Assessing the Impact of Occupation on Death During Eleven Years of Follow-up Using Data from the National Longitudinal Mortality Study

[1]Jessie Ausman

[1]Epidemiology Department, University of North Texas
Health Science Center, Fort Worth 76107, Texas

1. Introduction

Social determinants of health (SDOH) are defined by the Centers for Disease Control and Prevention (CDC) as the "conditions in the places where people live, learn, work, and play that affect a wide range of health and quality-of life risks and outcomes" [2]. SDOH that have been previously recognized as determinants which influence an individual's lifespan include geography, socioeconomic status (SES), race/ethnicity, gender, and age [3]. Although employment status and type of occupation are recognized as aspects of an individual's SES, no large-scale studies have attempted to analyze differences in survival time by employment status and/or occupational category, to our knowledge. Previously conducted research has largely focused on analyzing mortality trends among specific workers (such as coal miners), but we intend to determine whether the type of occupation worked affects an individual's hazard of death for a broad group of workers [4]. Should we find that occupational category is statistically associated with diminished survival time in a large, generalizable sample, we hope to improve public health by 1) identifying potential health disparities among different categories of workers and 2) recommending interventions aimed at reducing the hazards associated with the highest risk occupational categories.

To address these current gaps in research, we have utilized mortality data from the National Longitudinal Mortality Study (NLMS). We plan to conduct a descriptive analysis of the demographic composition of our sample and utilize Extended Cox Survival Analysis and Kaplan Meier Modeling to answer the following research questions: 1) 'does survival time differ by employment status at the time of interview?' (i.e. 'is there a statistical association between employment status and survival time?'), 2) 'does survival time differ based on the type of occupation worked at time of interview?' (i.e. 'is there a statistical association between occupational category and survival time?'), and 3) 'which occupational category is associated with the greatest hazard of death (while controlling for other confounding factors)?'. We hypothesize that both employment status and occupational category are significant predictors of survival time. Additionally, we hypothesize that employed individuals will have a lesser hazard of death than those who are unemployed (since they likely have more expendable income to use towards healthcare). Finally, we hypothesize that, in accordance with this previous hypothesis, those who have never worked or worked without pay will have the greatest hazard of death across all occupational categories.

2. Overview of Data

Data for all analyses was extracted from the NLMS, a nation-wide study sponsored by various components of the National Institutes of Health (NIH), National Center for Health Statistics (part of the Centers for Disease Control and Prevention [CDC]), and the U.S. Census Bureau. The NLMS Public Use Microdata Sample (PUMS) dataset contains demographic, socioeconomic, and mortality data for 1,835,072 individuals who were followed-up with for a total of eleven years [1]. We chose to restrict our dataset to individuals aged 15+ years to ensure that all eligible subjects included in our study were of working age. After this restriction, our sample contained 1,477,743 individuals.

Although all individuals are followed up for a maximum of eleven years in the PUMS dataset, the actual starting point of observation may vary. Follow-up data was collected in the 1980s and 1990s, but a theoretical 'starting point' for all study subjects was set as April 1, 1990 [1]. All data included in the PUMS dataset was extracted from the U.S. Census Bureau and death certificates [1].

Within the restricted PUMS dataset, a weighting variable was available to account for population totals for the non-institutionalized U.S. population [1]. While this variable was available, the SAS Statistical Software® does not offer a feasible weighting option for survival analysis procedures, so we were not able to include weighting in our analyses. The implications of this exclusion will be further addressed in the discussion section.

The outcome variable of interest in this study is time until death (follow). The censoring variable (inddean) describes whether or not the individual experienced death during the eleven-year follow-up period. No participants were lost to follow-up during the NLMS, so only those who did not experience death by the end of the study were censored in our analyses. The main exposures of interest are employment status (emp), which is divided into two categories (employed vs. unemployed), and occupational category (rcow), which is divided into five categories (never worked, private industry, government, self-employed, worked without pay). Other predictor variables that will be assessed for confounding on the relationship between exposure and outcome include biological sex (sex), age (age), urban-rural status (urban), highest level of education (edu), race (race), and Hispanic origin (hisp1).

Employment status (emp) is a binary variable that we created from the occupational category (rcow) variable. Employment status was classified as unemployed if rcow was equal to 5 (never worked) and was classified as employed otherwise. Biological sex (sex) was retained as a binary variable representing either male (1) or female (2). Age was reported as a continuous variable and urban-rural status was reported as a binary variable representing either urban (1) or rural (2) geographical location. Race was retained as a categorical variable representing 5 potential categories: White (1), Black (2), American Indian or Alaskan Native (3), Asian or Pacific Islander (4), or other nonwhite (5). We recoded the NLMS Hispanic origin variable to include Mexicans and other Hispanics in the same category (1) and non-Hispanics as a separate category (0) in a variable called hisp1. We also recoded the NLMS educational attainment variable by forming new categories representing no or less than one year of education (0), elementary education (1), high school education (2), or college education (4) in a variable called 'edu'. Finally, we recoded the occupational category variable (rcow) to combine the categories 'never worked' and 'worked without pay' under a single category (4) in the new variable 'rcow2'. Other 'rcow2' categories included private industry (1), government (2), and self-employed (3).

