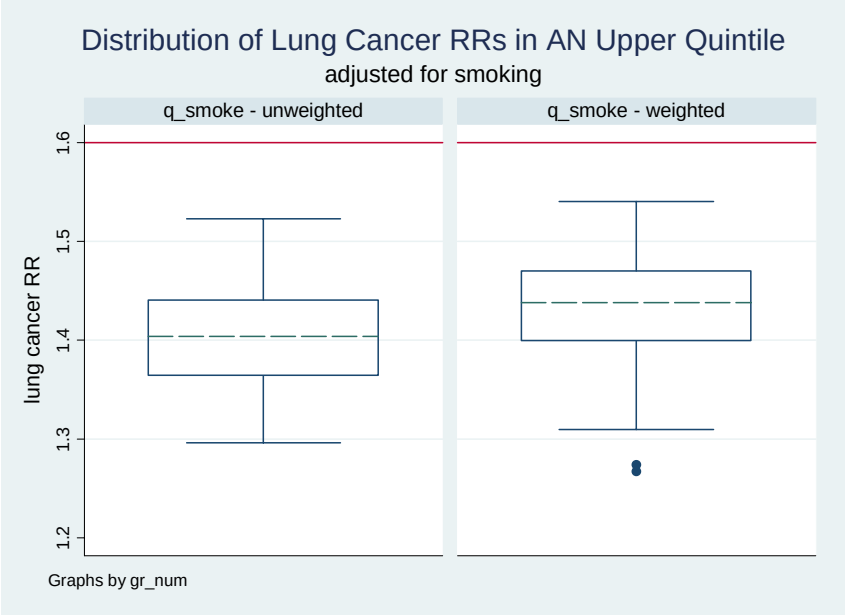

Figure 14: Boxplots of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted q_smoke variable

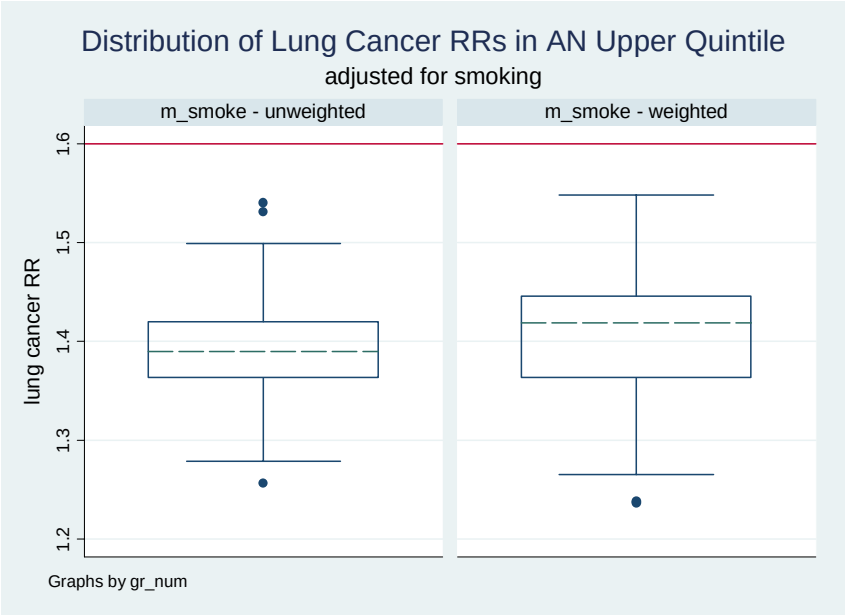

Figure 15: Boxplots of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted m_smoke variable

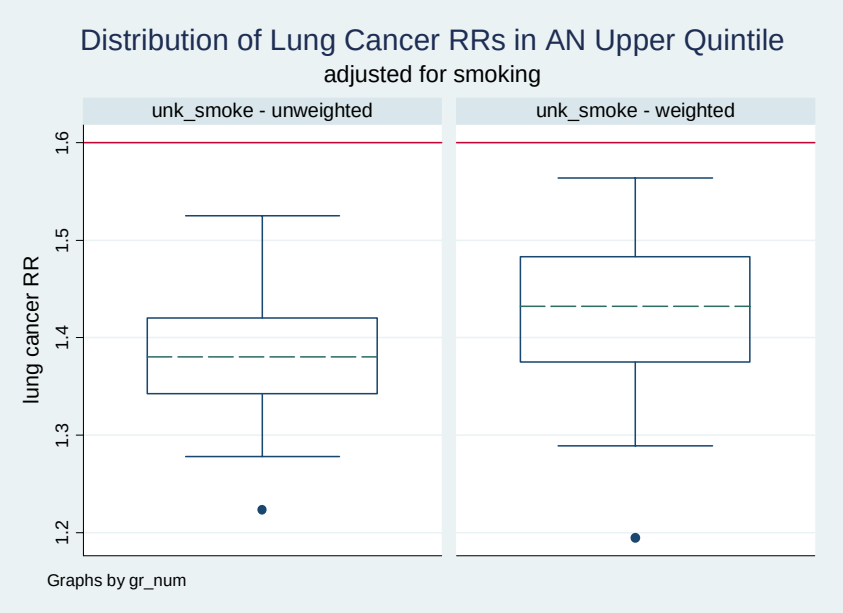

Figure 16: Boxplots of 100 imputed smoking adjusted RRs in the upper quintile of AN exposure for unweighted and weighted unk_smoke variable

For all of the lung cancer RRs due to AN exposure in the upper quintile, the unweighted RRs are slightly lower than the weighted RRs. None of the values approach the NCI's value of 1.6. Table 37 shows the mean, minimum, maximum, and percentiles for the 100 imputed relative risks by allocation for each of the exposure quintiles.

Table 37: Descriptive statistics for the 100 imputed RRs by quintile of AN exposure and allocation

