**Examining the Correlation between Metabolic Syndrome and Diabetes across Ethnicities between 2015 and 2018**

/\*Creating permanent SAS datasets\*/

libname MetS 'D:\6. Courses\Summer semester 2020\PH7525 Statistical Computing\SAS Data Analysis Project';

\*\*STEP 1\*\*;

/\*Cleaning the data i.e appending, merging, subsetting, etc.\*/

\*Demographic characteristics;

\*\* 2015-2016;

**data** Demo1;

set MetS.DEMO\_I;

**run**;

\*\* 2017-2018;

**data** Demo2;

set MetS.DEMO\_J;

**run**;

\*Combining the two demographic data from 2015-2016 and 2017-2018;

**proc** **sort** data = Demo1;

by SEQN;

**run**;

**proc** **sort** data = Demo2;

by SEQN;

**run**;

\*\*Formatting the values for Gender, AgeGroup, Race, Education level;

**proc** **format**;

value RIAGENDR\_ **1** = 'Male' **2** = 'Female';

value AgeGroup\_ **1** = '< 20' **2** = '20-39' **3** = '40-59' **4** = '60+';

value RIDRETH3\_ **1** = 'Mexican American' **2** = 'Other Hispanic' **3** = 'NH White' **4** = 'NH Black' **6** = 'NH Asian' **7** = 'Other';

value DMDEDUC2\_ **1** = 'Primary' **2** = 'Secondary' **3** = 'High School' **4** = 'College' **5** = 'Graduate';

**run**;

\*\*Merging Demo1 and Demo2;

**data** Main;

merge Demo1 Demo2;

by SEQN;

label SEQN = 'ID' RIAGENDR = 'Gender' RIDAGEYR = 'Age' RIDRETH3 = 'Race' DMDEDUC2 = 'Education level' AgeGroup ='Age group';

select;

when (missing(RIDAGEYR)) AgeGroup = **.**;

when (RIDAGEYR lt **20**) AgeGroup = **1**;

when (RIDAGEYR lt **40**) AgeGroup = **2**;

when (RIDAGEYR lt **60**) AgeGroup = **3**;

when (RIDAGEYR ge **60**) AgeGroup = **4**;

otherwise;

end;

if DMDEDUC2 in (**7**,**9**,**.**) then DMDEDUC2 = **.**;

format RIAGENDR RIAGENDR\_. RIDRETH3 RIDRETH3\_. DMDEDUC2 DMDEDUC2\_. AgeGroup AgeGroup\_.;

**run**;

\*Blood pressure status;

\*\* 2015-2016;

**data** BP1;

set MetS.BPX\_I;

**run**;

\*\* 2017-2018;

**data** BP2;

set MetS.BPX\_J;

**run**;

\*Combining the two blood pressure data;

**proc** **sort** data = BP1;

by SEQN;

**run**;

**proc** **sort** data = BP2;

by SEQN;

**run**;

\*\*Formating the values for BPstatus;

**proc** **format**;

value BPstatus\_ **1** = 'Hypertension' **0** = 'No hypertension';

**run**;

\*\*Merging BP1 and BP2;

**data** BP ;

merge BP1 BP2;

by SEQN;

bp\_sys = round(mean(BPXSY1, BPXSY2, BPXSY3));

bp\_dia = round(mean(BPXDI1, BPXDI2, BPXDI3));

label SEQN = 'ID' bp\_sys = 'Systolic BP' bp\_dia = 'Diastolic BP' BPstatus = 'BP status';

\*Hypertension criteria: Systolique BP >= 140 mmHg and Diastolic BP >=90 mmHg;

if missing(bp\_sys)or missing(bp\_dia)then BPstatus = **.**;

else if bp\_sys < **140** or bp\_dia < **90** then BPstatus = **0**;

else if bp\_sys >= **140** and bp\_dia >= **90** then BPstatus = **1**;

format BPstatus BPstatus\_.;

drop BPXSY1 BPXDI1 BPXSY2 BPXDI2 BPXSY3 BPXDI3;

**run**;

\*HDL\_cholesterol level;

\*\* 2015-2016;

**data** HDL1 (keep = SEQN LBDHDD);

set MetS.HDL\_I;

**run**;

\*\* 2017-2018;

**data** HDL2;

set MetS.HDL\_J;