3. Methods

First, descriptive analyses were performed to describe the composition of employed vs. unemployed study participants and to determine if the aforementioned predictors were statistically associated with employment status. To do so, SAS Statistical Software® was utilized to calculate the mean and standard deviation for continuous predictor(s) and a Kruskal-Wallis test of independence was performed to determine whether a statistically significant association existed between continuous predictor(s) and employment status. We decided to use the Kruskal-Wallis non-parametric test of independence for continuous predictor(s) since the assumption of normality was likely violated by our sample. For categorical predictors, the frequency and percentage of individuals who fit within each of the specified categories for each variable by employment status was calculated and a student's t-test was utilized to determine whether a statistically significant association existed between categorical predictor(s) and employment status.

We also graphed the frequency of individuals who fell within each of the occupational categories as a part of our descriptive analyses using a vertical bar chart. Next, we began our inferential analyses by testing the Cox proportional hazards (PH) assumption for each of the predictor and exposure variables using the Schoenfeld residuals approach. This approach entailed calculating the Schoenfeld residuals for each of the predictor/exposure variables and testing the correlation between these residuals and the ranked failure times. Since the p-values of this test of correlation were insignificant for the hisp1, edu, and age predictors, we considered the PH assumption to be violated and were unable to fit a typical Cox PH model. Instead, we fit an Extended Cox model by creating time-dependent interaction terms for the violating predictors. These interaction terms were created by multiplying the survival time variable (follow) by the violating variable(s) and including these interaction terms in the

Extended Cox model. From here, we fit a no-interaction and an interaction model and assessed which model was a better fit via a likelihood ratio test. We did not consider all possible interaction combinations, but rather only those that we found to be theoretically important.

We figured that it was likely that the likelihood of being employed would depend on sociodemographic factors such as race, educational attainment, and age. We also suspected that the occupational category may differ based on biological sex, as males and females may have different preferences or hiring rates for different types of jobs. Additionally, it is probable that educational attainment may vary based on age (since older individuals will have more chance to complete higher levels of schooling) and race (since individuals of different races would be likely to have different cultural expectations regarding educational attainment). For these purposes, we only considered interaction between occupational category and biological sex, employment status and race, employment status and educational attainment, employment status and age, educational attainment and age, and educational attainment and race. Using a Chi-Square likelihood ratio test, we determined that it was important to consider interaction in our final model. At this point, we used a one-at-a-time approach, dropping insignificant variables from the model one-at-a-time and using likelihood ratio tests to sequentially narrow down the number of variables included in our model until a single, best fitting model was settled on. Whether or not employment status and/or occupational category were included in our final model would help us answer our first two research questions. Once a final model was settled on, this model was used to assess the hazard associated with different employment statuses and different occupational categories to determine whether survival time was significantly associated with either of these variables. This would help us answer our final research question.

The final model consisted of a sample of 978,957 individuals. For this model, ‘unemployed’ was used as the reference group for the employment status variable (emp) and the ‘never worked/worked without pay’ category was used as the reference for the occupational category variable (rcow2).

Table 1. Descriptive analysis of employed vs. unemployed, NLMS, (N = 978,957).

	Employed (n= 972,567)	**Unemployed (n= 6,390)**	**P-value**
Age[α]	37.74±13.78	20.11±8.56	< 0.001
Sex[β]			< 0.001
Male	525,954 (54.08%)	2,994 (46.85%)	
Female	446,613 (45.92%)	3,396 (53.15%)	
Race[β]			< 0.001
White	855,959 (88.01%)	4,749 (74.32%)	
Black	81,938 (8.42%)	1,356 (21.22%)	
American Indian or Alaskan Native	8,198 (0.84%)	80 (1.25%)	
Asian or Pacific Islander	22,395 (2.30%)	183 (2.86%)	
Other Non-White	2,281 (0.23%)	22 (0.34%)	
Hispanic Origin[β]			< 0.001
Hispanic	87,918 (9.04%)	926 (14.49%)	
Non-Hispanic	863,768 (88.81%)	5,386 (84.29%)	
Highest Level of Education[β]			< 0.001
None/ Less Than 1 Year of School	2,289 (0.24%)	26 (0.41%)	
Elementary School	61,217 (6.29%)	858 (13.43%)	
High School	493,934 (50.79%)	4,938 (77.28%)	
College	415, 016 (42.67%)	566 (8.86%)	
Geographic Residence[β]			< 0.001
Urban	681,086 (70.03%)	4,715 (73.79%)	
Rural	291,166 (29.94%)	1,671 (26.15%)	

[α] Presented as mean±standard deviation. P-value determined via Kruskal-Wallis Chi-Square test.

[β] Presented as count (percentage). P-value determined via Student’s t-test (Chi-square test).

Finally, we took a simple random sample consisting of 25% of the total population used for our previous analyses (N = 369,436). We found it necessary to take a simple random sample since our available memory was not sufficient to support analyses using the entire sample size. We used this simple random sample to model the survival across groups stratified by employment status and occupational category (separately) using Kaplan-Meier curves. Additionally, we found it important to also model these survival curves using log negative-log plots, since it is often easier to discern between groups with very similar survival curves in the typical Kaplan-Meier plots.

