University of North Texas Health Science Center
UNTHSC Scholarly Repository

Theses and Dissertations

12-2009

Association between Lung Cancer/Multiple Myeloma Mortality and Exposure to Oncogenic Viruses- Statistical Analysis Using Non-model and Model-based Statistical Methods and Various Control Sampling Schemes for Cancer Mortality in Occupational Cohorts

Harrison Ndetan
University of North Texas Health Science Center at Fort Worth, ndetanharry@yahoo.co.uk

Follow this and additional works at: http://digitalcommons.hsc.unt.edu/theses

Recommended Citation
Ndetan, H. , "Association between Lung Cancer/Multiple Myeloma Mortality and Exposure to Oncogenic Viruses- Statistical Analysis Using Non-model and Model-based Statistical Methods and Various Control Sampling Schemes for Cancer Mortality in Occupational Cohorts" Fort Worth, Tx: University of North Texas Health Science Center; (2009).
http://digitalcommons.hsc.unt.edu/theses/35

This Dissertation is brought to you for free and open access by UNTHSC Scholarly Repository. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UNTHSC Scholarly Repository. For more information, please contact Tom.Lyons@unthsc.edu.

ABSTRACT

Ndetan, Harrison Tatandam, M.Sc., MPH. Association between Lung Cancer/Multiple Myeloma Mortality and Exposure to Oncogenic Viruses – Statistical Analysis Using Non-model- and Model-based Statistical Methods and Various Control Sampling Schemes for Cancer Mortality in Occupational Cohorts. Doctor of Public Health (Biostatistics), December 2009; 119 pp., 9 tables, 7 appendices, 38 titles.

This study was designed to compare non-model- and model- based statistical techniques typically applied in cohort mortality analyses, and various schemes for selecting controls in nested case-control studies to document risk for lung cancer and multiple myeloma mortality, among workers in poultry slaughtering/processing plants. These workers are conceived to have a high exposure to oncogenic viruses compared to the general public. Data from the ongoing Cancer Risk in Workers Exposed to Oncogenic Viruses (CRIWETOV) project for members in a local Union Pension Fund belonging to the United Food &Commercial Workers (UFCW) international union, and followed–up for mortality from January 1, 1972 to December 31, 2003 were used for analyses. This cohort comprised of two large groups: poultry slaughtering/processing and non-poultry workers. The statistical methods applied were direct and indirect standardizations, Poisson, Cox proportional hazards, and binary/multiple logistic regression models and the sampling schemes for selecting controls were the cumulative survival, cumulative incidence, case-cohort, and incidence density sampling schemes. The entire cohort and

sub groups of poultry and non-poultry separately had higher risks of mortality from both malignant diseases (statistically significant for lung cancer) compared to the United States' general population, but slightly lower (statistically not significant) risks among poultry compared to non-poultry workers. Results of comparative effect measures from the various statistical methods under consideration were similar with a very slight difference in variability/precision within the cohort analyses. The effect measures were also similar for nested case-control analyses that applied the cumulative survival, cumulative incidence and case-base sampling schemes in selecting controls. However, the incidence density sampling scheme led to markedly different results (both in magnitude and statistical significance), that were more profound with the Cox regression model. Where the Cox model was not appropriate the interval Poisson (exponential) model was used and predictions were similar to those obtained using other methods.

ASSOCIATION BETWEEN LUNG CANCER/MULTIPLE MYELOMA

MORTALITY AND EXPOSURE TO ONCOGENIC VIRUSES –

STATISTICAL ANALYSES USING NON-MODEL- AND

MODEL-BASED STATISTICAL METHODS AND

VARIOUS CONTROL SAMPLING SCHEMES

FOR CANCER MORTALITY IN

OCCUPATIONAL COHORTS

Harrison Tatandam Ndetan, M.Sc., MPH

APPROVED:

Major Professor

Committee Member

Committee Member

Departmental Chair

Dean, School of Public Health

ASSOCIATION BETWEEN LUNG CANCER/MULTIPLE MYELOMA MORTALITY AND EXPOSURE TO ONCOGENIC VIRUSES – STATISTICAL ANALYSES USING NON-MODEL- AND MODEL-BASED STATISTICAL METHODS AND VARIOUS CONTROL SAMPLING SCHEMES FOR CANCER MORTALITY IN OCCUPATIONAL COHORTS

DISSERTATION

Presented to the School of Public Health

University of North Texas

Health Science Center at Fort Worth

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Public Health

By

Harrison Tatandam Ndetan, M.Sc., M.P.H

Fort Worth, Texas

December 2009

Copyright by

Harrison Tatandam Ndetan

2009

ACKNOWLEDGEMENTS

My very special and sincere thanks go to my supervisor, Dr. Sejong Bae, Professor of Biostatistics, who found time from his tight schedule to supervise this work. I wish to extend my appreciation to Dr. Eric S. Johnson, Professor and Chair, Department of Epidemiology and Principal Investigator of the CRIWETOV project for allowing me to work on his NIH funded projects which resulted in this dissertation, his constant direction through out the study, as well as serving on my dissertation committee. I also wish to express my heartfelt appreciation to Dr. Karan Singh, Professor and Chair, Department of Biostatistics, and Dr. Martha Felini, Assistant Professor of Epidemiology, for serving on my dissertation committee. My regards to all the faculty/staff of the Department of Biostatistics for the training and support accorded me during my tenure as a graduate student at UNTHSC.

My profound gratitude also goes to the management of Parker College of Chiropractic, Dallas - Texas, especially the Dean of research, Dr. Ronald Rupert, and Mr. Corboy Michael, of the Corboy's Investment Company, for the financial support they provided me through out my MPH and Dr.PH studies. I also wish to acknowledge with gratitude my indebtedness to Ka-Ming Lo, a doctoral student in the department of Biostatistics, for his assistance in overcoming some huddles in SAS programming.

TABLE OF CONTENTS

Page
LIST OF TABLES……..vi
LIST OF FIGURES……...vii

Chapter
I. INTRODUCTION…….1
Research problem
Background/ Rational
Hypotheses and objectives
Limitations

II. LITERATURE REVIEW……11

III. STUDY DESIGN…….17
Definition of cohort
Follow-up mechanisms and processes
Information on exposure, study end points and coding of disease
Preliminary results

IV. NON-MODEL- BASED STATISTICAL METHODS……...……25
Direct standardization
Indirect standardization

V. MODEL- BASED STATISTICAL METHODS….……37
Poisson regression model
Cox proportional hazards regression model
Logistic regression model

VI. SAMPLING SCHEMES FOR NESTED CASE CONTROL STUDY...……51

VII. DISCUSSION, RECOMMENDATIONS, CONCLUSION……...57

REFERENCES……...……..75

APPENDICES……..….87
A. SAS source code for ICD conversion (from ICD-6, 7, 8 & 10 to ICD-9, and coding of OCMAP impossible ICDs)

B. SAS source code for computing directly standardized rate ratio, SMR, relative SMR, indirectly standardized rate ratio from SMR and rate ratio as SMR by using one of the comparison group (the non-poultry) as the standard or referent.
C. Abstract1: Cancer Mortality in Poultry Slaughtering/Processing Plant Workers Belonging to a Union Fund
D. Abstract2: Update of Cancer Mortality in the Missouri Poultry Union
E. Abstract3: Mortality from Malignant Diseases – Update of the Baltimore Union Poultry Cohort
F. Abstract4: Mortality in Baltimore Union Poultry Cohort - Non-malignant Diseases
G. Abstract5: A pilot case-cohort study of lung cancer in poultry and control workers

LIST OF TABLES

Tables Page

1 Summary statistics, standardized mortality ratio (SMR), and proportionate mortality ratio (PMR) for selected causes of death for members of a local Union Pension Fund of UFCW International Union (Full cohort, N=30,488 using US standard rates from 1972-2005) ……....66

2 Standardized mortality ratio (SMR), and proportionate mortality ratio(PMR) for selected causes of death for members of a local Union Pension Fund of UFCW International Union (Full cohort, N=30,488 using US standard rates from 1972-2001)……..…….67

3 Five-year interval (15 categories) age group distribution of number at risk, lung cancer and multiple myeloma deaths in a local Union Pension Fund of UFCW (N= 30,488)……68

4 Distribution of number at risk and number of lung cancer/ multiple myeloma deaths in a Union Pension Fund of UFCW (N=30,488) by age, race, and sex….69

5 Summary statistics, standardized mortality ratio (SMR), and proportionate mortality ratio (PMR) for selected causes of death for members of a local Union Pension Fund of UFCW International Union (Sub comparative cohort, N=20,712 using US standard rates from 1972-2005)……70

6 Directly standardized death rate (per 100,000) and rate ratio[RR (95% confidence interval)] for poultry versus non-poultry exposures (Union Pension Fund of UFCW)……...71

7 Standardized mortality ratio (SMR), relative SMR, indirectly standardized death rates (per 100,000) and rates ratio (RR) with 95% confidence interval for poultry versus non-poultry exposures (Union Pension Fund of UFCW)……...72

8 Summary of comparative analysis of risk estimation for lung cancer and multiple myeloma mortality due to poultry versus non-poultry exposures applying various analytical techniques on the full Union Pension Fund Cohort (N=30,488)……...73

9 Summary of relative effect measures for lung cancer mortality due to poultry versus non-poultry exposures from a nested case-control design based on different control sampling schemes and analytical techniques (a local Union Pension Fund of UFCW)……..74

LIST OF FIGURES

Figures Page

1 Evaluation of the proportionality assumption for the Cox proportional hazards regression model……..86

CHAPTER 1

INTRODUCTION

Research Problem

A lot of efforts are currently being invested in cancer research and new carcinogens are discovered continuously. Researchers continue to make improvements in research methods (study design, implementation and analytical techniques) yet the complex causal web of cancer morbidity/mortality in humans has not been disentangled. Frequently, findings from one type of study design (especially retrospective studies) cannot be confirmed in others (such as large scale prospective designs or randomized trials). The choice of study design methodology and analytical techniques applied to empirical data may well be contributing factors to the observed discrepancies. This study presents another opportunity to investigate such discrepancies by applying some theoretically based models to empirical data.

Background/ Rational

For a while chemical composition of animal food has seen increased focus as cancer-causing agents. Unfortunately, a lot of findings from retrospective studies have not been confirmed in large scale prospective studies or randomized trials (Johnson, 1994 & 2005). Johnson and others have long hypothesized that although chemical composition of animal foods may be a necessary piece in the causal constellation for many malignant/nonmalignant diseases it may not be a sufficient condition and proposed an alternative approach. Their proposals centered on investigating the role of disease-causing biological agents potentially transmitted by animal foods in the etiology of these

diseases (Johnson, 1986, 1987, & 2005; Johnson, Zhou, Macodou et al., 2007; Netto & Johnson, 2003). There are a myriad of transmissible agents present in animals used as food (poultry, cattle, pigs, sheep) such as viruses, prions, bacteria, and protozoa with historical pathways to malignancy (Diseases of Poultry, 2003; Johnson, 1986 & 2005; Johnson, Nicholson, & Durack, 1995a; Johnson, Overby, & Philpot, 1995b). Particularly, oncogenic agents such as Bovine leukemia virus (BLV), Bovine papilloma virus (BPV), reticuloendotheliosis viruses (REV), avian leukosis/sarcoma viruses (ALSV), Jaagsiekte Sheep Retrovirus (JSRV), and Marek's disease virus (MDV) have been associated with a wide variety of tumors such as leukemia, lymphoma, sarcoma and other cancers in these animals (Carbone, Pass, Miele, & Bocchetta, 2003; Johnson, 1994 & 2005).

Focusing on poultry, human exposure to these viruses is widespread. These may include through the consumption of infected chicken/turkey/eggs and their products by the general population; occupational exposures during raising, slaughtering, processing, and preparation of birds/poultry products; public vaccination with vaccines prepared from chicken embryo cells; and in potential future gene therapy using some of these viruses to transport genes into cells or activate specific host genes (Johnson, 1994; Johnson & Zhou, 2007; Netto & Johnson, 2003). Particularly, poultry slaughterhouse workers are conceivably exposed to oncogenic viruses at a rate higher than the general population (Johnson, Shorter, Rider, & Jiles, 1997; Johnson & Zhou, 2007; Pham, Spencer, & Johnson, 1999). Concerns have, therefore, been raised as to whether these viruses (known to be cancer causing agents in birds/poultry) also cause tumors in humans (Johnson et al., 1986; Johnson, 1994; Johnson et al., 1995; Johnson et al., 1997; Johnson et al., 2007).

Despite great improvements in research methods over recent decades, discrepancies still persist between results obtained or conclusions drawn from various epidemiological cancer studies involving similar exposure-outcome relationships even with the application of similar study designs (Breslow & Day, 1987; Johnson, 2005). The roles of study design and analytical methods in the value of research can never be overemphasized. Various study design and statistical approaches have been applied in occupational cancer mortality studies (Breslow & Day, 1980; Breslow et al., 1987). Frequently, an exploratory cohort design (prospective or retrospective) is applied to quantify the effects of exposure on disease incidence or death rates for a large number of causes in a large population for which detail analysis is not feasible (Breslow, Lubin, Marek, & Langholz, 1983). Extensive follow-up and detailed analyses are subsequently carried out only for a few identified causes of interest using nested case-control designs with the application of various control sampling schemes (Breslow et al., 1983; Metayer, Johnson, & Rice, 1998; Morabia, Have, & Landis, 1995). Some (a majority of cohort mortality) studies employ non-model-based statistical methods (direct and indirect standardization) while others apply model-based regression techniques such as the Poisson, Cox proportional hazards, and logistic regression models (Breslow et al., 1983; Breslow et al., 1987). Although most of these approaches have been touted to yield almost similar estimates of effect measures, some concerns have been raised on the efficiency and bias in the selection of a particular method of analysis, and on practical issues that arise in application with empirical data. In response, adjustments are constantly being made to existing methods and many new approaches are still being developed and applied (Greenland, 1991 & 2004; Zou, 2004; Joffe & Greenland, 1995;

Langholz & Jiao, 2007; McNutt, Wu, Xue, & Hafner, 2003; Modern epidemiology, 1998). These approaches have varying underlying assumptions and may be suitable in some scenarios more than others. Application of sub-optimal design/statistical methods to particular study design may well be a contributing factor to the observed discrepancies in epidemiological and other findings regarding cancer incidence and mortality (Breslow et al., 1987; Greenland, 1999; Greenland, Schwartzbaum, & Finkle, 2000; Zhang & Yu, 1998). Thus, this study is centered on the tradeoff between efficiency and bias in the selection of a particular design/analytical method, and on specific practical issues that arise in the conduct of cancer mortality studies.

Research (both epidemiological and laboratory work) has demonstrated that both lung cancer, which is more common, (Alberg & Samet, 2003; Brouchet, Valmary, Dahan, Didier, & Galateau-Salle, 2005; Chen, Chiou, Sheu, Hsieh, Chen, Chen et al., 2001; Giuliani, Jaxmar, Casadio, Gariglio, Manna, D'Antonio et al., 2007; Miyagi, Tsuhako, Kinjo, Iwamasa, & Hirayasu, 2000; Syrjanen, 2002) and multiple myeloma, a rare form of cancer, (Avet-Loiseau, Gerson, Magrangeas, Minvielle, Harousseau, & Bataille, 2001; Chesi, Bergsagel, Shonukan, Martelli**,** Brents**,** 1998; Dib, Gabrea, Glebov, Bergsagel, & Kuehl, 2008; Moore & Chang, 1998; Rettig, Ma, Vescio, Põld, Schiller, Belson et al, 1997) have viral components. As a consequence, therefore, the focus here is investigating the use of different study designs and analytical approaches in studying the relationship between lung cancer/multiple myeloma mortality and exposures to oncogenic viruses among poultry slaughtering/processing workers belonging to a particular pension fund.

Hypotheses and objectives

It has been determined that exposures to oncogenic viruses commonly associated with poultry/meat exposures may be responsible for elevated mortality from lung cancer/multiple myeloma (Johnson, Ndetan, & L; 2009c; Johnson, Ndetan, Sarda, Bankuru, & Felini, 2009d; Johnson, Yau, Zhou, Singh, & Ndetan, 2009b; Johnson, Zhou, Yau, Prabhakar, Ndetan, Singh et al., 2009a; Preacely, Felini, Shah, Christopher, Sarda, Elfaramawi et al., 2009). Accordingly, workers belonging to a local union pension fund were classified into a poultry and a non-poultry (mostly seafood) group and analyzed exclusively for mortality from the above causes in relation to these exposures. However, the relevant question is whether the ratios of lung cancer/ multiple myeloma death rates among poultry versus non-poultry workers vary systematically (taken consideration of age, racial and gender disparities) or stay consistent when applying different study designs, sampling schemes and statistical approaches bearing in mind design-/model-specifications in a practical situation. This study used data from the ongoing Cancer Risk in Workers Exposed to Oncogenic Viruses (CRIWETOV) project by Eric S. Johnson, M.B., B.S., Ph.D., to investigate lung cancer and multiple myeloma risks (mortality) among poultry slaughtering/processing workers belonging to a local Pension Fund of the United Food &Commercial Workers (UFCW) international Union, followed–up for mortality from January1, 1972 to December 31, 2003. The study applied two study design methodologies (cohort and nested case-control), four control sampling schemes for the nested case-control design (traditional cumulative survival, cumulative incidence, case-cohort or case-based, and matched concurrent sampling or incidence

density sampling schemes), and five (two non-model- and three model-based) statistical methods typically applied in cancer mortality studies.

Thus, the study had the following two hypotheses:

Hypothesis I: There is an excess risk of lung cancer and multiple myeloma among members of a local union pension fund belonging to the UFCW International Union who are conceivably exposed to oncogenic viruses at a rate higher than the general population.

Hypothesis II: The various study designs and statistical approaches used in the analysis of cancer mortality studies should invariably lead to the same general conclusions regarding the exposure-outcome relationships under investigation.

In the light of the aforementioned, the specific aims of this study were: to compare measures of comparative risk (rate ratio, odds ratio, risk ratio, and/or hazard ratio) for lung cancer and multiple myeloma mortality separately (adjusted for age, race and sex) among workers in a local pension fund belonging to the UFCW International Union, exposed to poultry versus non-poultry derived from:

I). a cohort study design applying two non-model- based (directly standardized and indirectly standardized) methods of statistical analysis and three model-based methods (Poison regression, Cox proportional hazards, and logistic regression models), and

II). a nested case-control design based on the 1) traditional cumulative survival, 2) traditional cumulative incidence, 3) case-cohort or case-based, and 4) incidence density or matched concurrent sampling schemes for selecting controls while applying the three model-based statistical methods for analysis. Each design and analytical approach respected the underlying assumptions, where possible, with the empirical data.

Limitations

Preliminary analyses indicated that exposures to poultry products could be important contributing factors to malignancy. In these analyses, the focus has been on lung cancer and multiple myeloma, for which we wish to examine in detail the relationship to exposures and other explanatory variables using stratified analyses and techniques of multivariate statistical modeling. These diseases (both of which have viral components) were first identified in two preliminary analyses that compared the number of deaths in this cohort from each cause/exposure with those expected from national vital statistics computed using the Occupational Mortality Analysis Program (OCMAP+) with the United States standard mortality rates up to 2001 and up to 2005. As such, the interpretation of the subsequent multivariate analyses should take cognizance of the multiplicity of comparisons that were made in the preliminary analyses.

A major goal of cohort analysis is to quantify the effects of exposure on disease incidence or death rates. Standardized mortality ratio (SMR) is a classical effect measure obtained from stratified (non-model based) analyses used for this purpose. While rate ratios, risk ratios (or relative risks) and hazard ratios are some classical effect measures from various multivariate regression modeling (model based techniques), these are not directly comparable. The goal of this study is to compare measures of comparative risk for specific cancer types in two exposure groups using the above techniques. To achieve this goal SMR from indirect standardization methods were converted to rate ratios as appropriate or interpreted directly where suitable. Both age and calendar period have a strong influence on death rates for the major human diseases and are used frequently in SMR analyses. Cause-specific death rates from national vital statistics are usually

published by 5-year intervals of age and calendar year and have been shown to be constant within quinquinquennia (Breslow et al, 1983). This methodology was replicated in the preliminary analyses with OCMAP+ that identified the two causes of death (lung cancer and multiple myeloma) under investigation. However, the subsequent analyses that compared study designs and analytical techniques did not make considerations to calendar time and had stratified age into categories other than 5-year intervals, where necessary, as suited in the model under consideration so as to avoid sparse data and violation of specific underlying assumptions.

Selection bias is a common phenomenon in epidemiological studies. The cohort used for these analyses was well-defined and complete, being derived from a local Union Pension Fund of UFCW. The investigators of the CRIWETOV project had independently checked and confirmed the completeness of the cohort by cross-checking application records with union dues payment records, and extensive methods of follow-up were employed. As such selection bias was unlikely.

The measures of effect obtained in this study were adjusted for race, sex and age. Because the follow-up methods used to ascertain deaths were very extensive, subjects whose vital status was unknown after the end of follow-up were assumed to be alive at the end of study. Date of birth information was missing for 259 subjects (0.8% of the entire cohort); however, it was available for all deceased workers. Dates of birth for these persons were imputed based on the median year of birth of workers with known date of birth joining the union in a particular year. This measure was deemed to be associated with negligible bias, since the total person-years will be affected to a negligible degree. Race and sex were also artificially assigned at random to each individual in the study

without a death certificate/known cause of death, based on the racial/ gender distribution of deceased persons with known race/sex. Standardized mortality ratio (SMR) and proportionate mortality ratio (PMR) analyses are usually comparable. SMR analyses included data with artificially assigned values. PMR analyses in which complete information were available were performed to provide check for the SMR. The results from these two sets of analyses were in very close agreement, indicating that no serious bias was introduced because of missing information or race, sex and date of birth. This approach has been applied in previously published studies (Johnson, et al., 2009a; Johnson, et al, 2009b; Johnson et al, 2009c).

In this study, follow-up and exposure periods overlapped and so individuals with the greatest exposure tend to be those who have lived the longest. Methods of analysis that compare SMR among categories of workers defined by total years of employment, wherein each person's entire contribution to the expected number of deaths is assigned to that category in which he finds himself at the end of the study, are well known to be fallacious (Breslow et al., 1983; Breslow et al., 1987). However, this analysis did not use years of employment as this information was not updated for most of the subjects, as such individuals may change their exposure classification as they progress through the study thus allowing one to avoid this common pitfall.

This study initially used a retrospective cohort design to document risk of lung cancer and multiple myeloma mortality among workers of poultry slaughtering & processing plants. Such a design is usually not suitable for investigating specific occupational causes. Especially, it is not possible to control for non-occupational factors such as tobacco smoking and alcohol consumption that could as well account for risk.

The idea was to apply a nested case-control design to further elucidate risk while investigating the role of various sampling schemes of selecting controls. However, the nested case-control analyses did not have sufficient statistical power to adequately investigate the role of the exposures. Moreover, we were still not able to adequately control for non-occupational confounding factors.

Also the underlying assumptions for some of the models under investigation were hardly met by virtue of the data collected. However, the potential effects that these may have had on the observed results have been examined and discussed.

