

Using Data Visualization to Explore the Association between Education and Mortality: Insights from the 2012-2018 National Health Interview Survey Linked Mortality Files

Cindy Zhang, Jessie Parker, Yeats Ye, and Cordell Golden

Data Linkage Methodology and Analysis Branch
Division of Analysis & Epidemiology
National Center of Health Statistics

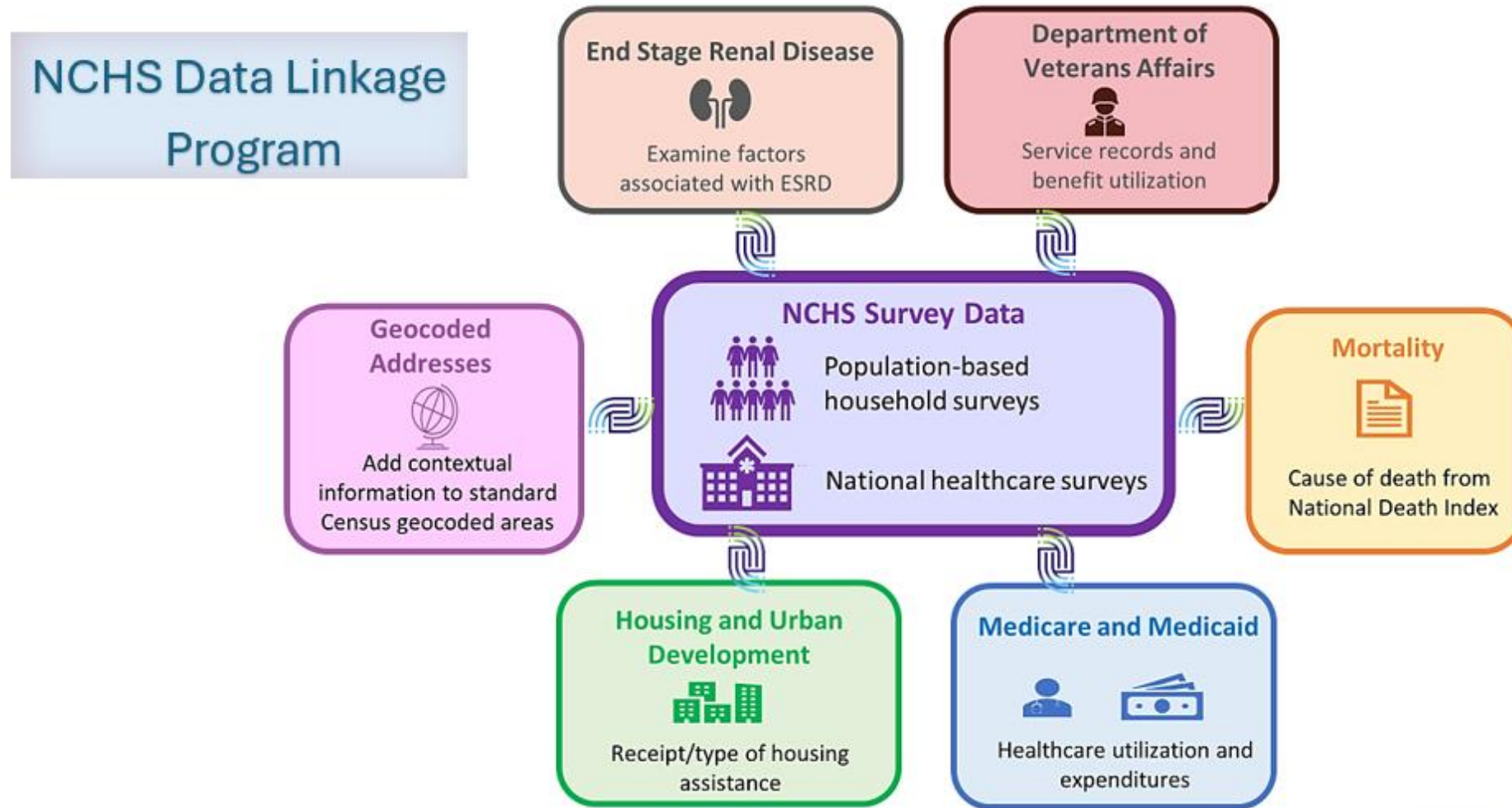
➤ Introduction – The National Center for Health Statistics (NCHS)

- The National Center for Health Statistics (NCHS) is the nation's principal health statistics agency.
- One of 13 principal federal statistical agencies.
- Mission: To provide timely, relevant, and accurate health data and statistics that inform and guide programs and policies to improve our nation's health.

➤ Introduction – NCHS Data Linkage Program

- The NCHS Data Linkage Program links survey data files with vital records and other administrative data.
- Linked surveys include:
 - ❖ National Health Interview Survey (NHIS)
 - ❖ National Health and Nutrition Examination Survey (NHANES)
 - ❖ National Health Care Surveys (NHCS)
 - ❖ National Ambulatory Medical Care Survey (NAMCS)

➤ Introduction – Linked Files by Data Linkage Program



➤ Introduction – Research Using Linked Data

Examples of epidemiologic research with linked data

*Over 1,200
publications based
on NCHS linked data*

Linked Mortality Data

- Deaths Associated with Underweight, Overweight, and Obesity
- Air Pollution Exposure and Heart Disease Mortality
- Educational Differentials in US Adult Mortality



Linked NCHS-CMS Data

- Characteristics of those who chose Medicare Advantage upon Medicare enrollment
- Health service use among the previously uninsured
- Concordance between survey reported childhood asthma and linked Medicaid



Linked NCHS-HUD Data

- Housing assistance and blood lead levels
- Cigarette smoking and adverse health outcomes among adults receiving federal housing assistance
- Housing assistance associated with insurance rates and unmet medical need