**run**;

\*Combining the two HDL-cholesterol data;

**proc** **sort** data = HDL1;

by SEQN;

**run**;

**proc** **sort** data = HDL2;

by SEQN;

**run**;

\*\* Merging HDL1 and HDL2;

**data** HDL;

merge HDL1 HDL2;

by SEQN;

**run**;

\*Albumin and creatinine ratio (ACR);

\*\* 2015-2016;

**data** Ratio1;

set MetS.ALB\_CR\_I;

**run**;

\*\* 2017-2018;

**data** Ratio2;

set MetS.ALB\_CR\_J;

**run**;

\*Combining the two albumin/creatinine data;

**proc** **sort** data = Ratio1;

by SEQN;

**run**;

**proc** **sort** data = Ratio2;

by SEQN;

**run**;

\*\*Formating values for ACRcat;

**proc** **format**;

value ACRcat\_ **1** = 'High' **2** = 'Low';

**run**;

\*\*Merging Ratio1 and Ratio2;

**data** Ratio;

merge Ratio1 Ratio2;

by SEQN;

label SEQN = 'ID' URXUMA = 'Albumine (ug/mL)' URXUCR = 'Creatinine (mg/dL)' ACratio = 'Albumin/creatinine ratio (ug/mg)' ACRcat = 'ACR category';

ACratio = (URXUMA/URXUCR)\***100**; \*ACR = (Albumin (ug/mL)/Creatinine (mg/dL))\*100;

select; \*Criteria for high albumin/creatinine ratio: ACR >= 30 mg/dL;

when (missing(ACratio)) ACRcat = **.**;

when (ACratio ge **30**) ACRcat = **1**;

when (ACratio lt **30**) ACRcat = **2**;

otherwise;

end;

format ACRcat ACRcat\_.;

drop URXUMA URXUCR;

**run**;

\*Body Mass Index (BMI);

\*\* 2015-2016;

**data** BMI1;

set MetS.BMX\_I;

**run**;

\*\* 2017-2018;

**data** BMI2 ;

set MetS.BMX\_J;

**run**;

\*Combining the two BMI data;

**proc** **sort** data = BMI1;

by SEQN;

**run**;

**proc** **sort** data = BMI2;

by SEQN;

**run**;

\*\*Formating values for BMIcateg;

**proc** **format**;

value BMIcateg\_ **1** = 'Underweight' **2** = 'Normal' **3** = 'Overweight' **4** = 'Obese';

**run**;

\*\*Merging BMI1 and BMI2;

**data** BMI;

merge BMI1 BMI2;

by SEQN;

label SEQN = 'ID' BMXWT = 'Weight (kg)' BMXHT = 'Height (cm)' BMI = 'Body Mass Index (kg/m\*\*2)' BMIcateg = 'BMI category';

BMI = BMXWT/(BMXHT/**100**)\*\***2**;\*To calculate BMI: BMI (kg/m\*\*2) = Weight (kg)/ Height\*\*2 (m\*\*2);

select;

when (missing (BMI)) BMIcateg = **.**;

when (BMI lt **18.5**) BMIcateg = **1**;

when (BMI ge **18.5** and (BMI lt **25**)) BMIcateg = **2**;

when (BMI ge **25** and (BMI lt **30**)) BMIcateg = **3**;

when (BMI ge **30**) BMIcateg = **4**;

otherwise;

end;

drop BMXWT BMXHT;

format BMIcateg BMIcateg\_.;

**run**;

/\*Combining all the datasets\*/

\*\* 1st: Main + BP;

**proc** **sort** data = Main;

by SEQN;

**run**;

**proc** **sort** data = BP;

by SEQN;

**run**;

**data** Pop1;

merge Main BP;

by SEQN;

**run**;

\*\* 2nd: Pop1 + HDL;

**proc** **sort** data = Pop1;

by SEQN;

**run**;

**proc** **sort** data = HDL;

by SEQN;

**run**;

\*\*Formating values for HDL\_level;

**proc** **format**;

value HDL\_level\_ **1** = 'Low' **2** = 'High';

**run**;

**data** Pop2;

merge Pop1 HDL;

by SEQN;

label HDL\_level = 'HDL-cholesterol level';