4. Results

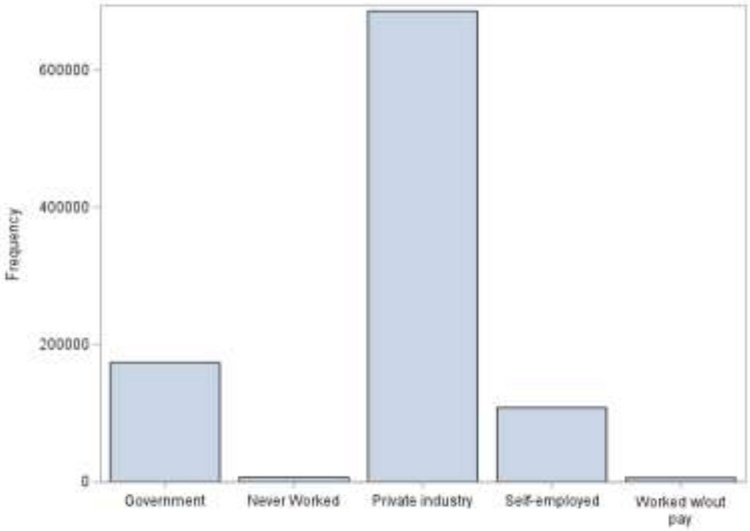
Descriptive analyses revealed that the employed group (n = 972,567) was statistically significantly older than the unemployed population (n = 6,390) in our study sample (p < 0.0001) (**Table 1**). The mean age for the employed subsample was about 38 years, whereas the mean age for the unemployed subsample was only about 20 years (**Table 1**).

Additionally, the employed group consisted of a greater proportion of males (54.1%), Whites (88.0%), and Non-Hispanic (88.8%) individuals (**Table 1**). Conversely, the unemployed group consisted of a greater proportion of females (53.2%) (**Table 1**). The most represented racial/ethnic categories among the unemployed was still White (74.3%) and Non-Hispanic (84.3%) individuals, but the proportion of Blacks among the unemployed group (21.2%) significantly differed from this proportion among the employed group (8.4%) (p < 0.0001) (**Table 1**).

While high school (50.8%) represented the most prominent category of educational attainment among employed individuals, the proportion of employed individuals whose highest level of educational attainment was college (42.7%) was significantly different from this proportion among the unemployed population (8.9%) (p < 0.0001) (**Table 1**).

As would be expected, a greater proportion of the study sample overall inhabited urban areas (70.0% of employed; 73.8% of unemployed) (**Table 1**). Interestingly, though, the percentage of individuals among the employed group who inhabited rural areas (29.9%) was greater than this percentage among the unemployed group (26.2%) (**Table 1**).

Graph 1. Bar chart describing the distribution of occupational categories worked by study subjects, NLMS, (N = 978,957).



We depicted the distribution of occupational categories worked by the study group by graphing this distribution using a bar chart. As can be seen in **Graph 1**, this distribution is not normally distributed in any way. The frequency of private industry workers is much greater than any other occupational category, followed by government workers and the self-employed (**Graph 1**). Those falling in the ‘never worked’ and ‘worked without pay’ categories were so few that we decided to combine these categories during subsequent inferential analyses.

When we tested the PH assumption using Schoenfeld’s residuals, multiple predictor variables violated this assumption. More specifically, educational attainment (p = 0.713), Hispanic origin (p = 0.208), and age (p = 0.626) all violated the PH assumption (**Table 2**). Thus, we found it necessary to create time-dependent variables for these predictors and fit an Extended Cox model instead of a typical Cox PH model. Multiple Extended Cox models were fit until a set of predictors representing the best fitting model was settled on.

The predictors and interaction terms included in this final model can be found in **Table 3**. All predictors in the final model are statistically significantly associated with survival time, except employment status (p = 0.061) (**Table 3**). However, interaction terms including employment status were statistically significant at the 95% level of confidence, so this predictor could not be dropped from the model. The model fit statistics (-LogL, AIC, SBC) for this model are also presented in **Table 4**.

Table 2. Pearson correlation coefficients and p-values for predictors violating the PH assumption.

Predictor Variable	Pearson Correlation Coeff.	P-value
Educational attainment (*edu*)	0.00175	0.7133
Hispanic Origin (*hisp1*)	0.00600	0.2081
Age (*age*)	-0.00233	0.6257

Table 3. Type 3 Chi-Square statistics for predictors included in the final Extended Cox Model.

Parameter	SAS Variable	Chi-Square	p-value[α]
Employment Status	*emp*	3.511	0.061
Occupational Category	*rcow2*	310.185	<0.0001
Race	*race1*	675.799	<0.0001
Hispanic Origin	*hisp1*	2286.151	<0.0001
Educational Attainment	*edu*	43497.580	<0.0001
Biological Sex	*sex*	24.456	<0.0001
Age	*age*	26689.715	<0.0001
Urban-rural Status	*urban1*	19.848	<0.0001
Hispanic Origin (Time-dependent variable)	*hisp_t*	3209.515	<0.0001
Educational Attainment (Time-dependent variable)	*edu_t*	1823.597	<0.0001
Age (Time-dependent variable)	*age_t*	46714.626	<0.0001

Interaction Parameters

Interaction Term	Chi-Square	p-value[α]
*edu_t * age_t*	45654.935	<0.0001
*edu_t * race1*	338.141	<0.0001
*edu_t * emp*	22.439	<0.0001
*sex1 * rcow2*	65.993	<0.0001

α. P-value is reported for Chi-Square joint tests. The level of confidence used to determine statistical significance of p-values is 95%.

