Dementia (OASIS data) - descriptive statistics, longitudinal analysis & modeling

Darius Alexandru Cocirta

2023-01-10

Introduction

Dementia is a term used to describe a range of cognitive and behavioral symptoms that can include memory loss, problems with reasoning and communication and change in personality, and a reduction in a person's ability to carry out daily activities, such as shopping, washing, dressing and cooking. The most common types of dementia are: Alzheimer's disease, vascular dementia, mixed dementia, dementia with Lewy bodies and frontotemporal dementia. Dementia is a progressive condition, which means that the symptoms will gradually get worse. This progression will vary from person to person and each will experience dementia in a different way – people may often have some of the same general symptoms, but the degree to which these affect each person will vary (Dementia Gateway, Social Care Institute for Excellence).

Context

The Open Access Series of Imaging Studies (OASIS) is a project aimed at making MRI data sets of the brain freely available to the scientific community. By compiling and freely distributing MRI data sets, they hope to facilitate future discoveries in basic and clinical neuroscience. OASIS is made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI)(at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN).

About data

The dataset consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

Assignment

- descriptive statistics
- longitudinal analysis
- modeling

Data quality assessment & preliminary exploration

```
## Rows: 373
## Columns: 15
## $ `Subject ID` <chr> "OAS2_0001", "OAS2_0001", "OAS2_0002", "OAS2_0002",
"OAS2~
## $ `MRI ID`     <chr> "OAS2_0001_MR1", "OAS2_0001_MR2", "OAS2_0002_MR1",
"OAS2_~
## $ Group        <chr> "Nondemented", "Nondemented", "Demented", "Demented",
"De~
## $ Visit        <dbl> 1, 2, 1, 2, 3, 1, 2, 1, 2, 3, 1, 3, 4, 1, 2, 1, 2, 1,
2, ~
## $ `MR Delay`   <dbl> 0, 457, 0, 560, 1895, 0, 538, 0, 1010, 1603, 0, 518,
1281~
## $ `M/F`        <chr> "M", "M", "M", "M", "M", "F", "F", "M", "M", "M",
"M", "M~
## $ Hand         <chr> "R", "R", "R", "R", "R", "R", "R", "R", "R", "R",
"R", "R~
## $ Age          <dbl> 87, 88, 75, 76, 80, 88, 90, 80, 83, 85, 71, 73, 75,
93, 9~
## $ EDUC         <dbl> 14, 14, 12, 12, 12, 18, 18, 12, 12, 12, 16, 16, 16,
14, 1~
## $ SES          <dbl> 2, 2, NA, NA, NA, 3, 3, 4, 4, 4, NA, NA, NA, 2, 2, 2,
2, ~
## $ MMSE         <dbl> 27, 30, 23, 28, 22, 28, 27, 28, 29, 30, 28, 27, 27,
30, 2~
## $ CDR          <dbl> 0.0, 0.0, 0.5, 0.5, 0.5, 0.0, 0.0, 0.0, 0.5, 0.0,
0.5, 1.~
## $ eTIV         <dbl> 1987, 2004, 1678, 1738, 1698, 1215, 1200, 1689, 1701,
169~
## $ nWBV         <dbl> 0.696, 0.681, 0.736, 0.713, 0.701, 0.710, 0.718,
0.712, 0~
## $ ASF          <dbl> 0.883, 0.876, 1.046, 1.010, 1.034, 1.444, 1.462,
1.039, 1~
```

Data summary

Data summary

Name	mydata
Number of rows	373
Number of columns	15
_______________________	
Column type frequency:	
character	5
numeric	10
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Subject ID	0	1	9	9	0	150	0
MRI ID	0	1	13	13	0	373	0
Group	0	1	8	11	0	3	0
M/F	0	1	1	1	0	2	0
Hand	0	1	1	1	0	1	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Visit	0	1.00	1.88	0.92	1.00	1.0	2.00	2.00	5.00
MR Delay	0	1.00	595.10	635.49	0.00	0.0	552.00	873.00	2639.00
Age	0	1.00	77.01	7.64	60.00	71.0	77.00	82.00	98.00
EDUC	0	1.00	14.60	2.88	6.00	12.0	15.00	16.00	23.00
SES	19	0.95	2.46	1.13	1.00	2.0	2.00	3.00	5.00
MMSE	2	0.99	27.34	3.68	4.00	27.0	29.00	30.00	30.00
CDR	0	1.00	0.29	0.37	0.00	0.0	0.00	0.50	2.00
eTIV	0	1.00	1488.13	176.14	1106.00	1357.0	1470.00	1597.00	2004.00
nWBV	0	1.00	0.73	0.04	0.64	0.7	0.73	0.76	0.84
ASF	0	1.00	1.20	0.14	0.88	1.1	1.19	1.29	1.59

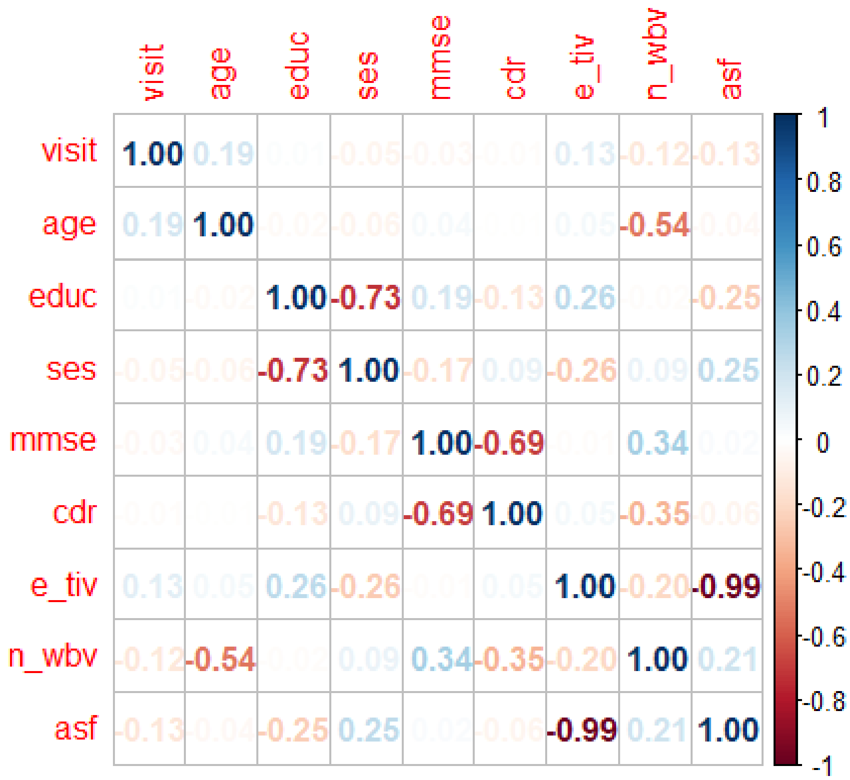
Visualizing dataset

```
## # A tibble: 6 x 15
##   Subjec~1 MRI I~2 Group Visit MR De~3 `M/F` Hand    Age  EDUC   SES  MMSE
CDR
##   <chr>    <chr>   <chr> <dbl>   <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl>
<dbl>
## 1 OAS2_00~ OAS2_0~ Nond~     1       0 M     R        87    14     2    27
0
## 2 OAS2_00~ OAS2_0~ Nond~     2     457 M     R        88    14     2    30
0
## 3 OAS2_00~ OAS2_0~ Deme~     1       0 M     R        75    12    NA    23
0.5
## 4 OAS2_00~ OAS2_0~ Deme~     2     560 M     R        76    12    NA    28
0.5
## 5 OAS2_00~ OAS2_0~ Deme~     3    1895 M     R        80    12    NA    22
0.5
## 6 OAS2_00~ OAS2_0~ Nond~     1       0 F     R        88    18     3    28
0
## # ... with 3 more variables: eTIV <dbl>, nWBV <dbl>, ASF <dbl>, and
abbreviated
## #   variable names 1: `Subject ID`, 2: `MRI ID`, 3: `MR Delay`
```

