Assignment 4 (Assignment 2 Revision), Epidem 207: Rebecca Fisher

Original Assignment 2 coauthored equally by Catherine Cortez, Rebecca Fisher, John Waggoner

1. Revise Assignment 2 based on instructor feedback and classmates' replication efforts and include an introduction (maximum 2 pages double-spaced) that (a) briefly summarizes revisions and (b) summarizes what you learned through this exercise.
2.Upload your final code for this assignment to your Epi 207 repo in your Assignment 4 folder. Paste a link to the repo:
https://github.com/rfisher2022/EPIDEM207-2022-winter

a. Introduction-Revisions

Upon review of Professor Mayeda's review and comments, as well as the review/comments from the anonymous reviewer, important clarifying revisions were made to describe the study more accurately as implemented and to improve reproducibility. The objective statement of the study now includes a description of the study population.

In methods, language that clarifies the baseline alcohol consumption measure has been added to labels and text. A considerable addition has been made to the section describe the outcome measure of time to mortality. Previously, the description did not clearly describe how mortality was assessed, how time to event was measured including both death and censorship at the end of the study period, or over what period the mortality assessment covered. While I was not involved in the original study or creation of the data, nor did the original authors clearly describe this, upon reviewing the data itself and making some assumptions, I have added additional detail. Furthermore, I have clarified that the mortality measure is not binary, but rather a time-to-event measure.

Also in methods, additional detail has been given related to why cox proportional hazards modeling was selected to leverage the time-to-event measurements available in the study. The specific models run were also more clearly clarified and labeled/referenced, accounting for both unadjusted crude models as well as the two adjusted models. Furthermore, I added additional detail to this section to clarify that the survival curves included in the results were based on the 3 cox proportional hazards model, rather than just crude Kaplan Meier curves.

In the results section, I took the reviewer's comments that only 2 decimals should be reported and adjusted the Table 2 results accordingly.

In the conclusion, I split the discussion of results from the limitations. I also closed this section with directions for future research that would address the limitations included.

Lastly, the anonymous reviewer noted that I did not include a link to the original study, so this has been added to the supplemental materials section.

b. Introduction-Learning:

I have learned a lot through this course and through this exercise. When my group for assignment 2 began planning our analysis and how we would work together and collaborate, we had to discuss and plan how we would work on the code together, as well as a shared document for the paper write-up and compiling results. Establishing a dedicated github repo made all of that so easy. Whenever any of us individually wanted to work on any part of the project, we just checked out the repository, did our work, documented our work, and then pushed our changes to the repo for review by our other collaborators.

Upon reviewing the anonymous reviewer's comments, I was very pleased and excited by the comment that the repository was a "lifesaver" and that the code was "well documented and ran smoothly". While my reviewer reflected that perhaps the paper alone wasn't detailed enough to allow another researcher to independently reproduce the results, the repo contained everything necessary to reproduce the results. This was exactly the intended goal.

The professor's and the reviewer's feedback on the write-up was also well-received. I think that the revisions that I have made based on their review has clarified points that needed clarification and has improved the quality and completeness of the paper/write-up.

All in all, I would say that I have learned that transparency and documentation of everything needed to reproduce a study's results and openness to reviews by external researchers can really improve the end-product. While reluctance for this and fear that transparency will bring unwanted feedback or harsh critiques, it is much better for the study in the end to lead with confident transparency and an openness to relevant feedback.

Revised Paper:

a. Objective

The objective of this analysis is to describe the association of mid-life alcohol use on mortality over a 20-year follow-up period from baseline among northern German adults aged 18-64.

b. Methods

Sample

Participants aged 18-64 years old were sampled from a population sample of adults in a northern German region using registry data. Of the 5,829 eligible individuals at baseline, 4,093 individuals completed interviews between July 1996 to March 1997. Baseline data was analyzed among 4075 individuals, and vital status at follow-up was obtained from April 2017 to April 2018 among a final analytic sample of 4028 individuals.

Measurement

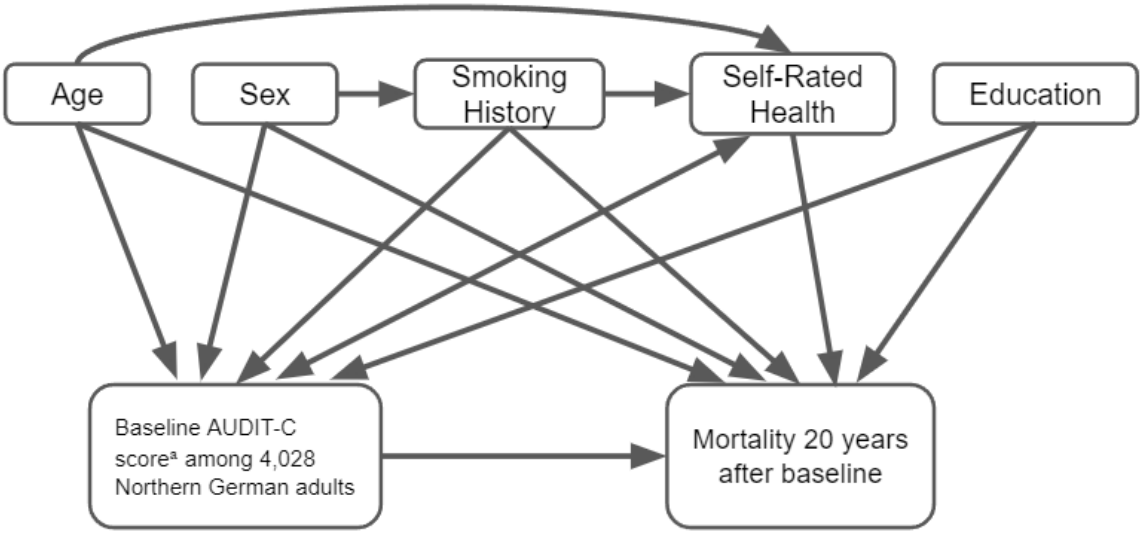
Baseline alcohol consumption
Alcohol abstinence and consumption 12 months prior to baseline was assessed using items from the Alcohol Use Disorder Identification Test-Consumption (AUDIT-C) questionnaire. This tool was selected due to ease of administration, common use, and standardization. A score was calculated to summarize frequency and quantity of alcohol consumption by the following categories: abstinence (0), low to moderate (1 to 3), moderate to high (4), high (5), very high (6 to 7), and extremely high alcohol consumption (8 to 12).