\*Criteria for low HDL\_level: HDL\_cholesterol < 35 mg/dl (male) and < 39 mg/dl (female);

if missing (RIAGENDR) | missing (LBDHDD) then HDL\_level = **.**;

else if RIAGENDR = **1** & LBDHDD < **35** then HDL\_level = **1**;

else if RIAGENDR = **1** & LBDHDD >= **35** then HDL\_level = **2**;

else if RIAGENDR = **2** & LBDHDD < **39** then HDL\_level = **1**;

else if RIAGENDR = **2** & LBDHDD >= **39** then HDL\_level = **2**;

format HDL\_level HDL\_level\_.;

**run**;

\*\* 3rd: Pop2 + ACR;

**proc** **sort** data = Pop2;

by SEQN;

**run**;

**proc** **sort** data = Ratio;

by SEQN;

**run**;

**data** Pop3;

merge Pop2 Ratio;

by SEQN;

**run**;

\*\* 4th Pop3 + BMI;

**proc** **sort** data = Pop3;

by SEQN;

**run**;

**proc** **sort** data = BMI;

by SEQN;

**run**;

**data** Pop4;

merge Pop3 BMI;

by SEQN;

**run**;

\*Diabetes;

\*\* 2015-2016;

**data** Diab1;

set MetS.DIQ\_I;

**run**;

\*\* 2017-2018;

**data** Diab2;

set MetS.DIQ\_J;

**run**;

\*Combining the two diabetes data;

\*\*Formating the values for Diabstatus;

**proc** **format**;

value Diabstatus\_ **1** = 'Diabetic' **2** = 'Non-diabetic';

**run**;

\*\*Merging Diab1 and Diab2;

**data** Diabetes;

merge Diab1 Diab2;

by SEQN;

label Diabstatus = 'Diabetic status';

select;

when (missing(DIQ010) | in (**3**,**7**,**9**)) Diabstatus = **.**;

when (DIQ010 = **1**) Diabstatus = **1**;

when (DIQ010 = **2**) Diabstatus = **2**;

otherwise;

end;

format Diabstatus Diabstatus\_.;

**run**;

/\*Forming the baseline data\*/

\*\*Adding Diabetes to Pop4;

**proc** **sort** data = Pop4;

by SEQN;

**run**;

**proc** **sort** data = Diabetes;

by SEQN;

**run**;

\*\*Merging Diabetes and Pop4;

**data** MetS.Baseline (keep = SEQN SDMVPSU SDMVSTRA WTINT2YR WTMEC2YR RIAGENDR RIDAGEYR AgeGroup RIDRETH3 DMDEDUC2 bp\_sys bp\_dia BPstatus LBDHDD HDL\_level ACratio ACRcat BMI BMIcateg Diabstatus);

merge Pop4 Diabetes;

by SEQN;

drop DIQ010;

**run**;

**proc** **contents** data = MetS.Baseline varnum;

**run**;

Title 'Baseline data';

**proc** **print** data = MetS.Baseline (obs = **10**) label;

var SEQN RIAGENDR RIDAGEYR AgeGroup RIDRETH3 DMDEDUC2 bp\_sys bp\_dia BPstatus LBDHDD HDL\_level ACratio ACRcat BMI BMIcateg Diabstatus;

**run**;

A screenshot of a computer

Description automatically generated

/\*Defining metabolic syndrome\*/

\*\* 2 or more of the criteria:

- High blood pressure: systolic >= 140 mmHg and diastolic >=90 mmHg

- Low HDL\_cholesterol < 35 mg/dl (male) and < 39 mg/dl (female)

- High Albumin/creatinine ratio >= 30 mg/mg

- BMI > 30 kg/m\*\*2;

**data** Metab;

set MetS.Baseline;

label NumCriteria = 'Number of criteria' MetS = 'Metabolic syndrome';

if BPstatus in (**1**) then Criteria1 = **1**;

else Criteria1 = **0**;

if HDL\_level in (**1**) then Criteria2 = **1**;

else Criteria2 = **0**;

if ACRcat in (**1**) then Criteria3 = **1**;

else Criteria3 = **0**;

if BMIcateg in (**4**) then Criteria4 = **1**;

else Criteria4 = **0**;