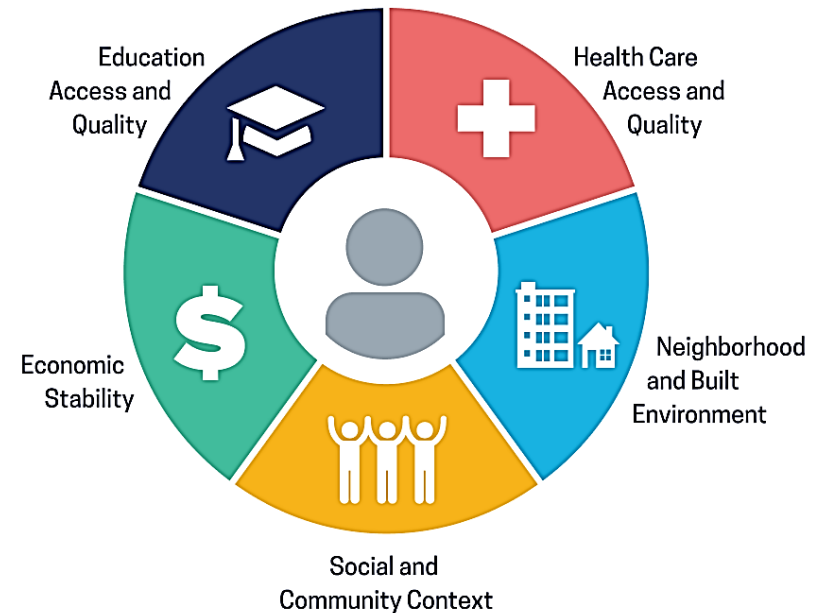


➤ Background – Examining the Association between Education Attainment and Mortality

Background:

- Social Determinants of Health (SDOH) such as access to education are associated with general health and mortality.
- A 2024 meta-analysis (*IHME; Lancet Public Health*) published a dose-response relationship between education and all-cause adult mortality, with an average reduction in mortality risk of 1.9% per additional year of education.

Social Determinants of Health



➤ Motivation – Examining the Association between Education Attainment and Mortality

Motivation:

- This work will examine the relationship between education and mortality using pooled data from the public-use 2012-2018 National Health Interview Survey (NHIS) linked to mortality information from the National Death Index (NDI).
- SAS-VIYA will be used in CDC's Enterprise Data, Analytics, and Visualization (EDAV) environment for data exploration, visualization and analysis, with steps including:
 - i. Checking variable distributions and conducting exploratory data analysis
 - ii. Calculating mortality rates by education attainment levels and age groups
 - iii. Visualizing distributions for leading causes of death by education attainment levels and age groups
 - iv. Visualizing the association results from Cox proportional hazards models

➤ Data Source – Public-use NHIS Linked Mortality Files

The image displays two screenshots of the CDC National Center for Health Statistics website. The left screenshot shows the 'National Health Interview Survey' page for 1997-2018, with a sidebar on the left and a main content area. The right screenshot shows the 'NCHS Data Linkage' page for '2019 Public-Use Linked Mortality Files', with a sidebar on the left and a main content area.

Left Screenshot: National Health Interview Survey

- Search bar:
- Section: National Health Interview Survey
- Sub-section: 1997-2018
- Print button
- Section: 2018 NHIS
- On This Page
- 2018 NHIS
- 2017 NHIS
- 2016 NHIS
- 2015 NHIS
- 2014 NHIS
- 2013 NHIS
- 2012 NHIS
- 2011 NHIS
- 2010 NHIS
- 2009 NHIS

Right Screenshot: NCHS Data Linkage

- Section: NCHS Data Linkage
- Section: 2019 Public-Use Linked Mortality Files
- Print button
- Text: Public-use Linked Mortality Files (LMF) are available for 1986-2018 NHIS, 1999-2018 NHANES, and NHANES III. The files include a limited set of mortality variables for adult participants only. The public-use versions of the NCHS Linked Mortality Files were subjected to data perturbation techniques to reduce the risk of participant re-identification. For select records, synthetic data were substituted for follow-up time or underlying cause of death. Information regarding vital status was not perturbed.
- Text: For information about accessing the restricted-use linked mortality files, please consult the [Restricted-Use Linked Mortality Data](#) page.
- Section: Data Files and Data Dictionaries
- File Description: [Public-Use 2019 LMF](#) [PDF - 180 KB]
- Public-Use Data Dictionary: [Public-Use 2019 LMF](#) [PDF - 109 KB]
- Public-Use Linked Mortality Files: [FTP Site](#)
- Software Programs:
 - [Sample SAS program](#) [SAS - 12 KB]
 - [Sample Stata program](#) [DO - 6 KB]
 - [Sample R program](#)

**Study
Sample**

Aged 25 years and
older at the
interview

Non-missing
demographic
variables including
Education

Eligible for linkage
(sufficient PII)