Outcome measures
Mortality records for the entire follow-up period from study enrollment to assessment date were reviewed during a 1-year period from April 2017 to April 2018. Mortality over the study period was assessed by death certificate information obtained from local health authorities where participants resided at baseline, and presumably covered any deaths occurring during the follow up period. The original study did not discuss in detail, however, based on the data I believe that total mortality was determined as of April 2017 regardless of when death assessment occurred in the 1-year period. I also believe that follow up time started at a single point in time because all non-deceased participants had an equivalent amount of follow up time (8035 days). This tells me follow up time start date was the same among all participants, and that for non-deceased participants the same censor date was used. For participants identified as having died during follow up time, their time to event was calculated through the date of death. The mortality outcome is a time-to-event measure.

Covariates
Smoking status at baseline was collected as a 5-level categorical variable, including never smokers, former limited smokers, former daily smokers, current lighter (less than 20 cigarettes per day) smokers and current heavy (20 or more cigarettes per day) smokers. Smoking is a clear confounder between alcohol use and mortality. Educational status was assessed as 12 or more, 10 to 11, or 9 or less years. Educational status has been demonstrated to confound many exposures and mortality, as it acts as a proxy for early life conditions and opportunity. Age and sex at baseline were also assessed and included as covariates. Self-rated health was collected as a 5-level categorical scale variable from excellent to poor, and re-coded as a 3-category variable by John et al. Self-rated health at baseline might have two different roles in

the causal path between alcohol consumption and mortality. At baseline, self-rated health might act as a proxy confounder in the relationship between alcohol consumption and mortality, as those with medical conditions that limit alcohol consumption might also rate their health as ‘poor’. Alternatively, self-rated health might mediate the relationship between alcohol consumption and mortality, as alcohol consumption over the previous 12-month period might directly impact the subject’s perception of health at the time of interview, especially among heavy drinkers.

Figure 1. DAG demonstrating proposed association between AUDIT-C and mortality


Age
Sex
Smoking History
Self-Rated Health
Education
Baseline AUDIT-C score[a] among 4,028 Northern German adults
Mortality 20 years after baseline

[a] AUDIT-C score (Alcohol Use Disorders Identification Test score) is used to measure alcohol consumption and abstinence at baseline

Data analysis

Descriptive summary statistics (frequencies, means) were performed to describe the study sample and follow-up time by study sample characteristics. To explore the association between baseline alcohol consumption as measured by the baseline AUDIT-C score and time to death, we selected cox proportional hazards modeling which can be used to investigate survival time by one or more predictors. Simple (model 1) and multivariate (model 2 and 3) Cox proportional hazards models were used to describe the hazard ratios and 95% confidence intervals for the association of baseline AUDIT-C score and time to death over the follow-up period. Model 1 was unadjusted. Model 2 adjusted for sex, age, education, and smoking history, but did not adjust for self-rated health at baseline under the assumption that it might be an intermediary between alcohol consumption and mortality during follow-up. Model 3 included self-rated health at baseline under the assumption that adjustment for self-rated health at baseline might de-confound the relationship between alcohol consumption and mortality during follow-up. Survival curves were also visualized based on the 3 cox proportional hazards models.

c. Results

We found that among 4028 study participants, 447 individuals (11%) were alcohol abstainers at baseline and slightly over half (54%) of individuals were low to moderate alcohol consumers (Table 1). The mean age was 41.7 years (SD ± 12.9), and roughly half of individuals were female (49.8%) and attained 9 years of education or less (47.8%). Self-rated health status was reported as very good to excellent among 35% of participants. Respective average and maximum follow-up time was 20.7 and 22.0 years.

Table 1. Demographic and baseline characteristics of population-based cohort of 4,028 northern German adults, 1996-1997.

Characteristic	**Total n (column %)**	**Person-years of follow-up Mean (SD)**
Total	4028 (100)	20.66 (4.0)
Demographic variables		
Age, mean (SD)	41.72 (12.9)	
Female	2006 (49.8)	21.0 (3.4)
School education, n(%)		
9 or less years	1927 (47.8)	20.1 (4.7)
10-11 years	1471 (36.5)	21.2 (3.0)
12 or more years	630 (15.6)	21.2 (3.2)
Smoking status, n(%)		
Never smoker	676 (16.8)	20.9 (3.6)
Ever less than daily	920 (22.8)	21.3 (2.9)
Former daily	839 (20.8)	20.5 (4.3)
Current daily <20 cpd	485 (12.0)	20.8 (3.6)
Current daily >= 20 cpd	1108 (27.5)	20.0 (4.7)
Self-rated health, n(%)		
Very good to excellent health	1427 (35.4)	21.2 (3.2)
Good health	1926 (47.8)	20.7 (3.9)
Poor to fair health	675 (16.8)	19.3 (5.3)
AUDIT-C sum score, n(%)		
Abstinent (AUDIT-C=0)	447 (11.1)	19.3 (5.4)

Low to Moderate (AUDIT-C=1-3)	2203 (54.7)	21.0 (3.5)
Moderate to High (AUDIT-C =4)	674 (16.7)	20.9 (3.5)
High (AUDIT-C=5)	383 (9.5)	20.7 (3.9)
Very High (AUDIT-C=6-7)	228 (5.7)	20.3 (4.6)
Extremely High (AUDIT-C =8-12)	93 (2.3)	18.9 (5.9)

Abbreviations: AUDIT-C, Alcohol Use Disorders Identification Test-Consumption

Compared to baseline alcohol abstainers, baseline alcohol consumers at all levels except for extremely high alcohol consumption had reduced risk of mortality over the follow-up time in crude models and both adjusted models (Table 2). Adjusting for sex, age, smoking status, and years of education (Model 2) and self-reported health (Model 3) did reduce the strength of the association but did not eliminate it to null.