NumCriteria = sum(of Criteria1, Criteria2, Criteria3, Criteria4);

drop Criteria1 Criteria2 Criteria3 Criteria4;

if NumCriteria >= **2** then MetS = 'Yes';

else MetS = 'No';

where RIDAGEYR >= **20**;

**run**;

Title 'Metabolic syndrome';

**proc** **print** data = Metab (obs = **10**) label;

**run**;

A screenshot of a computer

Description automatically generated

\*\*STEP 2\*\*;

/\*Descriptive statistics\*/

\*1. Frequency distribution of categorical variables (Gender, Race, Education level, Diabetic status, Metabolic syndrome);

**proc** **freq** data = Metab;

tables RIAGENDR RIDRETH3 DMDEDUC2 AgeGroup BPstatus HDL\_level ACRcat BMIcateg Diabstatus MetS;

**run**;

\*2. Summary statistics for quantitative variables (Age, Systolic BP, Diastolic BP, HDL-cholesterol, Albumin/creatinine ratio, BMI);

**proc** **means** data = Metab;

var RIDAGEYR bp\_sys bp\_dia LBDHDD ACratio BMI;

**run**;

\*3. Checking for missing values for each variable and excluding missing values for analysis;

\*Subsetting the sample to people with MetS and Diabetes;

**data** Complete;

set Work.Metab;

if cmiss(of \_all\_) gt **0** then delete;

**run**;

\*4. Running frequency distribution and summary statistics with weight variable;

**proc** **surveyfreq** data = Complete;

cluster SDMVPSU;

strata SDMVSTRA;

tables RIAGENDR RIDRETH3 DMDEDUC2 AgeGroup BPstatus HDL\_level ACRcat BMIcateg Diabstatus MetS;

weight WTINT2YR;

\*\*Comparing results with 1:

- Gender: The proportions of male (48%) and female (52%) are the same in both unweighted and weighted frequency

- Race: Non-Hispanic Whites (34%)are predominant in the unweighted analysis, and 15% more than Non-Hispanic Blacks (22%)

vs Non-Hispanic Whites (63%) are largely predominant in the weighted analysis, and with a 53% difference compared to Non-Hispanic Blacks (11%)

- Education level: Most respondents have college degree in both 31% in unweighted, and 32% in weighted

- Age group: Most of the respondents are 60+ years old (36%) in unweighted vs 40-59 years old in weighted;

**proc** **surveymeans** data = Complete;

cluster SDMVPSU;

strata SDMVSTRA;

var RIDAGEYR bp\_sys bp\_dia LBDHDD ACratio BMI;

weight WTMEC2YR;

**run**;

\*\*Comparing results with 2:

- Age: The mean age is 50 years old (unweighted) vs 48 years old (weighted)

- BP: The mean BP is 126/71 mmHg (unweighted) vs 124/71 mmHG

- HDL-cholesterol: The mean HDL-cholesterol is 54 mg/dL (unweighted) vs 55 mg/dL (weighted)

- Albumin/creatinine ratio : The mean ACratio is 53 ug/mg vs 34 ug/mg

- BMI: The mean BMI is the same 30 kg/m\*\*2

\*5. Statistical testing to examine relationship or association;

\*\* Metabolic syndrome and Race;

/\*H0: There is no association between Metabolic syndrome and Race

HA: There is an association between Metabolic syndrome and Race \*/

**proc** **surveyfreq** data = Complete;

cluster SDMVPSU;

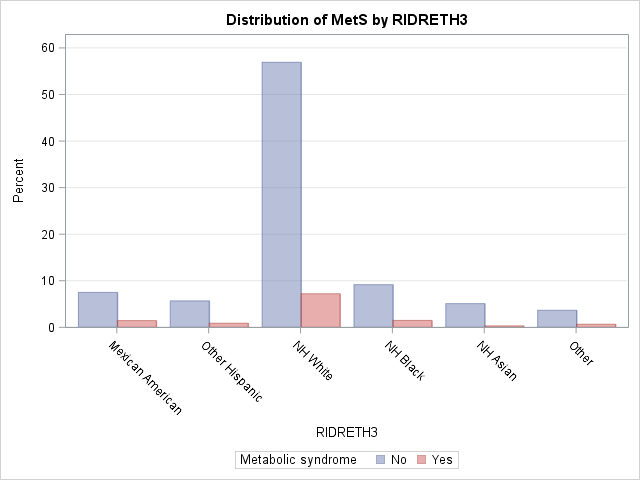
strata SDMVSTRA;

tables RIDRETH3\*MetS / plots = wtfreqplot (scale=percent groupby=ROW twoway=CLUSTER) chisq row nostd;

weight WTINT2YR;

format RIDRETH3 RIDRETH3\_.;