Quintile of AN Exposure	re - allocation	imputed smoking variable	type of imputation	mean	min	max	percentiles				
							10	25	50	75	90
1	1	q_smoke	unweighted	0.830	0.756	0.882	0.793	0.810	0.833	0.849	0.864
			weighted	0.834	0.734	0.924	0.792	0.813	0.836	0.858	0.878
	2	m_smoke	unweighted	0.791	0.713	0.854	0.751	0.772	0.794	0.812	0.829
			weighted	0.787	0.679	0.883	0.737	0.763	0.791	0.814	0.838
	3	unk_smoke	unweighted	0.806	0.740	0.858	0.768	0.786	0.804	0.830	0.841
			weighted	0.809	0.697	0.888	0.746	0.783	0.812	0.848	0.861
2	1	q_smoke	unweighted	1.056	0.976	1.124	1.011	1.029	1.059	1.080	1.097
			weighted	1.095	1.002	1.180	1.042	1.068	1.091	1.126	1.146
	2	m_smoke	unweighted	1.015	0.898	1.102	0.976	0.991	1.014	1.040	1.065
			weighted	1.039	0.934	1.145	0.985	1.005	1.045	1.065	1.093
	3	unk_smoke	unweighted	1.016	0.885	1.130	0.969	0.988	1.019	1.042	1.062
			weighted	1.075	0.947	1.206	1.004	1.034	1.082	1.106	1.139
3	1	q_smoke	unweighted	0.968	0.875	1.091	0.913	0.946	0.966	0.992	1.020
			weighted	1.069	0.928	1.308	0.990	1.025	1.061	1.110	1.143
	2	m_smoke	unweighted	0.952	0.843	1.063	0.902	0.915	0.951	0.985	1.014
			weighted	1.054	0.882	1.230	0.956	1.012	1.056	1.101	1.152
	3	unk_smoke	unweighted	0.951	0.840	1.072	0.886	0.919	0.951	0.981	1.006
			weighted	1.079	0.890	1.274	0.968	1.017	1.068	1.143	1.194
4	1	q_smoke	unweighted	0.839	0.762	0.908	0.798	0.819	0.838	0.862	0.877
			weighted	0.812	0.716	0.905	0.763	0.780	0.812	0.838	0.863
	2	m_smoke	unweighted	0.831	0.751	0.899	0.794	0.812	0.833	0.852	0.867
			weighted	0.796	0.686	0.912	0.741	0.767	0.800	0.828	0.842
	3	unk_smoke	unweighted	0.821	0.721	0.895	0.776	0.798	0.820	0.849	0.869
			weighted	0.817	0.662	0.897	0.754	0.786	0.828	0.858	0.874
5	1	q_smoke	unweighted	1.407	1.296	1.523	1.337	1.364	1.404	1.440	1.491
			weighted	1.431	1.267	1.541	1.357	1.399	1.438	1.470	1.497
	2	m_smoke	unweighted	1.390	1.256	1.540	1.321	1.363	1.390	1.420	1.451
			weighted	1.405	1.236	1.548	1.318	1.363	1.419	1.446	1.475
	3	unk_smoke	unweighted	1.382	1.223	1.525	1.316	1.343	1.380	1.420	1.446
			weighted	1.427	1.195	1.564	1.339	1.375	1.432	1.483	1.509

Figures 17-19 show the distriubution of the smoking adjusted RRs plotted across the AN exposure quintiles (1-5) for the different imputations and allocations. It appears that the greatest difference in the unweighted and weighted plots within each reallocation scheme occurs for AN exposure quintile 3, where the RRs increase and show slightly more variation (boxplots are more spread out).

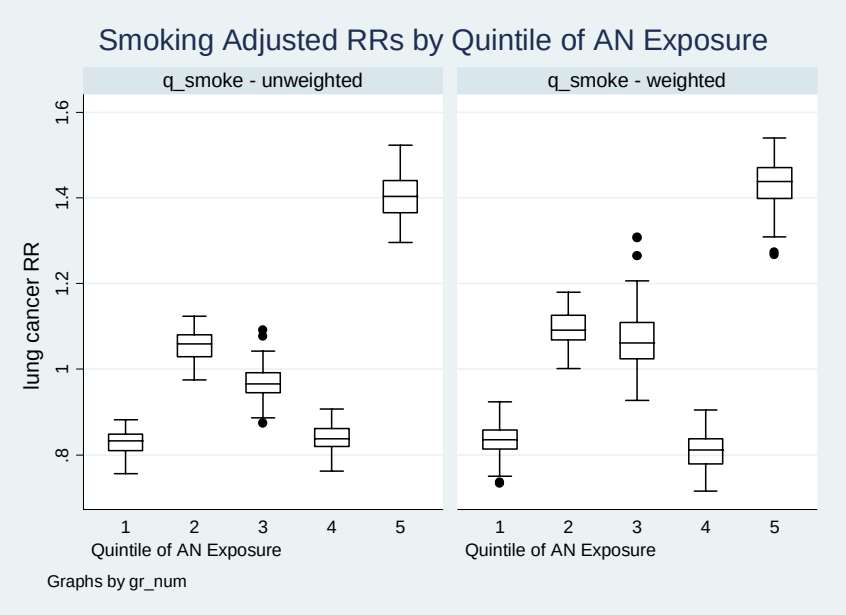

Figure 17: Boxplots of 100 imputed smoking adjusted RRs by quintile of AN exposure for unweighted and weighted q_smoke variable

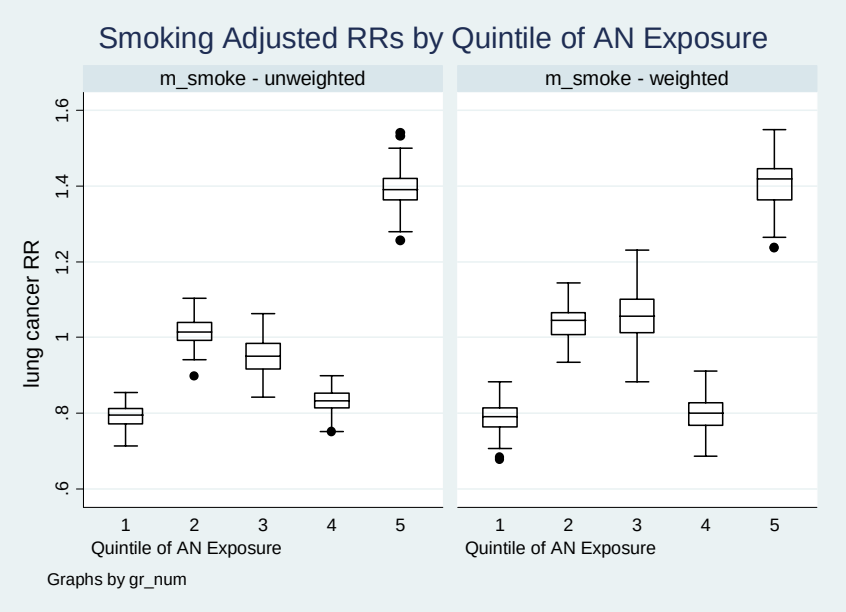

Figure 18: Boxplots of 100 imputed smoking adjusted RRs by quintile of AN exposure for unweighted and weighted m_smoke variable

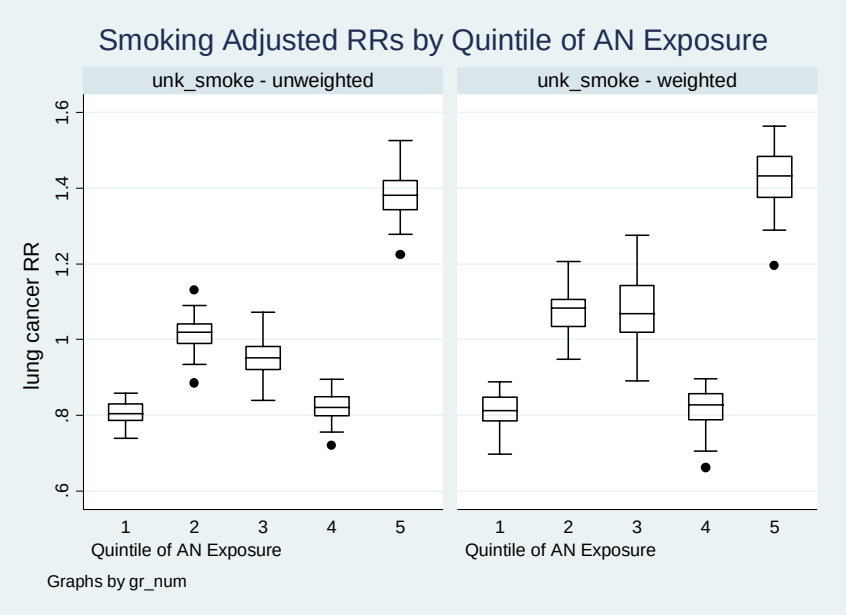

Figure 19: Boxplots of 100 imputed smoking adjusted RRs by quintile of AN exposure for unweighted and weighted unk_smoke variable