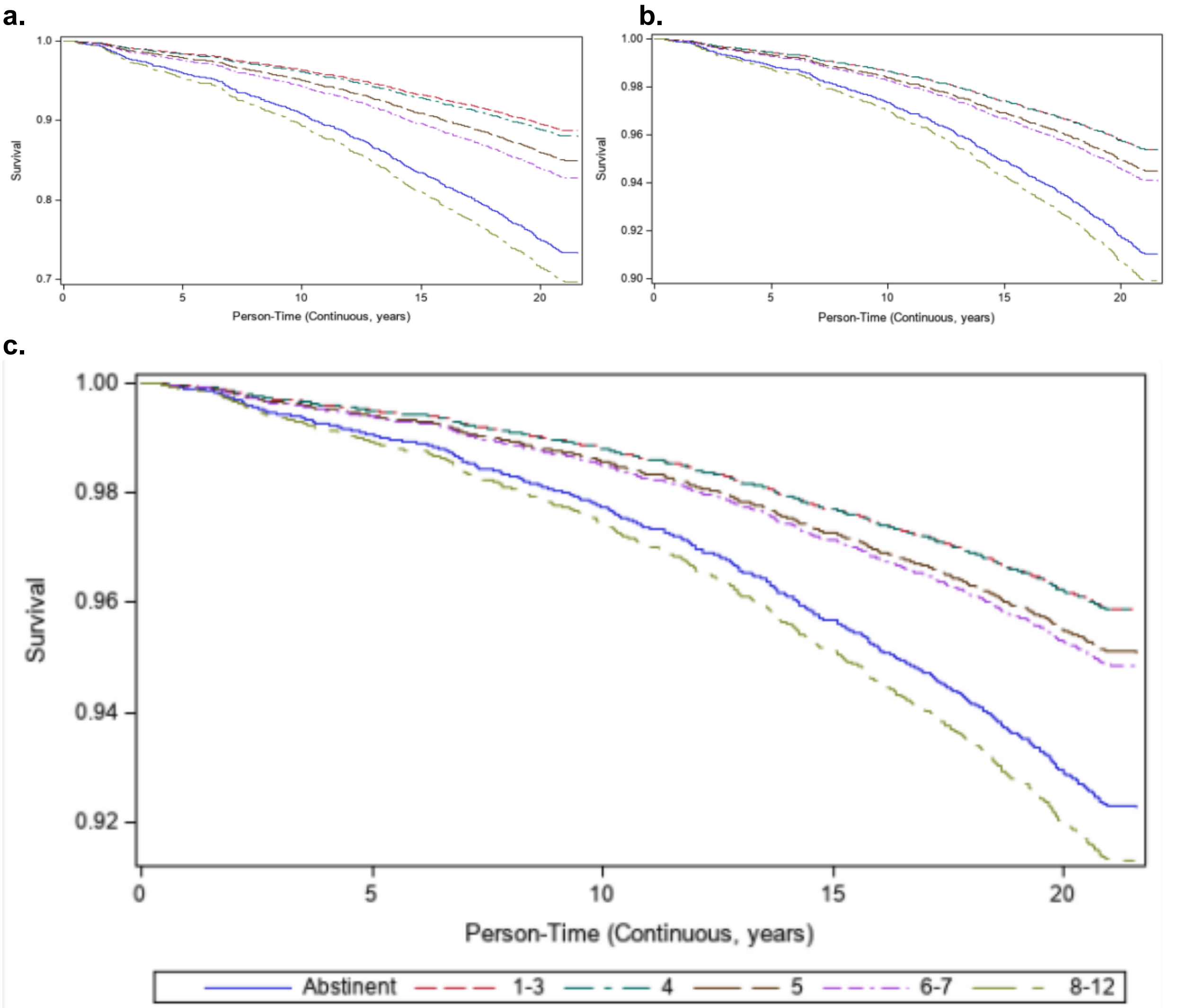
Table 2. Hazard of death 20 years after baseline by baseline alcohol consumption.

Alcohol Use At Baseline	**N**	**Deceased n(%)**	**Model 1 Unadjusted HR (95% CI)**	**Model 2[a] HR (95% CI)**	**Model 3[b] HR (95% CI)**
Abstinent (AUDIT-C=0)	447	119 (26.6)	1.00 [Reference]	1.00 [Reference]	1.00 [Reference]
Low to Moderate (AUDIT-C=1-3)	2203	248 (11.3)	0.38 (0.31-0.48)	0.50 (0.40-0.63)	0.52 (0.42-0.65)
Moderate to High (AUDIT-C =4)	674	81 (12.0)	0.41 (0.31-0.54)	0.50 (0.38-0.67)	0.52 (0.39-0.70)
High (AUDIT-C=5)	383	58 (15.1)	0.52 (0.38-0.72)	0.60 (0.43-0.83)	0.63 (0.45-0.87)
Very High (AUDIT-C=6-7)	228	39 (17.1)	0.61 (0.42-0.87)	0.65 (0.45-0.94)	0.66 (0.45-0.95)
Extremely High (AUDIT-C =8-12)	93	28 (30.1)	1.16 (0.77-1.76)	1.13 (0.74-1.73)	1.13 (0.73-1.73)

[a] Model adjusted for sex, age, smoking status, and years of education at baseline

[b] Model adjusted for sex, age, smoking status, years of education, and self-reported health at baseline

Figure 2. Survival over 20 years of follow-up after baseline by baseline alcohol consumption.



a. Unadjusted survival curves of study participants by baseline alcohol consumption over 20 years follow-up. Cox Proportional Hazards model 1. b. Survival curves of study participants by baseline alcohol consumption, adjusted for age, sex, education, and smoking history. Cox Proportional Hazards model 2. **c.** Survival curves of study participants by baseline alcohol consumption, adjusted for age, sex, education, smoking history, and self-rated health at baseline. Cox Proportional Hazards model 3.

d. Conclusion

Based on these study findings alone, it appears that low to high alcohol consumption over the prior 12 months, appears to reduce the risk of mortality over 20 years of follow up. These study findings are consistent with prior studies that have observed reduced risk of mortality among alcohol consumers over alcohol abstainers.

However, this study, as others, suffers from lack of measurement on potential confounding factors, including factors determining mid-life abstinence from alcohol such as former heavy drinking or alcohol use disorders. Assessment of mid-life abstinence from alcohol alone does not sufficiently characterize prior alcohol-related exposures.