**run**;



| **Rao-Scott Chi-Square Test** | |
| --- | --- |
| **Pearson Chi-Square** | 43.8795 |
| **Design Correction** | 0.9185 |
|  |  |
| **Rao-Scott Chi-Square** | 47.7718 |
| **DF** | 5 |
| **Pr > ChiSq** | <.0001 |
|  |  |
| **F Value** | 9.5544 |
| **Num DF** | 5 |
| **Den DF** | 150 |
| **Pr > F** | <.0001 |
| **Sample Size = 9247** | |

\*\*Diabetes and Metabolic syndrome;

/\*H0: There is no association between Diabetes and Metabolic syndrome

HA: There is an association between Diabetes and Metabolic syndrome\*/

**proc** **surveyfreq** data = Complete;

cluster SDMVPSU;

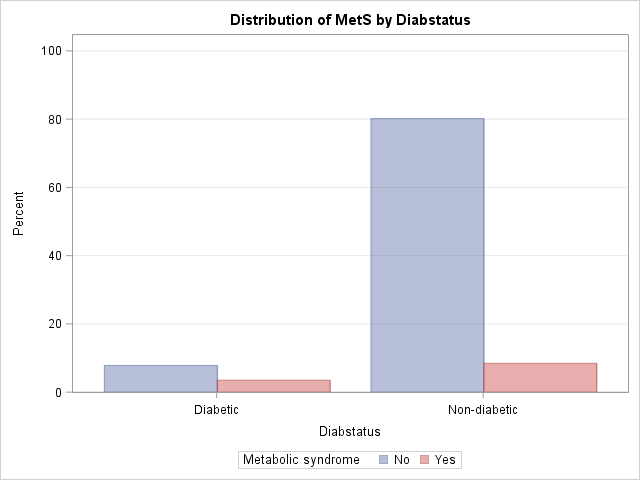
strata SDMVSTRA;

tables Diabstatus\*MetS/ plots = wtfreqplot (scale=percent groupby=ROW twoway=CLUSTER) chisq row nostd;

weight WTINT2YR;

format Diabstatus Diabstatus\_.;

**run**;



| **Rao-Scott Chi-Square Test** | |
| --- | --- |
| **Pearson Chi-Square** | 404.9941 |
| **Design Correction** | 1.6803 |
|  |  |
| **Rao-Scott Chi-Square** | 241.0272 |
| **DF** | 1 |
| **Pr > ChiSq** | <.0001 |
|  |  |
| **F Value** | 241.0272 |
| **Num DF** | 1 |
| **Den DF** | 30 |
| **Pr > F** | <.0001 |
| **Sample Size = 9247** | |

\*\*Diabetes and Race ;

/\*H0: There is no association between Diabetes and Race

HA: There is an association between Diabetes and Race\*/

**proc** **surveyfreq** data = Complete;

cluster SDMVPSU;

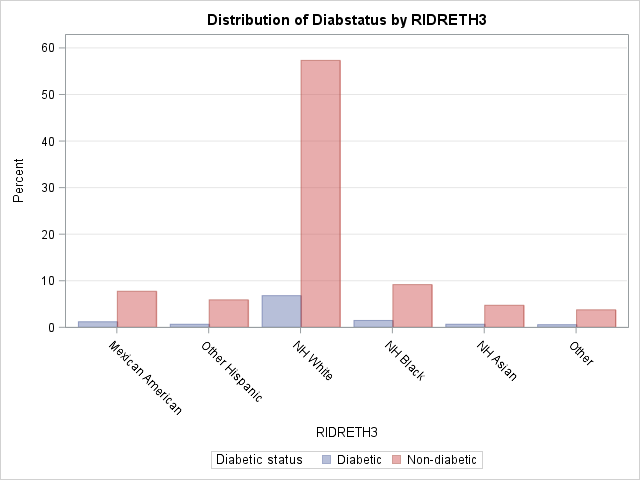
strata SDMVSTRA;

tables RIDRETH3\*Diabstatus/ plots = wtfreqplot (scale=percent groupby=ROW twoway=CLUSTER) chisq row nostd;

weight WTINT2YR;

format RIDRETH3 RIDRETH3\_. Diabstatus Diabstatus\_.;