5.2.3. Comparison of RRs from NCI study to imputed datasets

The smoking adjusted RRs from the imputed data are compared with the NCI study results in Table 38. The upper half of the table shows the results of the NCI study and the reanalysis using the q_smoke, m_smoke, and unk_smoke variables that were used to determine the smoking status. Recall that in all four of these analyses, there was still a certain amount of missing smoking information in the dataset. Therefore, subjects without smoking information dropped out of the analysis. The biggest difference between these models is the lung cancer RR in the upper quintile of AN exposure for the unk_smoke variable, which treated the 111 discordant observations as unknown. Two of these were cases, so they would have been dropped from the analysis which decreased the number of strata in the risk set by two. The other RRs for AN exposure quintiles 1-4 for unk_smoke are similar to the other three analyses with the missing smoking histories.

The bottom half of Table 38 shows the mean values of the imputed RRs by cumulative AN exposure group for the different imputations and allocations. The RR in the lowest quintile of AN exposure jumps from 0.3 to around 0.8 for all reallocations. Quintile 2's RRs also see a slight increase from 0.8 to 1.0. Exposure groups 3 and 4 are similar. The upper quintile values are slightly lower, ranging from a minimum of 1.382 for the unk_smoke unweighted scenario to a maximum of 1.431 in the q_smoke weighted. The values from Table 38 are plotted in Figures 20-23.

Table 38: Smoking adjusted lung cancer RRs by quintile of AN exposure and allocation

Description			lung cancer RR, adjusted for smoking				
			AN exposure quintile				
			1	2	3	4	5
Analysis with missings	NCI study	-	0.3	0.8	1	0.9	1.6
	q_smoke	-	0.308	0.847	0.921	0.839	1.555
	m_smoke	-	0.289	0.809	0.903	0.827	1.518
	unk_smoke	-	0.296	0.756	0.921	0.853	1.388
Imputed missing smoking information[a]	q_smoke	unweighted	0.830	1.056	0.968	0.839	1.407
	q_smoke	weighted	0.834	1.095	1.069	0.812	1.431
	m_smoke	unweighted	0.791	1.015	0.952	0.831	1.390
	m_smoke	weighted	0.787	1.039	1.054	0.796	1.405
	unk_smoke	unweighted	0.806	1.016	0.951	0.821	1.382
	unk_smoke	weighted	0.809	1.075	1.079	0.817	1.427

[a] Mean RR values shown for imputed data

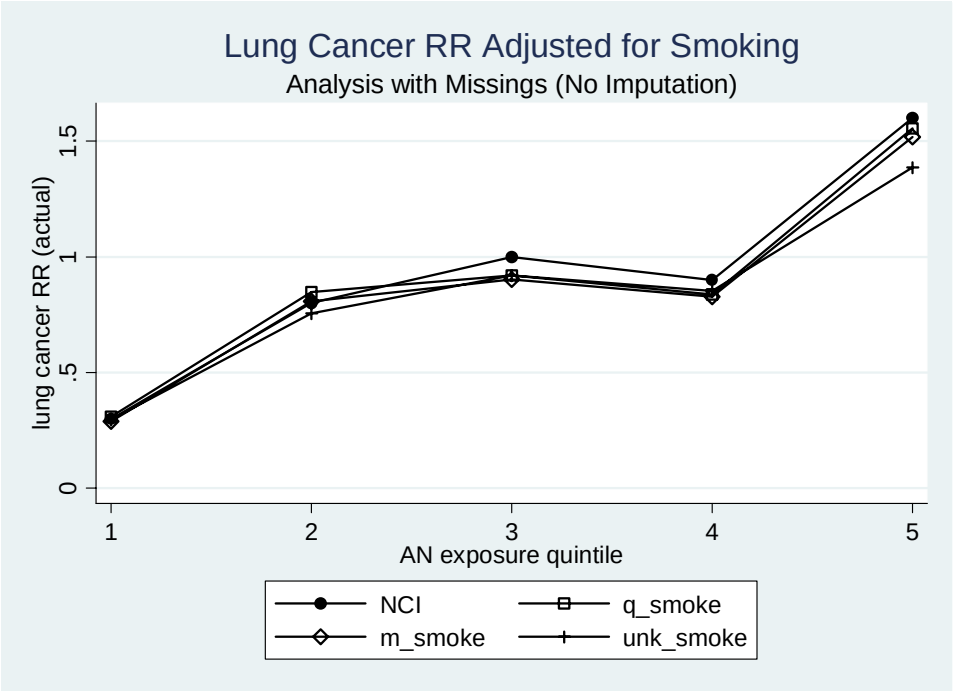

Figure 20: Smoking adjusted lung cancer RR by quintile of AN exposure for analysis models that included unknown smoking histories

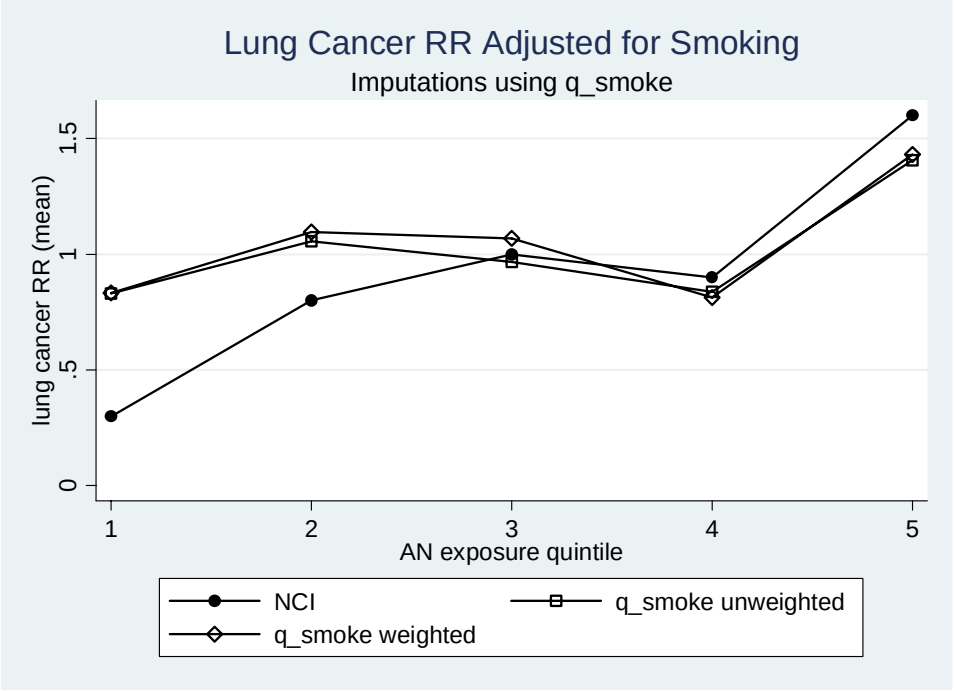

Figure 21: Smoking adjusted lung cancer RR by quintile of AN exposure for q_smoke imputations

