An Exploratory Analysis of Cancer Incidence and Mortality to Identify High-Risk Communities and Improve Survival

Ramiro Cadavid, Pri Nonis, Payman Roghani

September 24, 20168

```r
library(dplyr)
library(tidyr)
library(car)
library(fBasics)
library(xtable)
```

Introduction

In this project our efforts are focused on the analysis of data included in the csv file provided, to primarily understand the potential relationship between different parameters and the incidences of cancer across counties in the US. The main objectives are: 1. To understand factors that predict cancer mortality rate, with the ultimate aim of identifying communities for social interventions. 2. To determine which interventions are likely to have the most impact.

Cancer Data

```r
Cancer <- read.csv('cancer.csv')
```

```r
summary(Cancer) #summary statistics
       X            avgAnnCount         medIncome          popEst2015      
 Min.   :   1.0   Min.   :    6.0   Min.   : 22640   Min.   :     827  
 1st Qu.: 762.5   1st Qu.:   76.0   1st Qu.: 38882   1st Qu.:   11684  
 Median :1524.0   Median :  171.0   Median : 45207   Median :   26643  
 Mean   :1524.0   Mean   :  606.3   Mean   : 47063   Mean   :  102637  
 3rd Qu.:2285.5   3rd Qu.:  518.0   3rd Qu.: 52492   3rd Qu.:   68671  
 Max.   :3047.0   Max.   :38150.0   Max.   :125635   Max.   :10170292  
                                                                        
 povertyPercent               binnedInc      MedianAge     
 Min.   : 3.20   (45201, 48021.6]  : 306   Min.   : 22.30  
 1st Qu.:12.15   (54545.6, 61494.5]: 306   1st Qu.: 37.70  
 Median :15.90   [22640, 34218.1]  : 306   Median : 41.00  
 Mean   :16.88   (42724.4, 45201]  : 305   Mean   : 45.27  
 3rd Qu.:20.40   (48021.6, 51046.4]: 305   3rd Qu.: 44.00  
 Max.   :47.40   (51046.4, 54545.6]: 305   Max.   :624.00  
                 (Other)           :1214                   
 MedianAgeMale   MedianAgeFemale                                Geography   
 Min.   :22.40   Min.   :22.30   Abbeville County, South Carolina:   1  
 1st Qu.:36.35   1st Qu.:39.10   Acadia Parish, Louisiana        :   1  
 Median :39.60   Median :42.40   Accomack County, Virginia       :   1  
 Mean   :39.57   Mean   :42.15   Ada County, Idaho               :   1  
```

```
 3rd Qu.:42.50   3rd Qu.:45.30   Adair County, Iowa              :   1
 Max.   :64.70   Max.   :65.70   Adair County, Kentucky          :   1
                                 (Other)                         :3041
 AvgHouseholdSize PercentMarried   PctNoHS18_24     PctHS18_24
 Min.   :0.0221   Min.   :23.10   Min.   : 0.00   Min.   : 0.0
 1st Qu.:2.3700   1st Qu.:47.75   1st Qu.:12.80   1st Qu.:29.2
 Median :2.5000   Median :52.40   Median :17.10   Median :34.7
 Mean   :2.4797   Mean   :51.77   Mean   :18.22   Mean   :35.0
 3rd Qu.:2.6300   3rd Qu.:56.40   3rd Qu.:22.70   3rd Qu.:40.7
 Max.   :3.9700   Max.   :72.50   Max.   :64.10   Max.   :72.5

 PctSomeCol18_24 PctBachDeg18_24   PctHS25_Over   PctBachDeg25_Over
 Min.   : 7.10   Min.   : 0.000   Min.   : 7.50   Min.   : 2.50
 1st Qu.:34.00   1st Qu.: 3.100   1st Qu.:30.40   1st Qu.: 9.40
 Median :40.40   Median : 5.400   Median :35.30   Median :12.30
 Mean   :40.98   Mean   : 6.158   Mean   :34.80   Mean   :13.28
 3rd Qu.:46.40   3rd Qu.: 8.200   3rd Qu.:39.65   3rd Qu.:16.10
 Max.   :79.00   Max.   :51.800   Max.   :54.80   Max.   :42.20
 NA's   :2285
 PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage
 Min.   :17.60      Min.   : 0.400       Min.   :22.30
 1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20
 Median :54.50      Median : 7.600       Median :65.10
 Mean   :54.15      Mean   : 7.852       Mean   :64.35
 3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10
 Max.   :80.10      Max.   :29.400       Max.   :92.30
 NA's   :152
 PctEmpPrivCoverage PctPublicCoverage    PctWhite         PctBlack
 Min.   :13.5       Min.   :11.20     Min.   : 10.20   Min.   : 0.0000
 1st Qu.:34.5       1st Qu.:30.90     1st Qu.: 77.30   1st Qu.: 0.6207