CHAPTER TWO

LITERATURE REVIEW

The role of disease-causing biological agents potentially transmitted by animal food in the etiology of malignant and other diseases have long been hypothesized and investigated (Johnson, 1986; Johnson, 1987; Johnson, 2005; Johnson et al., 2007; Netto et al., 2003). A myriad of transmissible agents (viruses, prions, bacteria, and protozoa) present in animals used as food (poultry, cattle, pigs, sheep) have been associated with historical pathways to malignancy (Diseases of Poultry, 2003; Johnson, 1986; Johnson et al., 1995a; Johnson et al., 1995b; Johnson, 2005). Particularly, certain oncogenic agents such as Bovine leukemia virus (BLV), Bovine papilloma virus (BPV), reticuloendotheliosis viruses (REV), avian leukosis/sarcoma viruses (ALSV), Jaagsiekte Sheep Retrovirus (JSRV), and Marek's disease virus (MDV) have been associated with a wide variety of tumors such leukemia, lymphoma, sarcoma and other cancers in these animals (Carbone et al., 2003; Johnson, 1994; 2005). Among the most potent cancer-causing agents known are ALSV, REV, and MDV (Choudat, Dambrine, Delemotte, & Coudert, 1996; Johnson et al., 1995a; Johnson et al., 1995b).These are known to naturally/frequently infect chickens, turkeys and other birds destined for human consumption (Diseases of Poultry, 2003; Johnson, 1994). They have been shown to be present in raw poultry products, including raw or inadequately cooked poultry meat and eggs, especially endogenous /exogenous ALSV (Johnson et al., 2009; Pham et al., 1999), and in vaccines grown in eggs such as the measles and mumps vaccines (Johnson et al., 2009; Tsang, Switzer, Shanmugam, Johnson, Golsmith, Wright et al., 1999). These agents have been known to cause a variety of cancers in poultry, turkey, and birds

(Diseases of Poultry, 2003; Johnson, 1994). Johnson and others have demonstrated experimentally, that they can even cause cancer in primates, and can transform normal human cells into cancerous cells *in vitro* (Johnson, 1994; Johnson & Griswold, 1996).

Human exposure to these viruses is widespread. This may include through the consumption of infected chicken/turkey/eggs and their products by the general population; occupation exposures during raising, slaughtering, processing, and preparation of birds/poultry products; public vaccination with vaccines prepared from chicken embryo cells; and in potential future gene therapy using some of these viruses to transport genes into cells or activate specific host genes (Johnson, 1994; Johnson, 2005; Netto et al., 2003). Some studies have even documented infections from and the presence of antibodies in human blood against ALSV, REV, and MDV (Choudat et al., 1996; Johnson et al., 1995a; Johnson et al., 1995b). It is therefore an important public health concern whether these agents also cause cancer in humans.

A few analytic epidemiologic studies have been conducted to address this concern (Fritschi et al, 2003; Johnson et al., 1986a; Johnson et al., 1986b; Johnson, 1989; Johnson et al., 1997; Johnson et al., 2009a,b; Netto et al., 2003). This has been the entire focus of the CRIWETOV project by Johnson ES and others. They have previously conducted a number of mortality studies in workers who are employed in poultry slaughtering/processing plants (Johnson et al., 1986a; Johnson et al., 1986b; Johnson, 1989; Johnson et al., 1995a; Johnson et al., 1995b; Johnson et al., 1997; Johnson et al., 2009a,b; Preacely, et al., 2009) belonging in three large cohorts. This group, conceivably, has one of the highest human exposures to oncogenic agents through intimate contacts with blood, secretions and internal organs of chicken. “Frequent cuts/injury from sharp

knives/ bone splinters, and dermatitis from irritant enzymes and secretions make it easy for microorganisms to penetrate the skin and enter the body" (Johnson et al., 2009a; Johnson et al., 2009b; Johnson et al., 2009c).

One of these studies involved 2,639 members of Local 27, a United Food & Commercial Workers (UFCW) local union in Baltimore, Maryland who worked in poultry slaughtering & processing plants. This cohort was first followed from 1949 to 1989 (Johnson et al., 1986a; Johnson et al., 1986b; Johnson, 1989; Johnson et al., 1997) and follow-up later extended up to the end of 2003, registering a total of 790 deaths (Johnson et al., 2009a). Another was a cohort study of 7,700 workers who were members of Local 410A, a UFCW local poultry union located in Marshall, Missouri. They were followed up between 1969 and 1990 during which time a total of 459 deaths were recorded (Netto et al., 2003). Recently, an update of mortality in this cohort was conducted, extending follow-up to the end of 2003, within which time a total of 1,337 deaths had occurred (Johnson et al., 2009b). The present study is of 30,488 members of a Union Pension Fund who have been studied for cancer mortality from 1972 to 2003 (Johnson et al., 2009c).

From these studies it was of interest to see whether there is an increased risk of lung cancer, which is a more common type of cancer (all cancer are relatively rare), and multiple myeloma, a rarer form of cancer, (as well as many other causes of death) among poultry slaughterhouse workers who are conceivably exposed to oncogenic viruses at a rate higher than the general population (Johnson et al., 1997; Netto et al., 2003). This is a justified concern as a number of epidemiological and laboratory investigations have demonstrated that both of these cancer forms have viral components (Alberg & Samet,

2003; Avet-Loiseau et al., 2001; Brouchet et al., 2005; Chen et al., 2001; Chesi et al., 1998; Dib et al., 2008; Giuliani et al, 2007; Miyagi et al., 2000; Moore & Chang, 1998; Rettig et al., 1997; Syrjanen, 2002). While the Baltimore study was able to document increased risks for multiple myeloma consistently over time among white males only compared to the United States general population, it did indicate a significant risk for lung cancer only in the initial follow-up, which has decreased with time (Johnson et al., 2009a). On the other hand, a significantly higher risk was noted for lung cancer in the Missouri cohort but not for multiple myeloma (Johnson et al., 2009b). This study reports findings from the recently followed Union Pension Fund Cohort.

Various study design and statistical approaches have been applied in occupational cancer mortality studies. Typically, an exploratory cohort design (mainly retrospective) is applied to quantify the effects of exposure on disease incidence or death rates for a large number of causes in a large population for which detail analysis is not feasible. Extensive and detailed analyses are subsequently carried out only for a few identified causes of interest using nested case-control designs with the application of various control sampling schemes (Breslow & Day, 1980; Breslow et al., 1983; Breslow et al., 1987; Morabia et al., 1995). Two case-control studies nested within a cohort have been conducted in the CRIWETOV project. An earlier study involving poultry, cattle, pig, & sheep workers investigated whether occupational exposures were associated with death from tumors of the hemopoietic and lymphatic systems among members of a meatcutters' union in Baltimore, Maryland. Elevated risks were observed for butchers who killed animals, workers in chicken slaughtering plants, and workers in cattle/sheep/big abattoirs (Metayer et al., 1998). A recently completed pilot case-cohort study of lung cancer nested

within a cohort of poultry and non-poultry workers evaluated whether humans exposed to the oncogenic viruses of poultry have increased lung cancer risk. This study that was conducted within all the three poultry cohorts combined and within the cattle, pig, and sheep meat industry showed high risks of death from cancers of the lung (Preacely et al., 2009).

An indispensible element in documenting risk to a particular exposure is the analytical technique applied in analyzing the data. A majority of cohort mortality studies employ non-model-based statistical methods (particularly, indirect standardization by stratification). Classical effect measures from this approach are the standardized mortality ratio, SMR and the proportional mortality ratio, PMR (Breslow et al., 1987, Johnson et al., 1994; Weitkunat, Crispin, Grill, Fischer, Meyer, & Schotten, 2001). Others apply model-based regression techniques such as the Poisson, Cox proportional hazards, and logistic regression models (Breslow et al., 1983; Breslow et al., 1987, Chen, 1999; Greenland, 2004; Joffe & Greenland, 1995; McNutt, Wu, Xue, & Hafner, 2003; Modern epidemiology, 1998). Each of these methods display a different measure of effect to document risk (such as the hazard ratio, HR; odds ratio, OR; rate ratio, RR, risk ratio, respectively). Although most of these approaches have been suggested to yield almost equivalent estimates of effect measures (Onland-moret, Van der, Van Der schouw, Buschers, Elisa, Van Gils et al, 2007; Vonesh, Schaubel, Hao, & Collins, 2000), some concerns have been raised on the efficiency and bias in the selection of a particular method of analysis, and on practical issues that arise in application with empirical data (Breslow et al., 1983, Breslow et al., 1987). In response, adjustments are constantly being made to existing methods and many new approaches are still being developed and

applied, especially in the application of nested case-control studies (Greenland, 2004; Langholz & Jiao, 2007; Thomas, 1998; McNutt, Wu, Xue, & Hafner, 2003; Zou, 2004). These approaches have varying underlying assumptions and may be suitable in some scenarios as opposed to others. Very few studies have actually compared effect measures obtained from various methods while observing their underlying assumptions in empirical data.

Even with case control studies nested within a particular cohort concerns have been raised on the comparability of effect measures from different analytical approaches based on the control sampling schemes: cumulative survival, cumulative incidence, case base, or incidence density, sampling (Morabia et al., 1995; Wacholder, McLaughlin, Silverman, & Mandel, 1992). Thus, the application of sub-optimal design/statistical methods to particular study design may well be a contributing factor to the observed discrepancies in epidemiological and other findings regarding cancer incidence and mortality.

CHAPTER THREE

STUDY DESIGN

Definition of cohort

The study was a part of a larger project involving cancer and non-cancer risks in workers exposed to oncogenic virus (CRIWETOV) that started in 1979 by Eric Johnson, M.B., B.S., Ph.D., Chair and Professor, Department of Epidemiology, School of Public Health, University of North Texas Health Science (UNTHSC) at Fort Worth. It used a retrospective cohort epidemiological design to define members of the study cohort and identify cases after an extensive follow-up process. The study had two phases. Phase 1 was a cohort mortality study and phase 2 was a case-control study nested within the cohort. In the nested case-control phase four different sampling schemes were applied in selecting the controls from the cohort: cumulative survival sampling, cumulative incidence sampling, the case-based or case-cohort sampling, and concurrent or incidence density sampling. Although the entire CRIWETOV project involved three large cohorts of workers belonging to three different local union funds, the present analysis was based on members from only one union that had updated information for the two comparative groups (poultry & non-poultry) of interest.

Study population

The study population was derived from the cohort of workers employed in poultry slaughtering/processing plants belonging to a local Pension Fund of the United Food and Commercial Workers (UFCW) International Union, headquartered in Washington, DC. The study population consists of 20,132 subjects that worked in 11 poultry slaughtering/processing plants and were members of 7 local poultry unions located in 6

states (Alaska, Arkansas, Louisiana, Maine, Missouri, and Texas). Also included were 10,356 subjects who worked in non-poultry industries as a comparison occupationally-unexposed group, who were also members of the Pension Fund. They were derived mostly from 11 local seafood unions covering 21 seafood companies located in 8 states (Florida, Illinois, Indiana, Massachusetts, New Jersey, Ohio, Pennsylvania, and Texas). Thus, a total of 30,488 subjects altogether comprised the study population for this retrospective cohort.

Follow-up mechanisms and processes

The cohort was uniquely complete, in that records were kept meticulously. All workers who were ever members (even for only a few days) of the unions during the defined study period had a record. Mortality was studied for the period January 1, 1972 to December 31, 2003, during which time a total 4,119 workers had died. The follow-up mechanism employed was very extensive. The method included the National Death Index, Social Security Administration (SSA), Maryland State Department of Vital Records (MSDVR), Maryland State Department of Motor Vehicles, Health Care Financing Administration (HCFA), Veterans Administration, obituary notices, US Post Office, personal contact by telephone and mail, and internet tracing methods. The Pension Benefit Information Inc., a private company, was also used to identify deceased persons. This company matches subjects against US death records for all years from the 1800s to the present, also using information received from SSA, HCFA, & MSDVR, as well as the Civil Service Commission, Railroad Retirement Board, and the Department of Defense (Johnson et al., 2009a; Johnson et al., 2009b; Johnson et al., 2009c) .

Information on exposure, study end points, coding of disease

Death was studied in this cohort preliminarily for a wide variety of malignant and nonmalignant diseases. However, for this specific analysis, the study end point was death from cancer of the trachea, bronchus and lung (ICD 162, 9th Revision; here by referred to as lung cancer) or malignant immunoproliferative disease, multiple myeloma, and malignant plasma cell (ICD 203, 9th Revision; here by referred to as multiple myeloma). These diseases were identified by the international classification of disease (ICD) codes. Various vital status records had these diseases initially coded in the 6th, 7th, 8th, 9th and/or 10th revisions. All the revision codes other than the 9th were converted into the 9th revision (Appendix A) according to a rubric developed by Dr. John (Johnson et at, 2009a).

Death from these two diseases constituted the outcome variables. Subjects whose vital status was unknown after the end of follow-up were assumed to be alive at the end of study while those who died of causes other than those mentioned above before the end of the study period were considered censored. Person-years were accumulated from January 1972 for those who were already members of the union or employed in poultry or non-poultry plants before that date. For those who became members, or started employment later, person-years commenced on the date of membership or employment. Membership in the union was compulsory from the first day of employment, thus the date of hire was virtually the same as the date of membership for persons who were hired after the plant had been unionized. Person-years were enumerated up to the date of death, or

date of termination of the study on December 31, 2003, whichever was earlier (Johnson et al., 2009a; Johnson et al., 2009b; Johnson et al., 2009c).

Working in a poultry slaughtering/processing plant was the exposures of interest with working in seafood or other non-poultry industries as the comparative non-exposed group. Age, race and sex were considered potential confounders. Age was assessed from date of birth and date of death/date of end-of-study. Date of birth information was missing for 259 subjects (0.8% of the entire cohort); however, it was available for all deceased workers. Rather than excluding these 259 persons from the analysis, these subjects had their date of birth imputed based on the median year of birth of workers with known date of birth joining the union in a particular year. Thus, if a member without date of birth joined the union in 1975, he/she was assigned as his/her year of birth, the median year of birth for all persons with known date of birth who joined the union that particular year. Race and sex were also artificially assigned at random to each individual in the study without a death certificate/known cause of death, based on the racial/ gender distribution of deceased persons with known race/sex. These measures were deemed to be associated with negligible bias, since the total person-years will be affected to a negligible degree and has been applied in some previously published works (Johnson et al., 2009a; Johnson et al., 2009b; Johnson et al., 2009c).

Power and sensitivity analyses were performed to obtain suitable sample sizes for the nested case-control analyses. By the end of the study period, a total of 378 (1.2%) lung cancer and 20 (0.07%) multiple myeloma deaths had occurred. These constituted the cases and were not sufficient to provide enough power for our study to yield significance at the 5% level of even with the selection of 5 controls. Thus, nested case-control

analyses, applying various sampling schemes were performed only for lung cancer mortality that was a bit prevalent in this cohort compared to multiple myeloma. An optimal number of four controls was considered for each case in the various sampling schemes; except for the incidence density sampling scheme where only three controls were selected for each case.

Data management and final analyses were performed using the statistical analysis software version 9.1.3 (SAS Institute, Inc, Cary, NC). The data management process involved updating the dataset after follow-up was completed. The dataset was updated for vital status, date of birth, date of employment, study start date, date of termination of employment, date of death, causes of death (ICD), and study stop date, as well as race and sex from death certificates. As mentioned above, individuals with unknown vital status were coded as alive. The study start date was January 1, 1972 for all those who were employed on or before this date; otherwise date of employment was considered study start date (i.e., start of follow-up). For subjects who died while employed, date of termination of employment was considered the same as date of death. Subjects who died before study end date of December 31, 2003, had as study end date their date of death. The few subjects with missing information on date of birth, race, and sex were artificially assigned this information as previously described. Because employment information was not available for most of the subjects, this information was not included in these analyses. Age at entry to the study was defined from the study end date and date of birth while age at exit from the study was defined from the study end date and date of birth. The variable ‘age’ used in these analyses was age at exit.

Data was prepared to meet the underlying assumptions or model specifications for each model or statistical methods where possible as described in the various sections of the proceeding chapters. The diseases included in these analyses were first identified in preliminary analyses that compared the number of deaths in this cohort from each cause with those expected from national vital statistics computed using the Occupational Mortality Analysis Program -Plus (OCMAP+) with the United States standard death rates up to 2001 & 2003. As such, the interpretation of the subsequent multivariate analyses should take cognizance of the multiplicity of comparisons that were made in the preliminary analyses. The OCMAP software was developed by the University of Pittsburg and is widely distributed in the United States (Marsh, Youk, Stone, Sefcik, & Alcorn, 1998).

In order to perform OCMAP+ analyses, the ICDs for all subjects who were alive were coded ‘0000’ while those with missing ICDs, unknown causes of death, or with ICDs that were considered impossible by the OCMAP+ program were assigned the value ‘9999’. Vital status were coded ‘2’ for all subjects who were alive (and these included subjects who were actually alive and working, alive but separated from employment, unknown vital status but separated from employment, alive but retired from employment and unknown vital status but retired from employment), ‘3’ for those dead while separated from employment, ‘6’ for those dead after being retired from employment, and ‘8’ for those who died while employed, as required by the software. The exposure variable ‘plant’ was considered the occupational environment within the OCMAP+ program and was coded ‘01’ for poultry and ‘02’ for non-poultry. In addition to exposure (plant), race and sex, OCMAP+ analyses were performed with age and calendar time in 5-year intervals. Hence

these results are not directly comparable with those obtained from the subsequent analyses that did not consider calendar time and had stratified age in intervals other than 5-years.

Expected deaths were derived by multiplying the person-years in each cell by the corresponding gender-, calendar year-, age-specific mortality rate for the United States general population. Observed and expected deaths for each cell were summed over all ages and calendar years, and over all strata, and the SMR estimated as the total observed number of deaths divided by the total expected. The 95% confidence intervals for the SMR were calculated in OCMAP+ according to a simple exact method that links both the Poisson and chi-squared distributions (Ledell, 1984; Marsh et al., 1998).

Typical output from this program is provided for each exposure or occupational environment, as well as for the combined cohort (stratified according to white-male, white-female, nonwhite-male, and nonwhite-female). Abstracts of some recently completed analyses using this software within the CRIWETOV projects are reported as Appendix B, C, D, E, F, & G.

Preliminary Results

Table 1 presents the summary statistics for a few selected causes of deaths for the population as a whole and for sub-cohorts of poultry and non-poultry workers. The mean ± standard deviation of the duration of follow-up for the entire cohort was 24.3±5.7years. Mean age at entry was 28.8±11.1 years and 53.1± 11.7 years at exit of the study. Race and sex distributions of the base population are also shown in the table. As noted earlier a total of 4119 (13.5%) subjects died during the risk period. Of this number 2454 (59.6%) worked in poultry slaughtering/processing plants.

SMR and PMR were computed for a large number of diseases (causes) in the preliminary analyses using the OCMAP+ software separately with United States standard rates from 1972 to 2005 (Table1) & 1972 to 2001 (Table2). Only results for selected causes of death (all causes of death, all malignant neoplasm, lung cancer, and multiple myeloma) have been reported.

In comparison with the United States standard population, workers in this cohort had significantly higher mortality rates due to lung cancer. Multiple myeloma death rates were higher, however, not significant. While all cause mortality rates were also higher, mortality rates from all malignant neoplasms in general was lower across the general cohorts and sub groups of poultry and non-poultry workers. These results were in close agreement when using the United States standard mortality rates up to 2001 and rates up to 2005 in the OCMAP+ analyses.

CHAPTER FOUR

NON-MODEL-BASED METHODS

Introduction

One major area of focus in epidemiological research is the comparison of basic health indicators (health situation analyses) which allows one to define risk areas, define needs, and document inequalities in health among population subgroups. This process can sometimes be facilitated by the use of crude rates (of mortality, morbidity or other events) in comparative analyses. However, when the population distributions are not comparable for factors such as age, sex, race, or socioeconomic levels the interpretation of crude effect measures may be grossly distorted due to confounding and heterogeneity of effect. Particularly, age structure has an important impact on a population's overall mortality as crude rates are higher in older populations (Pan American Health Organization, 2002).

One classical epidemiological method for controlling or reducing the effect of confounding in comparative analyses is through standardization by stratification, which is the analysis of data within categories of covariates or potential confounders (Breslow et al., 1987; Pan American Health Organization, 2002; Szklo & Neito, 2007; Weitkunat et al., 2001). This is a non-model based approach of comparative analysis whereby a weighted average of an effect measure is obtained using the distribution of the structure of or rates from a standard or target population. This method provides an easy to use summary measure which is the estimate of what would have happened in the standard or target population if they had the same outcome (rates or risks) as the study population. The outcome (standardized rate), is a crude rate that has been adjusted for differences in

composition of alleged confounding factors between the region under study and the standard population (Pan American Health Organization, 2002). In a sense this is achieved through pooling (weighted average estimate) if rates or risks were constant across strata.

The issue as to whether to use an external standard population or rates to standardized rate in a study cohort or to use internal standard (i.e. part or all study) population has emerged as important controversy in the broader debate on risk standardization (Breslow et al., 1987; Breslow et al., 1983; Pan American Health Organization, 2002). Most techniques of cohort analysis have assumed that the underlying death rates as function of age and/or calendar year were known from vital statistics or other standard sources. Of course there are some consequences both for making this assumption and or doing without it. Draw backs to the use of an external standard surround homogeneity of population distribution in terms of the controlling variables or effect measures across different strata between the comparative groups/population (Breslow et al., 1987; Breslow et al., 1983; Pan American Health Organization, 2002). For example the epidemiology literature is replete with warnings against the uncritically use of SMR (obtained from standardization methods) as a summary index (Breslow et al., 1983). Many of these criticisms relate to the inadequacy of national vital statistics to represent the baseline mortality of occupational groups.

The typical phenomenon that workers tend to be healthier than the general population when first employed has been examined. Typically, the ratio of observed to expected number of deaths rises with years of employment and peaks around 15 years, in which time the effects of the initial selectivity are largely dissipated (Breslow et al.,

1983). In this study standardized effect measures were first obtained using rates from an external standard population (US standard rates from 1972 to 2005 as well as rates up to 2001; just for comparison). Then subsequently, the combined study population was used as an internal standard population for further analyses that compared different methods.

Two main standardization methods have been emphasized, characterized by whether the standard used is a population distribution (direct method) or a set of specific rates (indirect methods). These two methods are generally different both in the fundamental outcome of interest and the interpretation; yet pointing to a similar unifying goal (adjusting a measure of effect). In the sections that follow, both methods are introduced and the results of the corresponding analyses presented.

Direct Standardization

The direct standardization method uses the distribution of a standard or reference population, stratified according to the control variables (such as age, race and sex) and to which the specific rates of the corresponding strata in the study population is applied to obtain the expected cases (deaths) in each stratum if the population has the same composition. Typically, the outcome of interest here is a frequency measure (rate/risk). Thus, a directly standardized (adjusted) rate is a weighted mean event rate (e.g. mortality rate) for a study population, using the stratum sizes of a reference population as the weighting scheme. This is the rate that would be expected in the study populations should they all had the same composition according to the adjusted variables (age, race and sex, in the case of this study). It is obtained by dividing the total of the expected cases by the

standard population (Breslow et al., 1987; The analysis group, 2002; Szklo et al., 2007; Weitkunat et al., 2001). Thus, if we let

r_{ij} = age-specific rate (i-th population, j-th stratum) and

P_j= number of person years (or simply number of individuals) in the j-th stratum of the standard population, then the directly standardized rate R_i is given by

$R_i = \sum r_{ij}P_j/P$, j=1,2,…,m; where

$\sum r_{ij}P_j$ is the total estimated events (deaths);

P= total number of person time in the standard population i.e.

$P=\sum P_j$, j=1, 2,…, m; and m=number of strata.

The estimated number of events from each stratum is obtained by applying the rates in the j-th stratum of the i-th population to the number of persons in the corresponding stratum of the standard population. The sum of all such "events" produces the number of "events" that would be expected if the distribution of the controlling factors in the i-th population were identical to the standard population. In this way all the directly adjusted rates have the same distribution of the controlling factors.