We have identified important directions for future research. Future studies that are able to measure alcohol consumption over the lifecourse, as well as other confounders over the lifecourse, would be better able to disentangle potentially cumulative effects and/or effects that more clearly explain how alcohol consumption may or may not contribute to mortality risk over time.

e. Supplementary materials

Code, raw data and analytic dataset are available in the Github repository, linked below. The original study and raw dataset can be found here: John et al., Alcohol abstinence and mortality in a general population sample of adults in Germany: A cohort study: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003819

f. Dataset, data dictionary and codebook

All project materials are available in the Github repository, linked below.

g. Provide a link to github repository

https://github.com/rfisher2022/EPIDEM207-2022-winter

Figure 1: Survival of study participants after 20 years by baseline alcohol consumption

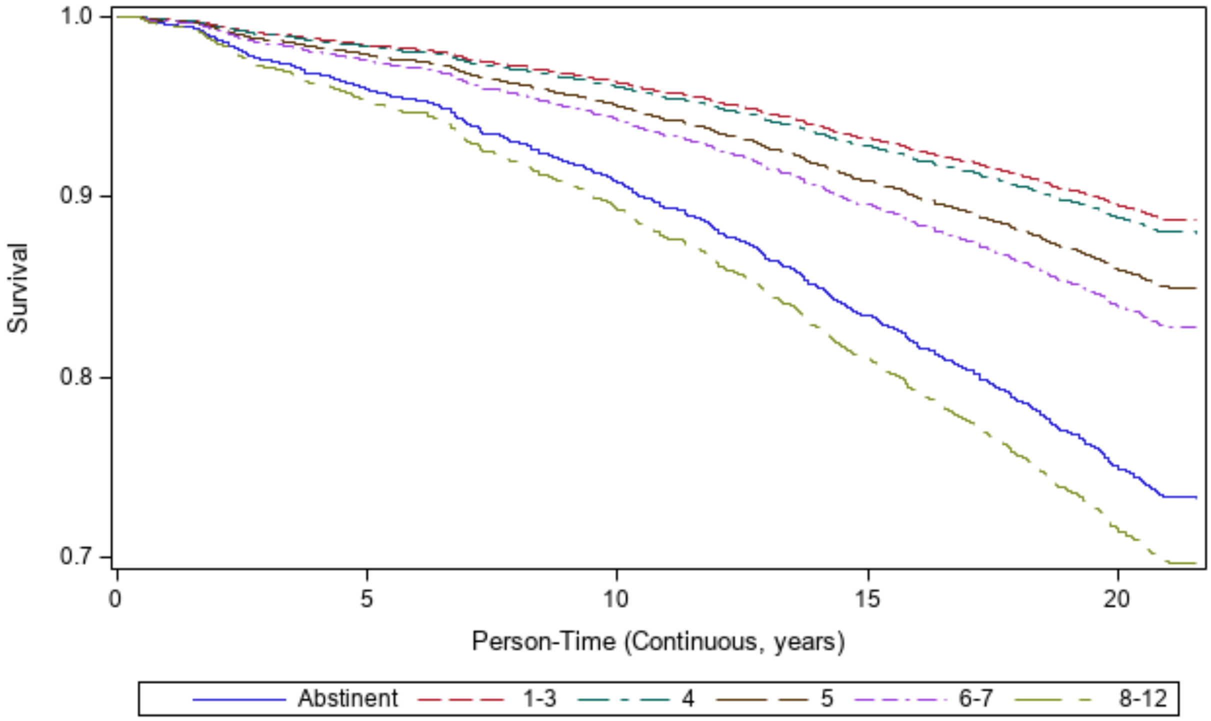
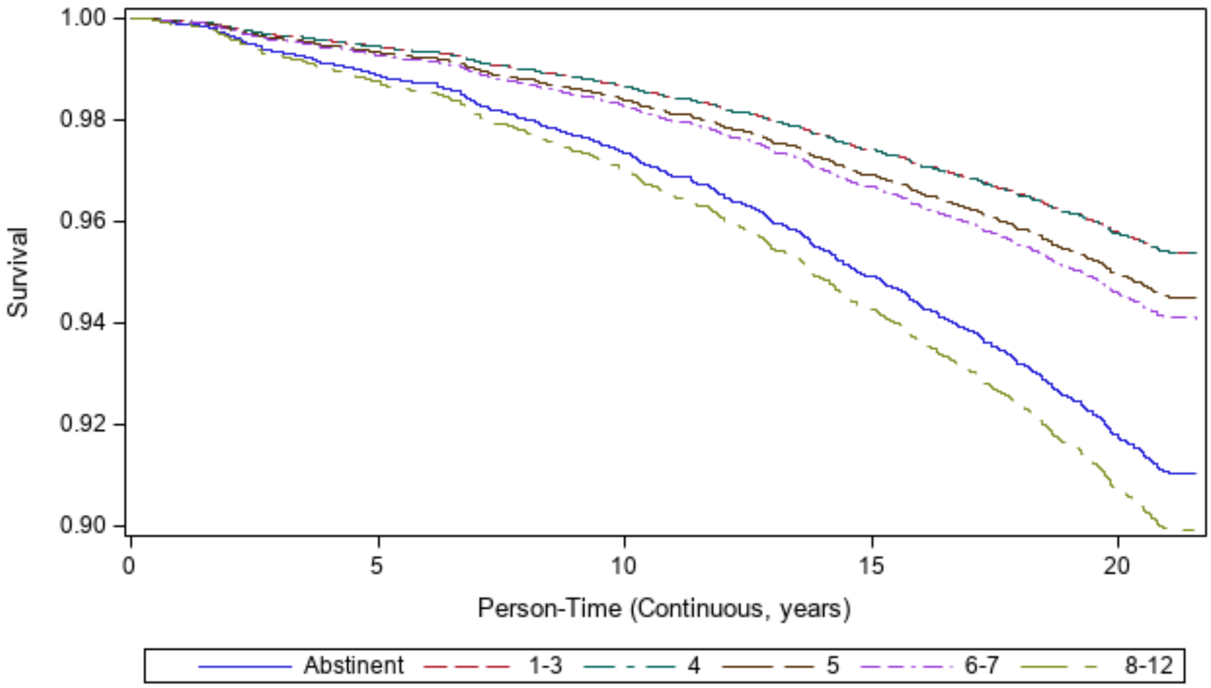


Figure 2: Survival of sutdy participants after 20 years by baseline alcohol consumption, adjusted for age, sex, education level, and smoking