 Median :41.1       Median :36.30     Median : 90.06   Median : 2.2476
 Mean   :41.2       Mean   :36.25     Mean   : 83.65   Mean   : 9.1080
 3rd Qu.:47.7       3rd Qu.:41.55     3rd Qu.: 95.45   3rd Qu.:10.5097
 Max.   :70.7       Max.   :65.10     Max.   :100.00   Max.   :85.9478

    PctAsian         PctOtherRace     PctMarriedHouseholds   BirthRate
 Min.   : 0.0000   Min.   : 0.0000   Min.   :22.99         Min.   : 0.000
 1st Qu.: 0.2542   1st Qu.: 0.2952   1st Qu.:47.76         1st Qu.: 4.521
 Median : 0.5498   Median : 0.8262   Median :51.67         Median : 5.381
 Mean   : 1.2540   Mean   : 1.9835   Mean   :51.24         Mean   : 5.640
 3rd Qu.: 1.2210   3rd Qu.: 2.1780   3rd Qu.:55.40         3rd Qu.: 6.494
 Max.   :42.6194   Max.   :41.9303   Max.   :78.08         Max.   :21.326

   deathRate
 Min.   : 59.7
 1st Qu.:161.2
 Median :178.1
 Mean   :178.7
 3rd Qu.:195.2
 Max.   :362.8
```

```
str(Cancer, max.level = 1, strict.width = "wrap")
'data.frame':   3047 obs. of  30 variables:
```

```
$ X : int 1 2 3 4 5 6 7 8 9 10 ...
$ avgAnnCount : num 1397 173 102 427 57 ...
$ medIncome : int 61898 48127 49348 44243 49955 52313 37782 40189 42579
   60397 ...
$ popEst2015 : int 260131 43269 21026 75882 10321 61023 41516 20848 13088
   843954 ...
$ povertyPercent : num 11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1
   ...
$ binnedInc : Factor w/ 10 levels "(34218.1, 37413.8]",..: 9 6 6 4 6 7 2 2
   3 8 ...
$ MedianAge : num 39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
$ MedianAgeMale : num 36.9 32.2 44 42.2 47.8 43.5 42.2 50.8 48.4 34.7 ...
$ MedianAgeFemale : num 41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
$ Geography : Factor w/ 3047 levels "Abbeville County, South Carolina",..:
   1459 1460 1464 1589 1618 1766 2051 2112 2143 2185 ...
$ AvgHouseholdSize : num 2.54 2.34 2.62 2.52 2.34 2.58 2.42 2.24 2.38 2.65
   ...
$ PercentMarried : num 52.5 44.5 54.2 52.7 57.8 50.4 54.1 52.7 55.9 50 ...
$ PctNoHS18_24 : num 11.5 6.1 24 20.2 14.9 29.9 26.1 27.3 34.7 15.6 ...
$ PctHS18_24 : num 39.5 22.4 36.6 41.2 43 35.1 41.4 33.9 39.4 36.3 ...
$ PctSomeCol18_24 : num 42.1 64 NA 36.1 40 NA NA 36.5 NA NA ...
$ PctBachDeg18_24 : num 6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4 7.1 ...
$ PctHS25_Over : num 23.2 26 29 31.6 33.4 30.4 29.8 31.6 32.2 28.8 ...
$ PctBachDeg25_Over : num 19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
$ PctEmployed16_Over : num 51.9 55.9 45.9 48.3 48.2 44.1 51.8 40.9 39.5
   56.6 ...
$ PctUnemployed16_Over: num 8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
$ PctPrivateCoverage : num 75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9
   ...
$ PctEmpPrivCoverage : num 41.6 43.6 34.9 35 35.1 32.6 28.3 25.9 29.9 44.4
   ...
$ PctPublicCoverage : num 32.9 31.1 42.1 45.3 44 43.2 46.4 50.9 48.1 31.4
   ...
$ PctWhite : num 81.8 89.2 90.9 91.7 94.1 ...
$ PctBlack : num 2.595 0.969 0.74 0.783 0.27 ...
$ PctAsian : num 4.822 2.246 0.466 1.161 0.666 ...
$ PctOtherRace : num 1.843 3.741 2.747 1.363 0.492 ...
$ PctMarriedHouseholds: num 52.9 45.4 54.4 51 54 ...
$ BirthRate : num 6.12 4.33 3.73 4.6 6.8 ...
$ deathRate : num 165 161 175 195 144 ...
```

```
colnames(Cancer)
 [1] "X"                    "avgAnnCount"          "medIncome"
 [4] "popEst2015"           "povertyPercent"       "binnedInc"
 [7] "MedianAge"            "MedianAgeMale"        "MedianAgeFemale"
[10] "Geography"            "AvgHouseholdSize"     "PercentMarried"
[13] "PctNoHS18_24"         "PctHS18_24"           "PctSomeCol18_24"
[16] "PctBachDeg18_24"      "PctHS25_Over"         "PctBachDeg25_Over"
[19] "PctEmployed16_Over"   "PctUnemployed16_Over" "PctPrivateCoverage"
[22] "PctEmpPrivCoverage"   "PctPublicCoverage"    "PctWhite"
[25] "PctBlack"             "PctAsian"             "PctOtherRace"
[28] "PctMarriedHouseholds" "BirthRate"            "deathRate"
cat("  \n")
```

```r
print(paste0('Number of rows: ', nrow(Cancer)))
[1] "Number of rows: 3047"
print(paste0('Number of columns: ', ncol(Cancer)))
[1] "Number of columns: 30"
```