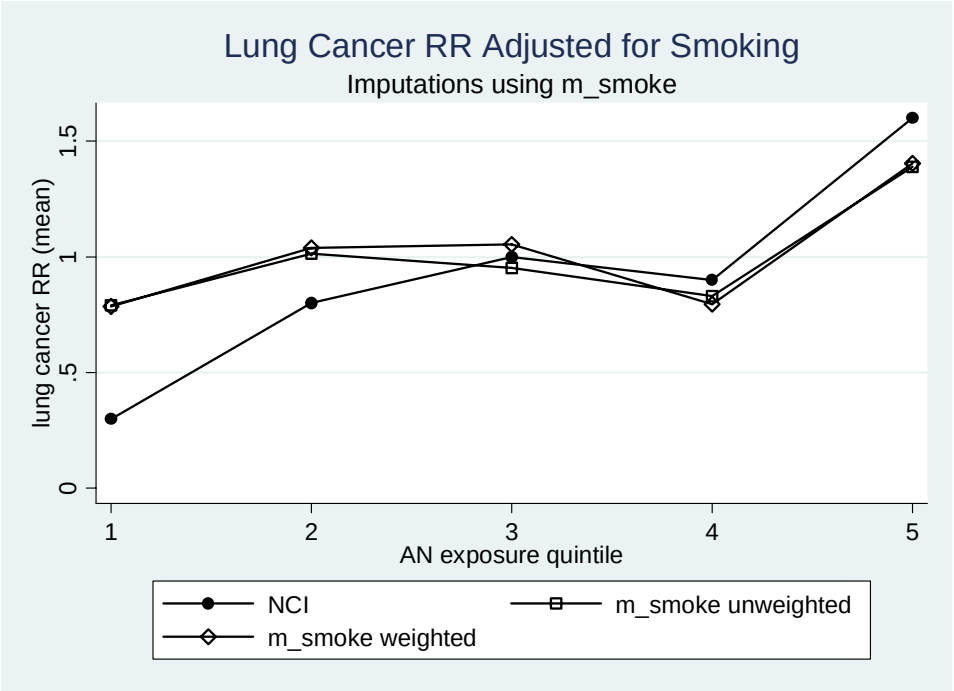

Figure 22: Smoking adjusted lung cancer RR by quintile of AN exposure for m_smoke imputations

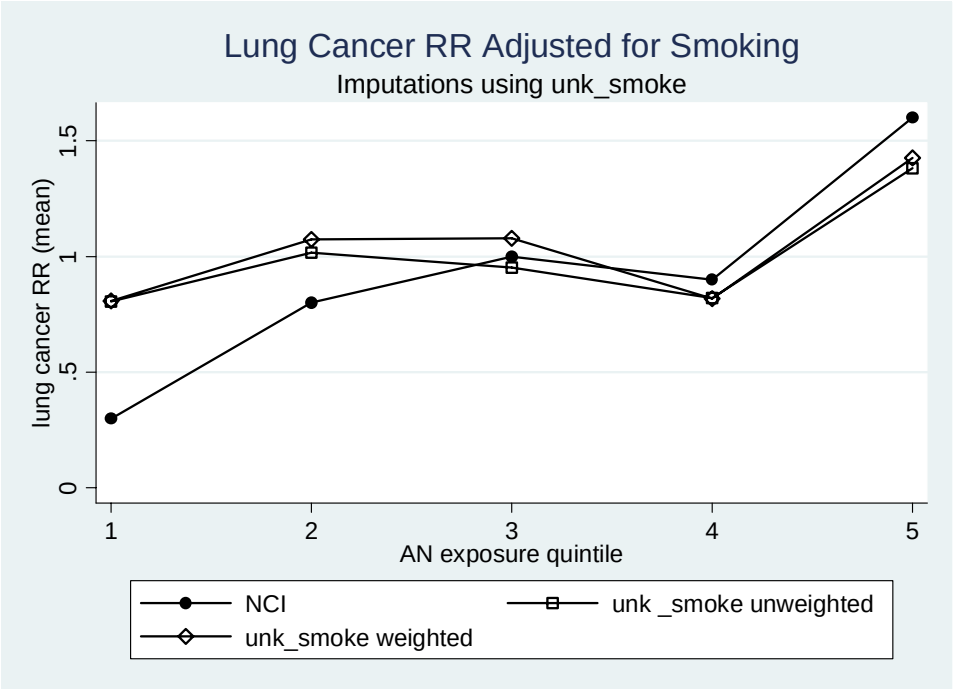

Figure 23: Smoking adjusted lung cancer RR by quintile of AN exposure for unk_smoke imputations

Figure 20 compares the three smoking status allocation schemes against the original NCI data. The missing values were included in these analyses, only the discordant smoking histories were reassigned. The plots are similar; the only notable difference is the slightly lower RR in the upper quintile of AN exposure for the unk_smoke allocation. Figures 21-23 show the mean values of the imputed lung cancer RRs (adjusted for smoking) by AN exposure quintile for each of the three allocation schemes. The original NCI study results are also shown in each figure. All plots of the imputed mean values show an increase in smoking adjusted RRs for quintiles 1 and 2 and a slight decrease in the upper quintile of AN exposure as compared to the NCI values. Lung cancer RRs in exposure quintiles 3 and 4 are similar for the imputations and the original study results.

5.2.4. Comparison of full cohort adjusted RRs from NCI study to imputed datasets

The final step in the reanalysis is to adjust the full cohort lung cancer RRs in the upper quintile of AN exposure. This adjustment involves applying the same proportional change observed in the sub-cohort smoking adjusted and unadjusted RRs to the full cohort unadjusted relative risk. In the NCI study, the change in the upper quintile of AN exposure for the full cohort, unadjusted to adjusted, was a drop of 1.5 to 1.4 (first and last rows of Table 1). The proportional change observed in the smoking sub-cohort, unadjusted to adjusted, was 1.7 to 1.6. Applying this same decrease to the full cohort resulted in the adjusted RR of 1.4 in the upper quintile of AN exposure. Table 39 shows the comparisons of the relative risks and the final adjustment for smoking.

