# Case-Study 9: Health-Related Behavioral Risk Factor using the Simulated Big Data

# (Version 2.0)

**Note:** In the original V.1.0 of this case-study, the variable **INCOMG** has an incorrect created category.

|  |  |  |
| --- | --- | --- |
| Variable name | V.1.0 Data | Corrected V.2 Data |
| INCOMG | 1,2,3,4,5,**9** | 1,2,3,4,5,**6** |

**V2.0 Data**: <https://umich.instructure.com/courses/38100/files/folder/Case_Studies/CaseStudy09_HealthBehaviorRisks_Data_V2>

**Overview:** This case study examines the associations between demographic factors, behavioral risk factors and chronic diseases among adults aged over 18 years old in Michigan in 2013.

# Driving Challenges:

* + Are there relations between demographic factors (age, gender, race, marital status, education, income and work) and chronic diseases?
  + Are there relations between behavioral risk factors (diet, physical activities, smoking, and alcohol use) and chronic diseases?
  + Can behavioral risk factors predict diabetes, stroke, and heart attack after controlling for those demographic factors?

# Meta-data:

* + **ID:** case subject identifier
  + **Demographic variables:** AGE\_G (1=Age 18 to 24, 2=Age 25 to 34, 3=Age 35 to 44, 4=Age 45 to 54, 5=Age 55 to 64, 6=Age 65 or older), SEX (1=Male, 2=Female), RACEGR3 (1=White only, Non-Hispanic, 2=Black only, Non-Hispanic, 3=Other race only, Non-Hispanic, 4=Multiracial, Non-Hispanic, 5=Hispanic, 9=Don’t know/Not sure/Refused), IMPEDUC (1=Never attended school or only kindergarten, 2=Grades 1 through 8 (Elementary), 3=Grades 9 through 11 (Some high school), 4=Grade 12 or GED (High school graduate), 5=College 1 year to 3 years (Some college or technical school), 6=College 4 years or more (College graduate)), IMPMRTL ( 1=Married, 2=Divorced, 3=Widowed, 4=Separated, 5=Never married, 6=A member of an unmarried couple), EMPLOY1 (1=Employed for wages, 2=Self-employed, 3=Out of work for 1 year or more,

4=Out of work for less than 1 year, 5=A homemaker, 6=A student, 7=Retired, 8=Unable to, 9=Refused), INCOMG (1=Less than $15,000, 2=$15,000 to less than $25,000, 3=$25,000 to less than $35,000, 4=$35,000 to less than $50,000, 5=$50,000 or more, 9=Don’t know/Not sure/Missing)

* + **Chronic diseases:** CVDINFR4 (Heart attack, 1=Yes, 2=No, 7=Don’t know/Not sure, 9=Refused), CVDCRHD4 (Angina or coronary heart disease, 1=Yes, 2=No, 7=Don’t know/Not sure, 9=Refused), CVDSTRK3 (Stroke, 1=Yes, 2=No, 7=Don’t know/Not sure, 9=Refused), DIABETE3 (Diabetes, 1=Yes, 2=Yes, but female told only during pregnancy, 3=No, 4=No, pre-diabetes or borderline diabetes, 7=Don’t know/Not sure, 9=Refused)
  + **Behavioral risk factors:** RFSMOK3 (Current smoking, 1=No, 2=Yes, 9=Don’t know/Refused/Missing), RFDRHV4 (Heavy alcohol consumption, 1=No, 2=Yes, 9=Don’t know/Refused/Missing), FRTLT1 (Consume fruits 1 or more times per day, 1=Yes, 2=No, 9=Don’t know/Refused/Missing), VEGLT1 (Consume vegetables 1 or more times per day, 1=Yes, 2=No, 9=Don’t know/Refused/Missing), TOTINDA (Leisure time physical activities per month, 1=Yes, 2=No, 9=Don’t know/Refused/Missing)

**Data Provenance:** This case study only uses simulated data. The complete R-script generating the data is included below. The entire case study is simulated from the Behavioral Risk Factor Surveillance System (BRFSS) survey (Seen from<http://www.cdc.gov/brfss/annual_data/annual_2013.html)> and can be used, updated, refactored and expanded by the entire community.

**R-Code:**

# Define number of subjects

NumSubj<-1000

set.seed(1234)

# Define data elements

# Cases

ID<-c(1:1000)

# Demographic variables

AGE\_G <- sample(c(1:6), NumSubj, replace = TRUE, prob = c(0.07,0.09,0.11,0.17,0.23,0.33))

RACEGR3 <- sample(c(1:5,9), NumSubj, replace = TRUE, prob = c(0.84,0.08,0.02,0.02,0.02,0.02))

IMPEDUC <- sample(c(2:6), NumSubj, replace = TRUE, prob = c(0.01,0.04,0.31,0.29,0.35))

IMPMRTL <- sample(c(1:6), NumSubj, replace = TRUE, prob = c(0.51,0.15,0.13,0.01,0.17,0.03))

SEX <- ifelse(runif(NumSubj)<.43,1,2)