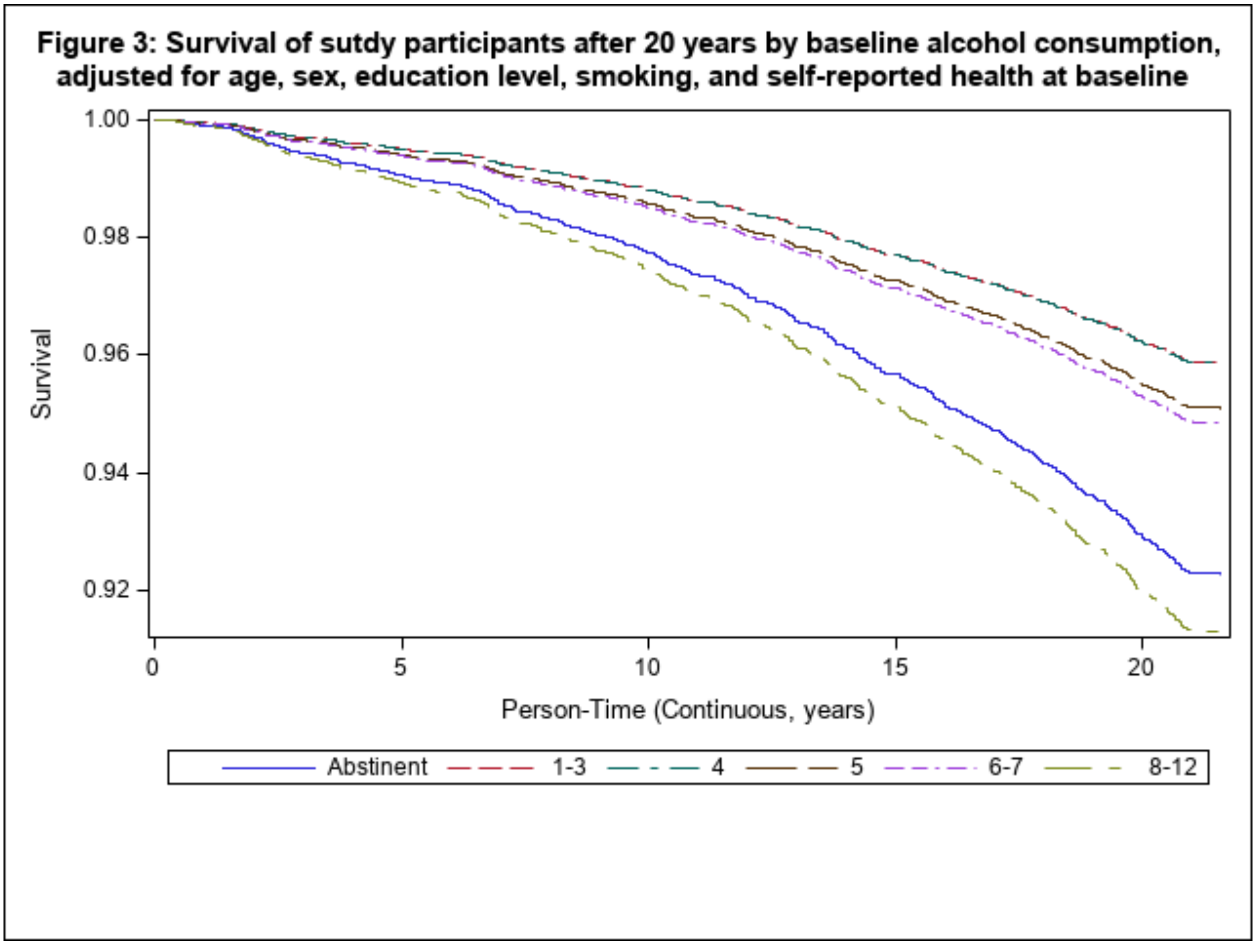


Figure 3: Survival of sutdy participants after 20 years by baseline alcohol consumption, adjusted for age, sex, education level, smoking, and self-reported health at baseline