Table 39: Final adjustments of the full cohort lung cancer RRs in the upper quintile of AN exposure

Design		Overall lung cancer RR due to smoking			Sub-cohort: Lung cancer RR in upper quintile of cumulative AN		Change (Adj/Unadj)	Full cohort: Lung cancer RR in upper quintile of cumulative AN	
		Actual	Mean	Median	Unadjusted for smoking[d]	Adjusted for smoking		Unadjusted for smoking	Adjusted for smoking[c]
Analysis with missing data	NCI[a]	3.6	-	-	1.7	1.6	0.94	1.5	1.4
	NCI reanalysis	3.87	-	-	1.73	1.55	0.896	1.5	1.34
	q_smoke	3.87	-	-	1.73	1.55	0.896	1.5	1.34
	m_smoke	4.22	-	-	1.73	1.52	0.875	1.5	1.31
	unk_smoke	4.40	-	-	1.58	1.39	0.876	1.5	1.31
Imputed missings[b]	q_smoke, unweighted	-	4.17	3.93	1.54	1.41	0.913	1.5	1.37
	q_smoke, weighted	-	5.29	4.45	1.54	1.43	0.928	1.5	1.39
	m_smoke, unweighted	-	4.64	4.33	1.54	1.39	0.902	1.5	1.35
	m_smoke, weighted	-	6.22	5.21	1.54	1.40	0.911	1.5	1.37
	unk_smoke, unweighted	-	4.88	4.41	1.54	1.38	0.896	1.5	1.34
	unk_smoke, weighted	-	6.37	5.62	1.54	1.43	0.926	1.5	1.39

[a] Results of original study only shown to one decimal place
[b] RRs shown are the mean values excpet as noted
[c] Adjusted value = (Unadjusted for smoking) x (change)
[d] With information on smoking

The first five rows show the results of the proportional hazards model run for the dataset that included individuals with missing smoking histories. The next portion of the table shows the mean and median of the overall lung cancer RRs for the six imputation schemes. The overall lung cancer RRs due to smoking do not achieve the desired increase to 8.0-10.0, with mean values ranging from 4.17 to 6.37 with the imputed missings. Of all designs, the overall smoking RR is highest in the un_smoke, weighted reallocation, with a mean of 6.37 and median of 5.62.

In order to apply the proportional adjustment to the full cohort, the unadjusted RRs in the upper quintile of AN exposure had to be found for the sub-cohort. These models are unadjusted for smoking, but are restricted to include only those individuals with observed smoking information. For the NCI analysis with the missing data and the smoking status allocations q_smoke, and m_smoke, this unadjusted RR is 1.73. These should be equal, because the number of missings is equal for these designs. The RR for unk_smoke, which had 111 more individuals with missing information, drops to 1.58. In the imputed designs, where smoking information

was available for all employees (imputed or originally known), the model only had to be run once to obtain an unadjusted RR of 1.54.

The 'change' column in Table 39 calculates the proportional change from the adjusted to the unadjusted relative risks. This is then multiplied (the method used by NCI) to the full cohort unadjusted RR in the upper quintile to obtain the full cohort adjusted RR due to AN exposure in the highest exposure group. The full cohort adjusted RRs in the upper quintile of AN exposure range from 1.31 to 1.39 compared to the NCI reported RR of 1.4. The lowest adjusted value of 1.31 actually occurs in the m_smoke and unk_smoke designs that did not involve any imputations.

6. DISCUSSION

The goal of increasing the overall lung cancer RR in ever- versus never-smokers was not achieved. The histograms in Figures 5, 7, and 9 do show some values that lie in the desired range of 8.0-10.0, so there are some imputations that resulted in the desired increase, although the mean and median RRs were not significantly greater than the original 3.6. The histograms also show that weighted imputations are not, except for a few outlying RRs, very much higher than the RRs for the unweighted imputations. This is most likely due to the frequency weights used for the controls. From Table 31, it can be seen that the weighted controls make up a much smaller proportion of the complete controls than do the cases. With only 64 complete cases, a frequency weight of 10 has much more of an effect than applying a frequency weight of 10 applied to 2039 complete controls. In order to force many more controls to be never-smokers, frequency weights of 10, 20, 40, 60, 80, and 100, for example, could be applied to the controls to

see if this overall relative risk could be increased even further. The case weights could also be increased, but because there are fewer cases and since most of the incomplete cases were assigned as ever-smokers already with the weights in Table 31, there would not be as much to gain by increasing the case weights as there would be with increasing the control weights.

The slight change to the lung cancer RR in the upper quintile of AN exposure after adjustment for smoking (1.5 to 1.4 reported by the NCI) was also not significantly different in any of the analyses when compared to the original study results. Although the sub-cohort smoking adjusted RRs decreased from 1.55 (reanalysis of NCI results) to 1.39-1.43 for the different allocations and imputation of the smoking histories, the full cohort smoking adjusted numbers were similar. This slight decrease in the sub-cohort adjusted RRs was counteracted by the decrease of the sub-cohort unadjusted RRs from 1.73 to 1.54. The proportional changes are therefore not as great, resulting in the similar full cohort adjusted values. For example, for the unk_smoke, weighted imputation design, the sub-cohort smoking adjusted mean RR was 1.38. The change is then 1.38/1.54, or 0.896. When multiplied to the unadjusted full cohort value of 1.5 (from NCI), the final adjusted RR is 1.34. This is not significantly less than the original NCI study result of 1.4.

From Table 39, the best combination of an increased RR overall and a lower adjusted RR in the upper quintile of exposure appears to be the m_smoke, weighted design. The mean overall RR is 6.22 and the full cohort RR adjusted for smoking is 1.37. But this value of 1.37 is very close to the NCI study (1.4) and could again be due to rounding.

Although the analysis of the imputed datasets did not yield results that were significantly different from the original NCI study, one very interesting result can be seen in Figures 21-23. The plots of the mean RRs for the three allocations appear to be much flatter than the plot of the

NCI study results, visually indicating less of an exposure-response trend for the imputed data than the analysis with the missing smoking histories. The effect is most pronounced in the lowest quintile of AN exposure, where the NCI RR of 0.3 increased to a range of 0.79 to 0.83 for the imputed analyses. The RR in exposure quintile 2 increased from 0.8 to slightly more than 1.00 for the imputed designs. The RRs in quintiles 3 and 4 are similar between the imputed and original designs. The upper quintile of AN exposure decreased from the NCI value of 1.6 to an average of 1.4 for the imputations. All analyses with the imputed missing smoking information have exposure-response plots that are pivoted clockwise around quintile 3, raising the RRs for quintiles 1 and 2 while lowering them for quintiles 4 and 5. When the NCI plot is compared with any of the imputed plots in Figures 21-23, the exposure-response trend is much less pronounced in the imputed designs than the original NCI results. And although the upper quintile of AN exposure RRs did not change markedly with the imputed data, analysis with the missing data appears to have had an effect on the lower quintiles of exposure, regardless of weighting or allocation.

Several areas could be considered for follow-up. Because the observed deaths by AN exposure quintile could not be reproduced exactly from the original study data, the authors could be contacted to try and resolve this issue. Additionally, the issue of rounding could be discussed. Because the NCI results are based on applying a proportionate change to a relative risk and then applying this change to an unadjusted RR, it would be important to have these values carried to 2 or possibly 3 decimal places so they could be more accurately compared with this reanalysis.