The cancer.csv file 29 variables (30 columns, including the first one that has the number of observations) and 3047 observations, where each observation (i.e. row) includes data for a county across the US. The variables are mostly numbers and integers, except for 2 that are factors (`binnedInc` and `Geography`). Below, we have explain the variables in detail and provide our assessment of the quality of the data.

> data on smoking and obesity and other cancer risk factors could've been very helpful

Variables

- Cancer data:
  - `avgAnnCount`: The average number of new cancer cases per year per county for years 2009-2013
  - `popEst2015`: Estimated population by county 2015
- Economic status:
  - `medIncome`: Median income per county
  - `povertyPercent`: Percent of population below poverty line
  - `binnedInc`: ???
- Population age and gender:
  - `MedianAge`: Median age per county
  - `MedianAgeMale`: Median age among males per county
  - `MedianAgeFemale`: Median age among females per county
- Location:
  - `Geography`: County, State
- Marital status:
  - `PercentMarried`: Percentage of married population
  - `PctMarriedHouseholds`: Percentage of married households per county
- Education:
  - `PctNoHS18_24`: Percentage of 18-24 year old population with no high school education
  - `PctHS18_24`: Percentage of 18-24 year old population with high school education
  - `PctSomeCol18_24`: Percentage of 18-24 year old population with some college education
  - `PctBachDeg18_24`: Percentage of 18-24 year old population with bachelor's degree
  - `PctHS25_Over`: Percentage of population above 24 years old with high school education
  - `PctBachDeg25_Over`: Percentage of population above 24 years old with bachelor's degree
- Household size:
  - `AvgHouseholdSize`: Average household size per county
- Employment status:
  - `PctEmployed16_Over`: Percentage of population above 15 years old who have jobs
  - `PctUnemployed16_Over`: Percentage of population above 15 years old with no jobs
- Health insurance coverage:
  - `PctPrivateCoverage`: Percentage of the population with private insurance coverage
  - `PctEmpPrivCoverage`: percentage of the population with employer-sponsored insurance coverage
  - `PctPublicCoverage`: Percentage of the population with public insurance coverage
- Race:
  - `PctWhite`: Percentage of white population by county
  - `PctBlack`: Percentage of African-American population by county
  - `PctAsian`: Percentage of Asian population by county
  - `PctOtherRace`: Percentage of other races by county
- Birth and death rates:
  - `BirthRate`: Birth rate per county