Number of subjects

```
n_distinct(mydata$'Subject ID')

## [1] 150
```

Correlation between variables




```
##                age        educ         mmse         cdr        e_tiv
n_wbv
## age    1.00000000 -0.02324642  0.03691943 -0.0055411  0.04776723 -
0.53554045
## educ  -0.02324642  1.00000000  0.18874246 -0.1316606  0.26374924 -
0.01808438
## mmse   0.03691943  0.18874246  1.00000000 -0.6945735 -0.01201046
0.33934614
## cdr   -0.00554110 -0.13166065 -0.69457354  1.0000000  0.04668680 -
0.34767971
## e_tiv  0.04776723  0.26374924 -0.01201046  0.0466868  1.00000000 -
0.20316779
## n_wbv -0.53554045 -0.01808438  0.33934614 -0.3476797 -0.20316779
1.00000000
```

Descriptive statistics

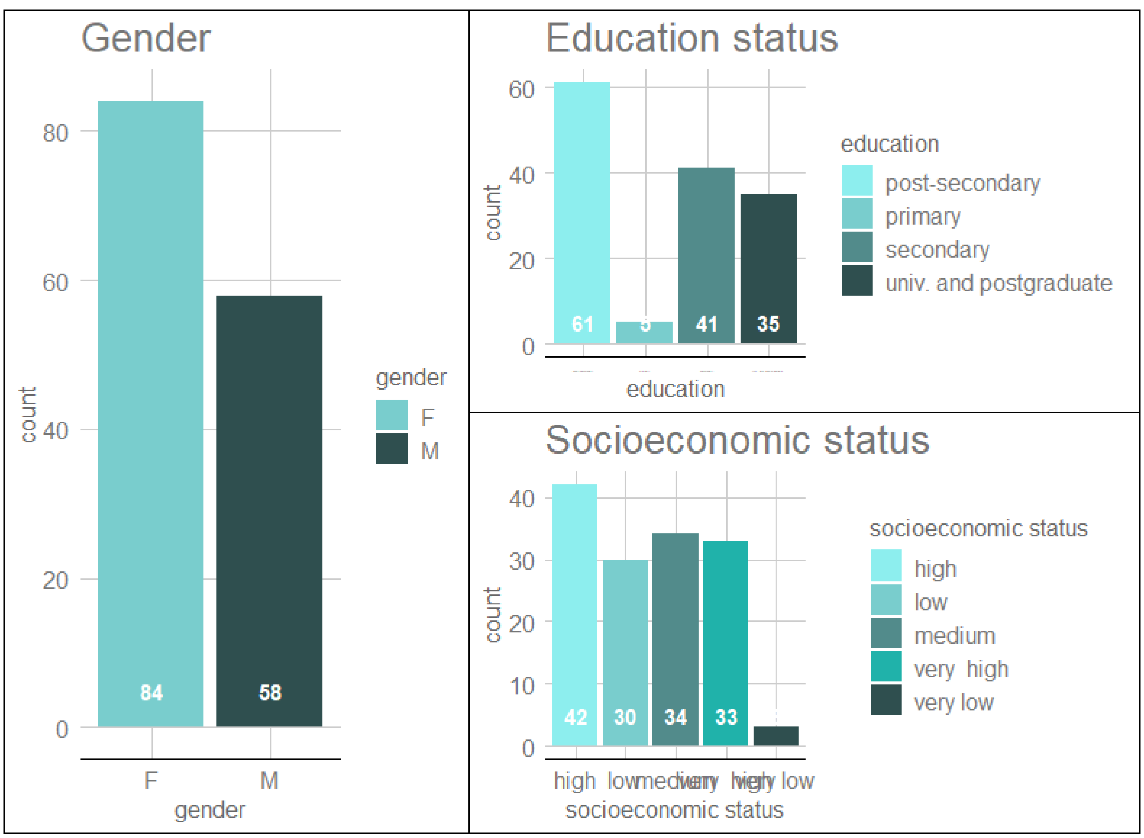
Gender, education and socioeconomic status

- Gender - Gender
- Educ - Years of education
- SES - Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status)

```
## # A tibble: 1 x 5
##   `mean(educ)` `median(educ)` `sd(educ)` `mean(ses)` `sd(ses)`
##          <dbl>          <dbl>      <dbl>       <dbl>     <dbl>
## 1         14.7             15       2.90        2.47      1.13
```

Mean years of schooling (MYS), the average number of completed years of education of a population, is a widely used measure of a country's stock of human capital. The global average is 8.7 years.

Males/Females ratio of dataset is 0.69. Global males/females ratio is 0.98.