Another possible area for further study might be to carry out the proportional hazards modeling with only those smoking status imputations that yielded the high overall RRs due to smoking. For example, from Table 34 the 75^{th} percentile RR for the unk_smoke, weighted

allocation was 8.53. Because these values could possibly occur outside of chance alone, they could be observed in a dataset and not be outside the realm of believability. The proportional hazards model could be run on just these observations, the estimates averaged, and the adjusted and unadjusted comparisons could be performed. But this selective process would defeat the purpose of doing the imputations and could cast doubt on any results.

One other approach would be to assume that there is no smoking information at all and perform a Monte-Carlo simulation as proposed by Steenland and Greenland (2004). They have developed a method of analyzing occupational cohorts where smoking information was not gathered to determine if any confounding has occurred. The method involves assuming a distribution of the workers' smoking habits (which could be obtained from the sub-cohort), the distribution of lung cancer risk due to smoking, and an added bias parameter. In the NCI study, this would lead to a distribution of the lung cancer relative risks due to AN exposure, hypothetically adjusted for smoking. This could be compared to the full cohort analysis that was unadjusted for smoking to determine if any confounding due to smoking was likely. This method could be applied to the entire cohort as if no smoking information was available and the results could be compared with the NCI sub-cohort analysis to see if they agree.

Another step might involve a more detailed analysis if the individuals with missing smoking information. The amount of follow-up, for example, might be different for the individuals with missing data than those who had smoking histories.

7. CONCLUSIONS

Because the reported overall lung cancer RR of 3.6 appeared to be much too low and also due to the large amount of missing smoking information, particularly in the cases, it would appear at first glance that the results of the NCI study could be questioned. There was disagreement between some of the smoking histories and the possibility of individuals being misclassified could also have occurred. But from the results of the reanalysis with the imputations, it appears that the adjusted lung cancer relative risk in the highest quintile of AN exposure is in the range of that reported NCI study. The adjusted RRs are lower, but may be closer (or lower) due to the rounding used by NCI. Differences in the observed deaths and some of the slight differences in RRs between the NCI study and the results presented here must also be considered. Although a smoking adjusted RR in the upper quintile that drops from 1.5 to 1.31 after adjustment is a 13% decrease, it would be difficult to say these results are more accurate than those presented in the NCI paper.

From the reanalysis of the designs with the three allocation schemes and the two types of imputations (unweighted and weighted), it appears that raising the overall lung cancer relative risk due to smoking and significantly lowering the smoking adjusted lung cancer RR in the upper quintile of AN exposure are not plausible given the scenarios and weighting used in this reanalysis. When the overall RR is increased by the weighting, the RRs in the upper quintile also increase. Unweighted imputation values have the desired lower RR in the upper quintile, adjusted for smoking, but when applied to the full cohort, this decrease is not as dramatic and is close to the adjusted value reported in the NCI study.

It does appear as if the original NCI results were biased in the lower quintiles of AN exposure, where the lung cancer RRs changed significantly when the imputed data were analyzed. This change in the lower quintiles changed the appearance of the exposure-response plots, indicating less of a trend than was observed using the original NCI results. This flattening of the exposure-response curve, when combined with the only slightly elevated lung cancer RR in the upper quintile of AN exposure, provide evidence against an association of elevated lung cancer risk due to high AN exposures.