EMPLOY1 <- sample(c(1:8), NumSubj, replace = TRUE, prob = c(0.4,0.07,0.03,0.03,0.07,0.03,0.3,0.7))

INCOMG <- sample(c(1:5,9), NumSubj, replace = TRUE, prob = c(0.1,0.16,0.11,0.13,0.36,0.14))

# Behavioral risk fctors

RFSMOK3 <- sample(c(1,2,9), NumSubj, replace = TRUE, prob = c(0.81,0.17,0.02))

RFDRHV4 <- sample(c(1,2,9), NumSubj, replace = TRUE, prob = c(0.91,0.06,0.03))

FRTLT1 <- sample(c(1,2,9), NumSubj, replace = TRUE, prob = c(0.62,0.32,0.06))

VEGLT1 <- sample(c(1,2,9), NumSubj, replace = TRUE, prob = c(0.73,0.2,0.07))

TOTINDA <- sample(c(1,2,9), NumSubj, replace = TRUE, prob = c(0.72,0.24,0.04))

# Chronic diseases

#Angina or coronary heart disease

CVDCRHD4 <- rep(0,NumSubj)

logit\_CVDCRHD4 <- 0.01-2.5\*AGE\_G+0.5\*(SEX-1)+1.5\*IMPEDUC-0.9\*(EMPLOY1==7)-1.8\*(RFSMOK3==2)-2.1\*(RFDRHV4==2)-0.7\*(FRTLT1==2)-1.2\*(VEGLT1==2)-1.9\*(TOTINDA==2)+rnorm(1,mean = 0,sd = 0.1)

prob\_CVDCRHD4 <- exp(logit\_CVDCRHD4)/(1+exp(logit\_CVDCRHD4))

CVDCRHD4 <- rbinom(NumSubj,1,prob\_CVDCRHD4)

CVDCRHD4 <- ifelse(CVDCRHD4==0,2,1)

#Stroke

CVDSTRK3 <- rep(0,NumSubj)

logit\_CVDSTRK3 <- 0.01-3\*AGE\_G+0.3\*(SEX-1)+1.2\*IMPEDUC-0.6\*(EMPLOY1==7)-3\*(RFSMOK3==2)-1.7\*(RFDRHV4==2)-0.4\*(FRTLT1==2)-0.8\*(VEGLT1==2)-2.5\*(TOTINDA==2)+rnorm(1,mean = 0,sd = 0.1)

prob\_CVDSTRK3 <- exp(logit\_CVDSTRK3)/(1+exp(logit\_CVDSTRK3))

CVDSTRK3 <- rbinom(NumSubj,1,prob\_CVDSTRK3)

CVDSTRK3 <- ifelse(CVDSTRK3==0,2,1)

#heart attack

CVDINFR4 <- rep(0,NumSubj)

logit\_CVDINFR4 <- 0.01-2.2\*AGE\_G+0.4\*(SEX-1)-2.4\*(RFSMOK3==2)-1.3\*(RFDRHV4==2)-0.9\*(FRTLT1==2)-0.6\*(VEGLT1==2)-1.8\*(TOTINDA==2)+rnorm(1,mean = 0,sd = 0.1)

prob\_CVDINFR4 <- exp(logit\_CVDINFR4)/(1+exp(logit\_CVDINFR4))

CVDINFR4 <- rbinom(NumSubj,1,prob\_CVDINFR4)

CVDINFR4 <- ifelse(CVDINFR4==0,2,1)

#Diabetes

DIABETE3 <- rep(0,NumSubj)

logit\_DIABETE3 <- 0.01-1.3\*AGE\_G+0.7\*(SEX-1)+2.1\*IMPEDUC-3\*(RFSMOK3==2)-2.3\*(RFDRHV4==2)-1.1\*(FRTLT1==2)-0.8\*(VEGLT1==2)-2.2\*(TOTINDA==2)+rnorm(1,mean = 0,sd = 0.1)

prob\_DIABETE3 <- exp(logit\_DIABETE3)/(1+exp(logit\_DIABETE3))

DIABETE3 <- rbinom(NumSubj,1,prob\_DIABETE3)

DIABETE3 <- ifelse(DIABETE3==0,3,1)

for(i in 1:NumSubj){

if(DIABETE3[i] == 1 & SEX[i] == 2){

DIABETE3[i] <- sample(c(1,2),1,prob = c(0.9,0.1))

}

else if(DIABETE3[i] == 3){

DIABETE3[i] <- sample(c(3,4),1,prob = c(0.9,0.1))

}

}

# Data (putting all components together)

data <- cbind(ID, AGE\_G, SEX, RACEGR3, IMPEDUC, IMPMRTL, EMPLOY1, INCOMG, CVDINFR4, CVDCRHD4, CVDSTRK3, DIABETE3, RFSMOK3, RFDRHV4, FRTLT1, VEGLT1, TOTINDA)

# Write out (save) the result to a file that can be shared

write.csv(data, "CaseStudy09\_HealthBehaviorRisks\_Data.csv", row.names=FALSE)