– `deathRate`: Death rate per county

Evaluation of Dataset and Variables

Based on the outputs from diagnostic and summary statistics functions that we used above and further analysis explained in later sections of this report, below we describe our evaluation of dataset and its variables. Since definitions of most variables were not provided to us, our first step was to ensure understanding of what such variables represent. We also evaluated the data to identify potentially erroneous values and determine what variables are key to our analysis and whether we the dataset has the right variables to help answer the project questions or we would need to create additional variables needed to achieve that goal.

> from the assignment document: Evaluate the data quality. Are there any issues with the data? Explain how you handled these potential issues. Explain whether any data processing or preparation is required for your data set. create references between bullet points below and analysis done to support our evaluation/assumptions

- Data time frame: While `avgAnnCount` represents the mean for years 2009-2013, the population by county is for 2015 and other variables do not have date stamps. Ideally all variables should have been from the same time period.

- `avgAnnCount`: There is no definition for incidence rate per county for the `avgAnnCount` variable. Since the sum of all values is 1,847,514 and based on cancer.gov data the average number of cases for years 2009-2013 is 1,617,144, we will assume this variable represents the actual count of new cases. Therefore, in our analysis we created a new variable called "…" to represent the incidence rate of cancer per county (number of new cases per 100,000 people).

```
#caclulating the total for avgAnnCount to compare with offical reports by cancer.gov
sum(Cancer$avgAnnCount)
[1] 1847514
```

```
#calculating the ofr mean cancer death count for years 2009-2013 based on cancer.gov data, in order to
incidence_cancer <- c(1660290, 1529560, 1596670, 1638910, 1660290)
mean(incidence_cancer)
[1] 1617144
```

- Through our assessment we realized that the number of cancer new cases (`avgAnnCount`) for 6 counties were greater than the those counties population (`popEst2015`). Looking at the 6 observations, we realized that the the new case count for all these 6 counties is exactly the same number (1962.667684). In fact there are a total of Y counties that have exactly the same average number of new cases, which is probably an erroneous value. We decided to replace all of them with NA in our analysis.

```
#cheking the number of observatios, where new case count is greater than the population
sum(Cancer$avgAnnCount > Cancer$popEst2015, na.rm = TRUE)
[1] 6
```

```
Cancer$avgAnnCount[Cancer$avgAnnCount == 1962.667684] <- NA #removing the potentially erroneous number
Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000 #creating a new variable: new c
```

- We checked the `Geography` variable to identify potential duplicates. Since the number of unique values in this column is equal to the total number of observations, there can not be any duplicates in this column.

```
length(unique(Cancer[["Geography"]]))
[1] 3047
```

- The `binnedInc` variable has 10 levels that seem arbitrary. It is not clear why the income bins have been defined this way.

- The maximum `MedianAge` shows a value of 624, which is clearly a wrong number. We actually identified a total of 30 values in this column that are above 100; therefore, we will replace such values with NA in our analysis.

```
age_error = subset(Cancer, MedianAge > 100) #checking the number of erroneous values
nrow(age_error)
[1] 30
```

```
Cancer$MedianAge[Cancer$MedianAge > 100] = NA #replacing erroneous values with NA
```

- The minimum `AvgHouseholdSize` is 0.0221, which does not make sense, since we don't expect a household size below 1. There are 61 values in this column that are below 1, which we will replace with NA in our analysis.

```
household_error = subset(Cancer, AvgHouseholdSize < 1) #checking the number of erroneous values
nrow(household_error)
[1] 61
```

```
Cancer$AvgHouseholdSize[Cancer$AvgHouseholdSize < 1] = NA #replacing erroneous values with NA
```