Table 4. Statistics to assess model fit for the chosen Extended Cox model.

Model Fit Criterion	Value (With Covariates)
Negative Log-Likelihood	806183.43
AIC	806237.43
SBC	806472.08

According to the results from the Extended Cox model described in **Tables 3** and **4**, those who were employed had a lesser hazard of death than those who were unemployed (HR = 0.502) (**Table 5**). The 'never worked/worked without pay' occupational category was used as the reference group for the remaining hazard ratios. According to the results in **Table 5**, those who worked in the private industry, in government positions, and who were self-employed (separately) have a greater hazard of death than those who never worked/worked without pay (HR[private industry] = 2.425; HR[government] = 2.719; HR[self-employed] = 2.041).

Overall, the results from our final Extended Cox model supported our hypotheses that employment status and occupational category are statistically significantly associated with survival time. Additionally, the employment status hazard ratio supported our hypothesis that employed individuals have a lesser hazard of death than unemployed individuals. Somewhat contradictorily though, the remaining hazard ratio results suggested that private industry, government, and self-employed workers had greater hazards of death than those belonging to the 'never worked/worked without pay' category, which did not support our original hypothesis.

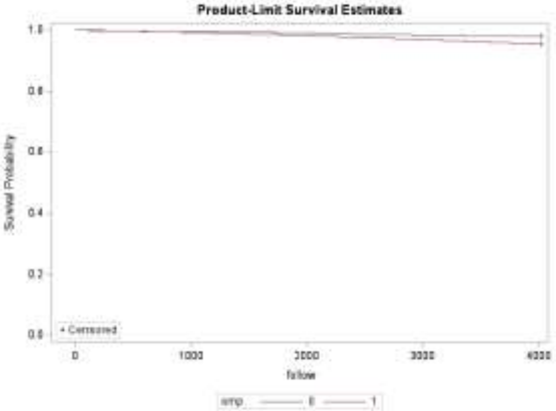
Table 5. Results of inferential analyses: Hazard ratios by employment status, NLMS, (N = 978,957).

Exposure	Hazard Ratio
Employment status *('employed' vs. 'unemployed')*	0.502
Occupational category *('private industry' vs. 'never worked/worked without pay')*	2.425
Occupational category *('government' vs. 'never worked/worked without pay')*	2.719
Occupational category *('self-employed' vs. 'never worked/worked without pay')*	2.041

Finally, the Kaplan-Meier survival curves were very similar between the employment status and occupational category strata, making it difficult to draw any real conclusions from this output (**Graph 2** and **3**). The log negative-log plots provided a more apparent contrast between the various strata. According to these plots, the unemployed group seemed to have a greater survival probability earlier in the study’s follow-up (**Graph 4**). However, later in the study, it seemed as if those who were employed had a greater survival probability than those who were unemployed (**Graph 4**). A Log-Rank test of equality over these strata revealed that the survival was significantly different across the strata (p < 0.0001).

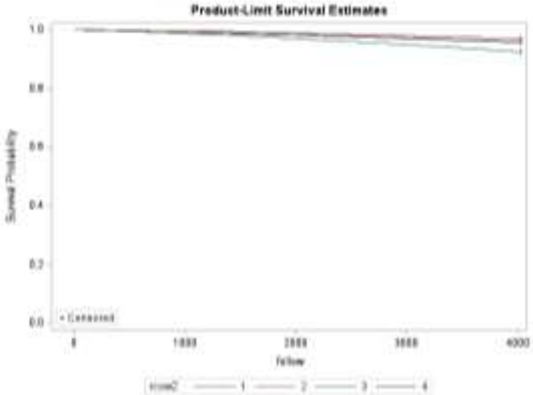
The survival probabilities of various strata defined by occupational category seem fairly similar later in the study’s follow-up, but early on there seemed to be much more discrepancy (**Graph 5**). Throughout majority of follow-up, it seemed that those who were self-employed had the greatest probability of survival, though these results should be interpreted with caution. A Log-Rank test of equality over these strata revealed that the survival was significantly different across the strata (p < 0.0001).

Graph 2. Kaplan-Meier survival curves stratified by employment status, (N = 369,436).