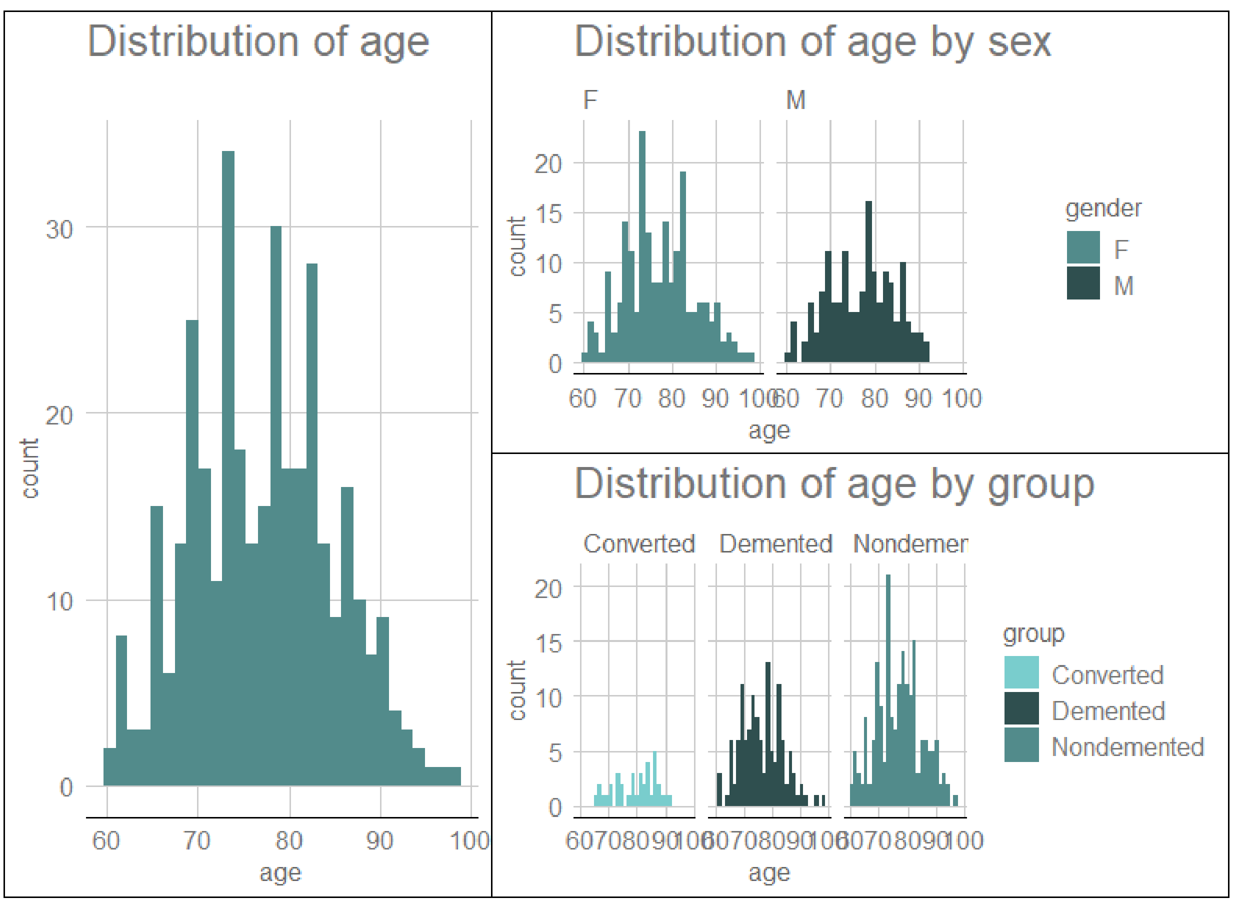


Age

Dementia is more common in people over the age of 65, but in some cases, it can also affect people in their 30s, 40s, or 50s.

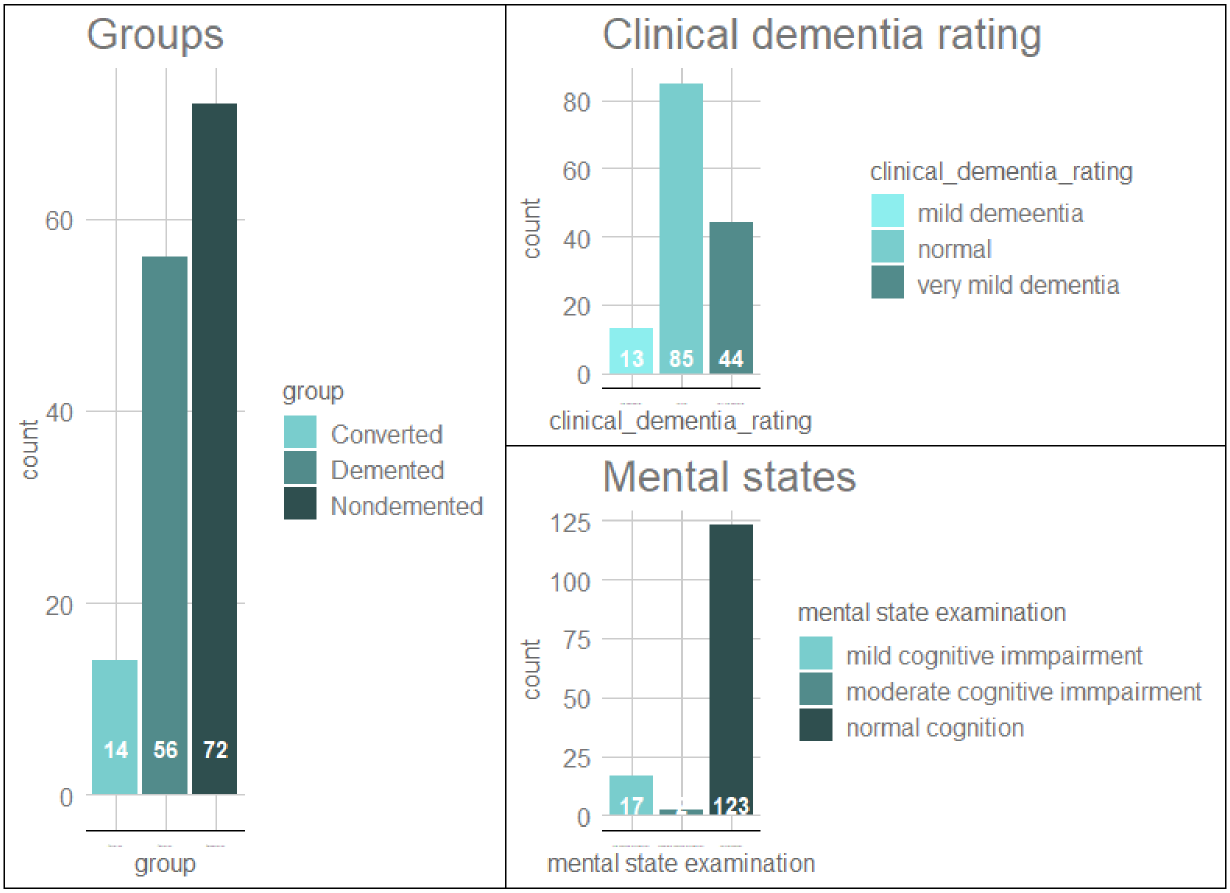
```
## # A tibble: 1 x 5
##   `mean(age)` `median(age)` `sd(age)` `max(age)` `min(age)`
##         <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1        77.1            77      7.80         98         60

## # A tibble: 2 x 6
##   gender `mean(age)` `median(age)` `sd(age)` `max(age)` `min(age)`
##   <chr>        <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1 F             77.2          77        8.05         98         60
## 2 M             76.9          77.5      7.48         92         60

## # A tibble: 3 x 6
##   group        `mean(age)` `median(age)` `sd(age)` `max(age)` `min(age)`
##   <chr>              <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1 Converted           79.8            81      7.43         92         65
## 2 Demented            76.4            76      7.34         98         61
## 3 Nondemented         77.1            77      8.10         97         60
```

Gaussian distribution of age.