CODE:

```
/*******************************************************
      filename: Epi207_Assignment3
      author(s): Catherine Cortez, Rebecca Fisher, John Waggoner
      created on: 2/13/2022
      purpose:
*******************************************************/

/**********Data read-in *********************/
*author: John Waggoner;
*date: 2/13/2022;
title 'Epi207_Assignment3';

/*    SET LIBNAMES, FILENAMES AND LOCATION MACRO VARIABLES:
      CHANGE FILE NAME TO FOLDER LOCAL FILE IS SAVED
      the download link to the actual data on PLOS is broken so
      until that is fixed, will have to use local files */

libname asgn3 'C:\Users\...';
filename johnetal 'C:\Users\...\johnetal.xls';
```

```
*also set up location of table 1 macro;
%let MacroDir = C:\Users\...\Table1;

*specify folder in which to store results - change path;

%let results=C:\Users\...\results;




*Import raw data from xls;

proc import out=asgn3.johnetalunclean datafile= johnetal
            dbms = xls replace;
            getnames=no; /*no column headers in original dataset*/
run;

/*********** Proc Contents ********************/

proc contents data=asgn3.johnetalunclean;
run;

/************ Data Cleaning *******************/
options fmtsearch=(asgn3);


/*   Formats for all variables of interest. I figured since
     we might want to use other variables that I'd keep this
     from Assignment 1, even though there might be extra var-
     -iables. We don't have to keep it like this though, and
     would be grateful for suggestions on better format labels
     or values - JOHN.   */

proc format library = asgn3;
    value    sexbin     0     =     "Female"
                              1     =     "Male";

    value    educat     1     =     "<=9 years"
                              2     =     "10-11 years"
                              3     =     ">=12 years";

    value    yesno      0     =     "No"
                              1     =     "Yes";

    value    hltcat     1     =     "Poor/Fair"
                              2     =     "Good"
                              3     =     "Excellent/Very Good";

    value    codcat     1     =     "Cardiovascular Disease"
                              2     =     "Cancer"
```

```
                                3       =       "Aerodigestive (?)"
                                4       =       "Acute (?)"
                                5       =       "Psychiatric (?)"
                                6       =       "Other or Unknown"
                                7       =       "No Cause Listed (?)";

    value   auditc      0       =       "Abstinent"
                                1       =       "1-3"
                                2       =       "4"
                                3       =       "5"
                                4       =       "6-7"
                                5       =       "8-12";

    value   smkcat                  0       =       "Never Smoker"
                                1       =       "Ever Smoker Less than daily"
                                2       =       "Former Smoker daily"
                                3       =       "Current Smoker <=19 cig/day"
                                4       =       "Current Smoker >19 cig/day";

    value   agecat      1       =       "17-39 years old"
                                2       =       "40-49 years old"
                                3       =       "50-64 years old";

    value   rskgrptwo   1       =       "Low to Moderate Alcohol (AUDIT-C 1-3)"
                                2       =       "Moderate to high Alcohol (AUDIT-C 4)"
                                3       =       "High Alcohol (AUDIT-C 5)"
                                4       =       "Very High Alcohol (AUDIT-C 6-7)"
                                5       =       "Extremely High Alcohol (AUDIT-C 8-12)"
                                6       =       "Abstinent At Baseline, Good Health"
                                7       =       "Abstinent At Baseline, History of
Alcohol/Drug Dependence"
                                8       =       "Abstinent At Baseline, History of Risk
Drinking"
                                9       =       "Abstinent At Baseline, Tried to Cut
Down or Stop"
                                10      =       "Abstinent At Baseline, Current Daily
Smoker (>19 cig/day)"
                                11      =       "Abstinent At Baseline, Current Daily
Smoker (<=19 cig/day)"
                                12      =       "Abstinent At Baseline, Former Smoker"
                                13      =       "Abstinent At Baseline, Poor Health";

    value   rskgrpthr   1       =       "Low to Moderate Alcohol, Never Smoker"
                                2       =       "Low to Moderate Alcohol, Former Daily
Smoker"
                                3       =       "Low to Moderate Alcohol, Current
Smoker (<=19 cig/day)"
                                4       =       "Low to Moderate Alcohol, Current
SMoker (?19 cig/day)"
                                5       =       "Moderate to High Alcohol, Never
Smoker"
```

```
                    6    =    "Moderate to High Alcohol, Former Daily
Smoker"
                    7    =    "Moderate to High Alcohol, Current
Smoker (<=19 cig/day)"
                    8    =    "Moderate to High Alcohol, Current
Smoker (>19 cig/day)"
                    9    =    "Very High Alcohol, Not Current Smoker"
                    10   =    "Very High Alcohol, Current Smoker
(<=19 cig/day)"
                    11   =    "Very High Alcohol, Current Smoker (>19
cig/day)"
                    12   =    "Extremely High Alcohol, Not Current
Smoker"
                    13   =    "Extremely High Alcohol, Current
Smoker"
                    14   =    "Absitnent at Baseline, Good Health"
                    15   =    "Absitnent at Baseline, History of
Alcohol/Drug Dependence"
                    16   =    "Abstinent At Baseline, History of Risk
Drinking"
                    17   =    "Abstinent At Baseline, Tried to Cut
Down or Stop"
                    18   =    "Abstinent At Baseline, Current Daily
Smoker (>19 cig/day)"
                    19   =    "Abstinent At Baseline, Current Daily
Smoker (<=19 cig/day)"
                    20   =    "Abstinent At Baseline, Former Smoker"
                    21   =    "Abstinent At Baseline, Poor Health";

  value    rskgrpfour   0   =   "Moderate to Heavy Drinking"
                    1    =    "Low to Moderate Alcohol, Good Health,
Never Smoker"
                    2    =    "No Alcohol, Good Health, Never
Smoker"
                    3    =    "Low to Moderate Alcohol, Poor Health,
Never Smoker"
                    4    =    "No Alcohol, Poor Health, Never
Smoker"
                    5    =    "Low to Moderate Alcohol, Ever Smoker
Less than daily"
                    6    =    "No Alcohol, Ever Smoker Less than
daily"
                    7    =    "Low to Moderate Alcohol, Former
Smoker daily"
                    8    =    "No Alcohol, Former Smoker daily"
                    9    =    "Low to Moderate Alcohol, Current
Smoker <=19 cig/day"
                    10   =    "No Alcohol, Current Smoker <=19
cig/day"
```

```
                                        11      =       "Low to Moderate Alcohol, Current