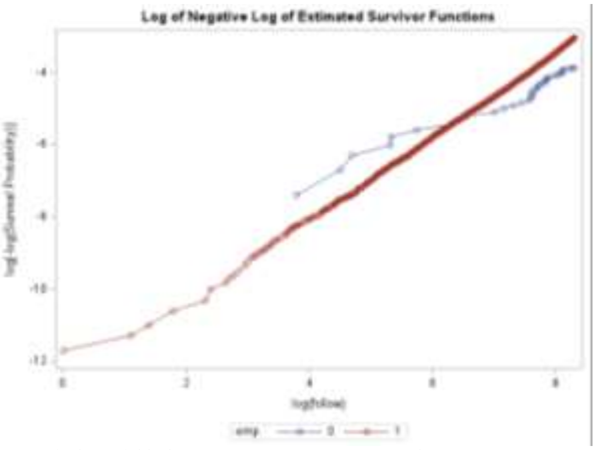
α. Emp is the variable for employment status. Emp = 0 if unemployed; Emp = 1 if employed.

Graph 3. Kaplan-Meier survival curves stratified by occupational category, (N = 369,436).



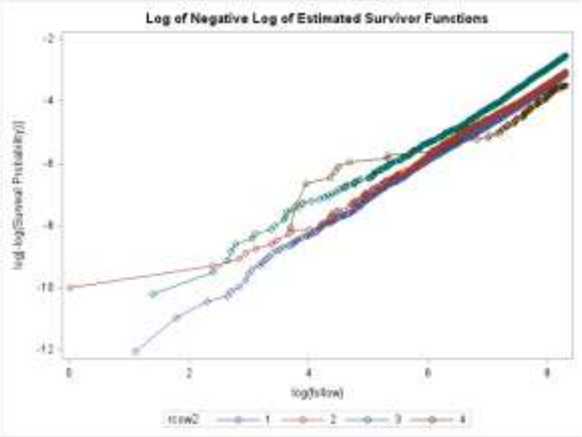
α. Rcow2 is the variable for occupational category. Rcow2 = 1 if private industry; Rcow2 = 2 if government, Rcow2 = 3 if self-employed, and Rcow2 =4 if worked without pay or never worked.

Graph 4. Log Negative-Log curves stratified by employment status, (N = 369,436).



α. Emp is the variable for employment status. Emp = 0 if unemployed; Emp = 1 if employed.

Graph 5. Log Negative-Log curves stratified by occupational category, (N = 369,436).



α. Rcow2 is the variable for occupational category. Rcow2 = 1 if private industry; Rcow2 = 2 if government, Rcow2 = 3 if self-employed, and Rcow2 =4 if worked without pay or never worked.

5. Discussion & Conclusions

In conclusion, our results suggest that in addition to pre-established determinants of health and survival (age, gender, race/ethnicity, and geographical location), employment status and occupational category are significant predictors of an individual's survival time, controlling for other aforementioned confounding variables. In the context of public health action and research, this means that it may not only be important to consider an individual's SES when assessing their health and survival probability. Instead, it may also be important to individually assess the effects of employment status and type of occupation on these outcomes as well. When designing public health interventions to target high-risk groups, it may be especially important to target certain groups defined by employment status and/or type of occupation.

Moreover, our results also suggest that employed individuals have a greater chance of survival than those who are unemployed. However, our hazard ratio results somewhat contradict this conclusion, as these results showed that those who were categorized as never having worked or only having worked without pay had a lower hazard of death. The Kaplan-Meier survival curves and Log Negative-Log plots only convoluted these results more, as it was very difficult to determine which groups truly had a greater survival probability given the limitations of our analyses and our study. Overall, we conclude that employment status and occupation type are significant predictors of survival, but, unfortunately, the results (combined with our knowledge of study limitations) are not conclusive enough to identify specific categories of workers that should be targeted for high-risk public health interventions.

As mentioned, this study included multiple limitations. First, dataset used for this study only included individuals that were not lost during the eleven years of follow-up, however the NLMS reference manual does not adequately specify whether these individuals were dropped from the study entirely without including information collected on them before loss to follow-up or whether there was truly no loss of this sort throughout the study duration. If individuals who were lost to follow-up were dropped, then this would introduce significant bias by failing to account for other forms of censoring in our analyses.

Secondly, the categories of occupation included in our analyses may be too broad to adequately assess risk of death across various types of occupation. Additionally, any results regarding the hazard of death across the occupational categories included in our analyses may not be very applicable in public health practice since the occupational categories are so broad and interventions cannot be feasibly targeted towards specific categories of workers.

Third, although parametric models may have produced better fitting models and more valid results for our analyses, we did not have sufficient memory to produce probability plots to assess parametric model fit. Even when we attempted to produce these plots using our simple random sample generated for the Kaplan-Meier and Log Negative-Log plots for parametric modeling, the SAS Statistical Software® was unable to produce the necessary probability plots. Thus, we only considered the Extended Cox model when choosing a final model to use for our inferential analyses.

Finally, the simple random sample that was generated for plotting the Kaplan-Meier survival curves and the Log Negative-Log curves was not stratified by any specific variables. Additionally, for lack of resources and/or time, we were unable to perform descriptive analyses for this smaller sample. Thus, although the total sample was determined to be generalizable to the US population, the results from the simple random sample cannot be confidently generalized.