Sample weights
adjusted for linkage
eligibility

➤ Data Source – Public-use NHIS Linked Mortality Files

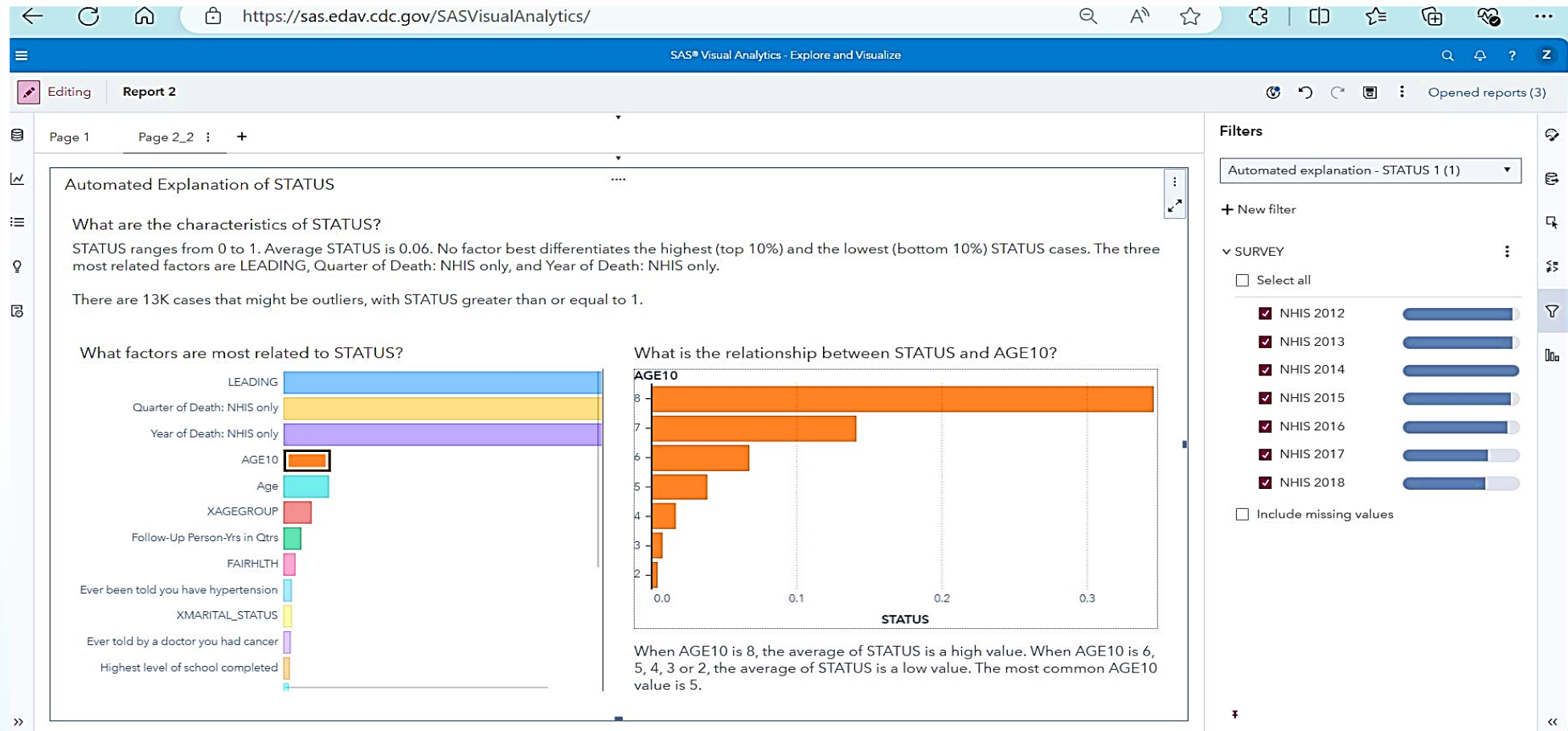
Base Variables	Means and Ranges						
	N	Mean	SUM	Minimum	25th Pctl	75th Pctl	Maximum
AGE_P	200,760	52.7	10,584,219	25.0	38.0	66.0	85.0
SAWT_1218	200,760	1,057.9	212,386,922	6.7	448.4	1,416.9	30,523.7
Person-Years, Follow-up	200,760	4.4	885,842	0.0	2.8	6.0	7.8

Demographic Variables		Frequencies and Percents, Unweighted	
Variables	Categories	Frequency (n)	Percent (%)
Gender	Female	111,130	55.35
	Male	89,630	44.65
Age Group	25-39 Years	54,660	27.23
	40-64 Years	91,301	45.48
	65+ Years	54,799	27.30
Race-Ethnicity	Hispanics	28,840	14.37
	Non-Hispanic White	132,012	65.76
	Non-Hispanic Black	26,484	13.19
	Non-Hispanic Others	13,424	6.69
Education Attainment	Less than high school	27,990	13.94
	High school diploma or GED	50,075	24.94
	Associate degree or some college	58,675	29.23
	Bachelor's degree or higher	63,184	31.47
	Unknown	836	0.42

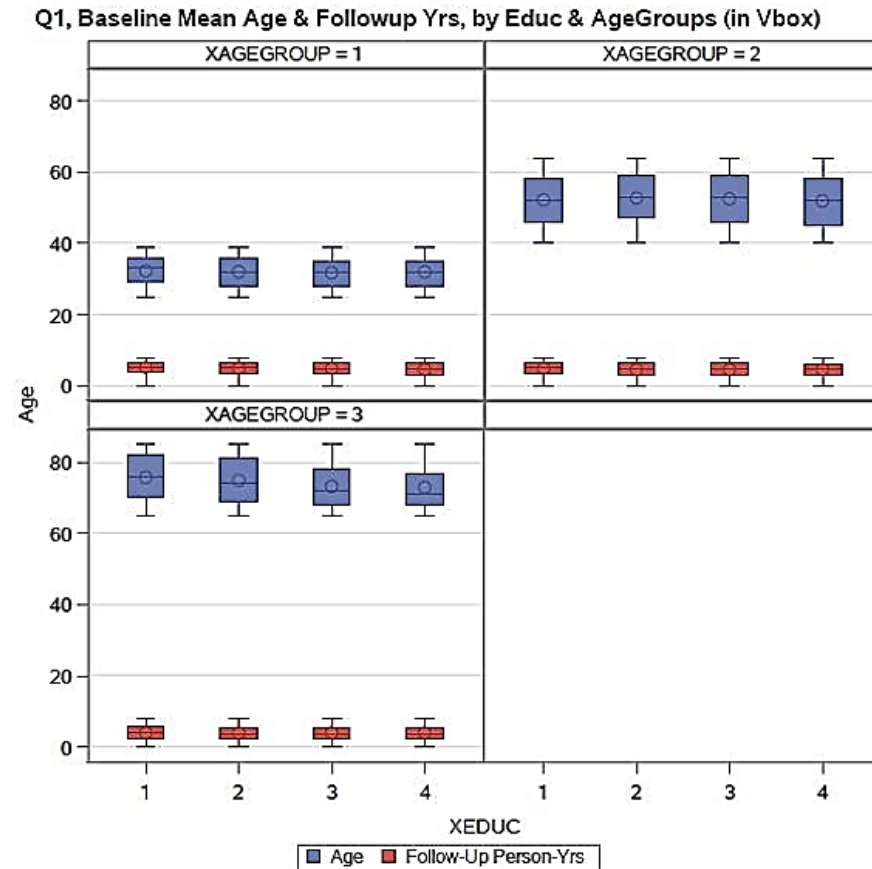
```
proc means
  data = cdemo_puf_x;
  n mean sum min p25 p75
  max maxdec=1;
  var age_p sawt_1218
  pery;
  where eligstat=1;
  title "Base Data Vars
        Check (Numeric)";
run;
```

```
proc freq
  data= cdemo_puf_x;
  tables sex xagegroup
         race_eth xeduc
         /missing;
  where eligstat=1;
  title "Base Data Vars
        Check (Character-
        Frequencies) ";
run;
```