Smoker >19 cig/day"
                                        12      =       "No Alcohol, Current Smoker >19
cig/day"
                                        13      =       "Low to Moderate Alcohol, Previous
Substance Abuse"
                                        14      =       "No Alcohol, Previous Substance
Abuse"
                                        15      =       "Low to Moderate Alcohol, Previous Risk
Drinking"
                                        16      =       "No Alcohol, Previous Risk Drinking"
                                        17      =       "Low to Moderate Alcohol, Tried to
Decrease Alcohol"
                                        18      =       "No Alcohol, Tried to Decrease Alcohol"
                                        19      =       "Low to Moderate Alcohol, AUDIT-C600 >
18 (?)"
                                        20      =       "No Alcohol, AUDIT-C600 > 18 (?)";

data asgn3.johnclean;
    set              asgn3.johnetalunclean;

    /*  Made arrays for faster conversion of categorical to numeric values
        The lettered variables in 'raw[*]'are the original variables from
        the original dataset. The variables in 'cln[*]' are their trans-
        -formed counterparts. If they contain a '0' at the end, they are
        not the final form of the variable and will undergo further trans-
        -formation within the data step   */

    array     raw[*]  B E M N O P Q R;
    array     cln[*]  life_abst_bin_0
                            age_cat_0
                            cod_cat
                            auditc_cat_0
                            rsk_grp_tab2
                            rsk_grp_tab3
                            rsk_grp_tab4_0
                            smk_cat;

    /*  This do-loop takes the array input from 'raw[*]' in the substr() fx.
        and strips all but the first two characters from the variable. For
        the variables in the array, the first two characters are either two
        numbers or a number followed by a period. The compress() fx. takes
        the output of substr() and removes the period character. The input()
        fx. takes the string output of compress() and reformats as numerical
        with the 8. format. Missing values coded as '.' in new variables.  */

    do z=1 to 8;
        cln[z] = input(compress(substr(raw[z], 1, 2), "."), 8.);
        if missing(raw[z]) then cln[z]=.;
    end;
```

```
/*  Numerical age, Subject ID, person time, and death status do not need
    transformation. They are renamed in the final data set. */

age_num=A;
deceased_bin=I;
id_num=S;
prsn_time=J;

/*  Person time (years) */
prsn_time_yrs=prsn_time/365;

/*  lifelong abstinence was coded in the original dataset as either
    '1. Ja 1.' or '5. Nein 5.' This codes the new variable as either
    1  or 0, where 1 means yes and 0 means no (see 'yesno.' format) */

life_abst_bin=.;
if (life_abst_bin_0=1) then life_abst_bin=1;
if (life_abst_bin_0=5) then life_abst_bin=0;

/* Recodes sex from character to numeric, see 'sexbin. format. */
sex_bin=1;
if find(C, "female")>0 then sex_bin=0;

/*  For the edu_cat, age_cat, and hlt_cat, find() fx. to identifies
    character strings for recoding into numeric values in new var.
    For hlt_cat, there was not a unique character identifier for the
    middle category 'good'. Because of this, it is coded first, so
    that the following line will only identify those that are 'very
    good' and overwrite their value from 2 to 3. */

edu_cat=.;
if find(D, "less")>0 then edu_cat=1;
if find(D, "10-11")>0 then edu_cat=2;
if find(D, "more")>0 then edu_cat=3;

age_cat=.;
if (age_cat_0=39) then age_cat=1;
if (age_cat_0=40) then age_cat=2;
if (age_cat_0=50) then age_cat=3;

hlt_cat=.;
if find(L, "poor") then hlt_cat=1;
if find(L, "good") then hlt_cat=2;
if find(L, "very") then hlt_cat=3;

/*  Fixes weird coding of 'abstient' AUDIT-C category. Was originally '5'
    and is now coded as '0', while extreme drinking is now coded as '5'
    in new variable */

auditc_cat=auditc_cat_0;
if (auditc_cat_0=5) then auditc_cat=0;
if (auditc_cat_0=6) then auditc_cat=5;

/*  Recodes missing values from the Risk Groups in Table 4 to indicate
    that they are those with moderate to heavy drinking. */

rsk_grp_tab4=rsk_grp_tab4_0;
if missing(rsk_grp_tab4_0) then rsk_grp_tab4=0;

label     age_num                      =      "Age (Continuous)"
          life_abst_bin =      "Lifelong Abstainer (Y/N)"
          sex_bin                      =      "Sex"
          edu_cat                      =      "Educational Attainment (years)"
          age_cat                      =      "Age (Categorical)"
          deceased_bin         =      "Deceased (Y/N)"
          cod_cat                      =      "Cause of Death"
          prsn_time            =      "Person-Time (Continuous, days)"
          prsn_time_yrs        =      "Person-Time (Continuous, years)"
          hlt_cat              =      "Self-Reported Health"
          auditc_cat           =      "AUDIT-C Score Sum (Categorical)"
          smk_cat                      =      "Smoking History"
          rsk_grp_tab2 =       "Risk Group II"
          rsk_grp_tab3 =       "Risk Group III"
          rsk_grp_tab4 =       "Risk Group IV"
          id_num                       =      "Subject ID";

format    life_abst_bin
          deceased_bin                 yesno.
          age_num
          prsn_time
          prsn_time_yrs
          id_num                              8.
          sex_bin                             sexbin.
          edu_cat                             educat.
          age_cat                             agecat.
          cod_cat                             codcat.
          hlt_cat                      hltcat.
          auditc_cat                   auditc.
          smk_cat                             smkcat.
          rsk_grp_tab2         rskgrptwo.
          rsk_grp_tab3         rskgrpthr.
          rsk_grp_tab4         rskgrpfour.;

/*  Drops all original variables from dataset - all variables of
    interest are renamed. Also drops all intermediaries, variables
    that end in '0' */


drop      A B C D E F G J H I J K L M N O P Q R S
          life_abst_bin_0
          age_cat_0 auditc_cat_0
          rsk_grp_tab4_0
```

```
                    z;

run;

proc contents data=asgn3.johnclean;
run;

/*check distribution of variables of interest to confirm if coded correctly */

proc freq data=asgn3.johnclean;
      tables auditc_cat /missing;
      tables sex_bin / missing;
      tables smk_cat / missing;
      tables deceased_bin / missing;
      tables edu_cat / missing;
      tables hlt_cat / missing;
      run;

/*    Compared with table 1 in John et al., frequencies of new dataset match
      those from the article. */

/******************** END OF CLEANING *********************/



/******************* Table 1 ****************************/
      *author: Catherine;
      *date: 2 15 2022;

options nodate nocenter ls = 147 ps = 47 orientation = landscape;

      /*  Create macro to file path with Table 1 SAS programs - change path to
          where Table 1 macro files are saved locally        */


filename tab1  "&MacroDir./Table1.sas";
%include tab1;

/************************/
/****UTILITY SASJOBS****/
/************************/