Groups: demented, nondemented, converted

72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.



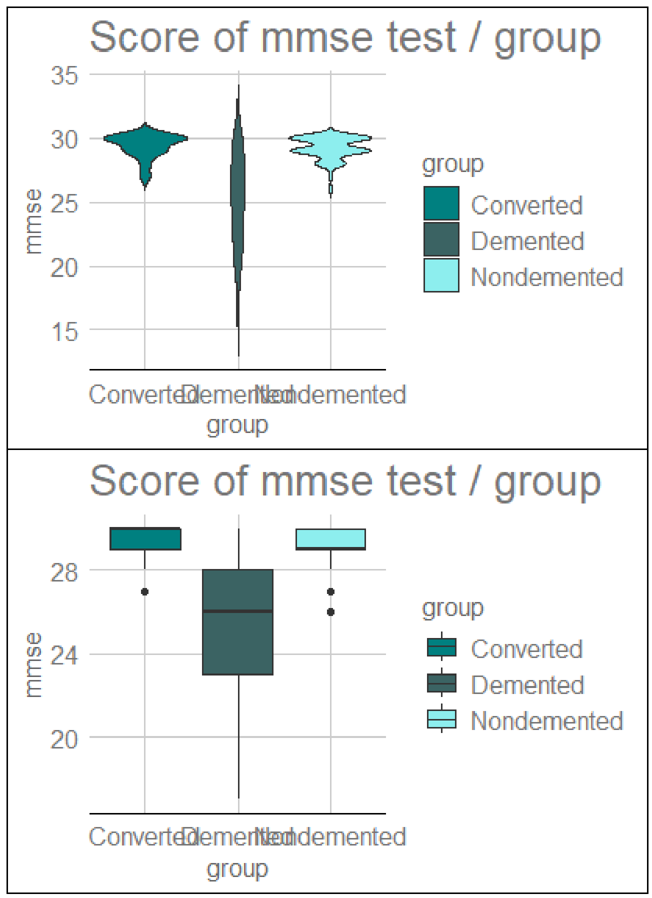
Mini-Mental State Examination (MMSE - test)

```
## # A tibble: 1 x 5
##   `mean(mmse)` `median(mmse)` `sd(mmse)` `max(mmse)` `min(mmse)`
##          <dbl>          <dbl>      <dbl>       <dbl>       <dbl>
## 1         27.5             29       3.65          30           4
```

A Mini-Mental State Examination (MMSE) is a set of 11 questions that doctors and other healthcare professionals commonly use to check for cognitive impairment (problems with thinking, communication, understanding and memory) and is used as part of the process for determining if someone has dementia.

What abilities does the MMSE check? The MMSE can be used to assess 6 areas of mental abilities, including:

- orientation to time and place — knowing the date and where you are;
- attention / concentration;
- short-term memory (recall);
- language skills;
- visuospatial abilities — visual and spatial relationships between objects;
- ability to understand and follow instructions.



CDR - clinical dementia rating

The CDR is a global rating scale for staging patients diagnosed with dementia. The CDR evaluates cognitive, behavioral, and functional aspects of Alzheimer disease and other dementias. Rather than a mental status examination or inventory, the rater simply makes a judgment on six categories based on all the information available. The scoring system for the CDR is somewhat complicated and heavily dependent on the memory scores, but the CDR has good interrater reliability in staging dementia. This instrument is a widely used scale in both Alzheimer disease centers and dementia research;

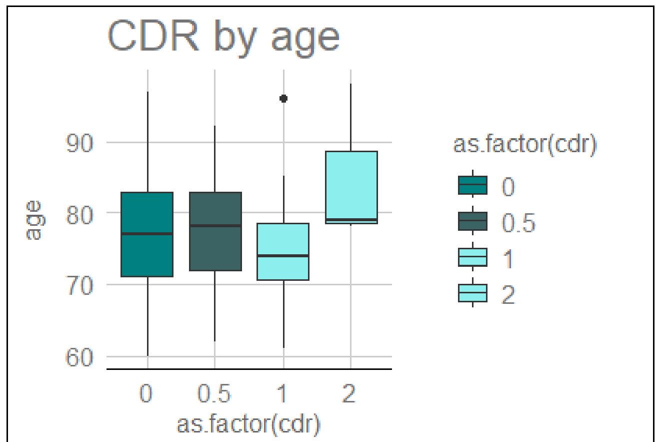
Structured interview with both patient and informant;

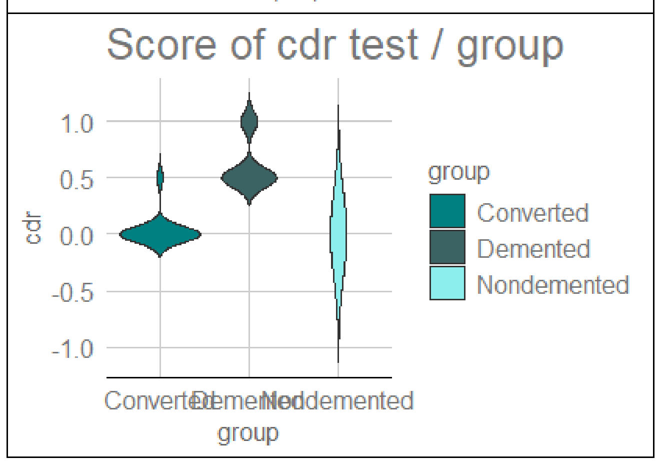
Performance is rated in six domains: memory, orientation, judgment and problem solving, community activities, home and hobbies, and personal care.

Clinical Dementia Rating scale (CDR):

- 0 = no impairment;
- 0.5 = questionable;
- 1 = mild;
- 2 = moderated;
- 3 = severe dementia.