APPENDIX A

EXAMPLE SAS PROGRAM USED IN ANALYSIS

```
* The following program is typical of the programs used to compute the * RRs
for the reproduction of the NCI results and to calculate the RRs * for the
various imputations and weighting schemes. ;

* This program opens the original NCI dataset, smkanal, and reads in
* the appropriate text file for the smoking status information from
* Stata. ;

* The program below shows the weighted q_smoke allocation, but the
* program is identical for q_smoke, m_smoke, unk_smoke analysis with
* the missing data, and also for the the q_smoke, m_smoke, unk_smoke
* unweighted and weighted imputations.  The only differences are the
* name of the Stata text file read into SAS and the variable names.

* Also, the imputed results are sent to tab delimited text files for
* further analysis, whereas the analysis with the missing data did not
* have to iterate through 100 imputations. ;

* The risk set creation and proportional hazards procedure were taken
* from the proportional hazards macro from Ichikawa and Barlow (1998);
/*
================================================================
  Name: CASECOH.SAS (Survival Analysis with Robust Variance Matrix)
  Version: 1.0.2
  Authors: Laura Ichikawa and William Barlow
           Center for Health Studies, Group Health Cooperative
           Seattle, WA
  Origin date: January 1996
  Revision date(s): March 1998

  The IML portion of the program is derived from the SAS sample
  program PHR610EX.SAS in the SAS Sample Library.

  ================================================================
*/
```

```
* Open NCI dataset, SMKANAL ;
data nci_subco ;
      set 'c:\mikes files\thesis\smkanal'
            (drop = avg_an cigsday cigsday2 cumderml cuminhal dermal
            dermconc dermfreq drm16-drm68 drmdays3 dydrm16-dydrm68
            dyphy16-dyphy68 inhaled max16-max68 phy16-phy68 phydays2
            phys_exp cum_an ) ;
* Unneccessary variables were dropped ;
* Create id variable ;
      id = substr(cohort,2,5) ;
* Create variable to match with SMOKECUM - NCI Excel dataset ;
      plnt_id = substr(cohort,1,6) ;
run ;

* Read in the data from the stata imputed file, in this case the
q_smoke weighted imputations.  Only the employee ID and the imputed smoking
statuses are needed from Stata. ;
data from_stata ;
      infile 'c:\mikes
      files\thesis\Thesis_Jul_05\imputations\imputed_smoking_from_stata_qsmk_
      mc02.txt'
            dlm = '09'X firstobs = 2 missover lrecl = 3000 ;
      input cohort $ smk_q1-smk_q100 ;
run ;

* Sort both datasets to merge the imputed smoking statuses with the original
NCI data ;
proc sort data = from_stata ;
      by cohort ;
run ;
proc sort data = nci_subco ;
      by cohort ;
run ;
data nci_subco_imp ;
      merge from_stata nci_subco ;
run ;

* Will now redo the risk sets with the imputed info ;

* Create logic to get the event times ;
data events_imp ;
      set nci_subco_imp ;
      if original = 1 and vital in (1 2) then eagedays =
datdif(dob,dod,'actual') ;
      if vital in (1 2 3 ) and original ne 1 then eagedays =
datdif(dob,dod,'actual') ;
      if vital = 0 then eagedays = datdif(dob,exitdate,'actual') ;
      eventage = eagedays/365.25 ;
      age_strt = datdif(dob,hiredate,'actual')/365.25 ;
      age_stop = datdif(dob,exitdate,'actual')/365.25 ;
run ;
proc sort ;
      by ca_case original vital ;
run ;
proc sort ;
      by eventage ;
run ;
```

```
* Create the risk set ;
 data one_imp ;
    retain strat 0 ;
    set events_imp ;
    if vital in (1 3) then do ;
       strat + 1 ;
       evnttime= eventage ;
       t=1 ;
       wt=1 ;
       output ;
       do i=_n_+1 to n ;
          set events_imp point=i nobs=n ;
          if original=1 and (age_strt lt evnttime le age_stop)
          then do ;
             t=2 ;
             wt=10 ;  /* 10% subcohort */
             output ;
          end ;
       end ;
    end ;
 run ;

/*
   The risk set formation above was taken from the CASECOH.SAS macro
   cited at the beginning of the program (Ichikawa and Barlow 1998).
   Barlow (1994) described the weights that reflect subcohort
   membership.
   Controls in the subcohort are weighted inversely proportional to the
   sampling fraction. In this case the sampling fraction is 10%, so the
   weight is 10.
 */

* No have to get the AN exposure for the employees ;
data dates_imp ;
      set one_imp ;
      evage = evnttime ;
      evdays = evnttime*365.25 ;
      if expdate ne . then exage =
            datdif(dob,expdate,'actual')/365.25 ;
      evdate = dob+evdays ;
      rndage = floor(evage) ;
* need to be able to handle the events past 68 years ;
      if rndage ge 68 then rndage = 67 ;
      ev_yr_exp_days = floor((evage-rndage)*365.25) ;
      dempd = datdif(hiredate,evdate,'actual') ;
      if expdate ne . then days = datdif(expdate,evdate,'actual') ;
run ;
```

```
* Create arrays to allow for caclulation of exposure.  The original NCI
      data had exposure by year for each employee, which has to be
      converted to the exact days elasped within a year. ;


* Exposure calculations follow the method outlined by Marsh et al. (1998) ;

data calc_exp_imp ;
      set dates_imp ;
      array wrk {16:68} wrk16-wrk68 ;
      array dyexp {16:68} dyexp16-dyexp68 ;
      array expo {16:68} exp16-exp68 ;
      array aie {16:68} aie16-aie68 ;
      cum_an = 0 ;
      cum_wrk = 0 ;
      cum_exp = 0 ;
* Calculate aie from given exposures  ;
      do i = 16 to 68 ;
            if dyexp{i} = 0 then dyexp{i} = -1 ;
            aie{i} = expo{i}/dyexp{i} ;
            if aie{i} le 0 then aie{i} = 0 ;
            if dyexp{i} = -1 then dyexp{i} = 0 ;
      end ;
run ;
* Get actual exposure at event time ;
data actual_exp_imp ;
      set calc_exp_imp ;
      array wrk {16:68} wrk16-wrk68 ;
      array dyexp {16:68} dyexp16-dyexp68 ;
      array expo {16:68} exp16-exp68 ;
      array aie {16:68} aie16-aie68 ;
      array act_exp {16:68} act_exp16-act_exp68 ;
      exp_days = 0 ;
      wrk_days = 0 ;
      cum_an = 0 ;
      do i = 16 to (rndage-1) ;
            act_exp{i} = aie{i}*dyexp{i} ;
      end ;
* do loop if exposure age is before the event ;
      if exage = 0 or exage lt evage then do ;
      if dyexp{rndage} = 0 or dyexp{rndage} lt ev_yr_exp_days
            then dyexp{rndage} = dyexp{rndage} ;
            else dyexp{rndage} = ev_yr_exp_days ;
      end ;
* need logic for the first exposure year, if the exp is
      after the event ;
      if exage ne 0 and exage gt evage then dyexp{rndage} = 0 ;
      if evage gt exage and floor(exage) = floor(evage) then
            dyexp{rndage} = days ;
* now get act exp for the event year ;
      act_exp{rndage} = dyexp{rndage}*aie{rndage} ;
* make the rest of the exposures zero ;
      do i = (rndage+1) to 68 ;
            dyexp{i} = 0 ;
            act_exp{i} = aie{i}*dyexp{i} ;
      end ;
      do i = 16 to 68 ;
```

```
            cum_an = cum_an + act_exp{i} ;
            exp_days = exp_days + dyexp{i} ;
            wrk_days = wrk_days + wrk{i} ;
      end ;
run ;

* Create exposure quintiles ;
data risk_imp ;
      set actual_exp_imp
            drop = dyexp16-dyexp68 exp16-exp68 wrk16-wrk68
            act_exp16-act_exp68 aie16-aie68) ;
      ancum =(5/7)*cum_an/365.25 ;
* create dummy variables for ancum ;
      ancum1= 0 ;
      ancum2= 0 ;
      ancum3= 0 ;
      ancum4= 0 ;
      ancum5= 0 ;
* NCI original quintiles ;
      if ancum eq 0 then ancum1 = 0 ;
      if ancum gt 0 and ancum lt .13 then ancum1 = 1 ;
      if ancum ge .13 and ancum lt .57 then ancum2 = 1 ;
      if ancum ge .57 and ancum lt 1.50 then ancum3 = 1 ;
      if ancum ge 1.50 and ancum lt 8.0 then ancum4 = 1 ;
      if ancum ge 8.0 then ancum5 = 1 ;
run ;

* Calculate the Overall lung cancer RR due to smoking ;
* need to iterate by the number at the end to get all estimates in one
      file ;
%macro impute ;
%do i = 1 %to 100 ;
proc phreg data = risk_imp outest = q_imp&i noprint ;
   model t*t(2) = smk_q&i ;
   freq wt ;
   strata strat ;
   id cohort ;
run ;
%end ;
%mend impute ;
%impute ;

%macro comb_imp ;
 %do n= 2 %to 100 ;
data q_imp1 ;
      set q_imp1 q_imp&n ;
run ;
%end ;
%mend ;
%comb_imp ;
proc print data = q_imp1 ;
run ;

* Combine results of imputations ;
data unadj_rr ;
      set q_imp1 ;
      array smk_q {1:100} smk_q1-smk_q100 ;
```

```
      unadj_lung = 0 ;
      smk_q_imp = _n_ ;
      do i = 1 to 100  ;
            if smk_q {i} = . then smk_q {i} = 0 ;
            unadj_lung = unadj_lung + smk_q {i} ;
      end ;
      unadj_lung_rr = exp(unadj_lung) ;
      drop i _ties_ _type_ _status_ _name_ smk_q1-smk_q100 _lnlike_  ;
run ;
proc print data = unadj_rr ;
run ;
proc sort ;
      by unadj_lung_rr ;
run ;
* Get descriptives for the 100 imputed RRs ;
proc univariate ;
      histogram unadj_lung_rr ;
run ;
* Export the values to a text file ;
proc export data = unadj_rr outfile = 'c:\mikes
      files\thesis\Thesis_Jul_05\imputations\q_smk_overallRR_imps.txt'
      dbms = tab replace ;
run ;

* NCI paper reproductions ;
* Create variables needed in model ;
data tab9 ;
      set risk_imp ;
* create new race var, put all unknowns into 0 ;
      if race2 = 9 then race2 = 0 ;
      caltime = dobyr + evage ;
run ;

* table 9 line 2, cum exp for full smoking subcohort not adjusted for
smoking ;

proc phreg data = tab9 outest = est1 ;
   model t*t(2) = caltime ancum1 ancum2 ancum3 ancum4 ancum5 gender race2 ;
   freq wt ;
   strata strat ;
   id cohort ;
run ;
proc print data = est1 ;
run ;

* table 9 line 3, cum exp for full smoking subcohort with
smoking info, not adjusted for smoking ;

proc phreg data = tab9 outest = est2 ;
   model t*t(2) = caltime ancum1 ancum2 ancum3 ancum4 ancum5 gender race2 ;
   freq wt ;
   strata strat ;
   id cohort ;
   where smkstat ne 9 ;
run ;
proc print data = est2 ;
run ;
```

```
* table 9 line 4, cum exp for full smoking subcohort adjusted for smoking ;

proc phreg data = tab9 outest = est3 ;
   model t*t(2) = smkstat caltime ancum1 ancum2 ancum3 ancum4 ancum5 gender
race2 ;
   freq wt ;
   strata strat ;
   id cohort ;
   where smkstat ne 9 ;
run ;
proc print data = est3 ;
run ;

* Rerun the models using the 100 imputed smoking status variables ;
* Use the different smoking imputations - NCI table 9 line 4 ;

%macro tab9_q ;
%do i = 1 %to 100 ;
proc phreg data = tab9 outest = q_t9&i  ;
   model t*t(2) = smk_q&i caltime ancum1 ancum2 ancum3 ancum4 ancum5 gender
race2 ;
   freq wt ;
   strata strat ;
   id cohort ;
run ;
%end ;
%mend ;
%tab9_q  ;

%macro comb_t9_q ;
 %do n= 2 %to 100 ;
data q_t91 ;
      set q_t91 q_t9&n ;
run ;
%end ;
%mend ;
%comb_t9_q ;
proc print data = q_t91 ;
run ;

* array to get results in columns ;
data comb ;
      set q_t91 ;
      array smk_q {1:100} smk_q1-smk_q100 ;
      lung_beta = 0 ;
      smk_q_imp = _n_ ;
      ancum_1 = exp(ancum1) ;
      ancum_2 = exp(ancum2) ;
      ancum_3 = exp(ancum3) ;
      ancum_4 = exp(ancum4) ;
      ancum_5 = exp(ancum5) ;
      do i = 1 to 100  ;
            if smk_q {i} = . then smk_q {i} = 0 ;
            lung_beta = lung_beta + smk_q {i} ;
      end ;
      lung_rr = exp(lung_beta) ;
```

```
      drop i _ties_ _type_ _status_ _name_ caltime smk_q1-smk_q100
            gender race2 _lnlike_ ancum1-ancum5 lung_beta ;
run ;
proc print data = comb ;
run ;
proc sort ;
      by ancum_5 ;
run ;

* Print descriptives for upper quintile of AN exposure only ;
proc univariate ;
      histogram ancum_5 ;
run ;

* Send data to tab delimited file ;
proc export data = comb
      outfile = 'c:\mikes
      files\thesis\Thesis_Jul_05\imputations\q_smk_table9_imps.txt'
            dbms = tab replace ;
run ;
```