A comparative study of adjusted rates may be carried out in different ways: through the use of absolute difference between the rates, their ratio, or the percentage difference between them, the later being valid only when same standards are used (The analysis group, 2002). In this analysis the ratio of two population rates (RR) was used as a comparative effect measure. The standardized rate ratio is the weighted averaged of the stratum-specific rate ratio. The causal parameter estimate then is the ratio of the number of cases which would have occurred if everyone exposed is compared to the number of cases which would have occurred if everyone was not exposed. Thus,

$RR = R_1/R_0 = \sum P_j r_{1j} / \sum P_j r_{0j}$, summing across the j strata, where

r_{ij}=crude frequency measure (rates) with in the j-th stratum of i-th study (index)population, i=0,1.

Underlying assumption

The underlying assumption of the direct method of standardization is the comparability within strata of controlling factors. The choice of a reference or standard population is also important. The method is not appropriate if there is not a consistent relationship between stratum-specific rates in different populations being compared. The distribution with regards to the adjustment factors should not be radically different in the populations compared (Breslow et al., 1987; Pan American Health Organization, 2002). Thus, a population that relates naturally to the group under study such as coming from the study population (average or sum, say) may be more appropriate. Some concern had been raised with this approach based on the size of the comparative sub groups making up the entire study population. There is a claim that the sizes of the population do not need to differ substantially as the larger population may unduly influence the adjusted rates (Kramer, 1988). However, this may be unlikely except in a situation where the distribution of the death rates, say, according to age in the comparative groups are in a reverse order, i.e, one group reporting higher death rates among the older subjects while the other group reports higher rates among the younger subjects. Another important issue with this method relates to the size of each stratum. This method is unreliable with small numbers. It is appropriate only for at least 25 overall observed events and at least one

event in each stratum. If the number of events is small, the option is to aggregate strata (Pan American Health Organization, 2002).

Data preparation and analysis

All the variables of interest were categorized. The two out come variables: death due to lung cancer, ‘Lung’ and death due to multiple myeloma, ‘Myeloma’ were binary categorized as ‘1’ if the subject died of any of these causes and ‘0’ otherwise. The interest was to compare mortality rates from each of the above diseases among union members exposed to poultry versus non-poultry. The exposure variable ‘plant’ was dichotomized into ‘1’ for ‘poultry slaughtering/processing plant workers’ and ‘0’ otherwise. These rates were adjusted for race, sex and age and standardized based on the structure of the combined study (base) population. Race and sex were jointly categorized in a 4-level ‘Rsex’ variable as follows: 1, 2, 3, 4, respectively for white male, white female, nonwhite male and nonwhite female. Age was initially categorized into 15-level-5year interval ‘Agrp’ variable. Due to sparse data and violation of underlying assumption (many categories with zero death) ‘Agrp’ was aggregated (some levels collapsed) into a 3-level categorical variable with levels ‘≤ 40’, ‘40-60’, and ‘60+’ years, respectively. Although these levels do not depict equal interval age ranges, they provided an evenly distributed base samples and allowed for at least one death per plant/race/sex/age category or stratum.

The entire cohort (both poultry and non-poultry) was used as standard population to adjust rates from the poultry and non-poultry groups separately. However, the size of the poultry group was about two times as large as the non-poultry. In order to study the influence of the population size of the comparative on the standardized rates and/or

groups avoid the common pitfall of the larger sample unduly influencing the rates, a sub-cohort was constructed. This sub-cohort contained all subjects from the non-poultry plants and a random sample of poultry workers of size equal to the non-poultry (n=10,356). Thus, the sub-cohort had the same number of subjects from each group. Analyses were performed separately using the full cohort and the sub-cohort. The results helped shed light on the actual role of the population sizes on the rates or rates ratio.

This data was analyzed using a self-generated SAS macro (Appendix G). The 95% confidence interval for the ratio of mortality rates in the poultry versus non-poultry groups was obtained using percentile bootstrap based on the case resampling bootstrapping method (Efron, 1982). Ten thousands bootstrap samples of equal size as the original cohort (resampling with replacement) were obtained (Chung & Lee, 2001). The bootstrap distribution was symmetrical and centered (revealing no bias). Thus, the 2.5 and the 97.5 percentile of the bootstrap distribution provided the limits of the 95% confidence interval (Davison & Hinkley, 2006; Efron, 1982).

Results

Five-year interval age categorization is a common phenomenon in the presentation of mortality data. Table 3 shows the distribution of lung cancer and multiple myeloma deaths distributed in 5 year interval age groups (15-levels) according to plants (for the full cohort). Many cells indicated zero deaths, thus severely violating the sparse data assumption for direct standardization procedures. This was exacerbated when race/sex stratification was included. Table 4 shows race, sex, and age, distribution of deaths in each plant but this time with the levels of age group collapsed into 3-levels.

Although a few cells still have no observed deaths the violation of assumption was not as severe.

The distribution of the stratum-specific death rates was similar among the sub groups of poultry & non-poultry across age strata. This suggested that although the poultry group was larger (two times) than the non-poultry group, it may not have any serious effect on the standardized rates when the combined study cohort is used as the reference or standard population. In fact, results from preliminary analyses using OCMAP+, with the sub-cohort of equal size of both comparative groups (Table 5) were in close agreement with those of the full cohort (Table 1), when the US standard rates up to 2005 were used.

Applying the direct standardization technique, the rate ratio (95% confidence interval, CI) for lung caner mortality was 0.92(0.78, 1.17), comparing the poultry to the non-poultry group based on the full cohort as standard. This rate ratio was similar (up to 1 decimal place) when the sub-cohort was used as standard for analyses (Table 6). Rate ratios for multiple myeloma mortality was 0.94 (0.44, 2.36) with the full cohort and 0.68 (0.17, 1.83) with the sub-cohort. These results suggested that working in non-poultry (which were dominantly seafood) industries may expose workers to a greater risk of dying from both of these forms of malignant diseases compared to working in a poultry slaughtering/processing plant but certainly not to statistical significance.

Indirect Standardization

The indirect standardization method utilizes specific rates from a standard population and applies them to the study population (previously stratified by the variables to be controlled) to compute the expected number of deaths in each stratum. A classical output from this method is a measure of association known as the standardized mortality ratio (SMR). SMR is calculated by dividing the total observed number of deaths by the total expected (Breslow et al., 1987; Szklo & Neito, 2007; Pan American Health Organization, 2002; Weitkunat et al, 2001). Let:

P_{ij} = person time (or simply number of individuals) in the j-th stratum in the i-th group

d_{ij}= number of observed deaths in the j-th stratum in the i-th group and

R_j =death rate from the j-th stratum in the standard population, the SMR in the i-th group is given by

$SMR_i= D_i/E_i$, where

$D_i=\sum d_{ij}$, j=1, 2, …,m is the total number of observed deaths in the i-th group and

Ei= $\sum$PijRj, j=1,…,m is the total expected number of deaths.

Thus, SMR allows for the comparison of deaths in each population under investigation to a standard population. A conclusion can be reached by simply calculating and looking at the SMR. An SMR higher than 1 (or 100%) indicates that the risk of dying in the observed population is higher than what would be expected if it had the same experience or risk as the standard population and vise versa (Pan American Health Organization, 2002).

Another possible output from indirect standardization is the proportionate mortality ratio (PMR). This estimate is obtained using only information from the cases (e.g. death persons only). In computing PMR, the study and standard populations are stratified according to the controlling variables. The proportion of all deaths due to a given cause in a given cell in the standard population is multiplied by the total number of deaths (all cause mortality) in the corresponding cell of the study population to get the expected number of deaths. The ratio of the corresponding observed to expected deaths gives the PMR. There have been some discussions on the use of PMR as an estimation of relative risk with concerns that this is a biased measure (Wong, Decoufle, 1982; Wong, Morgan, Kheifets, Larson, 1985). It is of essence to remain conscious of the underlying theory that suggests PMR will provide an unbiased estimate for cause specific relative risk provided the all cause SMR approximate unity. It will provide an under estimate and correspondingly an over estimate if the all cause SMR is greater than and less than unity, respectively (Johnson, 1986)

PMR analysis is not among the statistical methods compared in this study. However, PMR results from the preliminary analyses using OCMAP+, have been reported along side SMR. Since in the PMR analyses date of birth and race information was available for all subjects (who were death), the results may provide a check on the SMR analyses for which race, sex, and date of birth were artificially assigned to some subjects.

Comparative analyses for two populations can be performed using the ratio of their respective SMRs known as the relative SMR (RSMR) (Breslow, et al., 1987). Alternatively, adjusted or standardized rates can be calculated using the indirect method by multiplying the crude rate of every population by its SMR (Pagano & Gauvreau,

1993). This can, thus, provide a single value for each population (though only hypothetical representation), which takes into account the differences in the compositions of the populations. If one of the study population is used as a standard population to provide rates for which the other population is standardized, the resulting SMR gives the rates ratio (RR) for the two population adjusted for the confounding factors. Thus, rates ratio was computed from the indirect standardization methods using these three approaches: RSMR, RR (for standard rates obtained from products of crude rates and corresponding SMR), and SMR for poultry group when standardized with rates from the non-poultry group.

Underlying Assumptions

Underlying the calculation and interpretation of SMR is the assumption that the stratum-specific death rates are a constant multiple of the corresponding standard stratum-specific rates (homogeneity). If the stratum-specific rate ratios are not constant, SMR represents an "average" of a series of heterogeneous quantities. Thus, it does not provide a meaningful single summary of the underlying relationship between a comparison group and the standard population (Pan American Health Organization, 2002).

Data preparation and analyses

The same set and format of data used for direct standardization was used for indirect standardization. In order to compare the risk estimation across methods SMR was computed using a self generated SAS macro (Appendix G). Comparative risk estimation was computed using the three approaches described above: RSMR, Indirectly standardized rates ratio, RR, and SMR for the poultry group with the non-poultry group

serving as the standard population. For each of these methods, the 95% confidence intervals of the ratios were obtained by the bootstrap method previously described.

Results

The RSMR (95%CI) for lung cancer comparing the poultry to the non-poultry exposures was 0.96 (0.82, 1.17). RR was 0.71 (0.51, 1.03) while the SMR for poultry (with non-poultry as standard) was 0.87 (0.71, 1.11) (Table 7). The corresponding values for multiple myeloma are also stated in Table 7. The values are not markedly different between the full and sub-cohorts. SMR for the full as well as the sub-cohorts as standard populations are also presented. Note that the SMR from this analysis should not be compared with those from the OCMAP+ analyses (Table 1, 2, and 5) as those made considerations of calendar time (whereas this did not) and 5-year interval age categorization (Table 3) as opposed to the broad 3-level categorization used here (Table 4).

CHAPTER FIVE

MODEL-BASED METHODS

Introduction

The major goal of the analysis of cohort data is to quantify the effects of exposure on disease incidence or death rates. Stratified analysis especially with SMR as a classical effect measure has been used most commonly with cohort studies published in medical literature. Rate ratios calculated directly from data are typically subject to rather extreme sampling variation due to small numbers of observed events. Also, standardization techniques are not popular in multivariate settings. For example, while SMR is typically calculated for a large number of different diseases, and for sub-cohorts having particular types of exposure, relatively little attention has been paid to statistical modeling of ratio as a function of measured dose levels, with a number of covariables (Breslow et al, 1983).

However, Breslow et al. (1983) have attempted a statistical modeling for an extension of SMR into a multivariate multiplicative domain previously available only with regression techniques and considered several methods of cohort analysis in a unified conceptual framework (Breslow et al., 1983). Several regression techniques have evolved for multivariate modeling of risk which differ essentially in the overall objective and nature of data to be analyzed. Even within each method sub techniques have been developed to deal with different ways of obtaining variance (exact/asymptotic) and likelihood (partial- /pseudo-likelihood) estimators (Breslow et al., 1983; Greenland,

2004; Langholz & Jiao, 2007; McNutt et al., 2003; Zou, 2004;). In its simplest form a regression model consists of a function of the dependent variable modeled as a linear combination of a number of independent variables (Kutner, Nachtsheim, Neter, Li, 2005). This report presents and compares outputs from three commonly used regression techniques: the Poisson, Cox proportional hazards, and logistic regression models.

Poisson regression model

In analyzing mortality data, one can elect to analyze either aggregated subject mortality rates or individual subject survival times. Analyses based on aggregated mortality rates are common in large scale epidemiological or registry type studies where information on mortality is presented in summary form. In analyzing aggregated mortality rates, the response or outcome variable is the number of deaths that occur divided by the number of accumulated person- time at risk for death. A typical way in which such deaths rates have been compared between exposure groups is by using the Poisson regression model such as a multiplicative loglinear model of rates (Kutner et al., 2005; Loomis, Richardson, & Elliott, 2005; McNutt et al., 2003; Vonesh et al., 2000). This is a generalized linear model with log as the link function and Poisson as the distribution. Typically, if rates at which an event occurs in a cohort after a period of follow-up are available, Poisson regression model allow for the analysis of these rates such that: average rate= number of events during a specified time interval / total person-time accumulated during the interval; where the total person-time is the sum of the accumulated time for those individuals who experienced the event as well as those who could have but did not experience the event. The model equation can be written such that rate is a product of influences associated with a series of independent variables. For

example, in the case of death rates for a specific cause associated with exposures of a particular plant (or industry), while adjusting for case-mix differences in sex, race, and age the stratum specific rates (R_{ij}) may be described by the model:

$\ln(R_{ij}) = \beta_0 + \beta_1(\text{plant}) + \gamma G_j$

or

$R_{ij} = \exp[\beta_0 + \beta_1(\text{plant}) + \gamma G_j]$,

where γ and G_j are vectors; G_j being a vector of all the controlling variables: sex, race, and age (categorized in q+1 groups) dummy coded into 0 and1. If for each group the rates are constant across strata, the model is simplified such that

$\ln(R_{ij}) = \beta_0 + \beta_1(\text{Plant})$ and $R_{ij} = \exp[(\beta_0 + \beta_1(\text{Plant})]$.

The rate (R_{ij}) involves two component: the count of events and person-years of exposure i.e,

rate= count/ person-year such that count= rate*person-year

Such that

ln(count)= ln(rate) + ln(person-year).

Or

$\ln(\text{count}) = \beta_0 + \beta_1(\text{plant}) + \gamma G_j + \ln(\text{person-year})$

The coefficient of ln(person-years) is always 1, thus, an “offset”. Thus, the Poisson regression model is appropriate whenever the dependent variable is a count (i.e. number of cases of disease or death) within a series of subdivisions of the sample data.

The expected number of deaths in each stratum is

$E_{ij} = (\exp[\beta_0 + \beta_1(\text{plant}) + \gamma G_j])*(\text{number of person years in stratum } i,j)$.

The difference between the observed and expected number of deaths can be calculated by a Pearson chi-square statistics:

$X^2 = \sum(O_{ij}-E_{ij})2/E_{ij}$,

summing over all strata; and, assuming the model is correct, X^2 is distributed as $\chi^2_{(IJ-s)}$, where s= number of parameters estimated in the model and (IJ-s)= degree of freedom for the distribution.

An additive Poisson model implies that each category has a separate multiplicative influence on the rate. These are easily expressed as a ratio of two rates when a specific category is used as a reference. Thus, following the above discussion, the major components of the Poisson model are as follows:

1. Dependent variable (count of death cases) that has a Poisson distribution:

   $\Pr(Y_{ij} \text{ dying}) = [\exp(-\lambda_{ij})*(\lambda_{ij})^{Y_{ij}}]/Y_{ij}!$
2. The expected number of deaths in each stratum, represented by the parameter λ_{ij}, may be expressed as: $\lambda_{ij}=P_{ij}*\exp[\beta_0 + \beta_1(\text{plant}) + \gamma G_j]$
3. Finally $\ln(\text{rate})=\ln(R_{ij})=\ln(\lambda_{ij}/P_{ij})= \beta_0 + \beta_1(\text{plant}) + \gamma G_j$

It is, thus, worth of note that one may also elect to model instead the mean number of death to be expected in each plant at a given time period using the generalized linear model:

$\lambda_{ij}= \exp[\beta_0 + \beta_1(\text{plant}) + \gamma G_j]$.

The Poisson regression coefficients are estimated by maximizing the log-likelihood function $L(\xi)$ for the Poisson distribution:

$L(\xi) = \sum Y_{ij} \log\lambda_{ij} - \lambda_{ij} - \log Y_{ij}!$,

where

$\xi = (\beta_0, \beta_1, \gamma)$

is the vector of the regression coefficients (Kutner et al., 2005; Loomis et al., 2005; McNutt et al., 2003; Vonesh et al., 2000). Other approaches have evolved whereby a pseudo-likelihood or partial-likelihood estimators have been applied in estimating the coefficients as well as applying a robust sandwich or asymptotic variance estimators (Breslow et al., 1983; Greenland, 2004; Langholz & Jiao, 2007; McNutt et al., 2003; Thomas, 1998; Zou, 2004).

Underlying assumption

This model requires that the difference in the ln(rate) is the same in each stratum such that we expect $\ln(R_{i1}) - \ln(R_{i0}) = \beta_0 + \beta_1(\text{plant}) + \gamma(1) - [\ \beta_0 + \beta_1(\text{plant}) + \gamma(0)) = \gamma$ In the case of this study, plant has two groups: poultry (coded as 1) and non-poultry (coded 0). Of course, it is assumed that mortality (from lung cancer/multiple myeloma) occurring in a particular group are independent of one another and that a certain mean number of deaths per unit time is characteristic of the given set of exposure variables (plant, race, sex and age group). The mean itself is assumed to depend on these variables and always greater than zero.

The decision as to whether a Poisson model is appropriate can be based on one of several statistics. A commonly used statistic is the deviance statistics D, which is the

difference of -2log-likelihood. D is approximately a chi-squared random variable with degrees of freedom (n-p) for n number of observations and p parameters. A ratio D/(n-p) significantly larger than 1 may indicate model misspecification or an over-dispersed response variable; ratios less than one may also indicate model misspecification or an under-dispersed response variable (Timm & Mieczkowski, 1997). One approach to analyze an overdispersed model is by using the negative binomial distribution. This distribution adds a quadratic term to the variance representing overdispersion. Thus allowing for extra-Poisson variation of other variables not included in the model. Poisson or negative binomial models are ordinary count models that are appropriate only when there are not excess zeros in the data. If there are excess zeros, a zero-inflated Poisson model is used (Greene, 1994; Lambert, 1992).

In addition to a plausible basis for the underlying distributional assumptions, a goodness-of fit test is needed to validate the Poisson model. One important test is to ensure that the estimated regression coefficient for each covariate should be statistically significant, i.e., one should be able to reject the null hypothesis that the coefficient is zero. Of course there are other possible tests that can be performed (Timm & Mieczkowski, 1997).

If the non-poultry group is treated as the standard population in adjusting the poultry group then we can use model parameters to get SMR as follows:

$R_{i1}=\exp[\beta_0 + \beta_1(1)]$; for poultry and

$R_{i0}=\exp[\beta_0 + \beta_1(0)]$ for non-poultry (standard)

such that

rate ratio = $R_{i1}/R_{i0} = \exp[\beta_1(1) - \beta_1(0)]$ = SMR

The results obtained from this model should be similar, given that the underlying assumptions are met, with those obtained from the third option of the indirect standardization approach.

Data preparation and analyses

Poisson regression techniques are designed to be particularly effective when data are collected in a specific pattern. Our model of interest is

$\ln(\text{count}) = \beta_0 + \beta_1(\text{plant}) + \gamma G_j + \ln(\text{person-year})$.

Age was categorized according to 5 year interval resulting in 15 groups (Table 3). The number of observed deaths for each cause (Lung cancer and multiple myeloma), and the total person-year was computed for each plant, sex, race, and age group. This resulted in a new dataset with 120 observations as a consequence of the various permutations of plant, sex, race, and age groups. Analysis was perform using the SAS 'PROC GENMOD' with count or number of cases as the outcome, plant the predictor, race, sex, and age (dummy coded) as covariates, and ln(person years) as the offset. A model with the 'repeated' statement to obtain the robust standard errors for the Poisson regression coefficient was also explored.

Results

There was an average of 254 observations in each plant/sex/race/age group category or stratum, with a mean accumulated person-years of 6163.6 years. The ratio D/(n-p) was 0.9 for lung cancer and 0.7 for multiple myeloma. This does not suggest model misspecification or over-/under- dispersion. RR (95% CI) for poultry compared to non-poultry was 1.00 (0.81, 1.24) for lung cancer and 0.98 (0.37, 2.58) for multiple myeloma (Table 8).

Cox proportional hazards regression model

A more common, alternative approach in mortality analysis is to examine trends in mortality on the basis of individually determined patient survival times. Here, the response or outcome variable is the length of time until the event of interest takes place (e.g., death) or until some point in time when the patient is no longer followed (censored). Typically, the analysis of individual patient survival times is carried out using a Cox proportional hazards regression model (Kutner, 2005; Vonesh et al., 2000). This is a semiparametric regression model that describes survival time in a comparative sense where the complete description of survival time is not of primary importance. Instead, the focus is on how a particular risk factor modifies survival experience relative to not having the risk factor.

The model equation of Cox proportional hazards regression model can be written as

$h(t: x) = \lambda_0(t)\exp(\beta_i x_i)$,

where the baseline hazard, $\lambda_0(t)$, characterizes the hazard function when x=0 at time t. The baseline hazard rate $\lambda_0(t)$ does not have to be specified. In fact, the actual form of baseline function is of little importance if the focus is on relative comparisons (e.g. hazard ratio, rate ratio). The hazard ratio is defined by

$h(t : x+1)/ h(t : x) = \lambda_0 (t) \exp(\beta(x+1)/ \lambda_0 (t) \exp(\beta x) = \exp(\beta)$.

This describes how the hazard function changes as a function of covariate vector x. This is a semiparametric multiplicative model with no intercept (i.e. the price one pays for semi-parametric model). If an intercept were present, it would correspond to the log baseline hazard function. The implication is that one cannot reconstruct group specific rates; only ratios can be estimated.

The Cox proportional hazards regression model is well suited in modeling a situation with unequal observational time. If there is a clear idea about the distribution of the survival data, parametric models such as the exponential regression model, log-logistic regression model, or the Weibull regression model are preferred. Hazard ratio is considered an estimate for the relative risk of death. However, the Cox model makes no assumptions regarding what the shape the underlying hazard or death rate takes. As such it is reasoned that estimates of relative risk under the Cox model are more robust than what might otherwise be obtained using a fully parametric model (Kutner, 2005; Vonesh et al., 2000).

Underlying assumption

The Cox model assumes the death rate for a comparative group of subjects will be proportional to the death rate for the reference group within each specified interval of time. This is equivalent to assuming the relative risk of death between the two comparison groups will be constant over time. This assumption does not require that the death rates themselves be constant in time; it merely requires that their ratio be constant over time.

A number of options are available for assessing this assumption. One way is by using graphs to examine whether a considerable interaction exists between the hazard functions for each group. Crossing or touching lines may be indicative of an invalid model. A second option is to assess whether there is an interaction between time and the exposure of interest. This essentially is a trend test (hypothesis testing) on whether there is an increasing or decreasing trend over time in the hazard function. A significant interaction would imply the hazard function changes with time, and thus, the proportional

hazards model assumption violated. Yet still, one can evaluate this assumption by examining the residuals.