In future studies, we would advise researchers to use hardware with more memory or statistical software/procedures that do not have such restrictive hardware requirements to complete survival analyses on the basis of employment status and/or occupational category. We would also improve the results of this study by assessing the generalizability of the findings from the analyses performed on the simple random sample. While our results strongly support the hypothesis that employment status is significantly predictive of survival and that employed individuals have a greater chance of survival than unemployed individuals, further studies will be needed to more accurately determine whether different types of occupation are associated with greater chances of survival across time. In future studies, we would suggest that researchers use less broad occupational categories in their analyses, if possible. This will improve the chances that study findings can be applied in actual public health practice and interventions.

6. Appendix

```
/*******************************************
*****************************************
 *Course: Survival Analysis - BIOS 6324   *
 *Assignment Final Project NLMS Data     *
 *    Programmer(s): Jessie Ausman       *
 *    Program Name: Final Project        *
 *    Save Program/Log/Output:
C:\users\jessa\desktop\bios
6324\project\what i need      *
 *    Save Data Files:
C:\users\jessa\desktop\bios
6324\project\what i need                 *

********************************************
*****************************************/


/*************************************
      Part 0
      ***********************************/

*read in data;
libname bios "C:\users\jessa\desktop\bios
6324\project\what i need";

*create temporary dataset;
data Proj;
      set bios.eleven_new;
run;



*need to create censoring variable:
      censored if...
            - lost to follow up without
inddea recorded;

proc freq data=proj;
      tables cause113 follow inddea
follow*inddea age/ missing;
run;

*restrict sample to only those 15+ years;
data proj1;
      set proj;
      if age GE 15;
run;

*create variable indicating employment
status:
      0 = unemployed
      1 = employed;
data proj2;
      set proj1;
      if rcow = 5 then emp = 0;
      else if rcow in (1,2,3,4) then emp =
1;
      else if missing(rcow) then emp = .;
run;
```

```
proc freq data=proj2;
      tables rcow*emp emp;
run;

*recode hispanic origin - hispanic vs. non-
hispanic;
data proj3;
      set proj2;
      if hisp in (1,2) then hisp1 = 1;
      else if hisp = 3 then hisp1 = 0;
      else if missing(hisp) then hisp1 = .;
run;

*recode educ - none/elementary/high/college;
data proj4;
      set proj3;
      if educ = 01 then edu = 0;
             *none;
      else if educ in (02,03,04) then edu =
1; *elementary;
      else if educ in (05,06,07,08) then edu
= 2; *highschool;
      else if educ in (09,10,11,12,13,14)
then edu = 3; *college;
      else if missing(educ) then edu = .;
run;

data proj4_1;
      set proj4;
      inddean = input(inddea, 8.);
      race1 = input(race, 8.);
      urban1 = input(urban, 8.);
      sex1 = input(sex, 8.);
      rcow1 = input(rcow, 8.);
run;

proc contents data=proj4_1 varnum;
run;

/*****************************************
      Part 1
      ***********************************/

/*Descriptive analysis;*/


      **************************************
*****************************************
      *urban --> 1 = urban / 2 = rural

                    *
      *sex --> 1 = male / 2 = female

                          *
      *race --> 1 = white / 2 = black / 3 =
american indian / 4 = asian / 5 = other *
      *hisp1 --> 0 = non-hispanic / 1 =
hispanic
                    *
      *edu --> 0 = none / 1 = elementary / 2
= highschool / 3 = college
*
      **************************************
****************************************;


*includes chi-sq test for categorical vars;
proc freq data=proj4;
      tables emp*sex emp*race emp*hisp1
emp*edu emp*urban/ chisq missing;
run;

*mean & std dev for age;
proc sort data=proj4;
      by emp;
run;

proc means data=proj2;
      var age;
      by emp;
run;

*KW Chi-square test for cont. var;
proc npar1way data=proj4 wilcoxon;
      class emp;
      var age;
run;

data proj5;
      set proj4;
      if rcow = 01 then cat = "Private
industry";
      else if rcow = 02 then cat =
"Government";
      else if rcow = 03 then cat = "Self-
employed";
      else if rcow = 04 then cat = "Worked
w/out pay";
      else if rcow = 05 then cat = "Never
Worked";
      else if missing(rcow) then cat = " ";
run;

*bar chart for rcow;
proc sgplot data=proj5;
      vbar cat;
run;


/*****************************************
      Part 2
      ************************************/

/*check ph assumption*/
PROC PHREG DATA=proj4_1;
      title "No-interaction Cox PH";
      class       hisp1 (param=ref ref='0')
/*ref = nonhispanic*/
                  race1 (param=ref ref='1')
/*ref = white*/
                  edu (param=ref ref='0')
/*ref = none*/
                  emp (param=ref ref='0')
/*ref = unemployed*/
                  sex1 (param=ref ref='1')
       /*ref = male*/
                  urban1 (param=ref
ref='2')/*ref = rural*/
                  rcow1 (param=ref
ref='5')/*ref = never worked*/;
      model follow*inddean(0) = emp rcow1
race1 hisp1 edu sex1 age urban1 / type3;
      output out=resid ressch= remp rrcow1
rrace1 rhisp1 redu rage rurban1;
RUN;

data events;
      set resid;
      if inddean = 1;
run;

proc rank data=events out=ranked ties=mean;
      var follow;
      ranks timerank;
run;


proc corr data=ranked nosimple;
      var remp rrcow1 rrace1 rhisp1 redu
rage rurban1;
      with timerank;
run;




/*******************************************
      Part 3
      ************************************/

/*test interaction vs. no interaction model
- Extended Cox PH*/


data proj6;
      set proj4_1;

      hisp_t = hisp1*follow;
      edu_t = edu*follow;
      age_t = age*follow;

      hisp_log = hisp1*log(follow);
      edu_log = edu*log(follow);
      age_log = age*log(follow);

      if rcow1 = 01 then rcow2 = 1;
      else if rcow1 = 02 then rcow2 = 2;
      else if rcow1 = 03 then rcow2 = 3;
      else if rcow1 in (04, 05) then rcow2 =
4;
```

```
      else if missing(rcow1) then rcow2 = .;
run;


proc freq data=proj6;
      table rcow2 rcow1;
run;



/*MODEL 1*/
PROC PHREG DATA=proj6;
      title "No-interaction Extended Cox
PH";
      class       hisp1  (param=ref ref='0')
/*ref = nonhispanic*/
                  race1  (param=ref ref='1')
/*ref = white*/
                  edu  (param=ref ref='0')
/*ref = none*/
                  emp  (param=ref ref='0')
/*ref = unemployed*/
                  sex1  (param=ref ref='1')
       /*ref = male*/
                  urban1  (param=ref
ref='2')/*ref = rural*/
                  rcow2  (param=ref
ref='4')/*ref = never worked/worked without
pay*/;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1 hisp_t edu_t
age_t/ type3;
RUN;


/*MODEL 2*/
PROC PHREG DATA=proj6;
      title "No-interaction Extended Cox PH
- logtime";