➤ Data Exploratory Visualization – Using SAS Cloud Analytic Services (CAS)



➤ Data Visualization & Analysis – Using SURVEYFREQ and PROC SGPanel for Box Plots



*Note: Values Unweighted

```
proc sort
  data=mycaslib.cdemo_puf
  out =cdemo_puf_x;
  by xagegroup xeduc;
run;

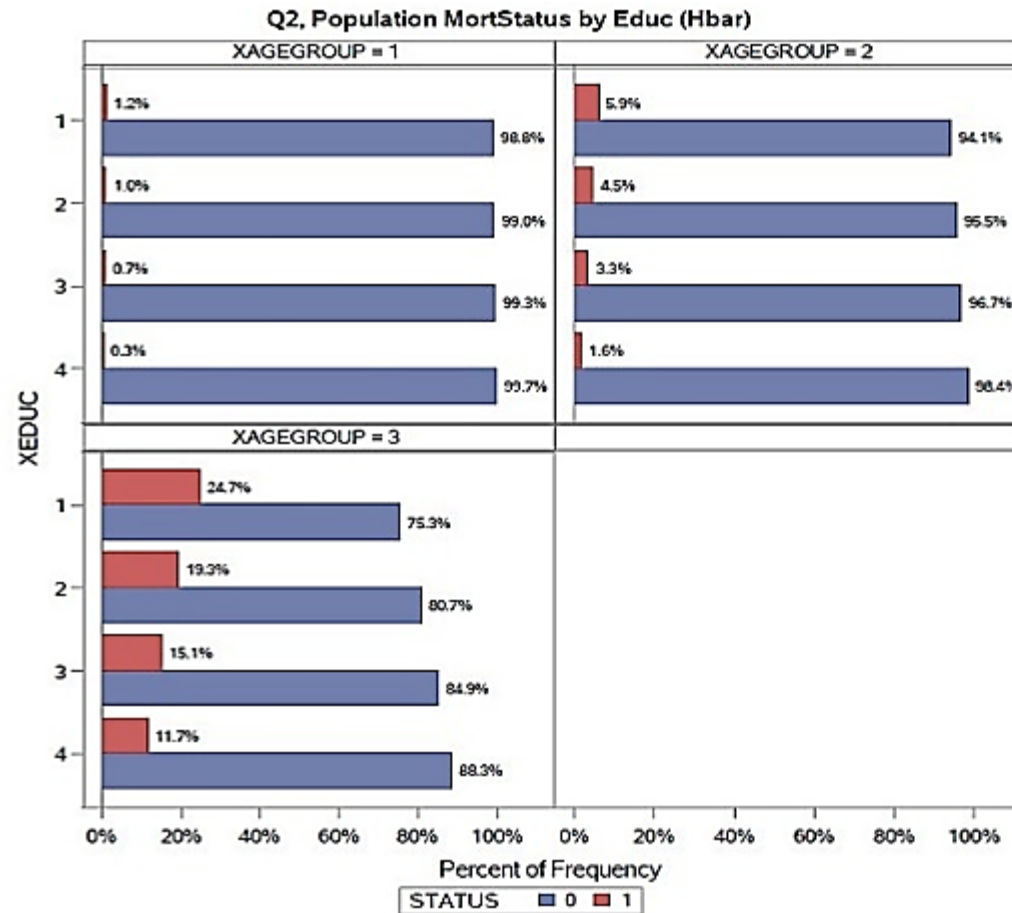
ods graphics / reset width=6.6in
  height=6.6in imagemap;

proc sgpanel data = cdemo_puf_x
  (where=(select=1));
  panelby xagegroup ;
  vbox age_p /category= status
  group = xeduc;
  vbox pery / category= status
  group = xeduc;
  format xeduc xedu.;
  title "Q1, Baseline Mean Age &
  Follow-Up Yrs, by Educ & Age-
  Groups (in Vbox)";
run;

ods graphics /reset;
```

```
proc surveyfreq
  data = cdemo_puf_x;
  * survey design variables ;
  strata pstrat;
  cluster ppsu;
  weight sawt_1218;
  * analysis variable(s);
  table xeduc mortstat /cl;
  * subpopulation(s);
  where select =1 and eligstat
  =1;
  title "01 Weighted Educ &
  Status (%), All Eligible of PUF
  NHIS 2012-18";
  format mortstat mortfmt.;
run;
```

➤ Data Visualization & Analysis – Using PROC SG PANEL for Bar Charts



*Note: Percent % Unweighted

```
ods graphics / reset width=6in
height=6in imagemap;
proc sgpanel data = cdemo_puf_x
pctlevel=group;
panelby xagegroup ;
title "Q2, Population
Mortstatus by Educ & AgeGrp
(Hbar)";
hbar xeduc /group = status
groupdisplay = cluster
stat=percent datalabel;
where select =1;
run; quit;
ods graphics /reset;
```

➤ Data Analysis – Using PROC SURVEYFREQ

Age Groups	Education Attainment	Frequency (n)	Deceased (%, weighted)	Std. Error	95% Lower CL	95% Upper CL
25-39 Years	Less than high school	74	1.04	0.17	0.71	1.37
	High school diploma or GED	109	0.68	0.09	0.51	0.85
	Associate degree or some college	117	0.67	0.08	0.51	0.82
	Bachelor's degree or higher	59	0.26	0.04	0.18	0.35
40-64 Years	Less than high school	678	4.80	0.25	4.30	5.29
	High school diploma or GED	1037	3.88	0.16	3.56	4.20
	Associate degree or some college	919	2.90	0.12	2.66	3.14
	Bachelor's degree or higher	454	1.28	0.08	1.13	1.42
65+ Years	Less than high school	2539	22.8	0.59	21.7	24.0
	High school diploma or GED	3045	17.9	0.42	17.0	18.7
	Associate degree or some college	2174	14.2	0.40	13.4	15.0
	Bachelor's degree or higher	1644	10.4	0.32	9.78	11.0

```
proc surveyfreq data = cdemo_puf;
  * survey design variables *;
  strata pstrat;
  cluster ppsu;
  weight sawt_1218;
  * analysis variable(s) *;
  table mortstat
    xagegroup * xeduc * mortstat
    /row cl wchisq;
  * specify subpopulation(s) *;
  where select =1 and eligstat=1;
  title "Weighted Status NDI-
  linkage (%) by Age-Grp &
  Education, NHIS 2012-18";
  format mortstat mortfmt.;
run;
```