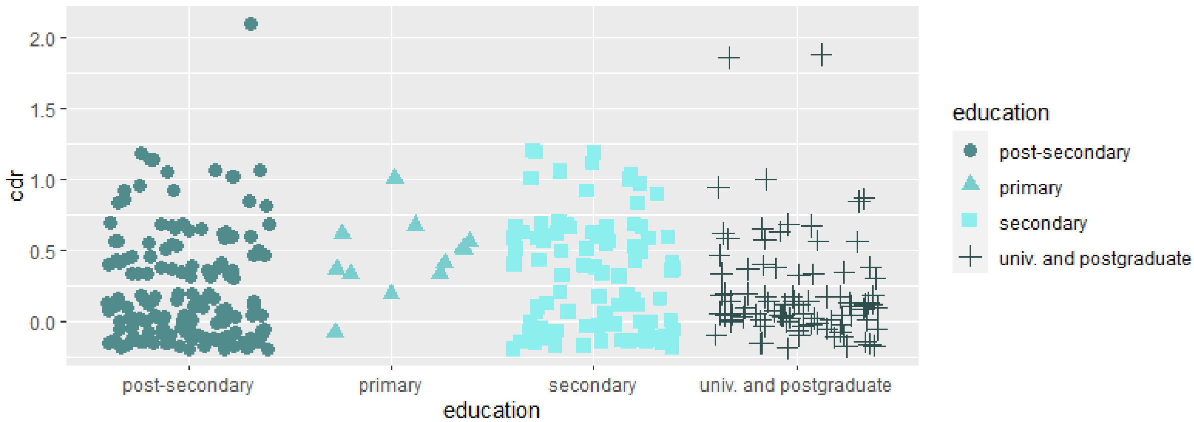
```
## # A tibble: 1 x 5
##   `mean(mmse)` `median(mmse)` `sd(mmse)` `max(mmse)` `min(mmse)`
##          <dbl>          <dbl>      <dbl>       <dbl>       <dbl>
## 1         27.5             29       3.65          30           4
```

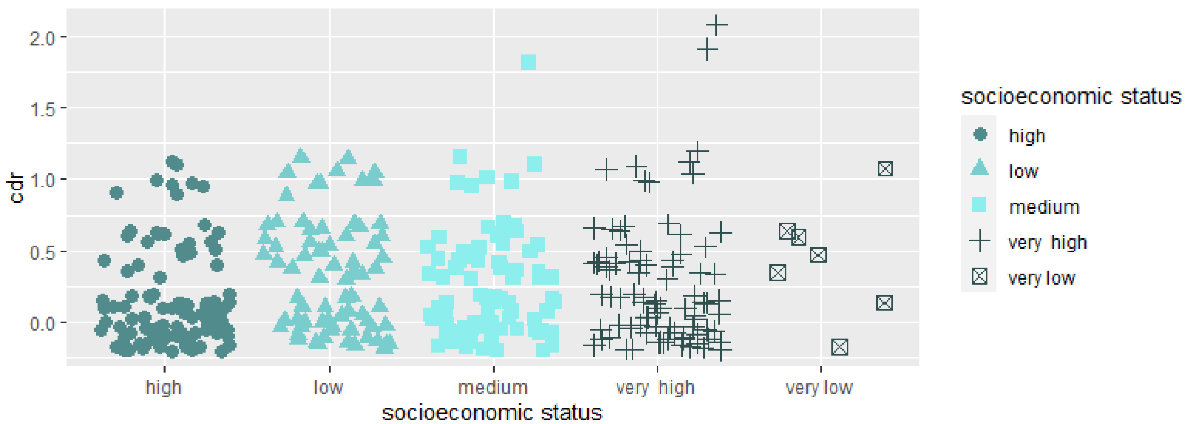




Education and dementia

Based on the data, I have observed that the CDR score is higher on subjects with primary education and also the CDR score is higher than 0 on subjects with a very low socioeconomic status.

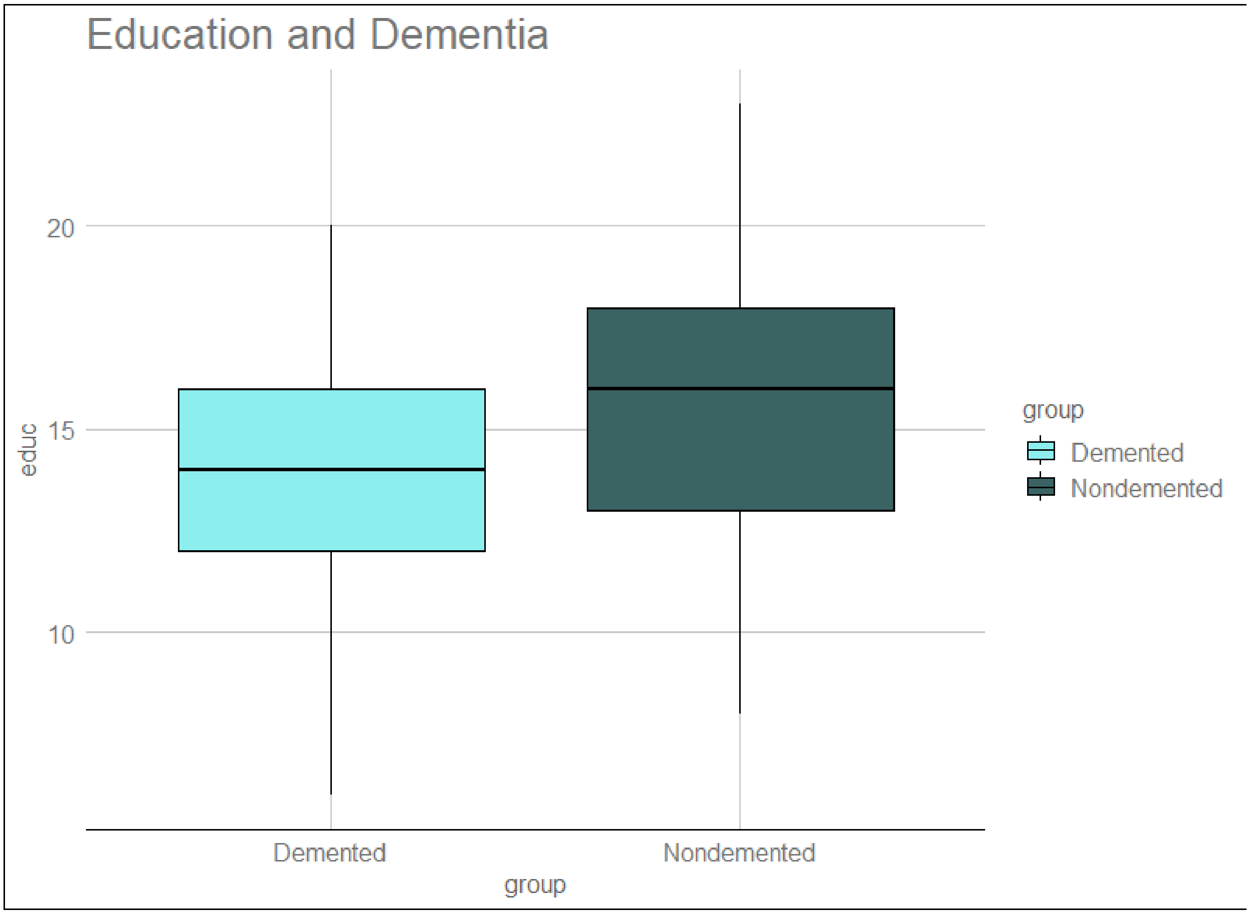




The plot below showed me that people with dementia have fewer years of education as background than healthy people.