However, it is worth noting that the Cox model can still be used even when the proportionality assumption is violated by simply introducing an appropriate set of time-dependent covariates into the regression (Kutner, 2005; Vonesh et al., 2000). The idea is, when the assumption of proportional death rates (i.e., constant relative risk) is violated, application of the standard Cox proportional hazards model yields an average relative risk. In some cases, this average risk may mislead investigators into thinking one type of exposure is more causally related to the outcome compared to another when in fact there are periods of time when the opposite is true. An alternative is to use an interval Poisson model (also referred to a piecewise exponential model). This avoids this pitfall by enabling the user to model the relative risk as a function of time which can be accomplished by including an interaction term between the interval follow-up times and exposures(Kutner, 2005; Vonesh et al., 2000).

Data preparation and analyses

The focus in this analysis was to study how the survival experience changes with age, race and sex across the different plants (poultry versus non-poultry). Thus, the outcome of interest was the time to event, which in essence was the time of follow-up previously defined in chapter three. This continuous outcome was further categorized in two ways: in 10 and 5 year intervals. An exposure-time interaction term was also defined using plant as the exposure and the time of follow-up. These new variables served in assessing the proportionality assumption for the hazard function and for subsequent analyses there after in case of model violation. Final analyses were performed using the

SAS ‘PROC PHREG’ option with the Breslow sandwich covariance estimator as the default.

Result

Figure1 shows the plotted hazard function over time for the Plant (exposure) groups. Crossing lines over time was a clear depiction of an invalid Cox model for this data. However, further analysis of the trend test (hypothesis testing) was indicative of a non-significant exposure-time interaction (p=0.40); suggesting a valid model. To evaluate the extent to which a potential violation of the model will affect the comparative effect measure for death due to lung cancer and multiple myeloma, 4 different approaches were explored: a model with continuous time-to-event, a model with the exposure-time interaction term included, and two interval Cox models: with a 10 year and a 5 year interval time-to-event. The results were fairly close; a non-significant hazard ratio of approximately 1.0 (Table 8).

Binary logistic regression model

Regression models are common with continuous outcome. When the outcome is categorical, logistic regression models have been used. Logistic regression is a widely used technique to adjust for confounders, not only in case-control studies (with prevalent data) but also in cohort studies. The effect measure from this model or approach is estimated odds ratio; however, in cohort studies the desired effect measure is usually relative risk, which has been touted to approximate each other. In its simplest yet general sense, the logistic (nonlinear) regression model relates a dichotomous outcome variable y which, denotes whether (y=1) or not (y=0) the individual experience an event

(morbidity/mortality) during the study period, to a series of K regression variables $X=(x_1, \ldots, x_k)$ through the equation

$$Pr(y=1|x)= \exp(\alpha+\sum\beta_k x_k)/[1+ \exp(\alpha+\sum\beta_k x_k)]$$

or equivalently,

$$\text{logit } Pr(y=1|x)= \alpha+\sum\beta_k x_k = \ln(\text{odds}).$$

This formulation implies that the odds ratio (or correspondingly, the relative risk) for individuals having two different sets of exposure variables x^* and x is given by:

$$OR= \{P(x^*)[1-P(x)]\}/ \{P(x)[1-P(x^*)]\}= \exp\{\sum\beta_k(x_k^*-x_k)\},$$

where α represents the log odds of the risk for a person with a standard (x=0) set of regression variables, while $\exp(\beta_k)$ is the fraction by which this odd (or risk) is increased (or decreased) for every unit change in x_k. This model finds frequent application in epidemiology because the parameters are easily interpretable in terms of relative risk. Various approaches have been used to estimate the regression parameters in logistic regression depending on the application. The most popular and general approach involves the maximum likelihood estimation approaches which is available with the General linear models in SAS (Breslow et al., 1980; Breslow et al., 1987; Kutner, Li, 2005; Stokes, Davis, Koch, 1995).

However, concerns have been raised on some bias associated with using the odds ratio from a logistic regression model to approximate relative risk (RR) in a cohort study. It has been noted that when the outcome of interest is common (>10%) in the study population (though it could be rare in the general population), the adjusted odds ratio from logistic regression may exaggerate a risk association, i.e., an over estimate if OR>1 or underestimate if OR<1 (Mantel, Haenszel, 1959; Walcholder, 1986; Zhang,Yu, 1998).

A myriad of approaches have been proposed to correct the estimated relative risks obtained from logistic regression models (estimated from odds ratios). One of the simplest of these is the approach proposed by Zhang and Yu who hypothesized and demonstrated that $RR = OR/[(1-P_0) + (P_0 * OR)]$, where P_0= incidence of the outcome of interest in the unexposed group. This formula applies also to the confidence limits (Zhang et al., 1998).

Underlying assumption

In our application the response variables of interest (death due to lung cancer and multiple myeloma) have only two qualitative outcomes and they can be represented by binary indicator variables with values 0 and 1. This is a binary random variable following a Bernoulli distribution. The predictor variables do not necessarily need to be categorical (indicator variables). However, if they are not categorical, a monotonic sigmoidal relationship is assumed for the logit response function; between $Pr(y=1|x)$ and $\beta_k x_k$ (Kutner et al., 2005). When this assumption is not appropriate the rule is to convert all predictor variables into categorical variables and the employ log-linear models (Breslow et al., 1980; Breslow et al., 1987).

Data preparation/analyses

All the variables used in this model have been previously described: Lung and Myeloma (responses), plant (predictor), race, and sex (covariates) were dummy coded and age at exit retained as continuous. The SAS ‘PROC LOGISTICS’ was used in analyzing this data.

Results

The adjusted odds ratios for mortality in the poultry compared to the non-poultry cohort were not significant: OR (95%CI) = 0.96 (0.78, 1.20) for lung cancer and 0.93 (0.35, 2.44) for multiple myeloma (Table 8).

CHAPTER SIX

SAMPLING SCHEMES FOR NESTED CASE CONTROL STUDIES

Introduction

This study initially used a retrospective cohort design to document risk of lung cancer and multiple myeloma mortality among workers of poultry slaughtering & processing plants. The results, especially from the preliminary analyses using OCMAP+, seemed to support the notion that the finding of excess lung cancer and multiple myeloma in this occupational group of workers is probably real. However, such a design is usually not suitable for investigating specific occupational causes; especially it is not possible to control for non-occupational factors such as tobacco smoking and alcohol consumption that could as well account for risk. It is a common practice in epidemiology for case-control studies nested within a cohort to be conducted to better elucidate the effects of a particular exposure on an outcome (such as disease incidence or death rates). Thus, for this study larger case-control studies nested within very large cohorts of poultry workers that have sufficient statistical power to adequately investigate all the possible carcinogenic exposures within the industry, while controlling for non-occupational and occupational confounding factors, should be ideal (Johnson et al, 2009c).

However, discussions have emerged severally on the estimation of different measures of relative effect from case-control studies based on the way controls are selected or sampled and the way the data is analyzed (Morabia et al., 1995). Four common control selection or sampling schemes are the cumulative survival, cumulative incidence, case-cohort or case-based, and the incidence density samplings (Breslow et al., 1980; Encyclopedia of Epidemiological Methods, 2001; Morabia et al., 1995). These

control selection methods are reviewed and applied in a case control study nested within the cohort of workers belonging to the Union Pension Fund described above. For each of the control selection methods, the empirical data were analyzed by the three regression techniques discussed in chapter 5 and the results compared.

In a cumulative survival sampling scheme, controls are selected at the end of the risk period from members who remain at risk (Encyclopedia of Epidemiological Methods, 2001; Morabia et al., 1995). Thus, for this study, only subjects who were alive by December 31, 2003 were eligible as potential controls. On the other hand, in a cumulative incidence sampling scheme, controls are sampled at the end of the risk period from members who do not develop the outcome of interest. These members must not necessarily be at risk (Encyclopedia of Epidemiological Methods, 2001; Morabia et al., 1995; Szklo, Bartlett, 2007). For the present study, all alive subjects by study end date as well as those who died of diseases other than lung cancer and correspondingly, multiple myeloma were potential controls. Both of these methods are typical in a traditional case control study. Logistic regression have been used commonly in analyzing data obtained from this scheme and the typical effect measure is the odds ratio (OR) (Morabia et al., 1995; Szklo et al., 2007).

When controls are selected from the baseline population, regardless of their disease state at the time a new case occur, such a sampling scheme is known as the case-based or case-cohort sampling (Morabia et al., 1995; Szklo et al., 2007). The measure of effect commonly computed is an estimated relative risk (Rel. Risk). Poisson regression models have been commonly applied in analyzing data obtained from this scheme (Morabia et al., 1995).

In some situations, the case-control study is based only on incident cases occurring over predefined risk period. The controls are chosen concurrently, as the cases occur, from among those who are at risk (disease-free). Usually the controls match the cases in common demographic variables such as sex, race and age (with in a range). In this scheme, the number of controls for each incidence case is a function of the duration of follow-up of the study (Encyclopedia of Epidemiological Methods, 2001; Morabia et al., 1995; Szklo et al., 2007). Thus, a method of analysis that has been touted to be most appropriate is that which takes into account the accumulated person-time for each subject included in the analysis. The typical effect measure yield is the rates ratio (RR) or relative incidence rate. An example of an analytical technique that fairs well with this scheme is the Cox proportional hazards regression model (Morabia et al., 1995; Novikov, Oberman, Freedman, 2005).

Previous studies have shown that when the disease is rare (risk <10%) the estimated effect measures obtained from the above sampling schemes (OR, Rel R, and RR) should be similar given that the suitable analytical technique is applied. However, if the disease is common (risk ≥10%) the estimated effect measure from the cumulative survival and cumulative incidence sampling schemes (OR) will over estimate those from the case-base (Rel. R) and incidence density (RR) sampling (Encyclopedia of Epidemiological Methods, 2001; Morabia et al., 1995; Szklo et al., 2007). However, these would be subject to the certain underlying assumptions.

Underlying assumption

When the risk of the disease is low (rare occurrence) no major underlying assumptions are required. However, if the risk is high it would be expected that both the exposure of interest and incidence of the disease be stable over time for the estimated measure of effect especially from the incidence density sampling (RR) to be valid. While unstable incidence is of a significant practical concern, unstable exposure is not unless in situations in which there is a substantial decrease in prevalence (over 50%, say) (Green land, 1987; Morabia et al., 1995). The major issue with unstable RR centers on interpretability.

Data preparation

Analysis was performed only for lung cancer mortality as previously explained (chapter three). The analytical data set was prepared for each analytical technique as previously defined. Power and sensitivity analysis was performed to determine the sample size suitable for the nested case-control analyses. The prevalence of the exposure among the non cases (control) was 0.7. There were only 378 (1.2%) lung cancer deaths in the entire cohorts. These were not enough, even with case-control ratio of 1:5 to detect a significance odds ratio of 1.4 between the groups of interest. An optimal case-control ratio of 4 was used, thus requiring 1512 controls for all 378 cases. The full dynamic cohort was used in all the sampling schemes. For the cumulative survival sampling, all lung cancer cases were selected a priori and a simple random sample of 1512 controls selected from all subjects who were alive at the end of the study follow-up date (December 31, 2003). While for the cumulative incidence sampling scheme, the controls were selected from the remainder of the cohort regardless of their vital status.

For the case-base sampling scheme, 1512 controls were selected prior to follow-up start date, constituting a sub-cohort. Thirteen of these became failures (died from lung cancer). The remaining 365 lung cancer cases that were not members of the sub-cohort were also sampled constituting a sample of 1877 subjects for this analysis. A new variable (lung 2) was created with value 0 for sub-cohort non-failure (final controls), 1 for sub-cohort failures, and 2 for non sub-cohort failures. This variable was used in preparing the analytical dataset used in computing the exact case-cohort pseudolikelihood estimates for rates ratio by a method due to Langholz and Jiao (Langholz, Jiao, 2007) as an adjunctive analysis.

Cases were individually matched with controls each time the controls occurred by age (within 5 years), race and sex, for the concurrent or incidence density sampling schemes. Four cases did not have a match and most of the cases had at most 3 matches. Thus, 3 controls were selected randomly for each case with more than 3 matched controls, all matched controls selected for those with 3 matches and the 4 cases with no matches eliminated from the analysis. This resulted in a total of 539 unique controls of which 43 later became cases. A total of 870 subjects (including 374 cases) were analyzed with this selection scheme (Table 9). All four cases that were eliminated from the analysis belonged to the poultry group, all white, one male and three female, and age 60 and 73 years. A similar analytic data set with lung2, as previously described, was prepared for a similar adjunctive analysis. All model based methods discussed in chapter five above were applied in analyzing data from each scheme. Data was analyzed using SAS procedures. Due to the effect of matching (potentials of selection bias), the maximum likelihood estimates from regular logistic regression models may not be valid.

Thus, the conditional logistic regression model with SAS ‘PROC PHREG’ and the ‘ties=breslow’ option was applied instead of the regular logistic. This was deemed suitable to control for the effect of individual matching. To do this the response variable ‘lung’ was recoded so that the probability of being a case was modeled (Encyclopedia of Epidemiological Methods, 2001; Greenland, Schwartzbaum, Finkle, 2000; Stokes et al., 1995).

Result

The results of the statistical analyses from various schemes with each model have been summarized in Table 9. Effect measures from the cumulative survival, cumulative incidence and case-base sampling were fairly close with each analytical technique. These effect measures were non significant. However, output from the Cox model (with exact case-cohort pseudolikelihood estimators) showed significant risk of lung cancer mortality among the poultry workers compared to the non-poultry group. The results from concurrent or incidence density sampling were different from the above sampling schemes but as well similar across statistical methods. The results show a significantly decreased risk of lung cancer mortality among poultry workers from each of the analytical techniques. This study explored the application of nested-case control analyses with various control sampling schemes but with inadequate statistical power and very few controlling factors. A further study with enough statistical power would be needed to control the range of possible occupational/non-occupational confounding factors inherent in this cohort.

CHAPTER SEVEN

DISCUSSION, RECOMMENDATIONS, CONCLUSION

Preliminary results of these analyses show that workers belonging to a local Union Pension Fund of the UFCW international union are reportedly at higher risks of mortality from lung cancer (statistically significant) and multiple myeloma (not statistically significant) compared to the United States' general population (based on separate sets of standard rates from 1972 to 2001 and 1972 to 2005). These findings were observed studying the entire cohort as a group as well as in sub groups of poultry slaughtering/processing workers and non-poultry workers. It was of particular interest to note that the risks were slightly lower among poultry workers compared to non-poultry workers, though not of statistical significance. Although one can hardly incriminate any particular exposure based on retrospective cohort analyses, it was interesting to note that the non-poultry industries/plants were predominantly seafood. This thus, raised a very important question for future investigation. That is the question as to whether there is a viral component to seafood products. Specifically, whether there is an association between exposures to seafood products and oncogenic virus which was the causal element of interest in this study. A review of the literature indicates that billions of people may be exposed to certain chemical agents such as methyle mercury imbedded in the edible tissues of fish with many health consequences. Fishermen and their families have been known to suffer from neurological and other defects whose causes may still be elusive. (Clarkson, 2002; Hunter, 1969; Swedish Expert Group, 1971). However, the epidemiological literature is scanty regarding the roles of disease-causing biological agents that may be associated with seafood products/exposures.

The focus of this study was not so much on documenting which of these two groups (poultry versus non-poultry) had a higher risk of mortality from the diseases under investigation but rather to evaluate whether comparative effect measures obtained from various study design and analytical techniques will lead to the same general conclusions. Studies that report the development of new techniques typically test their hypothesis through simulations with hypothetical data and on very rare occasions with real data. Even when attempts to illustrate using real data are made, they are not of large scale and hardly involve a large number of techniques; understandably so because it is hard to find practical situations in which a single data set meets all or most of the underlying requirements of a set of analytical techniques and as such though the approaches are plausible, they do not demonstrate the extent of the challenges encounter in real life.

This study arrived at the same general conclusion regarding the risk of lung cancer and multiple myeloma mortality between poultry and non-poultry exposure when the various statistical methods were applied within a cohort design and for most of the sampling schemes for selecting controls with in the nested case-control design. Rate ratios obtained from direct standardization agreed closely with relative SMR (or rate ratio obtained from SMR) through indirect standardization. These effect measures also compared very closely with the logistic regression odds ratio, Cox proportional regression hazard ratio and Poisson regression relative risk or risk ratio estimates within the cohort analyses. These findings confirm previous hypotheses and investigations (Breslow et al, 1983; Breslow & Day, 1987; Greenland, 1991; Greenland, 1991; Szklo & Neito, 2007; Pan American Health Organization, 2002; Weitkunat et al, 2001).

It was observed that similar conclusions can still be achieved even with slight violations of the underlying requirements or assumptions of some techniques. For example, the direct standardization technique (stratified analysis) requires at least one observed event (death) in each category (stratum) of the exposure/confounding variable (Pan American Health Organization, 2002). It warns against the impending loss of precision with very small categories and the potentials of residual confounding with very large categories. In our data, very large age categories were applied to reduce the number of cells with no observed deaths. As a consequence, adjusted rates were very close to the crude yet rate ratios were still close to those obtained from other approaches. Although not clearly visible in this data, ratios of two directly standardized rates are often criticized as being subject to greater sampling variability due to small number of invents (Breslow et al, 1983; Greenland, 1991; Greenland, 2004). Of course, if the number of controlled variables (confounders) is large sparseness increases, making this technique very unpopular in multivariate settings. The same problem may also result in over dispersion in model-based analyses.

SMR from indirect standardization techniques are frequently used in epidemiology to compare different study groups. The method lend credence to the ease of computation, the fact that it provides an estimate of the relative risk between the standard population and the study population (easily interpretable), as well as the fact that a single standard may be used to calculate SMRs of different causes in a population (cost effectiveness). However, indirect adjustment is most appropriate when applied to data where the ratios of the stratum-specific rates in the comparison groups to those in the standard population are approximately constant (Breslow et al, 1983; Breslow & Day,

1987; Pan American Health Organization, 2002; Szklo & Neito, 2007; Weitkunat et al, 2001). This homogeneity requirement was hardly met in our data. The fact that the relative SMRs and rate ratios obtained from this approach were similar to the effect measures from the other techniques, may suggest that the comparison of each group and the population of reference may not always be relevant.

The use of SMR directly as a comparative measure of mortality risk has also been criticized because of the use of an external standard population. Critics to this approach hold that two standardized ratios may lie outside the interval (when using external standard). For example, when age/calendar year specific ratios for one or both subgroups are not constant, the ratio of the two standardized ratios may lie entirely outside the range of the ratios of age/year specific rates comparing the two subgroups directly (Breslow et al, 1983). Fortunately, our analysis did not make use of calendar year (an acknowledge limitation) and as noted by Breslow et al the conditions that lead to such aberrant behavior are extreme, and serious misleading inferences seem to occur rarely in practice. In addition, it is reasoned that a direct comparison of different exposure groups without reference to standard population is preferable if it can be effected without greatly increased computational costs or serious sacrifice in statistical efficiency (Breslow et al, 1983).

In general, non-model based techniques (stratified analyses) tend to produce weighted averages. Using such summary estimates alone to make causal inference is misleading in the presence of heterogeneity of effects not due to chance (mask effect modification). Hence the need for model based methods that can test hypotheses.

Even with the multivariate regression modeling techniques, some of the underlying assumptions were hard to achieve and when this occurred appropriate alternative approaches were used. For example, output from the Cox model (with exact case-cohort pseudolikelihood estimators) showed significant risk of lung cancer mortality among the poultry workers compared to the non-poultry group. However, the Cox regression model requires a proportional hazard ratio (relative risk) between the comparison groups over time. This assumption was violated graphical depiction while the hypothesis testing of the exposure-time interaction term was not statistically significant suggesting an interval Cox model was more appropriate with this data (Vonesh et al., 2000). Of course results of interval Cox model gave results that were similar to those obtained from the other model.

Apart from the underlying model specifications, the goal for a particular analysis should serve as a primary guide for selecting a particular analytic technique. Multiple logistic regression was applied to model the likelihood that individuals in one exposure would die from a particular disease compared to the other; with no regards to how much time an individual contribute to the study or was exposed. On the other hand, in analyzing aggregated mortality rates, the Poisson regression models are preferred. This requires information on mortality to be presented in summary form in a manner different from that required by the logistic model. While in analyzing individual subject's survival experience (or time) the Cox model is typical. One can use either Poisson or Cox regression to carry out subject survival analysis and still achieve similar results (Vonesh et al., 2000). In cases where one suspects that the relative risk of death between two exposure groups varies with time, an interval Poisson (a piecewise exponential model) or

interval Cox regression that includes a modality by time interaction term is recommended. The interval Poisson model is similar to the Cox model in that both account for censored data and assume the death rates between any two groups of subjects will be proportional to one another (Kutner et al., 2005; Loomis, Richardson, & Elliott, 2005; McNutt et al., 2003; Vonesh et al., 2000).

The Cox proportional hazards theory has been used to model the effects of quantitative exposures in a framework that treats the underlying death rates as nuisance parameters rather than known constants. Due to the arduous computational process, this approach has not been used for large sample explorative epidemiological studies (Breslow et al., 1983). However, the approach has found a very wide application with case-cohort designs especially with the development of new techniques for estimating the regression parameters such as the partial- and pseudo-likelihood estimators and the robust (exact)/ asymptotic variance estimators (Breslow et al., 1983; Greenland, 2004; Langholz & Jiao, 2007; McNutt et al., 2003; Zou, 2004).

In the nested case-control analyses, results from the cumulative survival, cumulative incidence and case-based sampling techniques were similar across the different statistical methods probably due to the rare outcome (1.2% prevalence of lung cancer mortality). This was consistent with previous reports (Morabia et al., 1995). However, results from concurrent or incidence density sampling were different. They showed a significantly decreased risk of lung cancer mortality among poultry workers from each of the analytical techniques. This was particularly more profound with the Cox model (with exact case-cohort pseudolikelihood estimators). The Cox proportional hazards model yields an average relative risk. In some cases, this average risk may

mislead investigators into thinking that one type of exposure is more causally related to the outcome compared to another when in fact there are periods of time when the opposite is true. An alternative is to use an interval Poisson model. This avoids this pitfall by enabling the user to model the relative risk as a function of time which can be accomplished by including an interaction term between the interval follow-up times and exposures (Kutner, 2005; Vonesh et al, 2000). Results of the interval Poisson model applied to the incidence density sampling data were closer to those from the other methods.

A major source of confusion about the proper application of the Cox model especially to cohort data relates to the notice of the appropriate time variable in the basic model. One approach, proposed by Breslow et al, is to treat follow-up time on the study as the fundamental time variable, controlling the effects of age and calendar year by stratification or covariable modeling based on a subject's status at entry into the study (Breslow et al, 1983). It ensures that risk sets formed when undertaking the analysis are decreasing in size as time increases. However, they claim that this time is often an inappropriate choice for time variable in cohort studies, for two reasons:

'First, since death rates from the major diseases of interest rise rapidly with age, age effects should be controlled as precisely as possible. Second, many exposures are measured rather imperfectly, so that duration of employment and therefore also time on study are highly correlated with, and may to some extent serve as surrogate measures for, the cumulative exposures. In such cases, controlling for time on the study in the analysis may mask the very effects that one is attempting to uncover' (Breslow et al, 1983). They also alluded to the need to accounts for the "healthy worker" selection phenomenon if

many deaths of interest occur during the period when it is operative. They suggested an approach that considered age as the underlying time variable and control for secular trends by time-dependent strata consisting of 5-year calendar periods. With such an approach then, workers dying of lung cancer among the poultry group are compared with others reaching the same age in the same calendar period, irrespective of duration of previous employment. Follow-up time per se may then be ignored on the grounds that very few lung cancer deaths occurred with in 15 years of the study entry. Also, although selection factors may be less important for cancer (compared to other diseases), these factors need to be carefully controlled because Lung cancer mortality increased sharply with both age and year during the study period (Breslow et al, 1983). In this regards, it appeals to reason to make these considerations in future analyses. Also, one may wish to stratify further by date of hire, duration of employment, or length of time on study, after considering the degree to which these variables may be confounded with the exposures and the consequent difficulties of interpretation.