➤ Data Visualization & Analysis – Using PROC GCHART (Pie Charts)

UCOD -LEADING

- 001 Diseases of heart (I00-I09, I11, I13, I20-I51)
- 002 Malignant neoplasms (C00-C97)
- 003 Chronic lower respiratory diseases (J40-J47)
- 004 Accidents (unintentional injuries)
(V01-X59, Y85-Y86)
- 005 Cerebrovascular diseases (I60-I69)
- 006 Alzheimer's disease (G30)
- 007 Diabetes mellitus (E10-E14)
- 008 Influenza and pneumonia (J09-J18)
- 009 Nephritis, nephrotic syndrome, and nephrosis
(N00-N07, N17-N19, N25-N27)
- 010 All other causes (residual)

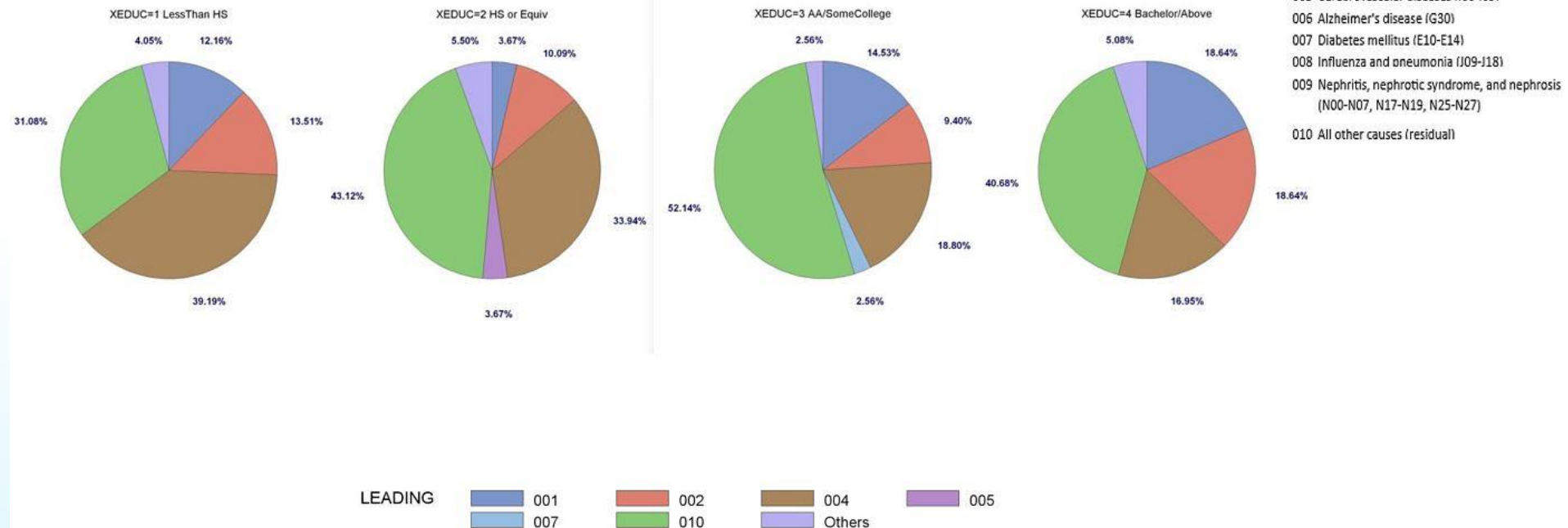
```
ods graphics / reset width=10in
height=10in imagemap;
*Pie Chart;
proc gchart data= cdemo_puf_x
  (where=(select=1));
  pie ucod_leading
    /group= xeduc noheading
    across=4
    other=2
    clockwise value=none
  percent=outside ;
  by xagegroup;
  format percent pctfmt.;
  title "Q3, Leading Causes of
  Death by Education & AgeGroups";
run; quit;
ods graphics / reset;
```