filename tab1prt  "&MacroDir./Table1Print.sas";
%include tab1prt;

filename npar1way  "&MacroDir./Npar1way.sas";
%include npar1way;

filename CheckVar  "&MacroDir./CheckVar.sas";
%include CheckVar;

filename Uni  "&MacroDir./Univariate.sas";
```

```
%include Uni;

filename Varlist  "&MacroDir./Varlist.sas";
%include Varlist;

filename Words  "&MacroDir./Words.sas";
%include Words;

filename Append  "&MacroDir./Append.sas";
%include Append;


    /*  macro call */

%Table1(DSName=asgn3.johnclean,
     Total=C,
       NumVars= prsn_time_yrs age_num,
     FreqVars= sex_bin age_cat edu_cat smk_cat hlt_cat auditc_cat,
     P=N,
     FreqCell=N(CP),
     Missing=Y,
     Print=N,
       Dec=2,
     Label=L,
     Out=test,
     Out1way=)
*options mprint  symbolgen mlogic;
run;

    /*  Generate excel file with output      */

ods excel file="&results.\John et al_Table1_output.xlsx";
%Table1Print(DSname=test,Space=Y)
ods excel close;
run;

    /*  Calculate values for person-year column */

proc tabulate data=asgn3.johnclean;
    class sex_bin edu_cat smk_cat hlt_cat auditc_cat;
    var prsn_time_yrs;
    table sex_bin edu_cat smk_cat hlt_cat auditc_cat, (mean std)*prsn_time_yrs;
run;

/******************** End of Table 1 code ******************************/




/******************** Table 2 ******************************/
    *author: Rebecca;
```

```
	*date: 2 20 2022;

title "Frequency of AUDIT-C levels and Death";
proc freq data=asgn3.johnclean;
	table auditc_cat*deceased_bin/ nocol nopercent;
run;

*Unadjusted HR;
title "Unadjusted Hazards of AUDIT-C levels and Death";

proc phreg data=asgn3.johnclean;
class auditc_cat(ref="Abstinent")/param=ref order=internal;
model prsn_time_yrs*deceased_bin(0)=auditc_cat/rl;
run;

*Adjusted HR for sex, age, smoking status and years of education at baseline;
title "Adjusted Hazards of AUDIT-C levels and Death, Model 1";

proc phreg data=asgn3.johnclean;
class auditc_cat(ref="Abstinent") sex_bin(ref="Female") smk_cat(ref="Never Smoker")
edu_cat(ref=">=12 years")/param=ref order=internal;
model prsn_time_yrs*deceased_bin(0)=auditc_cat age_num sex_bin smk_cat edu_cat/rl;
run;

*Adjusted HR for sex, age, smoking status, years of education and self-rated health at
baseline;
title "Adjusted Hazards of AUDIT-C levels and Death, Model 2";

proc phreg data=asgn3.johnclean;
class auditc_cat(ref="Abstinent") sex_bin(ref="Female") smk_cat(ref="Never Smoker")
edu_cat(ref=">=12 years") hlt_cat(ref="Excellent/Very Good")/param=ref order=internal;
model prsn_time_yrs*deceased_bin(0)=auditc_cat age_num sex_bin smk_cat edu_cat
hlt_cat/rl;
run;

/******************* End of Table 2 code ****************************/




/******************* Survival Graph Figures *************************/
	*author: John Waggoner;
	*date: 2/20/2022;
/******************* Survival Plot Unadjusted ***********************/

ods graphics on;

/*Create covariate values for graphing*/
data fig1_unadjusted;
	format	auditc_cat	auditc.;
```

```
    input       auditc_cat;
    datalines;
    0
    1
    2
    3
    4
    5
    ;
run;

/*Figure 1 Code*/
ods output survivalplot=_surv;
Title "Figure 1: Survival of study participants after 20 years by baseline alcohol
consumption";

proc phreg      data=asgn3.johnclean plots(overlay)=(survival);
    class               auditc_cat(ref="Abstinent")/param=ref order=internal;
    model               prsn_time_yrs*deceased_bin(0)=auditc_cat/rl;
    baseline    covariates=fig1_unadjusted/rowid=auditc_cat;
    where               auditc_cat=0|auditc_cat=1|auditc_cat=2|auditc_cat=3|
                        auditc_cat=4|auditc_cat=5;
run;

proc sgplot data=_surv;
    step x=time y=survival/group=auditc_cat;
    keylegend/title=" ";
run;


/****************** Plot for:Adjusted Model 1 ********************/

/*model for baseline values*/


proc phreg data=asgn3.johnclean plots=survival;
    class       sex_bin(ref='Female') smk_cat(ref='Never Smoker') edu_cat(ref='>=12
years')/param=ref;
    model       prsn_time_yrs*deceased_bin(0)=sex_bin smk_cat edu_cat age_num;
run;

/*Covariate Baseline Dataset for Graphing*/
data fig2_adjusted;
    format      auditc_cat      auditc.
                sex_bin                 sexbin.
                smk_cat                 smkcat.
                edu_cat                 educat.;
    input       auditc_cat
                sex_bin
                smk_cat
                edu_cat
                age_num;
      datalines;
      0 0 0 3 42
      1 0 0 3 42
      2 0 0 3 42
      3 0 0 3 42
      4 0 0 3 42
      5 0 0 3 42
      ;
run;

/*figure 2 Code*/
ods output survivalplot=_surv;
Title "Figure 2: Survival of sutdy participants after 20 years by baseline alcohol
consumption, adjusted for age, sex, education level, and smoking";

proc phreg     data=asgn3.johnclean plots(overlay)=(survival);
      class           auditc_cat(ref="Abstinent") sex_bin(ref="Female")
smk_cat(ref="Never Smoker") edu_cat(ref=">=12 years")/param=ref order=internal;
      model           prsn_time_yrs*deceased_bin(0)=auditc_cat age_num sex_bin
smk_cat edu_cat/rl;
      baseline  covariates=fig2_adjusted/rowid=auditc_cat;
      where           auditc_cat=0|auditc_cat=1|auditc_cat=2|auditc_cat=3|
                      auditc_cat=4|auditc_cat=5;
run;

proc sgplot data=_surv;
      step x=time y=survival/group=auditc_cat;
      keylegend/title=" ";
run;

/********************** Plot for Adjsuted Model 2 ****************/
/* model for baseline covariate values*/
proc phreg data=asgn3.johnclean plots=survival;
                class sex_bin(ref='Female') smk_cat(ref='Never Smoker') edu_cat(ref='>=12
years') hlt_cat(ref='Excellent/Very Good')/param=ref order=internal;
                model prsn_time_yrs*deceased_bin(0)= sex_bin age_num smk_cat hlt_cat
edu_cat;
run;

/*Covariate baseline dataset for graphing*/

data fig3_adjusted;
      format    auditc_cat     auditc.
                sex_bin                 sexbin.
                smk_cat                 smkcat.
                edu_cat                 educat.
                hlt_cat        hltcat.;
      input     auditc_cat
                sex_bin
                smk_cat
                edu_cat
                age_num
                hlt_cat;
     datalines;
     0 0 0 3 42 3
     1 0 0 3 42 3
     2 0 0 3 42 3
     3 0 0 3 42 3
     4 0 0 3 42 3
     5 0 0 3 42 3
     ;

/*figure 3 code*/
ods output survivalplot=_surv;
Title "Figure 3: Survival of sutdy participants after 20 years by baseline alcohol
consumption, adjusted for age, sex, education level, smoking, and self-reported health at
baseline";

proc phreg      data=asgn3.johnclean plots(overlay)=(survival);
     class              auditc_cat(ref="Abstinent") sex_bin(ref="Female")
smk_cat(ref="Never Smoker") edu_cat(ref=">=12 years") hlt_cat(ref='Excellent/Very
Good')/param=ref order=internal;
     model              prsn_time_yrs*deceased_bin(0)=auditc_cat age_num sex_bin
smk_cat edu_cat hlt_cat/rl;
     baseline   covariates=fig3_adjusted/rowid=auditc_cat;
     where              auditc_cat=0|auditc_cat=1|auditc_cat=2|auditc_cat=3|
                        auditc_cat=4|auditc_cat=5;
run;

proc sgplot data=_surv;
     step x=time y=survival/group=auditc_cat;
     keylegend/title=" ";
run;

ods graphics off;

/****** END FIGURE & GRAPH CODE ****************/
```