Conclusion

This study was designed to compare non-model and model based statistical techniques typically applied in cohort mortality analyses and various schemes for selecting controls in nested case-control studies to document risk for lung cancer and multiple myeloma mortality, among workers of poultry slaughtering/processing plants. The exposure of these workers to oncogenic virus is conceivably higher compared to the general public. The entire cohort and subgroups of poultry and non-poultry workers separately showed higher risks of mortality from both malignant diseases (statistically significant for lung

cancer) but slightly lower (statistically not significant) risks among poultry compared to non-poultry workers. Results of comparative effect measures from the various statistical methods under consideration were similar (up to one decimal place) with a very slight difference in variability/precision within the cohort analyses. The effect measures were also similar for nested case-control analyses that applied the cumulative survival, cumulative incidence and case-base sampling schemes in selecting controls. However, the incidence density sampling scheme led to markedly different results (both in magnitude and statistical significance), that were more profound with the Cox regression model. Where the Cox model was not appropriate the interval Poisson (exponential) model was used and predictions were similar to those obtained using other methods.

TABLES

Table 1: Summary statistics, standardized mortality ratio (SMR), and proportionate mortality ratio (PMR) for selected causes of death for members of a local Union Pension Fund of UFCW International Union (Full cohort, N=30,488 using US standard rates from 1972-2005)

		Poultry	**Non-poultry**	**Total**
Number of persons at risk N (%)		20132 (66.0)	10356 (34.0)	30488
White N (%)		13733 (68.2)	9511 (91.8)	23244 (76.2)
Nonwhite N (%)		6399 (31.8)	845 (8.2)	7244 (23.8)
Male N (%)		9631 (47.8)	5723 (55.3)	15354 (50.4)
Female N (%)		10501 (52.2)	4633(44.7)	15134 (49.6)
Mean duration of observation (yrs)		23.8± 5.2	25.2±6.5	24.3±5.7
Total person years		478708.5	261004.7	739710.2
Mean age at entry to study (yrs)		27.8±10.3	30.8±12.2	28.8±11.1
Mean age at exit of study		51.6±10.8	56.0±12.7	53.1±11.7
All causes of death	Observed	2454	1665	4119
	Expected*	2002.8	1658.7	366.2
	SMR	1.23(1.18, 1.28)	1.00 (0.96, 1.05)	1.13 (1.09, 1.16)
	PMR$	1.00	1.00	1.00
All Malignant neoplasms	Observed	560	402	962
	Expected	533.3	446.1	979.1
	SMR	1.05 (0.97, 1.14)	0.90 (0.82, 0.99)	0.98 (0.92, 1.05)
	PMR	0.81 (0.76, 0.87)	0.88 (0.81, 0.96)	0.84 (0.80, 0.88)
Lung cancer	Observed	217	161	378
	Expected	147.8	123.7	271.5
	SMR	1.47 (1.28, 1.68)	1.30 (1.11, 1.52)	1.39 (1.26, 1.54)
	PMR	1.14 (1.01, 1.30)	1.26 (1.09, 1.46)	1.19 (1.08, 1.31)
Multiple myeloma	Observed	12	8	20
	Expected	8.2	7.3	156
	SMR	1.46 (0.75, 2.55)	1.09 (0.47, 2.16)	1.29 (0.79, 1.99)
	PMR	1.16 (0.66, 2.04)	1.09 (0.55, 2.18)	1.13 (0.73, 1.75)

**Expected deaths were computed based on the United States standard rates from 1972 to 2005; output from OCMAP+*
$PMRs are shown just to provide a check for the SMRs. They are not computed from the expected deaths shown on this table
UFCW=United Food & Commercial Workers

Table 2: Standardized mortality ratio (SMR), and proportionate mortality ratio (PMR) for selected causes of death for members of a local Union Pension Fund of UFCW International Union (Full cohort, N=30,488 using US standard rates from 1972-2001)

		Poultry	**Non-poultry**	**Total**
Number of persons at risk N (%)		20132 (66.0)	10356 (34)	30488
Total person years		478708.5	261004.7	739710.2
All causes of death	Observed	2454	1665	4119
	Expected*	2010.6	1667.9	5443.1
	SMR	1.22 (1.17, 1.27)	1.00 (0.95, 1.05)	0.76 (0.73, 0.78)
	PMR	1.00	1.00	1.00
All Malignant neoplasms	Observed	560	402	962
	Expected	537.2	448.9	986.1
	SMR	1.04 (0.96, 1.13)	0.90 (0.81, 0.99)	0.98 (0.92, 1.04)
	PMR	0.81 (0.75, 0.86)	0.88 (0.81, 0.96)	0.84 (0.79, 0.88)
Lung cancer	Observed	217	161	378
	Expected	148.8	124.3	273.1
	SMR	1.46 (1.27, 1.67)	1.30 (1.10, 1.51)	1.38 (1.25, 1.53)
	PMR	1.14 (1.00, 1.29)	1.26 (1.09, 1.46)	1.19 (1.08, 1.31)
Multiple myeloma	Observed	12	8	20
	Expected	8.3	7.4	15.7
	SMR	1.44 (0.75, 2.52)	1.09 (0.47, 2.14)	1.28 (0.78, 1.97)
	PMR	1.15 (0.66, 2.03)	1.09 (0.55, 2.18)	1.13 (0.73, 1.75)

**Expected deaths were computed based on the United States standard rates from 1972 to 2001; output from OCMAP+*
UFCW=United Food & Commercial Workers

Table 3: Five-year interval (15 categories) age group distribution of number at risk, lung cancer and multiple myeloma deaths in a local Union Pension Fund of UFCW (N= 30,488)

	Poultry			Non-poultry			Total		
Age	At risk	Lung*	Myeloma$	At risk	Lung	Myeloma	At risk	Lung	Myeloma
<20	5	0	0	2	0	0	7	0	0
20-24	64	0	0	22	0	0	86	0	0
25-30	98	1	0	34	0	0	132	1	0
30-34	128	6	0	57	0	0	185	6	0
35-39	1149	6	0	386	4	1	1535	10	1
40-44	4127	17	0	1174	9	0	5301	26	0
45-49	4583	17	1	2093	14	0	6676	31	1
50-54	3902	30	1	1799	17	1	5701	47	2
55-59	2023	35	2	1406	19	1	3429	54	3
60-64	1388	30	3	951	27	1	2339	57	4
65-69	1064	34	1	709	22	2	1773	56	3
70-74	703	27	0	606	21	0	1309	48	0
75-79	474	6	3	499	19	1	973	25	4
80-84	293	6	1	349	8	1	642	14	2
85+	131	2	0	269	1	0	400	3	0
Total	20132	217	12	10356	161	8	30488	378	20

** Lung cancer refers ICD-9 code 162 for malignant neoplasm of the trachea, bronchus and lung*
$Multiple Myeloma refers to ICD-9 code203 for malignant immunoproliferative disease, multiple myeloma, and malignant plasma cell.
This table shows many empty cells with deaths for age and plant distributions only (sparseness would be exacerbated when distributed by race and sex are added
UFCW=United Food & Commercial Workers

Table4: Distribution of number at risk and number of lung cancer/ multiple myeloma deaths in a local Union Pension Fund of UFCW (N=30,488) by age, race, and sex

Age Group (years)	Poultry			Non-poultry			Total		
	At risk	**Lung***	**Myeloma$**	**At risk**	**Lung**	**Myeloma**	**At risk**	**Lung**	**Myeloma**
ALL RACE AND SEX									
All age groups	20132	217	12	10356	161	8	30488	378	20
≤40	2003	18	0	629	6	1	2632	24	1
40 - 60	14076	94	4	6344	57	2	20420	151	6
60+	4053	105	8	3383	98	5	7436	203	13
WHITE MALE									
All age groups	6782	96	2	5217	89	3	11999	185	5
≤40	883	10	0	405	3	0	1288	13	0
40 - 60	4835	42	1	3519	32	1	8354	74	2
60+	1064	44	1	1293	54	2	2357	98	3
WHITE FEMALE									
All age groups	6951	70	3	4294	57	5	11245	127	8
≤40	644	6	0	172	3	1	816	9	1
40 - 60	4580	27	1	2351	20	1	6931	47	2
60+	1727	37	2	1771	34	3	3498	71	5
NONWHITE MALE									
All age groups	2849	31	5	506	10	0	3355	41	5
≤40	269	1	0	42	0	0	311	1	0
40 - 60	2128	17	1	329	4	0	2457	21	0
60+	452	13	4	135	6	0	587	19	4
NONWHITE FEMALE									
All age groups	3550	20	2	339	5	0	3889	25	2
≤40	207	1	0	10	0	0	217	1	0
40 - 60	2533	8	1	145	1	0	2678	9	1
60+	810	11	1	184	4	0	994	15	1
ALL WHITE									
All age groups	13733	166	5	9511	146	8	23244	312	13
≤40	1527	16	0	577	6	1	2104	22	1
40 - 60	9415	69	2	5870	52	2	15285	121	4
60+	2791	81	3	3064	88	5	5855	169	8
ALL NONWHITE									
All age groups	6399	51	7	845	15	0	7244	66	7
≤40	476	2	0	52	9	0	528	2	0
40 - 60	4661	25	2	474	5	0	5135	30	2
60+	1262	24	5	319	10	0	1581	34	5
ALL MALE									
All age groups	9631	127	7	5723	99	3	15354	226	10
≤40	1152	11	0	447	3	0	1599	14	0
40 - 60	6963	59	2	3848	36	1	10811	95	3
60+	1516	57	5	1428	60	2	2944	117	7
ALL FEMALE									
All age groups	10501	90	5	4633	62	5	15134	152	10
≤40	851	7	0	182	3	1	1033	10	1
40 - 60	7113	35	2	2496	21	1	9609	56	3
60+	2537	48	3	1955	38	3	4492	86	6

** Lung cancer refers ICD-9 code 162 for malignant neoplasm of the trachea, bronchus and lung*

$Multiple Myeloma refers to ICD-9 code203 for malignant immunoproliferative disease, multiple myeloma, and malignant plasma cell

UFCW=United Food & Commercial Workers

Table 5: Summary statistics, standardized mortality ratio (SMR), and proportionate mortality ratio (PMR) for selected causes of death for members of a local Union Pension Fund of UFCW International Union (Sub comparative cohort, N=20,712 using US standard rates from 1972-2005)

		Poultry	**Non-poultry**	**Total**
Number of persons at risk N (%)		10356 (50.0)	10356 (50.0)	20712
White N (%)		7065 (68.2)	9511 (91.8)	16576 (80.0)
Nonwhite N (%)		3291 (31.8)	845 (7.2)	4136 (20.0)
Male N (%)		4972 (48.0)	5723 (55.3)	10696 (51.6)
Female N (%)		5384 (52.0)	4633 (44.7)	10017 (48.4)
Mean duration of observation (yrs)		23.8± 5.2	25.2±6.5	24.5±5.9
Total person years		246425.3	261004.7	507427.0
Mean age at entry to study (yrs)		27.7±10.2	30.8±12.2	29.3±11.4
Mean age at exit of study		51.5±10.8	60.0±12.7	53.8±12.0
All causes of death	Observed	1249	1665	2914
	Expected*	1022.5	1658.7	2681.2
	SMR	1.22 (1.16, 1.29)	1.00 (0.96, 1.05)	1.09 (1.05, 1.13)
	PMR	1.00	1.00	1.00
All Malignant neoplasms	Observed	282	402	684
	Expected	271.6	446.1	717.7
	SMR	1.03 (0.92, 1.17)	0.90 (0.82, 0.99)	0.95 (0.88, 1.03)
	PMR	0.80 (0.73, 0.88)	0.88 (0.81, 0.96)	0.84 (0.80, 0.90)
Lung cancer	Observed	114	161	275
	Expected	75.3	123.7	199.0
	SMR	1.52 (1.25, 1.82)	1.30 (1.11, 1.52)	1.38 (1.22, 1.56)
	PMR	1.18 (1.00, 1.41)	1.26 (1.09, 1.46)	1.23 (1.10, 1.37)
Multiple myeloma	Observed	5	8	13
	Expected	4.2	7.3	11.5
	SMR	1.19 (0.39, 2.78)	1.09 (0.47, 2.16)	1.13 (0.60, 1.93)
	PMR	0.96 (0.40, 2.30)	1.09 (0.55, 2.18)	1.04 (0.60, 1.78)

**Expected deaths were computed based on the United States standard rates from 1972 to 2005; output from OCMAP+*
UFCW=United Food & Commercial Workers

Table 6: Directly standardized death rate (per 100,000) and rate ratio [RR (95% confidence interval)[€]] for poultry versus non-poultry exposures (Union Pension Fund of UFCW)

	Poultry	Non-poultry	Combined
Lung cancer: full cohort (N=30,488)			
Observe deaths	217	161	378
Person years	478653.4	260973.4	739626.8
Crude Rate	45.3	61.7	51.1
Crude RR	0.73 (0.62, 0.89)		
Expected deaths	188.5	205.2	
Standardized rate*	25.5	27.7	
Standardized RR	0.92 (0.78, 1.17)		
Lung cancer: sub-cohort (N=20,712)			
Observe deaths	114	161	275
Person years	246396.9	260973.4	507370.3
Crude Rate	46.3	61.7	54.2
Crude RR	0.75 (0.63, 0.90)		
Expected deaths	140.5	144.5	
Standardized rate*	27.7	28.5	
Standardized RR	0.97 (0.79, 1.19)		
Multiple myeloma: full cohort (N=30,488)			
Observe deaths	12	8	20
Person years	478653.4	260973.4	739626.8
Crude Rate	2.5	3.1	2.7
Crude RR	0.82 (0.39, 1.89)		
Expected deaths	9.3	9.9	
Standardized rate*	1.3	1.3	
RR	0.94 (0.44, 2.36)		
Multiple myeloma: sub-cohort (N=20,712)			
Observe deaths	5	8	13
Person years	246396.9	260973.4	507370.3
Crude Rate	2.0	3.1	2.6
Crude RR	0.66 (0.18, 1.61)		
Expected deaths	4.8	7.0	
Standardized rate*	0.9	1.4	
RR	0.68 (0.17, 1.83)		

[€]*Confidence interval computed using the percentile bootstrap technique*

**Poultry and non-poultry rates were standardized using the structure of the combined population, serving as the standard population for this analysis*

UFCW=United Food & Commercial Workers

.

Table 7: Standardized mortality ratio (SMR), relative SMR, indirectly standardized death rates (per 100,000) and rates ratio (RR) with 95% confidence interval[€] for poultry versus non-poultry exposures (Union Pension Fund of UFCW)

	Poultry	Non-poultry	Poultry standardized by non-poultry rates[$]
Lung Cancer: full cohort (N=30,488)			
Observe deaths	217	161	217
Expected deaths[*]	220.5	157.5	249.4
SMR	0.98	1.02	0.87 (0.71, 1.11)
RSMR[ξ]	0.96 (0.82, 1.17)		
Rc x SMR[§]	44.6	63.1	
RR	0.71 (0.51, 1.03)		
Lung Cancer: sub-cohort (N=20,712)			
Observe deaths	114	161	114
Expected deaths[*]	115.1	159.9	128.1
SMR	0.99	1.01	0.89 (0.71, 1.13)
RSMR[ξ]	0.98 (0.82, 1.18)		
Rc x SMR[§]	45.8	62.1	
RR	0.74 (0.51, 1.05)		
Multiple myeloma: full cohort (N=30,488)			
Observe deaths	12	8	12
Expected deaths[*]	12.6	7.4	11.8
SMR	0.95	1.08	1.02 (0.47, 3.03)
RSMR[ξ]	0.89 (0.46, 1.89)		
Rc x SMR[§]	2.4	3.3	
RR	0.72 (0.19, 3.53)		
Multiple myeloma: sub-cohort (N=20,712)			
Observe deaths	5	8	5
Expected deaths[*]	6.0	7.0	6.0
SMR	0.83	1.15	0.83 (0.22, 2.75)
RSMR[ξ]	0.72 (0.23, 1.59)		
Rc x SMR[§]	1.7	3.5	
RR	0.54 (0.04, 2.62)		

[€]*Confidence interval computed using the percentile bootstrap technique*

*Expected deaths obtained using rates from the combined population serving as standard rates

[ξ]RSMR= Relative SMR =ratio of SMR for poultry to SMR for non-poultry

[§]Rc x SMR =standardized rate for each group obtained as product of crude rate (Rc) and SMR

$Expected deaths and SMR under this column are obtained for the poultry group using rates from the non-poultry (the comparative), serving as the standard population.

UFCW=United Food & Commercial Workers

Table 8: Summary of comparative analysis of risk estimation for lung cancer and multiple myeloma mortality due to poultry versus non-poultry exposures applying various analytical techniques on the full Union Pension Fund Cohort (N=30,488)

Methods	Effect measures	Lung cancer	Multiple myeloma
	Non-model-based		
Direct standardization	RR	0.92 (0.78, 1.17)	0.94 (0.44, 2.36)
Indirect standardization	RSMR	0.96 (0.82, 1.17)	0.89 (0.46, 1.89)
	RR	0.71 (0.51, 1.03)	0.72 (0.19, 3.53)
	SMR*	0.87 (0.71, 1.11)	1.02 (0.47, 3.03)
	Model-based		
Poisson regression	RR	1.00 (0.81, 1.24)	0.98 (0.37, 2.58)
Logistics regression	OR	0.96 (0.78, 1.20)	0.93 (0.35, 2.44)
Cox proportional hazards regression	HR[1]	1.08 (0.87, 1.34)	1.22 (0.45, 3.36)
	HR[2]	0.90 (0.52, 1.56)	0.33 (0.02, 4.55)
	HR[3]	1.01 (0.81, 1.25)	1.01 (0.38, 2.68)
	HR[4]	0.98 (0.79, 1.21)	0.95 (0.36, 2.51)

RR=rate ratio estimation of risk ratio, SMR= standardized mortality ratio, RSMR= relative SMR, OR=odds ratio, HR =hazard ratio

**SMR calculated for Poultry with the non-poultry group serving as the standard population*
[1]Hazard ratio from a Cox model with no exposure-time interaction
[2]Hazard ratio from a Cox model with exposure-time interaction
[3]Hazard ratio from an interval Cox model (5-year interval time to event)
[4]Hazard ratio from an interval Cox model (10-year interval time to event)

Table 9: Summary of relative effect measures for lung cancer mortality due to poultry versus non-poultry exposures from a nested case-control design based on different control sampling schemes and analytical techniques (a local Union Pension Fund of UFCW)

Sampling scheme	Total subjects	Cases n (%)	Logistic OR (95%CI)	Poisson RR(95%CI)	Cox HR(95%CI)
Cumulative survival	1890	378 (20)	0.84 (0.66, 1.08)	0.88 (0.73, 1.06)	0.88 (0.71, 1.09)
Cumulative incidence	1890	378 (20)	0.84 (0.66, 1.08)	0.88 (0.71, 1.09)	0.88 (0.71, 1.09)
Case-cohort	1877	378 (20)	0.93 (0.72, 1.19)	0.94 (0.78, 1.14)	0.95 (0.77, 1.18)[*] 1.31 (1.02, 1.58)[§]
Incidence density	870	370 (43)	0.66 (0.56, 0.77)[ξ]	0.48 (0.41, 0.55)	0.55 (0.45, 0.68)[*] 0.52 (0.44, 0.60)[§]

OR=odds ratio, RR=risk ratio or relative risk, and HR=hazard ration estimation of rate ratio

[*]*Hazard ratio from an interval Cox model (5-year interval time to event)*

[§]*Hazard ratio from a Cox model (due to Langholz and Jiao) to compute exact case-cohort pseudolikelihood estimator for rate ratio*

[ξ]*Odds ratio from conditional logistic regression model with the 'ties=breslow' option deemed appropriate for a 1: m individually matched data*

UFCW=United Food & Commercial Workers

References

Alberg, A. J., Samet, J. M. (2003). Epidemiology of lung cancer. *Chest*.123:21-49.

Avet-Loiseau, H., Gerson, F., Magrangeas, F., Minvielle, S., Harousseau, J., Bataille, R for the Intergroupe Francophone du Myélome (2001). Rearrangements of the *c-myc* oncogene are present in 15% of primary human multiple myeloma tumors *Blood*. 98 (10):3082-3086

Breslow, N. E., Day, N. E. (1980). Statistical methods in Cancer Research. Volume I – The analysis of case-control studies. Lyon, International Agency for research on cancer; 32

Breslow, N. E., Day, N. E. (1987). Statistical methods in Cancer Research. Volume II – The design and analysis of cohort studies. Lyon, International Agency for research on cancer; 82.

Breslow, N. E., Lubin, J. H., Marek, P. Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*. 78(581):1-13.

Brouchet, L., Valmary, S., Dahan, M., Didier, A., Galateau-Salle, F. (2005). Detection of oncogenic virus genomes and gene products in lung carcinoma *British Journal of Cancer*. 92, 743–746.

Carbone, M., Pass, H. I., Miele, L., & Bocchetta, M. (2003). New developments about the association of SV40 with human mesothelioma. *Oncogene*. 22: 5173–5180.

Chen, K. (1999) Case-cohort and Case control analysis with Cox's model. *Biometrik.* 8(4), 755-764.

Cheng, Y.W., Chiou, H. L., Sheu, G.T., Hsieh, L. L., Chen, J. T., Chen, C. Y., Su J. M.,

Lee, H. (2001). The association of human papillomavirus 16/18 infection with lung cancer among nonsmoking Taiwanese women. *Cancer Res*earch. 61: 2799–2803.

Chesi, M. P., Bergsagel, L., Shonukan, O., Martelli**,** M., Brents**,** L., et al. (1998). Frequent Dysregulation of the c-*maf* Proto-Oncogene at 16q23 by Translocation to an Ig. *Locus in Multiple Myeloma Blood*. Vol. 91 No. 12 (June 15), pp. 4457-4463.

Choudat, D., Dambrine, G., Delemotte, B., & Coudert, F. (1996). Occupational exposure to poultry and prevalence of antibodies against Marek's disease virua and avian leukosis retroviruses. *Occupational and Environmental Medicine.* 53: 403-410.

Chung, K. & Lee, S. (2001). Optimal bootstrap sample size in construction of percentile confidence bounds. *Board of the Foundation of the Scandinavian Journal of Statistics.* 28:225-239.

Clarkson, TW. (2002). The three modern faces of mercury. *Environmental Health Perspectives;* 110 (1):11-23.

Davison, A. C. & Hinkley, D. (2006) Bootstrap methods and their applications. 8th ed. Cambridge: Cambridge Series in Statistical and Probabilitic Mathematics.

Dib, A., Gabrea, A., Glebov, O., Bergsagel, L., Kuehl, M. (2008). Characterization of MYC Translocations in Multiple Myeloma Cell Lines. *Journal of the National Cancer Institute Monographs.* (39):25-31

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. 38 Society of Industrial and Applied Mathematics CBMS-NSF Monographs.

Encyclopedia of Epidemiological methods. (2001) Gail, M. & Benichou, J. Wiley reference series in Biostatistics. John Wiley & Sons, Ltd. Chichester, NY.