BIBLIOGRAPHY

Barlow, W.E. (1999). Robust variance estimation for the case-cohort design. *Biometrics* 50: 1064-1072.

Barlow, W.E., Ichikawa, L., Rosner, D., Izumi, S. (1999). Analysis of case-cohort designs. *Journal of Clinical Epidemiology* 52: 1165-1172.

Blair, A., Stewart, P.A., Zaebst, D.D., Pottern, L., Zey, J.N., Bloom, T.F., Miller, B., Ward, E., Lubin, J. (1998). Mortality of industrial workers exposed to acrylonitrile. *Scandinavian Journal of Work, Environment and Health* 24 suppl 2:25-41.

Ichikawa, L., Barlow, W. (1998). *User's guide to the survival analysis macro with robust variance.* Center for Health Studies, Group Health Cooperative; Seattle, WA, version 1.0.2.

International Agency for Research on Cancer (IARC). (1999). *Re-evaluation of some organic chemicals, hydrazine and hydrogen peroxide*. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 71. Lyon, France: IARC.

Little, R.J.A., Rubin, D.B. (2002). *Statistical analysis with missing data.* 2nd Ed. Hoboken, NJ: Wiley.

Marsh, G.M., Youk, A.O., Stone, R.A., Sefcik, S., Alcorn, C. (1998). OCMAP-PLUS: A Program for the comprehensive analysis of occupational cohort data. *Journal of Occupational and Environmental Medicine* 40(4):351-362.

Marsh, G.M., Youk, A.O., Collins, J.J. (2001). Reevaluation of lung cancer risk in the acrylonitrile cohort study of the National Cancer Institute and the National Institute for Occupational Safety and Health. *Scandinavian Journal of Work, Environment and Health* 27(1):5-13.

Occupational Safety and Health Administration (OSHA). "Acrylonitrile – (Organic Method #37)," Occupational Safety and Health Administration, U.S. Department of Labor. Available from http://www.osha.gov/dts/sltc/methods/organic/org037/org037.html.

Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 73: 1-11.

Rosner, B.A. (2000). *Fundamentals of biostatistics.* 5th Ed. Pacific Grove, CA: Duxbury.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal* 4(3): 227-241.

Royston, P. (2005). Multiple imputation of missing values: update. *Stata Journal* 5(2): 1-14.

Royston, P. (2005). "MICE for multiple imputation of missing values," 11th London Stata Users' Meeting, Stata Users Group. Available from www.stata.com/meeting/11uk/abstracts.html

Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63: 581-592.

SAS Institute Inc. (1999). *SAS OnlineDoc®, Version 8*. Cary, NC: SAS Institute Inc.

Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* 8: 3-15.

StataCorp. (2005). *Stata Statistical Software: Release 9.* College Station, TX: StataCorp Lp.

Steenland, K., Greenland, S. (2004). Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology* 160: 384-392.

Stewart, P.A., Zaebst, D.D., Pottern, L., Zey, J.N., Herrick, R., Dosemeci, M., Hornung, R., Bloom, T., Pottern, L., Miller, B.A., Blair, A. (1998). Exposure assessment for a study of workers exposed to acrylonitrile. *Scandinavian Journal of Work, Environment and Health* 24 suppl 2:42-53.

U.S. Department of Health and Human Services. (2004). *The health consequences of smoking: a report of the surgeon general.* U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

U.S. Environmental Protection Agency (EPA). "Technology Transfer Network Air Toxics Website," U.S. Environmental Protection Agency. Available from http://www.epa.gov/ttn/atw/hlthef/acryloni.html.

Van Buuren, S., Boshuizen, H.C., Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18: 681-694.

Van Buuren, S., Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Technical Report PG/VGZ/99.054, TNO Prevention and Health, Public Health, POB 2215, 2301 CE Leiden. Available from http://www.multiple-imputation.com.