- The `PctSomeCol18_24` variable has too many NA values (2285 out 3047). We will need to take this into account during our analysis.
- It is not clear how the birth rate is calculated and what exactly `BirthRate` represents. Often, the birth rate is defined as childbirths per 1,000 people each year, but applying that here would not give us the right number. For example in Los Angeles County with the population of 10,170,292, there were 124,641 live births in 2015, which translates into a birth rate of 12.25 (BR = (b ÷ p) X 1,000). However, the birth rate in our data shows a value of 4.7, which is probably the ratio of women aged 15-50 years old who gave birth in 2015 (source: http://www.towncharts.com/California/Demographics/Los-Angeles-County-CA-Demographics-data.html). As a result we didn't use this variable in our analysis.

```
#checking the BirthRate value for Los Angeles County
Cancer[1000,'BirthRate']
[1] 4.705281
```

```
#Calculating LA County birth rate based on official figures. Formula: BR = (b ÷ p) X 1,000
124641/10170292*1000
[1] 12.2554
```

- Based on our assessment, we believe the `deathRate` should represent the number of deaths due to cancer per 100,000 population per county. We looked at the figure for Kings County, NY (173.6) and the number in our data is closer to cancer death rate (140.3), as opposed to overall death rate (603.1). Based on this assumption, we also calculated the total death in a new column, calling the variable death_count (`deathRate` * `popEst2015`/100000) and total is 525,347, which is close to the figure reported by cancer.gov (589,430), further confirming our assumption regarding `deathRate` is most probably correct.

```
#checking the deathRate for Kings County, NY
Cancer[388, 'deathRate']
[1] 173.6
```

*Kings County, NY statistics: 2015 population: 2,673,000 2015 death rate (per 100,000 population): 603.1 2015 Cancer death rate (per 100,000 population): 140.3 Sources: DATA USA https://datausa.io/, NY State Dpt of Health https://www.health.ny.gov/

```
#comparing total death count in our dataset with official stats reported by officials
Cancer$death_count <- Cancer$deathRate * Cancer$popEst2015/100000
sum(Cancer$death_count)
[1] 525347.7
```

```
# 2015 cancer mortality reported by Cancer.gov: 589,430
```

- We assume that the values in `PctEmpPrivCoverage` column represent a subset of values The sum of values in `PctPrivateCoverage` column, since the sum of these two variables in some rows is above 100.
- Also, we assume that there is an overlap between people that have public health insurance and those with private health insurance, since the sum of `PctPrivateCoverage` and `PctPublicCoverage` in some rows is above 100.

```
#adding up health insurance coverage variables, to makes sence of such variables
Cancer$Pct_insured <- Cancer$PctPrivateCoverage + Cancer$PctPublicCoverage
Cancer$Pct_PersonalIsure <- Cancer$PctPrivateCoverage + Cancer$PctEmpPrivCoverage
print('Cancer$Pct_insured')
[1] "Cancer$Pct_insured"
summary(Cancer$Pct_insured)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65.40   96.25  101.30  100.61  105.80  131.70
print('Cancer$Pct_PersonalIsure')
[1] "Cancer$Pct_PersonalIsure"
summary(Cancer$Pct_PersonalIsure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.8    92.2   106.3   105.6   118.9   163.0
```

As seen in the summary statistics above, the Max for the 2 variables are above 100.