➤ Data Visualization & Analysis – Using PROC GCHART (Pie Charts) – Cont'd

Q3. Leading Causes of Death by Education & Age Groups

XAGEGROUP=Age 25-39 Years

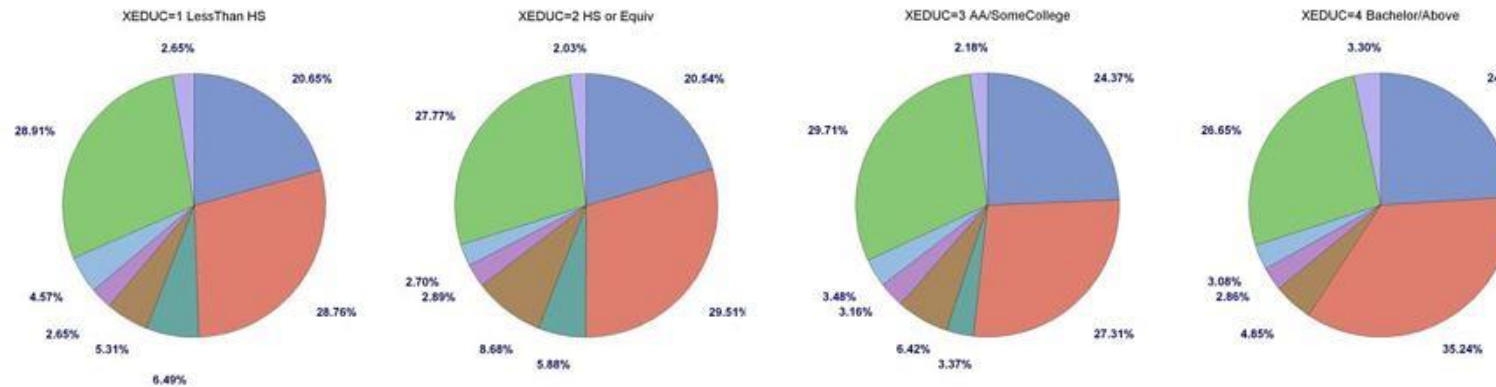
FREQUENCY of LEADING



➤ Data Visualization & Analysis – Using PROC GCHART (Pie Charts) – Cont'd

Q3. Leading Causes of Death by Education & Age Groups
XAGEGROUP=Age 40-64 Years

FREQUENCY of LEADING



UCOD -LEADING

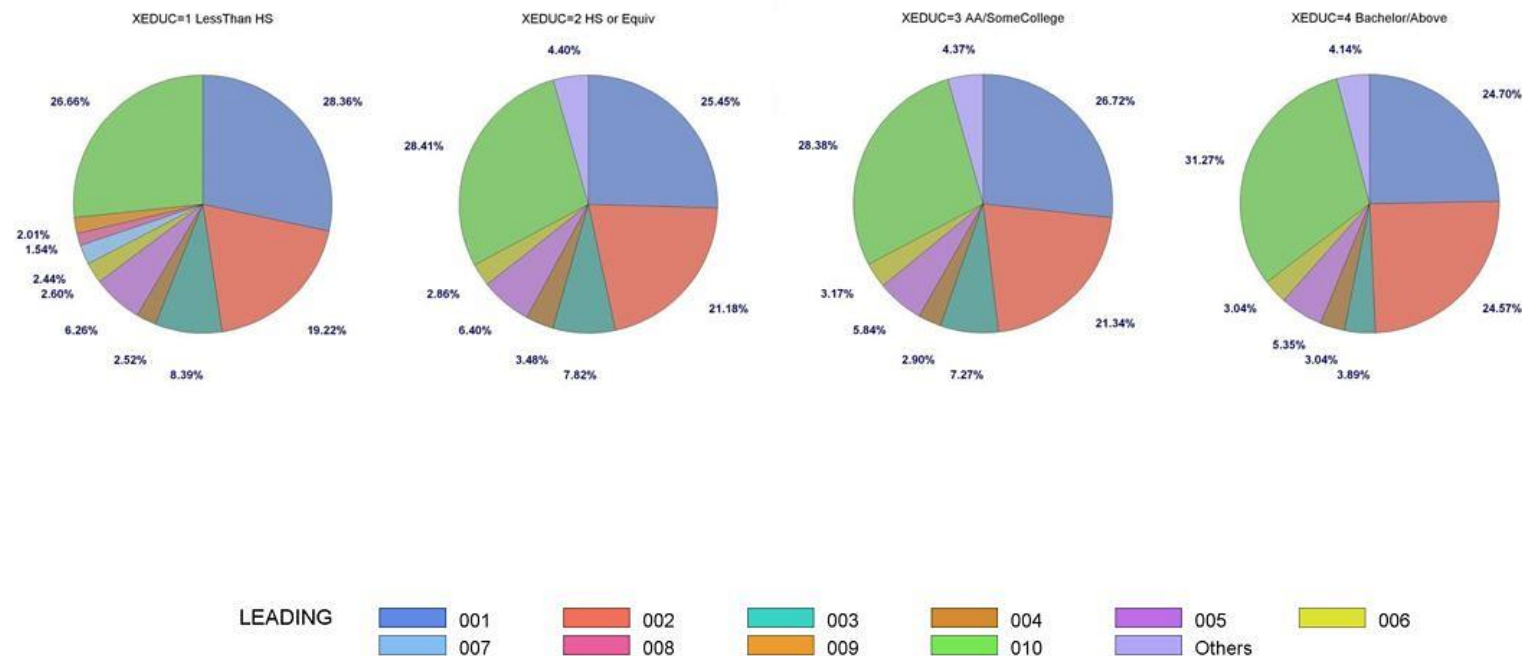
- 001 Diseases of heart (I00-I09, I11, I13, I20-I51)
- 002 Malignant neoplasms (C00-C97)
- 003 Chronic lower respiratory diseases (J40-J47)
- 004 Accidents (unintentional injuries) (V01-X59, Y85-Y86)
- 005 Cerebrovascular diseases (I60-I69)
- 006 Alzheimer's disease (G30)
- 007 Diabetes mellitus (E10-E14)
- 008 Influenza and pneumonia (J09-J18)
- 009 Nephritis, nephrotic syndrome, and nephrosis (N00-N07, N17-N19, N25-N27)
- 010 All other causes (residual)

LEADING

001	002	003	004
005	007	010	Others

➤ Data Visualization & Analysis – Using PROC GCHART (Pie Charts) – Cont'd

Q3. Leading Causes of Death by Education & Age Groups
XAGEGROUP=Age 65+ Years
FREQUENCY of LEADING



UCOD -LEADING
 001 Diseases of heart (I00-I09, I11, I13, I20-I51)
 002 Malignant neoplasms (C00-C97)
 003 Chronic lower respiratory diseases (J40-J47)
 004 Accidents (unintentional injuries) (V01-X59, Y85-Y86)
 005 Cerebrovascular diseases (I60-I69)
 006 Alzheimer's disease (G30)
 007 Diabetes mellitus (E10-E14)
 008 Influenza and pneumonia (J09-J18)
 009 Nephritis, nephrotic syndrome, and nephrosis (N00-N07, N17-N19, N25-N27)
 010 All other causes (residual)