Giuliani, L., Jaxmar, T., Casadio, C., Gariglio, M., Manna, A., D'Antonio, D., Syrjanen, K., Favalli, C., Ciotti, M. (2007). Detection of oncogenic viruses SV40, BKV, JCV, HCMV, HPV and p53 codon 72 polymorphism in lung carcinoma. *Lung Cancer*. Sep;57(3):273-81.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poison and negative binomial regression models. *Tenique report*

Greenland, S., Schwartzbaum, J., Finkle, W. (2000). Problems due to small samples and sparse data in conditional logistic Regression Analysis. *American Journal of Epidemiology*. 151,(5), 531-539.

Greenland, S. (1991). Estimating standardized parameters from generalized linear models. *Stat Med*;10:1069–74.

Greenland, S., Schwartzbaum, J. A., Finkle, W. D. (2000). Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*. 151:531–9.

Greenland, S. (1999). The relation of the probability of causation to the relative risk and the doubling dose: a methodological error that has become a social problem. *American Journal of Public Health*. 89:1166–9.

Greenland, S. (2004) Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology*. 160(4):301-305.

Greenland, S. Interval estimation by simulation as an alternative to and extension

of confidence intervals. *International Journal of Epidemiology* (in press).

Hunter D. (1969). Diseases of occupations. Boston: Little brown; 314-328

Hussain, A. I., Shanmugam, V., Switzer, W. M., Tsang, S. X., Fadly, A., Thea, D., Helfand, R., Bellini, W.J., Folks, T.M., Heneine, W. (2001). Lack of evidence of endogenous avian leukosis virus and endogenous avian retrovirus transmission to measles mumps rubella vaccine recipients. *Emerging Infectious Diseases*. 7(1).

Hussain, A. I., Johnson, J. A., da Silva-Freire, M., Heneine, W. (2003). Identification and characterization of avian retroviruses in chicken embryo-derived yellow fever vaccines. Investigation of transmission to vaccine recipients. *Journal of Virology.* 77(2): 1105-1111.

Joffe, M.M., Greenland, S. (1995). Estimation of standardized parameters from categorical regression models. *Statistics in Medicine.* 14:2131–41.

Johnson, E. S., Zhou, Y., Yau, C. L., Prabhakar, D., Ndetan, H., Singh, K, Preacely, N. (2009a). Mortality from Malignant Diseases - Update of the Baltimore Union Poultry Cohort. *Cancer Causes & Control*.

Johnson, E. S., Yau, L., Zhou, Y., Singh, K., Ndetan, H. (2009b). Mortality in Baltimore Union Poultry Cohort: non-malignant diseases. *International Archive of Occupational & Environmental Health;*

Johnson, E. S., Griswold, C. (1996). REV - Oncogenic retrovirus of cattle, chickens and turkeys -Potential infectivity and oncogenicity for humans. *Medical Hypotheses.* 46: 354-356.

Johnson, E. S., Ndetan H, Sarda V, Bankuru S, Felini M. Update of Cancer Mortality in

the Missouri Poultry Union. (Manuscript being prepared for submission)

Johnson, E. S., Ndetan, H., Lo, K-M.(2009c) Cancer Mortality in Poultry Slaughtering/Processing Plant Workers Belonging to a Union Pension Fund. (Manuscript being prepared for submission)

Johnson, ES., Ndetan, H., Sarda V., Bankuru, S., Felini, M.(2009d) Update of Cancer Mortality in the Missouri Poultry Union (Manuscript being prepared for submission)

Johnson, E. S. (2005). Assessing the role of transmissible agents in human disease by studying meat workers. *Cellscience Reviews.* 2 (1), 1-15.

Johnson, E. S., Yi, Zhou.(2007). Non-Cancer Mortality in Supermarket Meat Workers. *Journal occupational and environmental Medicine.* 49, 1-7.

Johnson, E. S., Shorter, C., Rider, B., Jiles, R. (1997). Mortality from Cancer and Other Diseases in Poultry Slaughtering Processing Plants. *International Journal of Epidemiology.* 26:1142-1150.

Johnson, E. S., Yi, Zhou., Macodou, sall., Mohammed El Faramawi., Nihita Shah., Anitha Christopher., Nigel Lewis.(2007). Non-Malignant disease mortality in meat workers: a model for studying the role of zoonotic transmissible agents in non malignant chronic diseases in humans. *Occupational and Environmental Medicine.* 000,1-7 .

Johnson, E. S. (1986). Correspondence : PMR and relative risk. *British Journal of Industrial Medicine.* 43:214-216.

Johnson, E. S., Fischman, H. R., Matanoski, G. M., & Diamond, E. (1986a). Cancer

occurrence in women in the meat industry. *British Journal of Industrial Medicine* 43: 597-604.

Johnson, E. S., Fischman, H. R., Matanoski, G. M. & Diamond, E. (1986b). Cancer mortality among white males in the meat industry. *Journal Occupational Medicine* 28(1):23-32.

Johnson, E. S. (1987a). Noncancer mortality in the meat industry: white males. *Journal of Occupational Medicine*, 29(4):330-4.

Johnson, E. S. (1987b). Mortality from non-malignant diseases among women in the meat industry. *British Journal of Industrial Medicine*, 44(1):60-3.

Johnson, E. S., Nicholson, L.G., Durack, D. T. (1995b). Detection of Antibodies to Avian Leukosis/Sarcoma Viruses (ALSV) and Reticuloendotheliosis Viruses (REV) in Humans by ELISA. *Cancer Detection and Prevention.* 19(5):394-404.

Johnson, E. S., Overby, L., Philpot, R. (1995a). Detection of Antibodies to Avian Leukosis/Sarcoma Viruses (ALSV) and Reticuloendotheliosis Viruses (REV) in Humans by Western Blot Assay. *Cancer Detection and Prevention.* 19(6):472-486.

Johnson, E. S. (1994). Poultry oncogenic retroviruses and humans. *Cancer Detection and Prevention.* 18(1):9-30.

Kramer, S. (1988). Clinical Epidemiology and Biostatistics. A primer for clinical in investigators and decision-makers. Berlin Heidelberg, German: Springer-Verlag.

Kutner, Nachtsheim, Neter, Li, (2005). Applied linear statistical models. 5th ed. Mc Graw- Hill Irwin.

Lambert, D. (1992). Zero-inflated Poisson regression models with an application to defects in manufacturing. *Technometrics*. 34:1-14.

Langholz, B., & Jiao, J. (2007). Computaional methods for case-cohort studies. *Computational Statististics & Data Analysis*. 51: 3737-3748.

Liddell, F. D. (1984). Simple exact analysis of the standardized mortality ratio. *Journal of Epidemiological Community Health*. 38:85-88.

Loomis, D., Richardson, D., Elliott, L. (2005). Poisson regression analysis of ungrouped data. *Occupational Environment Medicine*. 62,325-329.

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 Environmental Medicine*. 40(4):351-62.

McNutt, L. A., Wu, C., Xue, X., Hafner, J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*. 157:940–943.

Metayer, C. Johnson, E. S., Rice, J. C. (1998). Nested case-control study of tumors of the hemopoietic and lymphatic systems among workers in the meat industry. *American Journal of Epidemiology*. 147(8):727-738.

Miyagi, J., Tsuhako, K., Kinjo, T., Iwamasa, T., Hirayasu, T. (2000). Recent striking changes in histological differentiation and rate of human papillomavirus infection in squamous cell carcinoma of the lung in Okinawa, a subtropical island in southern Japan. *Journal of Clinical Pathology*. 53: 676–684

Morabia, A., Have, T. T., Landis, J. R. (1995). Empirical evaluation of the influence of

control selection schemes on relative risk estimation: the Welsh nickel workers study. *Occupational and Environmental Medicine.* 52:489-493.

Moore, P. S., Chang, Y. (1998). Kaposi's Sarcoma-Associated Herpesvirus-Encoded Oncogenes and Oncogenesis. *Journal of the National Cancer Institute Monographs*. 23:65-71.

Netto, G. F., Johnson, E. S. (2003) Mortality in workers in poultry slaughtering/processing plants: the Missouri poultry cohort study. *Journal of Occupational and Environmental Medicine.* 60:784-788.

Novikov, I., Oberman, B., Freedman, L. (2005). Modification of the computational procedure in parker and Bregman's method of calculating sample size from matched case-control studies with a dichotomous exposure. *Biometrics.* 61, 1123-1127.

Onland-moret, C., Van der, D., Van Der schouw, Y., Buschers, W., Elisa, S., Van Gils, et al., (2007). Analysis of case-cohort data: A comparison of different methods. *Journal of Clinical Epidemiology.* 60,350-355.

Pagano, M. & Gauvreau, K. (1993). Principles of biostatistics. Belmont, California: Wadsworth, Inc.

Pham, T. D., Spencer, L. J., Johnson, E. S. (1999). Detection of avian leukosis virus in albumen of chicken eggs using reverse transcription polymerase chain reaction. *Journal of Virological Methods*. 78:1-11.

Pham, T. D., Spencer, J. L., Johnson, E. S. (1999). Detection of avian Leukosis Virus in albumen of chicken eggs using reverse transcription polymerase chain reaction. *Journal of Virological Methods.,* 78,1-11.

Preacely, N., Felini, M., Shah, N., Christopher, A., Sarda, V., Elfaramawi, M., Sall, M., Bangara, S., Gandhi, S., Johnson, E. S., Ndetan, H. A pilot case-cohort study of lung cancer in poultry & control workers. (Submitted)

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

Rettig, M. B., Ma, H. J., Vescio, R. A., Põld, M., Schiller, G., Belson, G., et al. (1997). Sarcoma-Associated Herpesvirus Infection of Bone Marrow Dendritic Cells from Multiple Myeloma Patients. *Science*. 276 (5320):1851 – 1854

Rothman, K. J., Greenland, S. (1998). Modern epidemiology. 2nd edition. Philadelphia, PA: Lippincott-Raven.

Saif, Y. M., Barnes J. H., Fadly, A. M., Swayne, D., & Glisson, J. R. (2003). Diseases of Poultry. 11th Edition. *Iowa State Press*.

Stokes, M., Davis, C., Koch, G. (1995). Categorical data analysis using the SAS system. SAS Institute Inc, Cary, NC.

Swedish Expert Group (1971). Methylemercury in fish. A toxicological epidemiological evaluation of risk. Nord Hyg Tidskr 4(suppl):19-364

Syrjanen KJ. (2002). HPV infections and lung cancer. *Journal of Clinical Pathology*. 55: 885–891.

Szklo, M. & Neito, J. (2007). Epidemiology: beyond the basics. 2nd Ed. Jones and Bartlett Publishers Sudbury, Massachusetts.

The analysis group, Pan American Health Organization, PAHO. (2002). Special program for health analysis. *Epidemiological Bulletin*; 23(3): 1-5.

Thomas, D. (1998). New techniques for the Analysis of Cohort Studies. *Epidemiology Reviews*. 20(1), 122-131.

Timm, N. & Mieczkowski, T. (1997). Univariate & multivariate general linear models: theory and applications using SAS software. SAS Institute Inc, Cary, NC.

Tsang, S. X., Switzer, W. M., Shanmugam, V., Johnson, J. A., Golsmith, C., Wright, A., Fadly, A., Thea, D., Jaffe, H., Folks, T. M., & Heneine, W. (1999) Evidence of avian leucosis virus subgroup E and endogenous avian virus in measles and mumps vaccines derived from chicken cells: investigation of transmission to vaccine recipients. *Journal of Virology.* 73(7): 5843-5851.

Vonesh, E. F., Schaubel, D. E., Hao, W., Collins, A. J. (2000). Statistical methods for comparing mortality among ESRD patients: examples of regional/international variations. *Kidney International*. 57, 19-27.

Wacholder, S., McLaughlin, J. K., Silverman, D. T., Mandel, J. S. (1992). Selection of controls in case control studies. *American journal of Epidemiology*., 135(9), 1019-1028.

Weitkunat, R., Crispin, A., Grill, E., Fischer, R., Meyer, N., Schotten, K. (2001). Standardization of non-aggregated data: theory and practice. *Computer Methods and Programs in Biomedicine.* 65:207-227.

Wong, O, Decoufle, P. (1982). Methodological issues involving the standardized mortality ratio and proportionate mortality ratio occupational studies. *British Journal of Industrial Medicine.* 24:299-304.

Wong, O., Morgan, R. W., Kheifets, L., Larson, S. R. (1985). Comparison of SMR,

PMR, and PCMR in a cohort of union members potentially exposed to diesel exhaust emissions. *British Journal of Industrial Medicine*. 42:449-60.

Zhang, J., Yu, K. (1998a). A method of correcting the odds ratio in cohort studies of common outcomes. *American Medical Association.* 280,(19), 1690-1691.

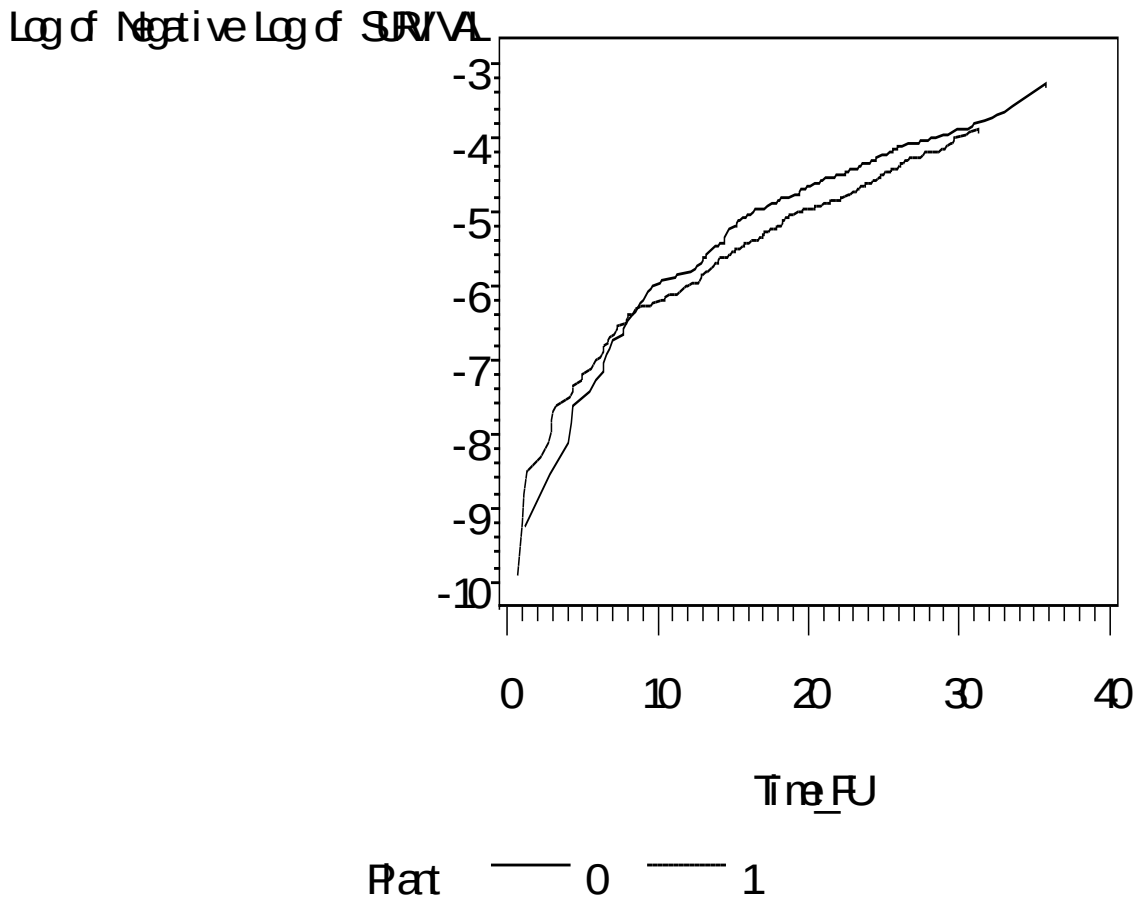
Zhang, J., Yu, K. (1998b). What's a relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*. 280:1690–1691.

Zheng, H., Abdel Aziz, H., Nakanishi, Y., Masuda, S., Saito, H., et al. (2007). Oncogenic role of JC virus in lung cancer Pathological. *Society of Great Britain and Ireland.* 212 (3): 306 – 315.

Zou, G. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology.* 159:702–6.

FIGURES

Figure1: Evaluation of the proportionality assumption for the Cox proportional hazard regression model



**The proportionality assumption seems to be grossly violated from this curve, however the exposure (plant)-time interaction term is not significant, suggesting a non-violation of the assumption.*
Plant 1=poultry, plat 0=non-poultry