**run**;



| **Rao-Scott Chi-Square Test** | |
| --- | --- |
| **Pearson Chi-Square** | 15.3145 |
| **Design Correction** | 0.8011 |
|  |  |
| **Rao-Scott Chi-Square** | 19.1174 |
| **DF** | 5 |
| **Pr > ChiSq** | 0.0018 |
|  |  |
| **F Value** | 3.8235 |
| **Num DF** | 5 |
| **Den DF** | 150 |
| **Pr > F** | 0.0027 |
| **Sample Size = 9247** | |

\*Subsetting to individuals with metabolic syndrome and diabetes;

**data** MD;

set Work.Complete;

label MetsD = 'Metabolic syndrome and diabetes';

if MetS = 'Yes' & Diabstatus = **1** then MetsD = 'Yes';

else MetsD = 'No';

**run**;

/\*H0: There is no association between Metabolic syndrome + Diabetes and Ethnicity

HA: There is an association between Metabolic syndrome + Diabetes and Ethnicity\*/

**proc** **surveyfreq** data = MD;

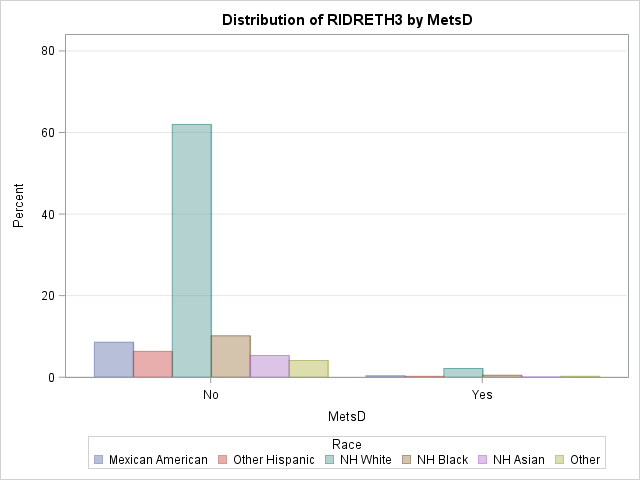
cluster SDMVPSU;

strata SDMVSTRA;

tables MetsD\*RIDRETH3/ plots = wtfreqplot (scale=percent groupby=ROW twoway=CLUSTER) chisq row nostd;

weight WTINT2YR;

**run**;



| **Rao-Scott Chi-Square Test** | |
| --- | --- |
| **Pearson Chi-Square** | 14.2226 |
| **Design Correction** | 0.8481 |
|  |  |
| **Rao-Scott Chi-Square** | 16.7693 |
| **DF** | 5 |
| **Pr > ChiSq** | 0.0050 |
|  |  |
| **F Value** | 3.3539 |
| **Num DF** | 5 |
| **Den DF** | 150 |
| **Pr > F** | 0.0067 |
| **Sample Size = 9247** | |

/\*Assessing which ethnic group is more likely to have MetS and Diabetes\*/

\*Odds ratio Mets syndrome & Ethnicity\*;

**proc** **Surveylogistic** data = Complete;

class RIDRETH3 (ref='NH White')/param=ref;

cluster SDMVPSU;

Strata SDMVSTRA;

Model Mets (event= 'yes') = RIDRETH3;

Weight WTINT2YR;

**run**;

\*Odds ratio Diabetes & Ethnicity\*;

**proc** **Surveylogistic** data = Complete;

class RIDRETH3 (ref='NH White')/param=ref;

cluster SDMVPSU;

Strata SDMVSTRA;

Model Diabstatus (event='diabetic') = RIDRETH3;

Weight WTINT2YR;

**run**;

\*Odds ratio MetS + Diabetes & Ethnicity\*;

**proc** **Surveylogistic** data = MD;

class RIDRETH3 (ref='NH White')/param=ref;

cluster SDMVPSU;

Strata SDMVSTRA;

Model MetsD (event='Yes') = RIDRETH3;

Weight WTINT2YR;

**run**;