➤ Data Analysis – Using PROC SURVEYPHERG

```
options nofmterr;
%macro hzrt(xld=);
  data cdemo_puf_model;
    format _all_ informat _all_;
    set mycaslib.cdemo_puf;
    %if "&xld"="000" %then %do;
      if mortstat in (0 1) then xlead000= mortstat; %end;
    %else %do;
      if ucod_leading ="&xld." then xlead&xld.=1;
      else if mortstat in (0 1) then xlead&xld.=0; %end;
    format sex sexgrp.;
  run;
  ods output ParameterEstimates= hrpm&xld.;
  proc surveyphreg data=cdemo_puf_model;
    domain select;
    strata pstrat;
    cluster ppsu ;
    weight sawt_1218;
    class sex race_eth xagegroup xmarital_status region &covars./ ref=first;
    class xeduc /ref=last;
    model pery * xlead&xld. (0)= age_p sex race_eth xmarital_status region &covars. xeduc /risklimits;
    title "04 SurveyPhreg Education for Leading Causes of Death &xld.";
  run;
  data hzrt&xld.; set hrpm&xld.; if substr(Parameter,1,5) ='XEDUC';
  inset=&xld.;
  keep inset Parameter HazardRatio HRLowerCL HRUpperCL; run;
%mend;

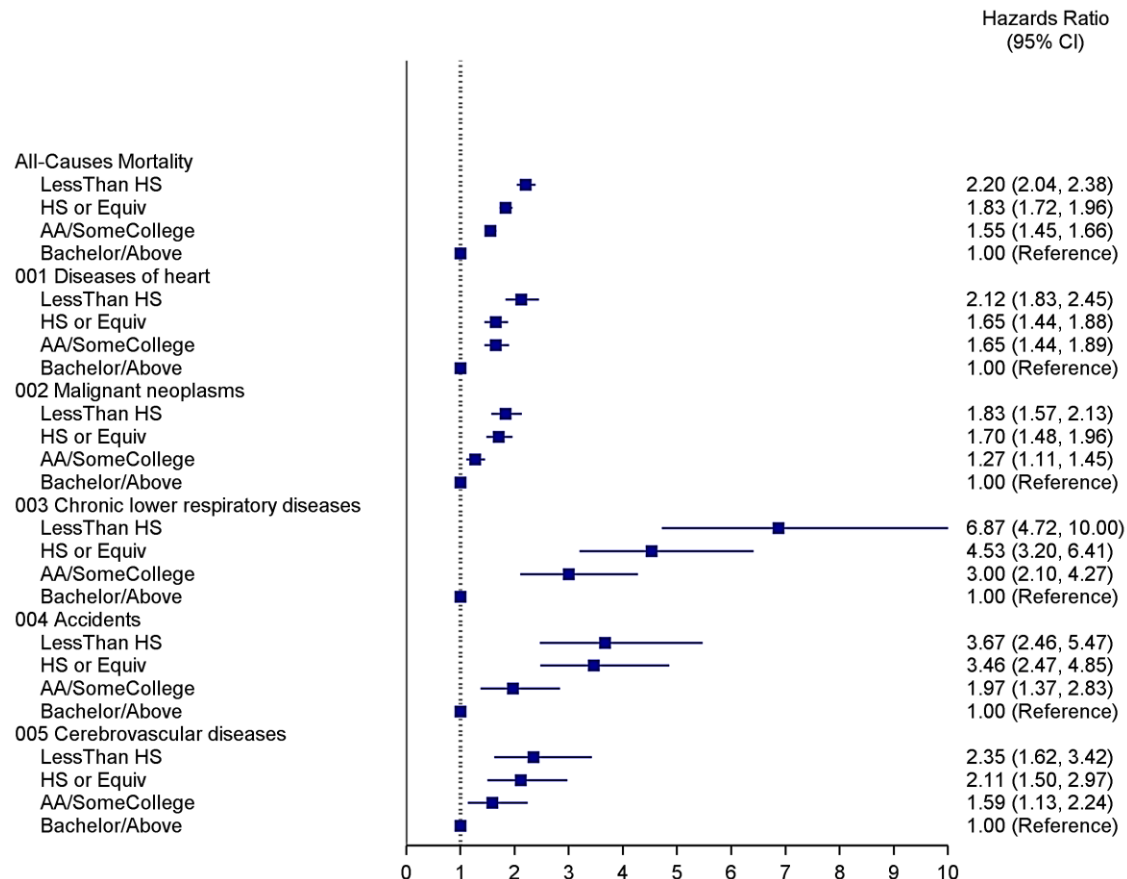
* Two Models Adjusting Variables: Basic Demographic, and Advanced OtherVars;
%let covars =;
%hzrt(xld=000) %hzrt(xld=001) %hzrt(xld=002)
%hzrt(xld=003) %hzrt(xld=004) %hzrt(xld=005)

%let covars = fairhlth nohins smkst alcohol obesity;
%hzrt(xld=000) %hzrt(xld=001) %hzrt(xld=002)
%hzrt(xld=003) %hzrt(xld=004) %hzrt(xld=005)
```

- Procedure SurveyPhreg can be used to create a hazards ratio model, including a cox hazards model with the macro %HZRT
- Two sample code parts incorporated:
 - ❖ a base model -- adjusting the demographic variables,
 - ❖ an advanced model -- adjusting multiple confounding variables: general health, insurance coverage, smoking, alcohol, and obesity