APPENDICES

APPENDIX A

SAS Source code for ICD conversion (from ICD-6, 7, 8 &10 to ICD-9, and coding of OCMAP impossible ICDs)

```
*-----------------------------------------------------------------------*
| STUDY      LUNG CANCER/MULTIPLE MYELOMA RISK IN CHICAGO POULTRY      |
|            WORKERS EXPOSED TO ONCOGENIC VIRUS                        |
|                                                                      |
| JOB        COMPARING STATISTICAL METHODS IN DOCUMENTING C. MORTALITY |
| TASK       PREPARATION OF COHORT AND DEFINITION OF VARIABLES         |
|                                                                      |
| NAME       Dr.PH DISSERTATION                                        |
|                                                                      |
| FUNCTION   ICD CONVERSION FROM ICD-8 and ICD-10 to ICD-9             |
|                                                                      |
| FILES      Input: updated1.sas7bdat' (combined Chicagodata file)     |
|            Output: CHICAGO.DAT                                       |
|                                                                      |
| LANGUAGE   SAS 9.1.2                                                 |
|                                                                      |
| AUTHOR     HARRISON NDETAN                                           |
|                                                                      |
| DATE       JULY 22, 2009                                             |
|            Modified & re-run 7/23/2009, 7/24/09                      |
|                                                                      |
*-----------------------------------------------------------------------*;


LIBNAME JUN09 'H:\UNTHSC\DrPH DISSERTATION\JUN09\Dissertation';

/*ORIGINAL DATA FROM MING*/
Data Chic;
Set 'F:\JUN09\Dissertation\updated3.sas7bdat';
run;

Data Chic1;
Format icd1 $char3.;
Set JUN09.updated1;
ICD2=left(ICD_UP);

icd2=left(ICD_2003);

icd1=substr(icd2,1,3);

If icd1='A00' or icd1='A01' or icd1='A02' or icd1='A03' or icd1='A04' or
icd1='A05'
or icd1='A06' or icd1='A07' or icd1='A08' or icd1='A09' then ICD_2003='001';

else If (icd1='A15' or icd1='A16' or icd1='A16' or icd1='A17' or icd1='A18' or
icd1='A19') then ICD_2003='010';

else If (icd1='A20' or icd1='A21' or icd1='A22' or icd1='A23' or icd1='A24' or
icd1='A25' or icd1='A26'
or icd1='A27' or icd1='A28') then ICD_2003='020';


else If icd1='A30' or icd1='A31' or icd1='A32' or icd1='A33' or icd1='A34' or
icd1='A35' or icd1='A36' or icd1='A37' or icd1='A38' or icd1='A39'
```

```
or icd1='A40' or icd1='A41' or icd1='A42' or icd1='A43' or icd1='A44' or
icd1='A45' or icd1='A46' or icd1='A47' or icd1='A48' or icd1='A49' then
ICD_2003='030';

else If icd1='A81' then ICD_2003='046';

else If icd1='A90' or icd1='A91' or icd1='A92' or icd1='A93' or icd1='A94' or
icd1='A95' or icd1='A96' or icd1='A97' or icd1='A98' or icd1='A99' then
ICD_2003='060';

else If icd1='B35' or icd1='B36' or icd1='B37' or icd1='B38' or icd1='B39' or
icd1='B40' or icd1='B41' or icd1='B42' or icd1='B43' or icd1='B44' or
icd1='B45' or icd1='B46' or icd1='B47' or icd1='B48' or icd1='B49' then
ICD_2003='110';

else If icd1='B65' or icd1='B66' or icd1='B67' or icd1='B68' or icd1='B69' or
icd1='B70' or icd1='B71' or icd1='B72' or icd1='B73' or icd1='B74' or
icd1='B75' or icd1='B76' or icd1='B77' or icd1='B78' or icd1='B79' or
icd1='B80' or icd1='B81' or icd1='B82' or icd1='B83' then ICD_2003='120';

else If icd1='B50' or icd1='B51' or icd1='B52' or icd1='B53' or icd1='B54' or
icd1='B55' or icd1='B56' or icd1='B57' or icd1='B58' or icd1='B59' or
icd1='B60' or icd1='B61' or icd1='B62' or icd1='B63' or icd1='B64' then
ICD_2003='006';

else If icd1='C00' then ICD_2003='140';

else If icd1='C01' or icd1='C02' then ICD_2003='141';

else If icd1='C03' then ICD_2003='143';

else If icd1='C04' then ICD_2003='144';

else If icd1='C05' or icd1='C06' then ICD_2003='145';

else If icd1='C07' or icd1='C08' then ICD_2003='142';

else If icd1='C09' or icd1='C10' then ICD_2003='146';

else If icd1='C11' then ICD_2003='147';

else If icd1='C12' or icd1='C13' then ICD_2003='148';

else If icd1='C14' then ICD_2003='149';

else If icd1='C15' then ICD_2003='150';

else If icd1='C16' then ICD_2003='151';

else If icd1='C17' then ICD_2003='152';

else If icd1='C18' then ICD_2003='153';

else If icd1='C19' or icd1='C20' or icd1='C21' then ICD_2003='154';

else If icd1='C22' then ICD_2003='155';

else If icd1='C23' or icd1='C24' then ICD_2003='156';

else If icd1='C25' then ICD_2003='157';

else If icd1='C30' or icd1='C31' then ICD_2003='160';
```

```
else If icd1='C32' then ICD_2003='161';

else If icd1='C33' or icd1='C34' then ICD_2003='162';

else If icd1='C35' or icd1='C36' or icd1='C37' or icd1='C45' then
ICD_2003='163';

else If icd1='C40' or icd1='C41' then ICD_2003='170';

else If icd1='C43' then ICD_2003='172';

else If icd1='C44' then ICD_2003='173';

else If icd1='C48' then ICD_2003='158';

else If icd1='C46' or icd1='C49' then ICD_2003='171';

else If icd1='C50' then ICD_2003='174';

else If icd1='C53' then ICD_2003='180';

else If icd1='C54' or icd1='C55' or icd1='C58' then ICD_2003='179';

else If icd1='C56' then ICD_2003='183';

else If icd1='C60' then ICD_2003='187';

else If icd1='C61' then ICD_2003='185';

else If icd1='C62' then ICD_2003='186';

else If icd1='C64' or icd1='C65' or icd1='C66' then ICD_2003='189';

else If icd1='C67' then ICD_2003='188';

else If icd1='C69' then ICD_2003='190';

else If icd1='C47' or icd1='C70' or icd1='C72' then ICD_2003='191';

else If icd1='C73' then ICD_2003='193';

else If icd1='C74' or icd1='C75' then ICD_2003='194';

else If icd1='C76' or icd1='C80' then ICD_2003='195';

else If icd1='C81' then ICD_2003='201';

else If icd1='C82' or icd1='C83' or icd1='C84' or icd1='C85' then
ICD_2003='200';

else If icd1='C88' or icd1='C90' then ICD_2003='203';

else If icd1='C91' then ICD_2003='204';

else If icd1='C92' then ICD_2003='205';

else If icd1='C93' then ICD_2003='206';

else If icd1='C94' or icd1='C95' then ICD_2003='204';

else If icd1='D18' then ICD_2003='228';
```

```
else If icd1='D10' or icd1='D11' or icd1='D12' or icd1='D13' then
ICD_2003='210';

else If icd1='D14' or icd1='D15' then ICD_2003='212';

else If icd1='D17'  then ICD_2003='214';

else If icd1='D20'  then ICD_2003='2118';

else If icd1='D21'  then ICD_2003='215';

else If icd1='D24' then ICD_2003='217';

else If icd1='D25' then ICD_2003='218';

else If icd1='D27' then ICD_2003='220';

else If icd1='D32' or icd1='D32' or icd1='D33' then ICD_2003='226';

else If icd1='D34' or icd1='D35' then ICD_2003='226';

else If icd1='D45' then ICD_2003='2384';

else If icd1='Q06' then ICD_2003='7425';
else If icd1='Q82' then ICD_2003='7573';

else If icd1='D51' or icd1='D52' then ICD_2003='2810';

else If icd1='D59' then ICD_2003='283';

else If icd1='D60' or icd1='D61' then ICD_2003='284';

else If icd1='D69' then ICD_2003='287';

else If icd1='D70' or icd1='D71' or icd1='D72' then ICD_2003='288';

else If icd1='E00' or icd1='E01' or icd1='E02' or icd1='E03' or icd1='E04' or
icd1='E05' or icd1='E06' or icd1='E07' then ICD_2003='240';

else If icd1='E10' or icd1='E11' or icd1='E12' or icd1='E13' or icd1='E14' then
ICD_2003='250';

else If icd1='E16' then ICD_2003='251';

else If icd1='E21' then ICD_2003='252';

else If icd1='E22' or icd1='E23' or icd1='E24' then ICD_2003='253';

else If icd1='E32' then ICD_2003='254';

else If icd1='E25' or icd1='E26' or icd1='E27' then ICD_2003='255';

else If icd1='E28' then ICD_2003='256';

else If icd1='E29' then ICD_2003='257';

else If icd1='E34' or icd1='E35' then ICD_2003='258';

else If icd1='F00' or icd1='F01' or icd1='F02' or icd1='F03' then
ICD_2003='290';
```

```
else If icd1='F10' then ICD_2003='291';

else If icd1='F20' then ICD_2003='295';

else If icd1='F30' or icd1='F31' then ICD_2003='296';

else If icd1='F22' then ICD_2003='297';

else If icd1='F40' or icd1='F41' or icd1='F42' then ICD_2003='300';

else If icd1='F60' then ICD_2003='301';

else If icd1='K58' or icd1='K59' then ICD_2003='564';

else If icd1='G00' or icd1='G01' or icd1='G02' or icd1='G03' then
ICD_2003='320';

else If icd1='G04' or icd1='G05' then ICD_2003='323';

else If icd1='G06' or icd1='G07' then ICD_2003='324';

else If icd1='G08' then ICD_2003='325';

else If icd1='G09' then ICD_2003='326';

else If icd1='G20' or icd1='G21' or icd1='G22' then ICD_2003='332';

else If icd1='G35' then ICD_2003='340';

else If icd1='G81' or icd1='G82' then ICD_2003='342';

else If icd1='G40' then ICD_2003='345';

else If icd1='G12' then ICD_2003='335';

else If icd1='G11' or icd1='G95' then ICD_2003='334';

else If icd1='G90' then ICD_2003='337';

else If icd1='G71' or icd1='G72' then ICD_2003='359';

else If icd1='M30' or icd1='M31' or icd1='M32' or icd1='M33' or icd1='M34' or
icd1='M35' then ICD_2003='446';

else If icd1='H00' or icd1='H01' or icd1='H02' or icd1='H03' or icd1='H04' or
icd1='H05' or icd1='H06' or icd1='H07' or icd1='H08' or icd1='H09' or
icd1='H10' or icd1='H11' or icd1='H12' or icd1='H13' or icd1='H14' or
icd1='H15' or icd1='H16' or icd1='H17' or
icd1='H18' or icd1='H19' or icd1='H20' or icd1='H21' or icd1='H22' or
icd1='H23' or icd1='H24' or icd1='H25' or icd1='H26' or icd1='H27' or
icd1='H28' or icd1='H29' or icd1='H30' or icd1='H31' or icd1='H32' or
icd1='H33' or icd1='H34' or icd1='H35' or icd1='H36'
or icd1='H37' or icd1='H38' or icd1='H39' or icd1='H40' or icd1='H41' or
icd1='H42'
or icd1='H43' or icd1='H44' or icd1='H45' or icd1='H46' or icd1='H47' or
icd1='H48' or icd1='H49' or icd1='H50' or icd1='H51' or icd1='H52' or
icd1='H53' or icd1='H54' or icd1='H55' or icd1='H56' or icd1='H57' or
icd1='H58' or icd1='H59' then ICD_2003='360';
else If icd1='H60' or icd1='H61' or icd1='H62' or icd1='H63' or icd1='H64' or
icd1='H65' or icd1='H66' or icd1='H67' or icd1='H68' or icd1='H69' or
icd1='H70' or icd1='H71' or icd1='H72' or icd1='H73' or icd1='H74' or
icd1='H75' or icd1='H76' or icd1='H77' or icd1='H78' or icd1='H79' or
```

```
icd1='H80' or icd1='H81' or icd1='H82' or icd1='H83' or icd1='H84' or
icd1='H85' or icd1='H86' or icd1='H87' or icd1='H88' or icd1='H89' or
icd1='H890' or icd1='H91' or icd1='H92' or icd1='H93' or icd1='H94' or
icd1='H95' then ICD_2003='380';

else If icd1='I00' or icd1='I01' or icd1='I02' then ICD_2003='390';

else If icd1='I05' or icd1='I06' or icd1='I07' or icd1='I08' or icd1='I09' then
ICD_2003='393';

else If icd1='I10' or icd1='I11' or icd1='I12' or icd1='I13' or icd1='I14' or
icd1='I15' then ICD_2003='401';

else If icd1='I20' or icd1='I21' or icd1='I22' or icd1='I23' or icd1='I24' or
icd1='I25' then ICD_2003='410';

else If icd1='I30' then ICD_2003='420';

else If icd1='I33' then ICD_2003='421';

else If icd1='I40' then ICD_2003='422';

else If icd1='I44' or icd1='I45' or icd1='I45' or icd1='I46' or icd1='I47' or
icd1='I48' or icd1='I49' then ICD_2003='426';

else If icd1='I60' then ICD_2003='430';

else If icd1='I61' or icd1='I62' then ICD_2003='431';

else If icd1='I63' or icd1='I64' or icd1='I65' then ICD_2003='433';

else If icd1='I66' then ICD_2003='435';

else If icd1='I26' then ICD_2003='415';

else If icd1='I71' then ICD_2003='441';

else If icd1='J12' or icd1='J13' or icd1='J14' or icd1='J15' or icd1='J16' or
icd1='J17' or
icd1='J18' then ICD_2003='480';

else If icd1='J41' or icd1='J42' then ICD_2003='491';

else If icd1='J45' then ICD_2003='493';

else If icd1='J60' or icd1='J61' or icd1='J62' or icd1='J63' or icd1='J64' or
icd1='J65' or
icd1='J66' then ICD_2003='500';

else If icd1='J68' then ICD_2003='506';

else If icd1='J85' then ICD_2003='513';

else If icd1='J47' then ICD_2003='494';

else If icd1='K02' or icd1='K03' or icd1='K04' or icd1='K05' or icd1='K06' then
ICD_2003='521';

else If icd1='K11' then ICD_2003='527';

else If icd1='K12' or icd1='K13' or icd1='K14' then ICD_2003='528';
```

```
else If icd1='K20' or icd1='K21' or icd1='K22' or icd1='K23' then
ICD_2003='530';

else If icd1='K25' then ICD_2003='531';

else If icd1='K26' then ICD_2003='532';

else If icd1='K28' then ICD_2003='534';

else If icd1='K30' or icd1='K31' then ICD_2003='536';

else If icd1='K35' or icd1='K36' or icd1='K37' or icd1='K38' then
ICD_2003='540';

else If icd1='K40' or icd1='K41' or icd1='K42' or icd1='K43' or icd1='K44' or
icd1='K45' or icd1='K46' then ICD_2003='550';

else If icd1='K50' or icd1='K51' or icd1='K52' then ICD_2003='555';

else If icd1='K65' then ICD_2003='567';

else If icd1='K72' then ICD_2003='570';

else If icd1='K72' or icd1='K73' or icd1='K74' or icd1='K75' or icd1='K76' or
icd1='K77' then ICD_2003='571';

else If icd1='K80' then ICD_2003='574';

else If icd1='K81' or icd1='K82' or icd1='K83' then ICD_2003='575';

else If icd1='K85' or icd1='K86' then ICD_2003='577';

else If icd1='N00' or icd1='N01' or icd1='N02' or icd1='N04' or icd1='N05' then
ICD_2003='580';

else If icd1='N03' then ICD_2003='582';

else If icd1='N05' then ICD_2003='583';

else If icd1='N10' or icd1='N12' then ICD_2003='590';


else If icd1='N20' then ICD_2003='592';

else If icd1='N28' then ICD_2003='593';

else If icd1='N21' or icd1='N22' then ICD_2003='594';

else If icd1='N30' or icd1='N31' or icd1='N32' or icd1='N33' or icd1='N34' or
icd1='N35' then ICD_2003='595';

else If icd1='N40' or icd1='N41' or icd1='N42' then ICD_2003='600';

else If icd1='N60' or icd1='N61' or icd1='N62' or icd1='N63' or icd1='N64' then
ICD_2003='610';

else If icd1='N70' or icd1='N73' or icd1='N83' then ICD_2003='614';

else If icd1='N71' or icd1='N72' or icd1='N75' or icd1='N76' or icd1='N80' then
ICD_2003='615';

else If icd1='N80' then ICD_2003='617';
```

```
else If icd1='N81' then ICD_2003='618';

else If icd1='O01' then ICD_2003='630';

else If icd1='O20' or icd1='O42' or icd1='O43' or icd1='O44' or icd1='O45' or
icd1='O46' then ICD_2003='640';

else If icd1='O10' or icd1='O11' or icd1='O12' or icd1='O13' or icd1='O14' or
icd1='O15' or icd1='O16' then ICD_2003='642';

else If icd1='O40' then ICD_2003='657';

else If icd1='M05' or icd1='M06' or icd1='M08' or icd1='M09' then
ICD_2003='714';

else  If icd1='M86' then ICD_2003='730';

else If icd1='Q00' then ICD_2003='740';
else If icd1='V01' or icd1='V02' or icd1='V03' or icd1='V04' or icd1='V05' or
icd1='V05' or icd1='V06' or icd1='V07' or icd1='V08' or icd1='V09' or
icd1='V10' or icd1='V11' or icd1='V12' or icd1='V13' or icd1='V14' or
icd1='V15' or icd1='V16' or
icd1='V17' or icd1='V18' or icd1='V19' or icd1='V20' or icd1='V21' or
icd1='V22' or icd1='V23' or icd1='V24' or icd1='V25' or icd1='V26' or
icd1='V27' or icd1='V28' or icd1='V29' or icd1='V30' or icd1='V31' or
icd1='V32' or icd1='V33' or icd1='V34'
or icd1='V35' or icd1='V36' or icd1='V37' or icd1='V38' or icd1='V39' or
icd1='V40' or icd1='V41' or icd1='V42' or icd1='V43' or icd1='V44' or
icd1='V45' or icd1='V46' or icd1='V47' or icd1='V48' or icd1='V49' or
icd1='V50' or icd1='V51' or icd1='V52'
or icd1='V53' or icd1='V54' or icd1='V55' or icd1='V56' or icd1='V57' or
icd1='V58' or icd1='V59' or icd1='V60' or icd1='V61' or icd1='V62' or
icd1='V63' or icd1='V64' or icd1='V65' or icd1='V66' or icd1='V67' or
icd1='V68' or icd1='V69' or icd1='V70'
or icd1='V71' or icd1='V72' or icd1='V73' or icd1='V74' or icd1='V75' or
icd1='V76' or icd1='V77' or icd1='V78' or icd1='V79' or icd1='V80' or
icd1='V81' or icd1='V82' or icd1='V83' or icd1='V84' or icd1='V85' or
icd1='V86' or icd1='V87' or icd1='V88' or icd1='V89' then ICD_2003='E800';

else If icd1='V90' or icd1='V91' or icd1='V92' or icd1='V93' or icd1='V94' then
ICD_2003='E830';

else If icd1='V95' or icd1='V96' or icd1='V97' then ICD_2003='E840';

else If icd1='X40' or icd1='X41' or icd1='X42' or icd1='X43' or icd1='X44' or
icd1='X45' or icd1='X46' or icd1='X47' or icd1='X48' or icd1='X49' then
ICD_2003='E850';

else If icd1='W00' or icd1='W01' or icd1='W02' or icd1='W03' or icd1='W04' or
icd1='W05' or icd1='W06' or icd1='W07' or icd1='W08' or icd1='W09' or
icd1='W10' or icd1='W11' or icd1='W12' or icd1='W13' or icd1='W14' or
icd1='W15' or icd1='W16' or icd1='W17' or icd1='W18' or icd1='W19' then
ICD_2003='E880';

else If icd1='X00' or icd1='X01' or icd1='X02' or icd1='X03' or icd1='X04' or
icd1='X05' or icd1='X06' or icd1='X07' or icd1='X08' or icd1='X09' or
icd1='X10' or icd1='X11' or icd1='X12' or icd1='X13' or icd1='X14' or
icd1='X15' or icd1='X16' or icd1='X17' or icd1='X18' or icd1='X19' or
icd1='X20' or icd1='X21' or icd1='X22' or icd1='X23'
or icd1='X24' or icd1='X25' or icd1='X26' or icd1='X27' or icd1='X28' or
icd1='X29' or icd1='X30' or icd1='X31' or icd1='X32' or icd1='X33' or
```

```
icd1='X34' or icd1='X35' or icd1='X36' or icd1='X37' or icd1='X38' or
icd1='X39' or icd1='X50' or icd1='X51' or icd1='X52' or icd1='X53' or
icd1='X54' or icd1='X55' or icd1='X56' or icd1='X57'
or icd1='X58' or icd1='X59' or icd1='W20' or icd1='W21' or icd1='W22' or
icd1='W23' or icd1='W24' or icd1='W25' or icd1='W26' or icd1='W27' or
icd1='W28' or icd1='W29' or icd1='W30' or icd1='W31' or icd1='W32' or
icd1='W33' or icd1='W34' or icd1='W35' or icd1='W36' or icd1='W37' or
icd1='W38' or icd1='W39' or icd1='W40' or icd1='W41'
or icd1='W42' or icd1='W43' or icd1='W44' or icd1='W45' or icd1='W46' or
icd1='W47' or icd1='W48' or icd1='W49' or icd1='W50' or icd1='W51' or
icd1='W52' or icd1='W53' or icd1='W54' or icd1='W55' or icd1='W56' or
icd1='W57' or icd1='W58' or icd1='W59' or icd1='W60' or icd1='W61' or
icd1='W62' or icd1='W63' or icd1='W64' or icd1='W65'
or icd1='W66' or icd1='W67' or icd1='W68' or icd1='W69' or icd1='W70' or
icd1='W71' or icd1='W72' or icd1='W73' or icd1='W74' or icd1='W75' or
icd1='W76' or icd1='W77' or icd1='W78' or icd1='W79' or icd1='W80' or
icd1='W81' or icd1='W82' or icd1='W83' or icd1='W84' or icd1='W85' or
icd1='W86' or icd1='W87' or icd1='W88' or icd1='W89'
or icd1='W90' or icd1='W91' or icd1='W92' or icd1='W93' or icd1='W94' or
icd1='W95' or icd1='W96' or icd1='W97' or icd1='W98' or icd1='W99' then
ICD_2003='E900';

else If icd1='W28' or icd1='W29' or icd1='W30' or icd1='W31' then
ICD_2003='E919';

else If icd1='W25' or icd1='W26' or icd1='W27' then ICD_2003='E920';

else If icd1='X60' or icd1='X61' or icd1='X62' or icd1='X63' or icd1='X64' or
icd1='X65' or icd1='X66' or icd1='X67' or icd1='X68' or icd1='X69' or
icd1='X70' or icd1='X71' or icd1='X72' or icd1='X73' or icd1='X74' or
icd1='X75' or icd1='X76' or icd1='X77' or icd1='X78' or icd1='X79' or
icd1='X80' or icd1='X81' or icd1='X82' or icd1='X83' or icd1='X84' then
ICD_2003='E950';

else If icd1='S02' or icd1='S12' or icd1='S22' or icd1='S32' or icd1='S42' or
icd1='S52' or icd1='S62' or icd1='S72' or icd1='S82' or icd1='S92' or
icd1='T02' or icd1='T08' or icd1='T10' or icd1='T12' or icd1='T14.2' then
ICD_2003='800';

else If icd1='S06' then ICD_2003='850';

else If icd1='S01' or icd1='S11' or icd1='S21' or icd1='S31' or icd1='S41' or
icd1='S51' or icd1='S61' or icd1='S71' or icd1='S81' or icd1='S91' or
icd1='ST01' then ICD_2003='870';

else If  icd1='S15' or icd1='S25' or icd1='S35' or icd1='S45' or icd1='S55' or
icd1='S65' or icd1='S75' or icd1='S85' or icd1='S95'  then ICD_2003='900';

else If icd1='S00' or icd1='S10' or icd1='S20' or icd1='S30' or icd1='S40' or
icd1='S50' or icd1='S60' or icd1='S70' or icd1='S80' or icd1='S90' or
icd1='T00' or icd1='T09.0' or icd1='T14' then ICD_2003='910';

else If icd1='S07' or icd1='S17' or icd1='S28' or icd1='S38' or icd1='S47' or
icd1='S57' or icd1='S67' or icd1='S77' or icd1='S87' or icd1='S97' or
icd1='T04'  then ICD_2003='925';

else If icd1='T15' or icd1='T16' or icd1='T17' or icd1='T18' or icd1='T19' then
ICD_2003='930';

else If icd1='T20' or icd1='T21' or icd1='T22' or icd1='T23' or icd1='T24' or
icd1='T25' or icd1='T26' or icd1='T27' or icd1='T28' or icd1='T29' or
icd1='T30' or icd1='T31' or icd1='T32' then ICD_2003='940';
```

```
else If icd1='S04' or icd1='S14' or icd1='S24' or icd1='S34' or icd1='S44' or
icd1='S54' or icd1='S64' or icd1='S74' or icd1='S84' or icd1='S94' then
ICD_2003='950';

/*Section II: Convet those ICDs which have four digits*/

If icd2='S001' or icd2='S051' or icd2='S062' or icd2='S100' or icd2='S200' or
icd2='S300' or icd2='S302'
or icd2='S700' or icd2='S701' or icd2='S800' or icd2='S801' then
ICD_2003='920';

else if icd2='S090' or icd2='T145' then ICD_2003='900';

else if icd2='T144' then ICD_2003='950';

else If icd2='C945' or icd2='D471' then ICD_2003='2898';

else if  icd2='T147' then ICD_2003='925';

else If icd2='G700' then ICD_2003='3580';

else If icd2='D640' or icd2='D641' or icd2='D642' or icd2='D643' then
ICD_2003='2850';

else If icd2='F102' then ICD_2003='303';

else If icd2='F112' or icd2='F122' or icd2='F132' or icd2='F142' or
icd2='F152' or icd2='F162' or icd2='F172' or icd2='F182' or icd2='F192'  then
ICD_2003='304';

else If icd2='N130' or icd2='N131' or icd2='N132' or icd2='N133' then
ICD_2003='591';

RUN;

data Chic2;
SET Chic1(drop=icd4);
icd4=substr(ICD_2003,1,1);
IF icd4='A' or icd4='B' or icd4='C' or icd4='D' or icd4='E' or icd4='F' or
icd4='G' or icd4='H'
or icd4='I' or icd4='J' or icd4='K' or icd4='L' or icd4='N' or icd4='M' or
icd4='O' or icd4='S' or icd4='Q' or icd4='P' or icd4='R'
or icd4='S' or icd4='T' or icd4='U' or icd4='V' or icd4='W' or icd4='X' or
icd4='Y' or icd4='Z' Then ICD_UP='9999';
ELSE ICD_UP=ICD_2003;
run;

/*New updates*/
Data chic3;
Set chic2;
 IF ICD_UP IN (      'B171' 'B182'  'B207' 'B220' 'B238' 'B24'
                     'C570' 'C787' 'C793' 'C97' 'D430' 'D431'
                     'D469' 'E46' 'E668' 'E780' 'E86' 'F171' 'G309'
                     'G319' 'G931' 'I279' 'I350' 'I38'
                     'I420' 'I422' 'I429' 'I500' 'I514' 'I516' 'I517'
                     'I519' 'I671' 'I674' 'I694' 'I698' 'I709' 'I729'
                     'I739' 'I802' 'J209' 'J439' 'J448' 'J449' 'J690'
                     'J840' 'J841' 'J849' 'J960' 'J969' 'J984'
                     'K550' 'K559' 'K562' 'K573' 'K631' 'K701' 'K703'
                     'K922' 'L930' 'M419' 'N151' 'N170' 'N179' 'N188'
                     'N189' 'N19' 'N390' 'Q231' 'Q403' 'R568' 'R579'
```

```
                    'R99' 'X93' 'X95' 'Y08' 'Y09' 'Y12' 'Y14' 'Y17' 'Y86')
Then ICD_UP='9999';
  ELSE IF ICD_UP ='B203' THEN ICD_UP='203'; /*Treated as Multiple Myeloma*/
  ELSE IF ICD_UP ='C719' THEN ICD_UP='191'; /*Cnfirmed with Dr. Johnson*/
  ELSE IF ICD_UP ='K296' THEN ICD_UP='535';
RUN;

/* These are 9th revision ICDs for which the 'E' needs to be dropped:

                    E800  E805 E810 E812 E814 E815 E816 E818 E819
                                E822 E826  E830 E831 E832 E835 E844 E850 E855
                    E859  E860 E869 E880 E882 E887 E890 E900 E910
                                E911 E913 E916 E917 E922 E923 E925 E927 E928
E931
                                E932 E950 E952 E953 E955 E956 E958 E963 E965
E966
                                E968 E980 E984 E985 E988

*/

Data Chic4;
Format ICDA $char4.;
Set Chic3;
ICD_A=left(ICD_UP);
IF ICD1='.' THEN ICD2='';
ELSE IF SUBSTR(ICD_A,1,1) IN ('E') THEN ICDA=SUBSTR(ICD_A,2,3);
ELSE IF 800<(SUBSTR(ICD_A,1,3))*1<999 THEN ICDA='9999';
ELSE ICDA=ICD_A;
IF DOD_UP>='01JAN2004'D THEN ICDA='';
Drop ICD_A;
RUN;

/*OCMAP IMPOSSIBLE ICDs************************/
DATA CHICAGO;
SET  Chic4
IF (SUBSTR(ICD_2003,1,3)*1) IN
(0,19,28,29,58,59,67,68,69,89,105,106,107,108,109,113,119,166,167,168,169,
176,177,178,209,247,249,327,329,338,339,399,400,406,407,408,409,418,419,
439,445,449,450,467,468,469,479,488,489,497,498,499,509,538,539,544,545,
546,547,548,549,554,559,561,563,609,612,613,649,677,678,679,687,688,689,
699,760,761,762,763,764,765,766,767,768,769,770,771,772,773,774,775,776,
777,778,779,808,809,839,859,877,889,979) THEN ICD_2003='9999';
RUN;
```