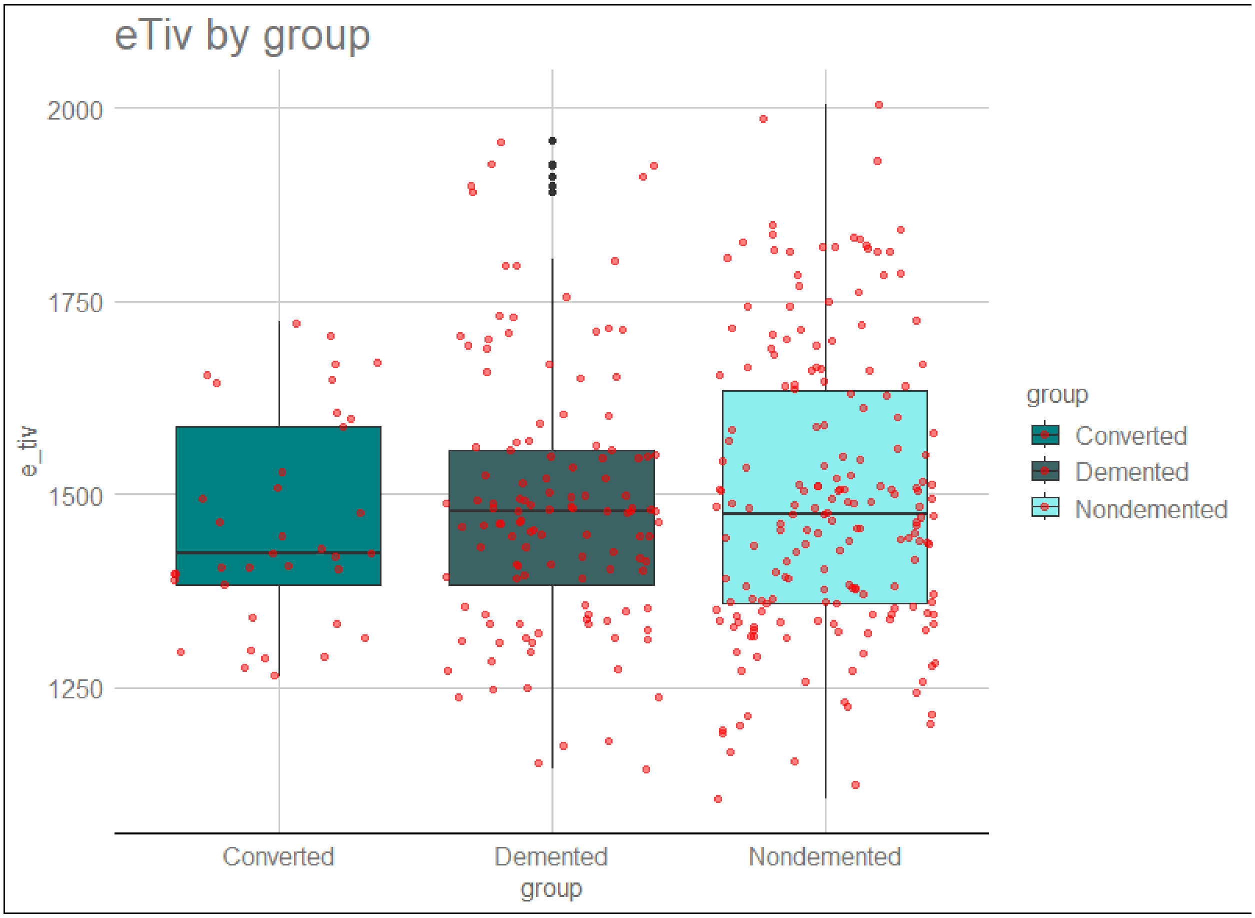
Searching for other scientific papers, I have found that over the past decade, studies on dementia have consistently showed that the more time you spend in education, the lower your risk of dementia. For each additional year of education there is an 11% decrease in risk of developing dementia, the studies reports.

However, these studies have been unable to determine whether or not education - which is linked to higher socioeconomic status and healthier lifestyles - protects the brain against dementia.


Education and Dementia
20
15
10
educ
Demented
Nondemented
group
group
Demented
Nondemented

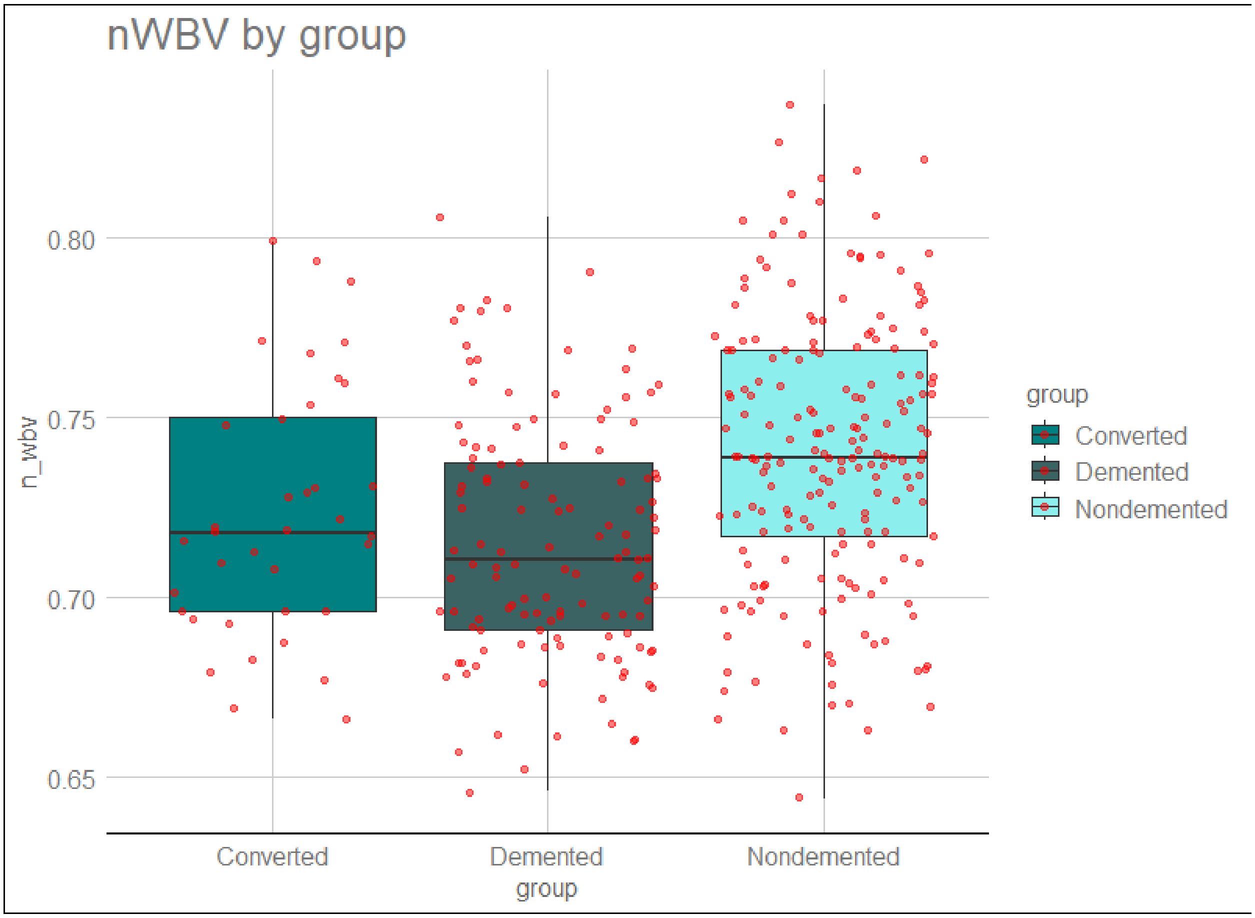
Estimated total intracranial volume (eTIV)