➤ Data Visualization & Analysis – Using PROC SGPLOT for Forest Plots

Q4 Hazards Ratio in Forest Plot Model1



```
ods graphics / reset width=8in height=6in imagemap;
proc sgplot data=anahzrt nocycleattrs noautolegend;
  styleattrs axisextent=data;
  refline 0 / axis=x;
  refline 1 / axis=x lineattrs=(pattern=35 color=gray
    thickness=2);

  highlow y=subgrpnr low=hrlowercl high= hruppercl
    /clipcap clipcapshape=openarrow
    lineattrs=(color=darkblue
    thickness=1) type=line;

  scatter y=subgrpnr x= hazardratio /markerattrs
    =(size=4pt symbol=squarefilled
    color=darkblue);

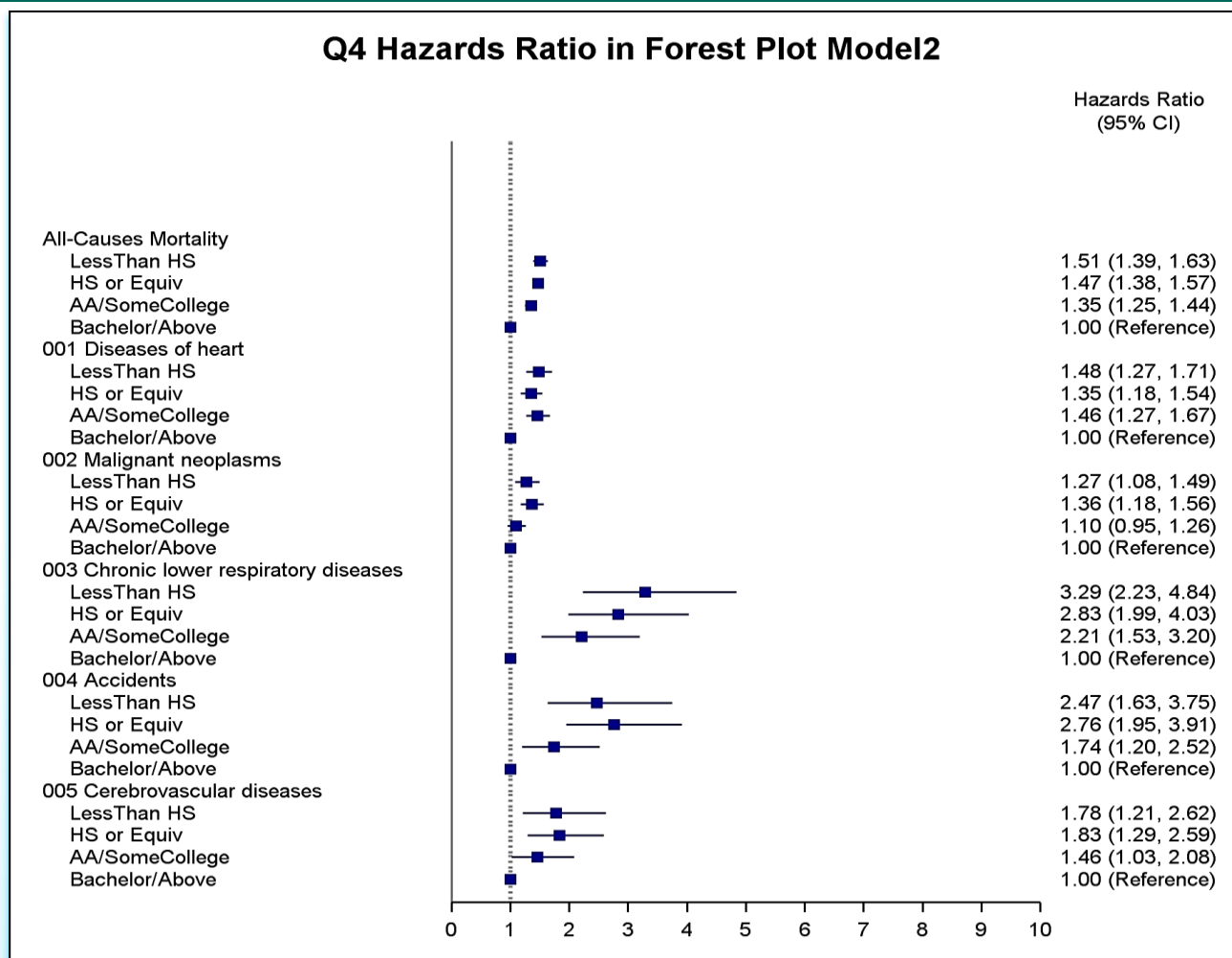
  yaxistable subgrpnr /nolabel location=inside
    position=left textgroup=id textgroupid=text
    labelattrs=(size=7)
    indentweight=indentfl;

  yaxis reverse display=none offsetmin=0.1
    colorbands=odd;

  xaxis display=(nolabel) valueattrs=(size=7)
    values=(0 to 10 by 1);

  title1 "Q4 Hazards Ratio in Forest Plot";
run;
ods graphics / reset;
```

➤ Data Visualization & Analysis – Using PROC SGPLOT for Forest Plots – Cont'd



➤ Conclusions



SAS VIYA provides an efficient platform for interactive data exploration and utilizing compatible legacy SAS code to generate data visualizations and conduct data analysis.	
This study used the public-use 2012-18 NHIS linked mortality files to explore the association between education attainment and risk of mortality.	• The NHIS LMFs provides a rich resource for exploring the association between selected socioeconomic factors and mortality • The visualizations show that mortality rates vary by cause of death, education level, and age group
Future related studies could include:	• Exploring the use of other SAS VIYA features for exploratory data analysis (EDA), including interactive visualizations • Using SAS Studio to integrate Python and R code into the data visualization and analysis process

➤ Acknowledgments & References

ACKNOWLEDGEMENTS

- Thanks to CDC SAS EDAV Team for providing the platform and SAS-Viya Q/A sessions.
- Thanks to DLMAB Linkage Team and our colleagues for their prior works on methods, publications and release contributions. Thanks for all the review/ suggestions related to this demonstration work.
- The findings and conclusions are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.
- SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

REFERENCES

- Healthy People 2030, U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. Accessed August 19, 2024. Available at <https://health.gov/healthypeople/objectives-and-data/social-determinants-health>
- IHME-CHAIN Collaborators. 2024. "Effects of Education on Adult Mortality: a Global Systematic Review and Meta-Analysis." Lancet Public Health. 9(3):e155-e165
- Integrated Public Use Microdata Series (IPUMS). 2024. "IPUMS Health Surveys." Accessed August 19, 2024. Available at <https://nhis.ipums.org/nhis>
- Mirel, L.B. El Bural Félix, S. Zhang, C. Golden, C. and Cox, C.S. 2020. "Comparative Analysis of the National Health Interview Survey Public-Use and Restricted-Use Linked Mortality Files." Natl Health Stat Report. 143:1-32.
- National Health Interview Survey (NHIS). 2024. "The National Health Interview Survey." Accessed August 19, 2024. Available at <https://www.cdc.gov/nchs/nhis/index.htm>.
- NCHS Data Linkage. 2024. "NCHS Data Linked to NDI Mortality Files." Accessed August 19, 2024. Available at <https://www.cdc.gov/nchs/data-linkage/mortality.htm>.
- SAS® Viya®. 2024. "SAS® Viya® Features List". Accessed August 19, 2024. Available at https://www.sas.com/en_us/software/visual-analytics/viya-features.html

Contact Us

Your comments and questions are valued and encouraged. Please contact us –

NCHS DAE Data Linkage Team

datalinkage@cdc.gov

<https://www.cdc.gov/nchs/data-linkage/index.htm>

<https://www.cdc.gov/rdc/index.htm> (for Restricted-Use Linked Mortality Data)



➤ **LISTSERV Information**

**Subscribe to the NCHS
Data Linkage Program
LISTSERV to receive
updates!**

**Email a message to
list@cdc.gov. Leave the
subject line blank. In the
body of the message, type:**

**SUBSCRIBE NCHS-DATA-
LINKAGE-PROGRAM
lastname, firstname**

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