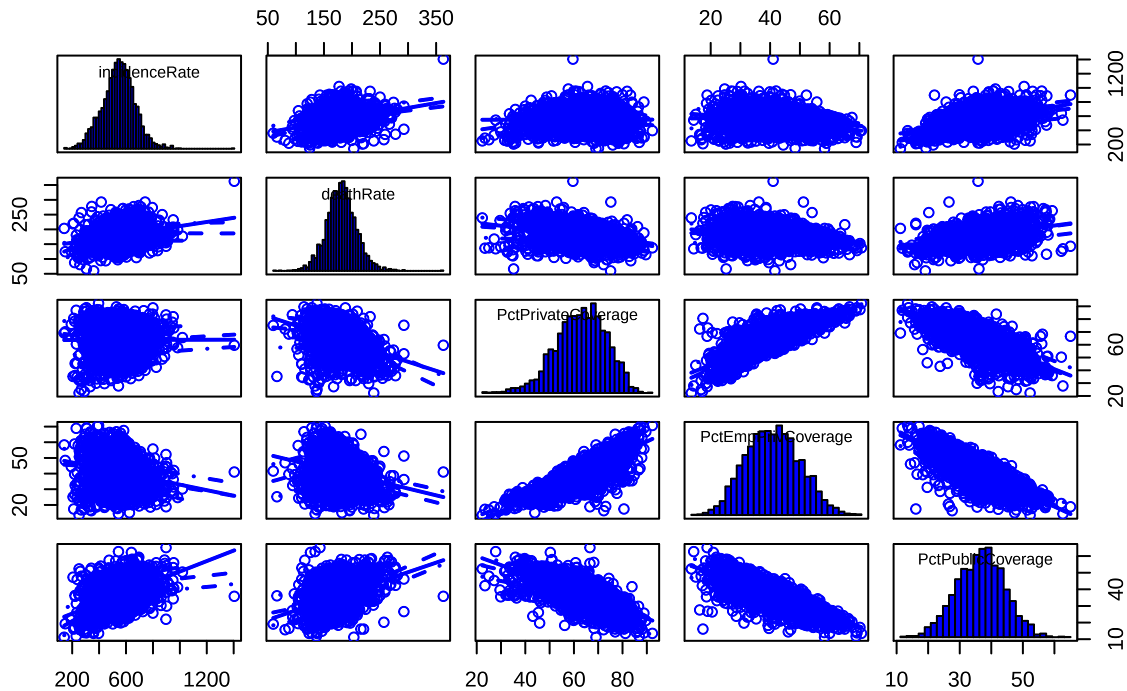
- other: removal of outliers? check with team

```
#adding 2 separate columns for County and State, in order to State-wide analysis of the data
Cancer <- Cancer %>% separate(Geography, c("County", "State"), sep = ",", remove = FALSE)
```

Multiavriate analysis

```
scatterplotMatrix( ~ incidenceRate + deathRate +
                   + PctPrivateCoverage + PctEmpPrivCoverage + PctPublicCoverage
                   ,diagonal=list(method="histogram"),
                   data = Cancer, main = "Scatterplot Matrix to Understand the Impact of Insrance Covera
```

Scatterplot Matrix to Understand the Impact of Insrance Coverage




```
cor(Cancer[ , c("incidenceRate", "deathRate",
                 "PctPrivateCoverage", "PctEmpPrivCoverage", "PctPublicCoverage")],
                  use = "complete.obs")
                   incidenceRate  deathRate PctPrivateCoverage
incidenceRate        1.000000000  0.3105464        0.002481271
deathRate            0.310546443  1.0000000       -0.369920199
PctPrivateCoverage   0.002481271 -0.3699202        1.000000000
PctEmpPrivCoverage  -0.228859552 -0.2534238        0.834285327
PctPublicCoverage    0.492747640  0.4040169       -0.722409606
                   PctEmpPrivCoverage PctPublicCoverage
incidenceRate              -0.2288596         0.4927476
deathRate                  -0.2534238         0.4040169
PctPrivateCoverage          0.8342853        -0.7224096
PctEmpPrivCoverage          1.0000000        -0.7757656
PctPublicCoverage          -0.7757656         1.0000000
```

Payman's note: * There is a positive correlation between incidenceRate and PctPublicCoverage (0.49), while the correlation between incidenceRate and PctPrivateCoverage is almost zero (-0.22 for PctEmpPrivCoverage) * There is a positive correlation between deathRate and PctPublicCoverage (0.40), while the correlation between incidenceRate and PctPrivateCoverage is negative (-0.36) * Based on this we can make a conclusion that public health insurance probably results in higher incidence of cancer and mortality * Caveat: the type of health insurance coverage (public vs private) is often affected by other factors. For example for geographic locations with low average employment/income, we can expect higher public insurance coverage. * Note: for future recommendations, we should also consider the major changes in public health insurance coverage due to Affordable Care Act, which aims to increase the quality of care through establishment of pay-for-performance and value-based healthcare policy.