APPENDIX B

SAS Source code for computing directly standardized rates ratio, SMR, relative SMR, indirectly standardized rates ratio from SMR and rate ratio as SMR by using one of the comparison group (the non-poultry) as the standard or referent.

```
*-------------------------------------------------------------------*
| STUDY      LUNG CANCER/MULTIPLE MYELOMA RISK IN CHICAGO POULTRY      |
|            WORKERS EXPOSED TO ONCOGENIC VIRUS                        |
|                                                                      |
| JOB        COMPARING STATISTICAL METHODS IN DOCUMENTING C. MORTALITY |
|                                                                      |
| NAME       Dr.PH DESSERTATION                                        |
|                                                                      |
| FUNCTION   DIRECT/INDIRECT STANDARDIZATION (FULL AND SUBCOHORTS)     |
|                                                                      |
| FILES      Input: JUN09.CHI_ANALYSIS AND JUN09.CHIC_SUBCOHORT        |
|                   USED IN DIFFERENT TIMES                            |
|            Output: CHICAGO.DAT                                       |
|                                                                      |
| LANGUAGE   SAS 9.1.2                                                 |
|                                                                      |
| AUTHOR     HARRISON NDETAN                                           |
|                                                                      |
| DATE       SEPTEMBER 2, 2009                                         |
|            Modified Nov 13, 2009                                     |
|                                                                      |
*-------------------------------------------------------------------*;

              /*MULTIPLE MYELOMA*/

%macro lsy;
%global RRc  RR  RSMR RRStd  RR2;

DATA Chi_analysis3;
SET JUN09.CHI_ANALYSIS; /*FULL COHORT*/
*SET JUN09.CHIC_SUBCOHORT; /*SUBCOHORT*/
If Race = 1 AND sex= 1 then RSex=1; /*White Male*/
Else If Race = 0 AND sex= 1 then RSex=2; /*NonWhite Male*/
Else If Race = 1 AND sex=0 then RSex=3; /*White Female*/
Else If Race = 0 AND sex=0 then Rsex=4; /*NonWhite Female*/
run;

proc surveyselect data=Chi_analysis3  method = urs sampsize = 30488
   rep=1 out=Chi_analysis2;
   *id id read write math science socst;
run;

*******************************************************
* X Count: find x count in each category and total sum *
*******************************************************;
%macro sqlx_all;
*%do g=0 %to 1;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global Xall_Cs&h.Rs&i.Ag&k;
%GLOBAL Xall&h&i&k;
title Xall_Cs&h.Rs&i.Ag&k;
proc sql;
```

```
select count(CASEID) into : Xall_Cs&h.Rs&i.Ag&k
from chi_analysis2
where MYELOMA=&h and RSex=&i and Agrp4=&k;
quit;
%let Xall&h&i&k=&&&Xall_Cs&h.Rs&i.Ag&k;
%put MYELOMA  specific death COUNT FOR Xall&h&i&k IS &&&Xall&h&i&k;
%end;%end;%end;
%mend;

%macro sqlx(plt);
%let g=&plt;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global X&plt._Jc&g.Cs&h.Rs&i.Ag&k;
%GLOBAL X&plt&g&h&i&k;
title X&plt._Jc&g.Cs&h.Rs&i.Ag&k;
proc sql;
select count(CASEID) into : X&plt._Jc&g.Cs&h.Rs&i.Ag&k
from chi_analysis2
where plant=&g and MYELOMA=&h and RSex=&i and Agrp4=&k;
quit;
%let X&plt&g&h&i&k=&&&X&plt._Jc&g.Cs&h.Rs&i.Ag&k;
%put  MYELOMA specific death COUNT FOR X&plt._&g&h&i&k IS &&&&X&plt&g&h&i&k;
%end;%end;
%mend;

%macro xall_sum;
%global xall_Sum;
%Let xall_sum=0;
*%do g=0 %to 1;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%let xall_sum = %SYSEVALF(&xall_sum + &&&Xall&h&i&k);
%put Xall value:&&&&Xall&h&i&k is (Sum of xall is &xall_sum);
%end;%end;%end;
%mend;

%macro x_sum(plt);
%global x&plt._Sum;
%Let x&plt._sum=0;
%let g=&plt;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%let x&plt._sum = %SYSEVALF(&&&x&plt._sum + &&&&X&plt&g&h&i&k);
%put X&plt._&g.&h.&i.&k value:&&&&X&plt&g&h&i&k is (Sum of x&plt.all is
&&&x&plt._sum);
%end;%end;
%mend;

*****************************************************************
* Person-Year: find person-year in each category and total sum *
*****************************************************************;
%macro sqlpy_all;
*%do g=0 %to 1;*jobcode(plant);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global PYall_Rs&i.Ag&k;
%GLOBAL PYall&i&k;
title PYall_Rs&i.Ag&k;
```

```
proc sql;
select sum(time_fu) into : PYall_Rs&i.Ag&k
from chi_analysis2
where RSex=&i and Agrp4=&k;
quit;
%let PYall&i&k=&&PYall_Rs&i.Ag&k;
%put Person-time sum FOR PYall_&i.&k IS &&&PYall&i&k;
%end;%end;%end;
%mend;

%macro sqlpy(plt);
%let g=&plt;*jobcode(plant);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global PY&plt._Jc&g.Rs&i.Ag&k;
%GLOBAL PY&plt&g&i&k;
title PY&plt._Jc&g.Rs&i.Ag&k;
proc sql;
select sum(time_fu) into : PY&plt._Jc&g.Rs&i.Ag&k
from chi_analysis2
where plant=&g and RSex=&i and Agrp4=&k;
quit;
%let PY&plt&g&i&k=&&&PY&plt._Jc&g.Rs&i.Ag&k;
%put Plant &plt person-time sum FOR PY&plt._&g.&i.&k IS &&&PY&plt&g&i&k;
%end;%end;
%mend;

%macro pyall_sum;
%global pyall_Sum;
%Let pyall_sum=0;
*%do g=0 %to 1;*jobcode(plant);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%let pyall_sum = %SYSEVALF(&pyall_sum + &&PYall&i&k);
%put pyall_&i.&k value:&&PYall&i&k is (Sum of pyall is &pyall_sum);
%end;%end;%end;
%mend;

%macro py_sum(plt);
%global py&plt._sum;
%Let py&plt._sum=0;
%let g=&plt;*jobcode(plant);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%let py&plt._sum = %SYSEVALF(&&&py&plt._sum + &&&PY&plt&g&i&k);
%put PY&plt._&g.&i.&k value:&&&PY&plt&g&i&k is (Sum of py&plt.all is
&&&py&plt._sum);
%end;%end;
%mend;

****************************************
* Stratum specific Crude rate : X / PY *
****************************************;
%macro Crude_all;
*%do g=0 %to 1;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global rall&h.&i.&k;
%LET rall&h.&i.&k = %SYSEVALF((&&&&Xall&h&i&k)/(&&PYall&i&k));
%put Crude_all for rall_Cs&h.Rs&i.Ag&k is &&rall&h.&i.&k
(Xall_Cs&h.Rs&i.Ag&k=&&&Xall&h&i&k, PYall_Cs&h.Rs&i.Ag&k=&&PYall&i&k);
```

```
%end;%end;%end;
%mend;

%macro Crude(plt);
%let g=&plt;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global r&plt&g.&h.&i.&k;
%LET r&plt&g.&h.&i.&k = %SYSEVALF((&&&&X&plt&g&h&i&k)/(&&&PY&plt&g&i&k));
%put Crude&plt for r&plt&g.&h.&i.&k is &&&r&plt&g.&h.&i.&k
(X&plt.Jc&g.Cs&h.Rs&i.Ag&k=&&&&X&plt&g&h&i&k,
PY&plt.Jc&g.Cs&h.Rs&i.Ag&k=&&&PY&plt&g&i&k);
%end;%end;
%mend;

*****************************************************
* Total Crude Rate: Sum of X / Sum of PY,  per plant*
*****************************************************;
%macro rc(plt);
%global rc&plt;
%let rc&plt = %SYSEVALF((&&&x&plt._sum)/(&&&py&plt._sum));
%put rc&plt = &&&rc&plt (x&plt._sum = &&&x&plt._sum, py&plt._sum =
&&&py&plt._sum);
%mend;
*****************************************************
* Expected: r_plt * PY_all for each stratum, then sum  *
*****************************************************;
%macro Expected(plt);
%let g=&plt;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global E&plt&g.&h.&i.&k;
%LET E&plt&g.&h.&i.&k = %SYSEVALF((&&&&r&plt&g.&h.&i.&k)*(&&&PYall&i&k));
%put PYall&i&k:&&&PYall&i&k, Crude rate:&&&&r&plt&g.&h.&i.&k ,
Expected:&&&E&plt&g.&h.&i.&k;
%end;%end;
%mend;

%macro Expectedsum(plt);
%global Expected&plt._sum;
%Let Expected&plt._sum=0;
%let g=&plt;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%let Expected&plt._sum = %SYSEVALF(&&&Expected&plt._sum + &&&E&plt&g.&h.&i.&k);
%put Expected value for E&plt.Jc&g.Cs&h.Rs&i.Ag&k is &&&E&plt&g.&h.&i.&k,
and sum of expected is &&&Expected&plt._sum;
%end;%end;
%mend;

*************************************************************************
* DIRECTLY STANDARDIZED RATES: Sum of Expected per plant / Sum of PY_ALL *
*************************************************************************;
%macro R(plt);
%global r&plt;
%let R&plt = %SYSEVALF((&&&Expected&plt._sum)/(&pyall_sum));
%put R&plt = &&&r&plt (Expected&plt._sum = &&&Expected&plt._sum, pyall_sum =
&pyall_sum);
%mend;
```

```
options nosymbolgen;
ods listing close;
%sqlx_all/*X count - both plants*/
%sqlx(1)/*x count - plant=1*/
%sqlx(0)/*x count - plant=0*/
%xall_sum/*X sum - both plants*/
%x_sum(1)/*x sum - plant=1*/
%x_sum(0)/*x sum - plant=0*/
%sqlpy_all/*Person Years - both plants*/
%sqlpy(1)/*Person Years - plant=1*/
%sqlpy(0)/*Person Years - plant=0*/
%pyall_sum/*PY sum - both plants*/
%py_sum(1)/*py sum - plant=1*/
%py_sum(0)/*py sum - plant=0*/
%Crude_all/*Crude - both plants*/
%Crude(1)/*Crude - plant=1*/
%Crude(0)/*Crude - plant=0*/
%Expected(1)/*Expected - plant=1*/
%Expected(0)/*Expected - plant=0*/
%Expectedsum(1)/*Expected sum - plant=1*/
%Expectedsum(0)/*Expected sum - plant=0*/
%rc(all)/*R crude - both plants*/
%rc(1)/*R crude - plant=1*/
%rc(0)/*R crude - plant=0*/
%r(1)/*R - plant=1*/
%r(0)/*R - plant=0*/
**********************************************
*Calculating Rates Ratio*
**********************************************;
/*CRUDE rr*/

%let RRc=%SYSEVALF((&rc1)/(&rc0));
%put RRc = &RRc;

/*Standardized RR*/

%let RR=%SYSEVALF((&r1)/(&r0));
%put RR = &RR;
/*END OF DIRECT STANDARDIZATION*/

****************************
* Indirect Standardization *
****************************;

*************************************************************************
* EXpected counts: (crude rate for joint pop * PY per plant) done per cell*
*************************************************************************;

/*Expected per cell per plant*/

%macro Exp_ind(plt);
%let g=&plt;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global Eind&plt&g.&h.&i.&k;
%LET Eind&plt&g.&h.&i.&k = %SYSEVALF((&&rall&h.&i.&k)*(&&&PY&plt&g&i&k));
%put rall_&h.&i.&k:&&rall&h.&i.&k, PY&plt._&g&i&k:&&&PY&plt&g&i&k,
Exp_ind:&&&Eind&plt&g.&h.&i.&k;
%end;%end;
%mend;
```

```
/*Sum of expected*/

%macro Exp_ind_sum(plt);
%global Eind&plt._sum;
%Let Eind&plt._sum=0;
%let g=&plt;*jobcode(plant);
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%let Eind&plt._sum = %SYSEVALF(&&&Eind&plt._sum + &&&Eind&plt&g.&h.&i.&k);
%put Expected(indirect) value for Eind&plt.Jc&g.Cs&h.Rs&i.Ag&k is
&&&Eind&plt&g.&h.&i.&k,
and sum of expected is &&&Eind&plt._sum;
%end;%end;
%mend;

****************************************
*SMR: SUM of Observed X / Sum of Expected
****************************************;
%macro SMR(plt);
%global smr&plt;
%let SMR&plt = %SYSEVALF((&&&x&plt._sum)/(&&&Eind&plt._sum));
%put SMR&plt = &&&smr&plt (Eind&plt._sum = &&&Eind&plt._sum, x&plt._sum =
&&&x&plt._sum);
%mend;

%Exp_ind(1)
%Exp_ind(0)
%Exp_ind_sum(1)
%Exp_ind_sum(0)
%smr(1)
%smr(0)

/*Calculate RSMR*/
%let RSMR=%SYSEVALF((&smr1)/(&smr0));
%put RSMR = &RSMR;


************************************************************
*ALTERNATIVELY: CALCULATE INDIRECT RATES BY: SMR*CRUDE RATE*
************************************************************;
/*Will change this to get a model that run both Rstd1 and Rstd0 at once*/

%let RStd1=%SYSEVALF((&rc1)*(&smr1));
%put RStd1 = &RStd1;

%let RStd0=%SYSEVALF((&rc0)*(&smr0));
%put RStd0 = &RStd0;

%let RRStd=%SYSEVALF((&RStd1)/(&RStd0));
%put RRStd = &RRStd;

*************************************************************
Poisson model of indirect standardization:
*Use plant0 as standard to standardize plant1 such that SMR=RR*
*************************************************************;

/*Expected deaths in plant1 based on rates in plant0*/

%macro ex;
%let h=1;*MYELOMA(cause);
```

```
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%global ex&h.&i.&k;
%LET ex&h.&i.&k = %SYSEVALF((&&&r00&h.&i.&k)*(&&&PY11&i&k));
%put r00&h.&i.&k:&&&r00&h.&i.&k, PY11&i&k:&&&PY11&i&k,
ex&h.&i.&k:&&&ex&h.&i.&k;
%end;%end;
%mend;

%ex/*Need to run sqlx(0), sqlpy(0), sqlpy(1), and Crude(0)first*/

%macro ex_sum;
%global ex_sum;
%Let ex_sum=0;
%let h=1;*MYELOMA(cause);
%do i=1 %to 4;*race sex;
%do k=1 %to 3;*age grp;
%let ex_sum = %SYSEVALF(&ex_sum + &&&ex&h.&i.&k);
%put Expected value for exCs&h.Rs&i.Ag&k is &&&ex&h.&i.&k,
and sum of expected is &ex_sum;
%end;%end;
%mend;

%ex_sum

/*Calculate RR2, need to run sqlx(0), sqlx(1), %x_sum(1), and Crude(0)first*/
%let RR2=%SYSEVALF((&x1_sum)/(&ex_sum));
%put RR2 = &RR2 (x1_sum:&x1_sum, ex_sum:&ex_sum;

*************************************************************************;

%mend lsy;

data jun09.store;
run;

PROC PRINTTO LOG="C:\BBB.LOG";
RUN;
%macro runitagain(N);
%do i=1 %to &N;
%lsy;
data jun09.aa;
RRc=&RRc;           /*Crude rate*/
RR=&RR;  /*Directly standardized*/
RSMR=&RSMR; /*Relative SMR*/
RRSTD=&RRSTD; /*Indirectly Standardized*/
RR2=&RR2; /*SMR for poultry standardized with rates from non-poultry*/
RUN;
DATA JUN09.STORE;
SET JUN09.STORE JUN09.AA;
RUN;
%END;
%MEND RUNITAGAIN;
%RUNITAGAIN(N=10000);
```

APPENDIX C

ABSTRACT 1

Cancer Mortality in Poultry Slaughtering/Processing Plant Workers Belonging to a Union Pension Fund

Eric S. Johnson, M.B.;B.S., Ph.D.[1], Harrison Ndetan, M.P.H., M.Sc.[2], Ka-Ming Lo, M.P.H.[2]

Background: Humans are commonly exposed to viruses that naturally infect and cause cancer in chickens & turkeys. It is not known if these viruses also cause cancer in humans. To find out, we studied cancer mortality in a cohort of 20,132 workers in poultry slaughtering & processing plants, an occupational group with one of the highest human exposures to these viruses.

Methods: Mortality in poultry workers was compared with that in the US general population through the estimation of proportional mortality ratios (PMR) and standardized mortality ratios (SMR) separately for each race/sex group, and for the whole cohort.

Results: Increased SMRs and PMRs were observed in the cohort as a whole or in subgroups, for several cancer sites, most of which had been previously reported to be in excess in two other poultry cohorts (cancers of the buccal, nasal and pharyngeal cavities, liver, pancreas, trachea/bronchus/lung, myeloma, and lymphoid leukemia). New sites observed to be in excess in this study were cancers of the uterine cervix and penis, and monocytic leukemia.

Conclusion: Exposure to poultry oncogenic viruses is probably responsible for the occurrence of the excess of some of these cancers in these workers. However, studies are needed that will consider the role of other occupational and non-occupational exposures in the occurrence of these cancers, before a definitive conclusion can be reached. The findings may have serious implications for the general population which is also exposed to these viruses.

Key Terms: Killing chickens & turkeys; oncogenic viruses; wrapping fumes; cooking, curing, & smoking meat.

APPENDIX D

ABSTRACT 2

Update of Cancer Mortality in the Missouri Poultry Union

Eric S. Johnson, M.B.;B.S., Ph.D.[1], Harrison Ndetan, M.Sc.,M.P.H. [1], Yi Zhou, M.S.[1,2], C. Lillian Yau, Ph.D.[2], Vishnu Sarda, M.B.;B.S., M.P.H. [1] Satish Bankuru, M.B.;B.S.,[1] Nykiconia Preacely, Dr.P.H.[1] Saritha Bangara, B.Sc., Martha Felini, Ph.D. [1]

Background: Workers in poultry slaughtering and processing plants have one of the highest human exposures to oncogenic viruses that cause cancer in chickens and turkeys, and also have other occupational carcinogenic exposures. The general population is also exposed to these viruses. We studied poultry workers because if these viruses cause cancer in humans it should be readily evident in them.

Methods: We investigated cancer mortality in workers who belong to a poultry union in Missouri, and estimated standardized mortality and proportional mortality ratios.

Results: Increased mortality was observed for cancers of the lung, thymus/heart/mediastinum/pleura, the adrenals & other endocrine organs, cervix, and other specified type/unspecified type of leukemia.

Conclusion: The findings are based on small numbers of deaths, but add to the growing evidence that suggests that subjects exposed to oncogenic viruses and other occupational carcinogens in the poultry industry, are at increased risk of dying from certain cancers.

Key Words: Oncogenic viruses, chickens, workers, occupational, carcinogenic

APPENDIX E

ABSTRACT 3

Mortality from Malignant Diseases - Update of the Baltimore Union Poultry Cohort

Eric S. Johnson, M.B.;B.S., Ph.D.[1], Yi Zhou, M.S.[1,2], C. Lillian Yau, Ph.D.[2],Deepak Prabhakar, M.D., M.P.H.[1], Harrison Ndetan, M.P.H., M.Sc.[1] Karan P. Singh, Ph.D.[1,] Nikiconia Preacely, Dr.P.H.[1]

We previously studied mortality up to 1989 in 2639 members of a local union who had *ever* worked in poultry slaughtering & processing plants, because they were exposed to oncogenic viruses present in poultry. In this report, cancer mortality was updated to the year 2003 for 2580 of the 2639 subjects who worked *exclusively* in poultry plants. Mortality in poultry workers was compared with that in the US general population through the estimation of proportional mortality and standardized mortality ratios separately for each race/sex group, and for the whole cohort. Compared to the US general population, an excess of cancers of the buccal and nasal cavities and pharynx (base of the tongue, palate and other unspecified mouth, tonsil & oropharynx, nasal cavity/middle ear/accessory sinus), esophagus, recto-sigmoid/rectum/anus, liver and intrabiliary system, myelofibrosis, lymphoid leukemia and multiple myeloma, was observed in particular subgroups of, or in the entire poultry cohort. We hypothesize that oncogenic viruses present in poultry, and exposure to fumes, are candidates for an etiologic role to explain the excess occurrence of at least some of these cancers in the poultry workers. Larger

studies which can control for confounding factors are urgently needed to determine the significance of these findings.

Key Terms: Poultry slaughtering/processing, cancer, viruses, fumes

APPENDIX F

ABSTRACT 4

Mortality in the Baltimore Union Poultry Cohort – Non-malignant Diseases

Eric S. Johnson, M.B.;B.S., Ph.D.,[1] Lillian C. Yau, Ph.D.[2] Yi Zhou, M.S.[2], Karan P. Singh, Ph.D., Harrison Ndetan, M.Sc., M.P.H.[1]

Objective: Workers in poultry plants have high exposure to a myriad of transmissible agents present in poultry and their products. Subjects in the general population are also exposed. It is not known whether many of these agents cause disease in humans. If they do, we reason this would be readily evident in a highly exposed group such as poultry workers. We report here on mortality from non-malignant diseases in a cohort of poultry workers.

Methods: Mortality was compared with that of the US general population, and with that of a comparison group from the same union. Risk was estimated by standardized mortality ratio, proportional mortality ratio, and directly standardized risk ratio.

Results: Poultry workers as a group had an overall excess of deaths from diabetes, anterior horn disease, and hypertensive disease, and a deficit of deaths from intracerebral hemorrhage. Deaths from zoonotic bacterial diseases, helminthiasis, myasthenia gravis, schizophrenia, other diseases of the spinal cord, diseases of the esophagus and peritonitis were non-significantly elevated overall by all analyses, and significantly so in particular race/sex subgroups.

Conclusions: Poultry workers may have excess occurrence of disease affecting several organs and systems, probably originating from widespread infection with a variety of

micro-organisms. The results for neurological diseases are very interesting, and could well represent important clues to the etiology of these diseases in humans. The small numbers of deaths involved in some cases limit interpretation.

APPENDIX G

ABSTRACT 5

A PILOT CASE-COHORT STUDY OF LUNG CANCER IN POULTRY & CONTROL WORKERS.

Preacely N,[1] Felini MJ,[1] Shah N,[1,2] Christopher A,[1,2] Sarda V,[1] Elfaramawi M,[1,2] Sall M,[2] Bangara S,[1] Gandhi S,[1] Johnson ES,[1] Ndetan, H.

To evaluate whether humans exposed to the oncogenic viruses of poultry have increased risk of lung cancer, we conducted a pilot case-cohort study of lung cancer nested within a cohort of poultry and non-poultry workers. Specifically we wanted to 1) examine if it is feasible to conduct a large-scale study; 2) provide preliminary information on exposures associated with excess lung cancer risk; 3) test a draft questionnaire. Risk ratios were estimated controlling for tobacco smoking by the Mantel-Haenszel method and by logistic regression. There was no evidence of any serious source of bias, and the results obtained for poultry, and non-poultry risk factors related to meat intake, were consistent with those reported in the literature, indicating that a full-scale study is feasible and will give valid results. Working in the stockyard where exposure to live animals occurred, and slaughtering of poultry birds which are associated with the greatest opportunity for exposure to the oncogenic viruses of poultry were associated with the highest occupational risks of lung cancer. This finding may have important public health implications, since the general population is also exposed to these viruses. Working in the deli and meat departments of supermarkets where exposure to fumes from the wrapping machine occurred also appeared to be associated with significant risks of lung cancer after controlling for tobacco smoking.