      class       hisp1  (param=ref ref='0')
/*ref = nonhispanic*/
                  race1  (param=ref ref='1')
/*ref = white*/
                  edu  (param=ref ref='0')
/*ref = none*/
                  emp  (param=ref ref='0')
/*ref = unemployed*/
                  sex1  (param=ref ref='1')
       /*ref = male*/
                  urban1  (param=ref
ref='2')/*ref = rural*/
                  rcow2  (param=ref
ref='4')/*ref = never worked/worked without
pay*/;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1 hisp_log
edu_log age_log/ type3;
RUN;
```

```
********************************************
**************

Compare AIC to determine which time function
to use:

time AIC = 841967.76

logtime AIC = 905281.93

-> Smaller AIC is better, so we will use
time rather than
logtime for time-dependent vars.
********************************************
**************;

/*MODEL 3*/
PROC PHREG DATA=proj6;
      title "Interaction Cox PH";
      class       hisp1  (param=ref ref='0')
/*ref = nonhispanic*/
                  race1  (param=ref ref='1')
/*ref = white*/
                  edu  (param=ref ref='0')
/*ref = none*/
                  emp  (param=ref ref='0')
/*ref = unemployed*/
                  sex1  (param=ref ref='1')
       /*ref = male*/
                  urban1  (param=ref
ref='2')/*ref = rural*/
                  rcow2  (param=ref ref='4')
       /*ref = never worked/worked without
pay*/;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1
                  hisp_t edu_t age_t
                  edu_t*age_t edu_t*race1
                  emp*edu_t emp*age_t
emp*race1
                  rcow2*sex1/ type3;
RUN;



*run LR test interaction vs no interaction;
DATA TEST;
REDUCED = 841931.76;
FULL = 806178.22;
DF = 14;
PVALUE = 1 - PROBCHI(REDUCED-FULL,DF);
RUN;

proc print data=test;
run;

*CONCLUSION: p-value significant --> Must
assess interaction;
```

```
/********************************************
      Part 4
      *************************************/

*drop any non-significant variables to find
best fitting model;

/*full model*/
PROC PHREG DATA=proj6;
      title "Interaction Cox PH";
      class       hisp1 (param=ref ref='0')
/*ref = nonhispanic*/
                  race1 (param=ref ref='1')
/*ref = white*/
                  edu (param=ref ref='0')
/*ref = none*/
                  emp (param=ref ref='0')
/*ref = unemployed*/
                  sex1 (param=ref ref='1')
       /*ref = male*/
                  urban1 (param=ref
ref='2')/*ref = rural*/
                  rcow2 (param=ref ref='4')
       /*ref = never worked/worked without
pay*/;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1
                  hisp_t edu_t age_t
                  edu_t*age_t edu_t*race1
                  emp*edu_t emp*age_t
emp*race1
                  rcow2*sex1/ type3;
RUN;

**************drop race1*emp since it is
most insignificant & perform LR
test****************;

/*ALT MODEL 1*/
PROC PHREG DATA=proj6;
      title "Interaction Cox PH";
      class       hisp1 (param=ref ref='0')
/*ref = nonhispanic*/
                  race1 (param=ref ref='1')
/*ref = white*/
                  edu (param=ref ref='0')
/*ref = none*/
                  emp (param=ref ref='0')
/*ref = unemployed*/
                  sex1 (param=ref ref='1')
       /*ref = male*/
                  urban1 (param=ref
ref='2')/*ref = rural*/
                  rcow2 (param=ref ref='4')
       /*ref = never worked/worked without
pay*/;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1
                  hisp_t edu_t age_t
                  edu_t*age_t edu_t*race1
                  emp*edu_t emp*age_t
                  rcow2*sex1/ type3;
RUN;

DATA TEST;
REDUCED = 806181.61;
FULL = 806178.22;
DF = 4;
PVALUE = 1 - PROBCHI(REDUCED-FULL,DF);
RUN;

proc print data=test;
run;

****insignificant p-value so we can drop
emp*race1*****;



**************drop age_t*emp since it is
most insignificant & perform LR
test****************;


/*ALT MODEL 2*/
PROC PHREG DATA=proj6;
      title "Interaction Cox PH";
      class       hisp1 (param=ref ref='0')
/*ref = nonhispanic*/
                  race1 (param=ref ref='1')
/*ref = white*/
                  edu (param=ref ref='0')
/*ref = none*/
                  emp (param=ref ref='0')
/*ref = unemployed*/
                  sex1 (param=ref ref='1')
       /*ref = male*/
                  urban1 (param=ref
ref='2')/*ref = rural*/
                  rcow2 (param=ref ref='4')
       /*ref = never worked/worked without
pay*/;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1
                  hisp_t edu_t age_t
                  edu_t*age_t edu_t*race1
                  emp*edu_t
                  rcow2*sex1/ type3;
RUN;


DATA TEST;
REDUCED = 806183.43;
FULL = 806181.61;
DF = 1;
PVALUE = 1 - PROBCHI(REDUCED-FULL,DF);
RUN;

proc print data=test;
```

```
run;

****insignificant p-value so we can drop
emp*age_t*****;

********************************************
***************************************
all remaining interaction terms are
significant at 95% level of confidence.
only predictor remaining in the model with
an insignificant p-value is
emp, however emp is a component of several
significant interaction terms so
it cannot be dropped from the model
                  --> ALT MODEL 2 is our
chosen model
********************************************
**************************************;



/*******************************************
      Part 5
      ************************************/

/*test parametric models for best fitting
model - only test parametric models that do
NOT
rely on the PH assumption (lognormal and
loglogistic)*/


/*ALT MODEL 3*/
PROC LIFEREG DATA=proj6;
      title "Parametric Cox - Lognormal";
      class       hisp1
                  race1
                  edu
                  emp
                  sex1
                  urban1
                  rcow2;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1
                  hisp_t edu_t age_t
                  edu_t*age_t edu_t*race1
                  emp*edu_t
                  rcow2*sex1/ dist=lnormal;
      probplot;
RUN;



/*ALT MODEL 4*/
PROC LIFEREG DATA=proj6;
      title "Parametric Cox - Log-logistic";
      class       hisp1
                  race1
                  edu
                  emp
                  sex1
                  urban1
                  rcow2;
      model follow*inddean(0) = emp rcow2
race1 hisp1 edu sex1 age urban1
                  hisp_t edu_t age_t
                  edu_t*age_t edu_t*race1
                  emp*edu_t
                  rcow2*sex1/
dist=llogistic;
      probplot;
RUN;

/*probplots could not be produced --> stick
with ALT MODEL 2 as final model*/



/*******************************************
      Part 6
      ************************************/

proc surveyselect data=proj6
      out = random_sample
      method = srs
      samprate = 0.25
      seed = 1;
run;

proc contents data=random_sample varnum;
run;

PROC LIFETEST DATA=random_sample METHOD=KM
PLOTS=(S,LLS);
TIME follow*inddean(0);
STRATA emp;
RUN;

PROC LIFETEST DATA=random_sample METHOD=KM
PLOTS=(S,LLS);
TIME follow*inddean(0);
STRATA rcow2;
RUN;
```