The ICV measure, sometimes referred to as total intracranial volume (TIV), refers to the estimated volume of the cranial cavity as outlined by the supratentorial dura matter or cerebral contour when dura is not clearly detectable. (Source: *PubMed Central® website.*)



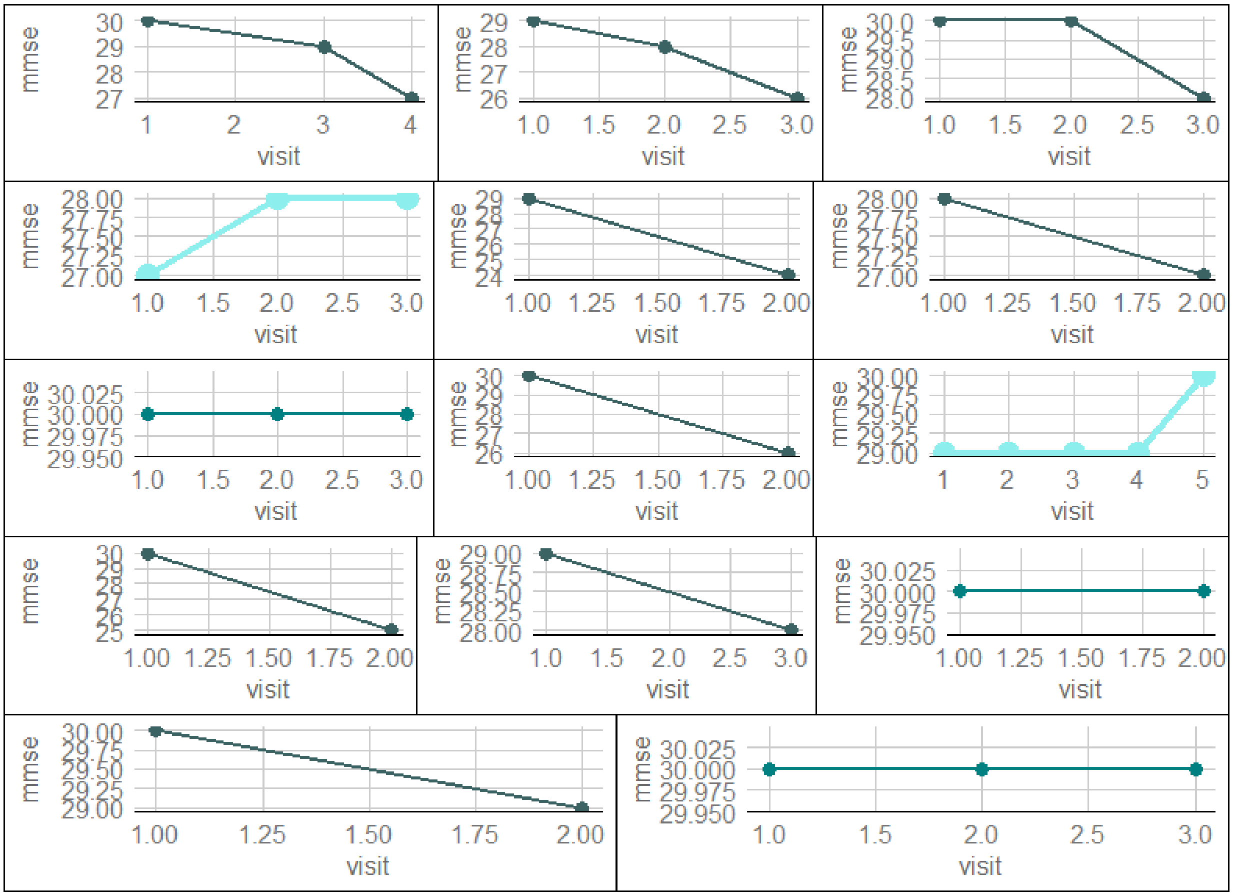
nWBV

Normalized whole brain volume (nWBV), reflecting the percentage of the intracranial cavity occupied by brain, was obtained using previously established methods. (Source: *PubMed Central® website.*)



Longitudinal analysis

In a longitudinal study, researchers repeatedly examine the same individuals to detect any changes that might occur over a period of time. The benefit of a longitudinal study is aiming to detect developments or changes in the characteristics of the target population at both the group and the individual level.



- 14 subjects
- two or more visits, separated by at least one year. max visits - > 5
- first visit median age: 78 (7.6 years), 5th visit median age: 86
- maximum age: 92 (4th visit), minimum age: 65 (1st visit)
- CDR median of first visit is 0 (subjects started developing dementia after 65 y.o)
- starting with 73, people start to develop mild to moderate symptoms of dementia
- attention / concentration and orientation continue to decrease over the visits (mmse test)
- there is a single subjects who went to 5 visits and also only 2 subjects with 4 visits
- in 9 out of 14 cases the mmse score decreased over time; in 3 cases, the score was constant during multiple visits
- there are two subjects that improved their mmse score over visits (from 29 points to 30 and from 27 to 28).

```
## # A tibble: 5 x 2
##   visit     n
##   <dbl> <int>
## 1     1    14
## 2     2    12
## 3     3     8
## 4     4     2
## 5     5     1
```

```
## # A tibble: 1 x 5
##   `mean(age)` `median(age)` `max(age)` `min(age)` `sd(age)`
##         <dbl>         <dbl>      <dbl>      <dbl>     <dbl>
## 1        79.8            81         92         65      7.43
```

```
## # A tibble: 1 x 8
##   `mean(cdr)` `median(cdr)` `max(cdr)` min(cdr~1 sd(cd~2 mean(~3 max(m~4
min(m~5
##         <dbl>         <dbl>      <dbl>     <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
## 1       0.257           0.5        0.5         0   0.253    28.7      30
24
## # ... with abbreviated variable names 1: `min(cdr)`, 2: `sd(cdr)`,
## #   3: `mean(mmse)`, 4: `max(mmse)`, 5: `min(mmse)`
```