7. Acknowledgements/References

[1] (2015). *National longitudinal mortality study: Public use microdata sample* [data set & manual]. https://biolincc.nhlbi.nih.gov/studies/nlms/

- The NLMS data used for this study was derived from this source. Additionally, information regarding how this data was collected, how variables were coded, and other information pertaining specifically to the raw dataset were derived from this source.

[2] Centers for Disease Control and Prevention [CDC]. (2021, September 30). *Social determinants of health: Know what affects health.* https://www.cdc.gov/socialdeterminants/index.htm

- This source was used to collect information regarding potentially confounding factors that have been established through previous research as significant determinants of health. Using this information, we were able to identify predictors that were necessary to account for in our statistical models.

[3] Penman-Aguilar, A., Talih, M., Huang, D., Moonesinghe, R., Bouye, K., & Beckles, G. (2016). Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity. *Journal of Public Health Management & Practice, 22*(1), S33-S42.

- This source was also used to collect information regarding pre-established determinants of health that were necessary to account for in our statistical models.

[4] Graber, J. M., Stayner, L. T., Cohen, R. A., Conroy, L. M., & Attfield, M. D. (2015). Respiratory disease mortality among US coal miners; results after 37 years of follow-up. *Occupational and Environmental Medicine, 71*(1), 30-39. https://dx.doi.org/10.1136%2Foemed-2013-101597

- This source was used to show that survival time and risk of mortality has been somewhat evaluated in previous research, but with smaller samples and for very restricted occupational categories. This helped us establish a need for the research we performed in this study.