Grouped by visit

```
## # A tibble: 5 x 6
##   visit `mean(age)` `max(age)` `min(age)` `median(age)` `sd(age)`
##   <dbl>       <dbl>      <dbl>      <dbl>         <dbl>     <dbl>
## 1     1        77.1         87         65          78.5      7.69
## 2     2        78.8         88         67          80        7.08
## 3     3        83           91         75          82.5      6.12
## 4     4        88           92         84          88        5.66
## 5     5        86           86         86          86       NA
```

```
## # A tibble: 5 x 9
##   visit `mean(cdr)` median(cdr~1 sd(cd~2 max(c~3 min(c~4 mean(~5 media~6
sd(mm~7
##   <dbl>       <dbl>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
## 1     1      0.0357          0     0.134     0.5     0      29.4    30
0.929
## 2     2      0.333           0.5   0.246     0.5     0      28      28.5
2.09
## 3     3      0.438           0.5   0.177     0.5     0      28.5    28.5
1.31
## 4     4      0.5             0.5   0         0.5     0.5    28      28
1.41
## 5     5      0.5             0.5  NA         0.5     0.5    30      30
NA
## # ... with abbreviated variable names 1: `median(cdr)`, 2: `sd(cdr)`,
```

```
## #   3: `max(cdr)`, 4: `min(cdr)`, 5: `mean(mmse)`, 6: `median(mmse)`,
## #   7: `sd(mmse)`
```

Modeling

Supervised machine learning algorithms uncover insights, patterns, and relationships from a labeled training dataset – that is, a dataset that already contains a known value for the target variable for each record. Because you provide the machine learning algorithm with the correct answers for a problem during training, the algorithm is able to "learn" how the rest of the features relate to the target, enabling you to uncover insights and make predictions about future outcomes based on historical data.

For developing a model that predicts if a person is demented or nondemented, I will use a classification algorithm called Support Vector Machine

Preprocessing

Because there are three target outputs and a limited number of instances to train for developing a model, I have decided to transform any "Converted" to "Demented" if CDR is greater or equal to 0.5 and to "Nondemented" if CDR is less then 0.5. The new dataset (mdata) will only have one target variable with two outputs: "Demented" and "Nondemented".

```
## # A tibble: 3 x 2
##   group           n
##   <chr>       <int>
## 1 Converted      37
## 2 Demented      124
## 3 Nondemented   190

mdata %>% count(group)

## # A tibble: 2 x 2
##   group           n
##   <chr>       <int>
## 1 Demented      145
## 2 Nondemented   206
```

Because the number of women was significantly higher than men, I have chosen to equalize the ratio between them. The number of men is 148, so I have chosen a random sample of 148 females and created a new dataset.

```
## # A tibble: 2 x 2
##   gender     n
##   <chr>  <int>
## 1 F        203
## 2 M        148

f_sample <- sample_n(females, 148) # 148 females samples
mdata_eq <- rbind(males, f_sample)
```

```
shuffled_data= mdata_eq[sample(1:nrow(mdata_eq)), ]
mdata_eq <- shuffled_data

mdata_eq %>% count(gender)
```

```
## # A tibble: 2 x 2
##   gender     n
##   <chr>  <int>
## 1 F        148
## 2 M        148
```

Scalling, centering, splitting training and testing

```
temp<-as.data.frame(cbind(numeric_variables, factor_variables))
temp<-temp[,c(-1,-8,-9,-10)]
train_set <- round(0.9 * nrow(temp)) # 90 % training 10% testing
indices <- sample(1:nrow(temp), train_set)
train <- temp[indices,]
test <- temp[-indices,]

head(train)
```

```
##             age       educ ses       mmse         n_wbv        asf gender.M
cdr
## 251 -1.0249443  0.4190924   2  0.1627892  0.213100144 -1.6737742        1
0
## 55  -0.6363232 -0.2569368   3  0.1627892  1.269677071 -0.1627711        0
0
## 294 -0.3772425  0.4190924   2  0.6981848  0.001784758 -0.6783740        1
0
## 248  0.1409189  1.0951216   1  0.6981848 -1.662323902  0.2525758        0
0
## 117  0.1409189 -0.9329659   3 -1.9787932 -1.054792169  0.9114018        0
0.5
## 209 -1.5431057 -0.5949514   2  0.6981848  1.903623227  0.9543687        0
0
```

```
head(test)
```

```
##             age       educ ses       mmse      n_wbv         asf gender.M
cdr
## 4   0.01137854  1.0951216   1 -0.6403042  1.3753348 -0.64256826        1
0.5
## 19  0.52953996  1.0951216   2  0.4304870 -1.1076210  0.07354697        0
0
## 26 -0.63632322  1.0951216   1  0.4304870 -1.1076210 -1.43745615        1
0
## 32  1.30678208 -0.2569368   2 -0.1049086 -0.8698912 -2.15357138        1
0
## 42  1.56586279 -0.9329659   4 -0.1049086  0.1074425  0.97585215        0
0
```

```
## 45 -0.63632322 -0.9329659   2 -0.1049086 -1.1868643  0.30270384        1
1
```

Support vector machine

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new inputs.

Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands NORMALLY).

```
set.seed(11)
ctrl<-trainControl(method = "cv", number = 12)
svm.mod<-train(cdr~., data = train, method = "svmRadial", metric =
"Accuracy", trControl = ctrl, tuneLength = 15)
pred1 <- predict(svm.mod, newdata = test[,-9], cost = 100, gamma = 1)
confusionMatrix(pred1, test$cdr)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0 0.5  1  2
##        0   18   4  0  0
##        0.5  1   4  1  0
##        1    0   0  2  0
##        2    0   0  0  0
##
## Overall Statistics
##
##                Accuracy : 0.8
##                  95% CI : (0.6143, 0.9229)
##     No Information Rate : 0.6333
##     P-Value [Acc > NIR] : 0.03992
##
##                   Kappa : 0.5794
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 0.5 Class: 1 Class: 2
## Sensitivity            0.9474     0.5000  0.66667       NA
## Specificity            0.6364     0.9091  1.00000        1
## Pos Pred Value         0.8182     0.6667  1.00000       NA
## Neg Pred Value         0.8750     0.8333  0.96429       NA
## Prevalence             0.6333     0.2667  0.10000        0
## Detection Rate         0.6000     0.1333  0.06667        0
```

Evaluation

Taking in consideration the small amount of labeled data:

- Accuracy: 80%
- Predicted 18 correct “CDR - 0” out of 22
- Predicted 4 correct “CDR – 0.5” out of 6
- Predicted 2 correct “CDR – 1” out of 2