Heart Disease Prediction and Logistic Regression

Rob van Mechelen

9/10/2022

Data

As first step, include libraries to run the code:

```
library(ggplot2)
library(ggpubr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.6      v dplyr   1.0.10
## v tidyr   1.1.4      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(table1)
```

```
##
## Attaching package: 'table1'
##
## The following objects are masked from 'package:base':
##
##     units, units<-
```

As a next step load the data set:

```
HD <- read.csv("Heart_Disease_Prediction.csv")
colnames(HD)  # 15 colnames
```

```
##  [1] "Age"                    "Sex"
##  [3] "Chest.pain.type"        "BP"
##  [5] "Cholesterol"            "FBS.over.120"
##  [7] "EKG.results"            "Max.HR"
##  [9] "Exercise.angina"        "ST.depression"
## [11] "Slope.of.ST"            "Number.of.vessels.fluro"
## [13] "Thallium"               "Heart.Disease"
```

- Colnames for risk factors: The colnames *Age*, *Sex*, *BP*, *Cholesterol*, and *FBS.over.120* (diabetes mellitus), all refer to well known risk factors for heart disease, in particular coronary artery disease (CAD). The Framingham study published more than 20 years ago, marked these variables as most important to predict the presence of CAD.

see: https://www.framinghamheartstudy.org/

- Colnames for CAD: The colnames *Chest.pain.type* and *EKG.results* refer to clinical characteristics and EKG abnormalities in patients suspected of CAD at rest.
- Colnames for the results of exercise tests: The colnames *Max.HR*, *Exercise.angina*, *ST.depression* and *slope.of.ST* and *Thallium* refer to the results of patients, who underwent an exercise test.
- Colname for severity of CAD: The colname *number.of.vessels.fluro* refers to the result of an invasive coronary angiography test done by fluroscopy.
- Colname for the outcome of the study: The colname *Heart.Disease* is the standard outcome of this study. Either heart disease was present or not in this study population. All other variables were tested against the *presence* or *absence* of this variable.

Type of study

This is a cross-sectional study, where 13 variables are tested against an outcome variable *Heart.Disease*. In a total of 270 patients, there were 150 patients without heart disease and 120 patients with heart disease.

```
table(HD$Heart.Disease)
```

```
##
##  Absence Presence
##      150      120
```

```
addmargins(table(HD$Heart.Disease))
```

```
##
##  Absence Presence      Sum
##      150      120      270
```

Analysis of risk factors

As mentioned, the colnames *Age*, *Sex*, *BP*, *Cholesterol*, and *FBS.over.120* (Diabetes Mellitus), all refer to well known risk factors for heart disease. The table gives an overview of the risk factors in the absence or presence of heart disease observed in the study.

```
# change the Sex variable to a factor variable
HD$Sex <- factor(HD$Sex,
                 levels=c(0,1),
                 labels=c("female",
                          "male"))

# add a status variable to show the outcome
status <- ifelse(HD$Heart.Disease == "Absence", 0,1)
```

```
HD$status <- status
HD$status <- factor(HD$status,
                    levels=c(0,1),
                    labels=c("no HD",
                             "HD"))

# change the FBS.over.120 variable to a factor variable
HD$FBS.over.120<- factor(HD$FBS.over.120,
                     levels=c(0,1),
                     labels=c("no", "yes"))
```

Main risk factors and outcome

```
table1(~ Age + Sex + BP + Cholesterol +
         FBS.over.120 | status, overall = "Total", data=HD)
```

	no HD	HD	Total
	(N=150)	(N=120)	(N=270)
Age			
Mean (SD)	52.7 (9.51)	56.6 (8.12)	54.4 (9.11)
Median [Min, Max]	52.0 [29.0, 76.0]	58.0 [35.0, 77.0]	55.0 [29.0, 77.0]
Sex			
female	67 (44.7%)	20 (16.7%)	87 (32.2%)
male	83 (55.3%)	100 (83.3%)	183 (67.8%)
BP			
Mean (SD)	129 (16.5)	134 (19.1)	131 (17.9)
Median [Min, Max]	130 [94.0, 180]	130 [100, 200]	130 [94.0, 200]
Cholesterol			
Mean (SD)	244 (54.0)	256 (48.0)	250 (51.7)
Median [Min, Max]	236 [126, 564]	256 [149, 409]	245 [126, 564]
FBS.over.120			
no	127 (84.7%)	103 (85.8%)	230 (85.2%)
yes	23 (15.3%)	17 (14.2%)	40 (14.8%)

This table gives an overview of the risk factors in the absence or presence of heart disease observed in the study.

What we miss are the p values. To create a function to compute the p-value for continuous or categorical variables.I took the *pvalue* function from the Cran website, see https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html

```
pvalue <- function(x, ...) {
  # Construct vectors of data y, and groups (strata) g
  y <- unlist(x)
  g <- factor(rep(1:length(x), times=sapply(x, length)))
  if (is.numeric(y)) {
    # For numeric variables, perform a standard 2-sample t-test
    p <- t.test(y ~ g)$p.value
```

```
  } else {
    # For categorical variables, perform a chi-squared test of independence
    p <- chisq.test(table(y, g))$p.value
  }
  # Format the p-value, using an HTML entity for the less-than sign.
  # The initial empty string places the output on the line below the variable label.
  c("", sub("<", "&lt;", format.pval(p, digits=3, eps=0.001)))
}
```

p value function code: Next, we put this function in the table1 code and put a *p values* column in place of the *total* column.

The result is a table with p values replacing the *total* column.

```
table1(~ Age + Sex + BP + Cholesterol +
         FBS.over.120 | status, data= HD,
         overall=F, extra.col=list(`P-value`=pvalue))
```

	no HD	HD	P-value
	(N=150)	(N=120)	
Age			
Mean (SD)	52.7 (9.51)	56.6 (8.12)	<0.001
Median [Min, Max]	52.0 [29.0, 76.0]	58.0 [35.0, 77.0]	
Sex			
female	67 (44.7%)	20 (16.7%)	<0.001
male	83 (55.3%)	100 (83.3%)	
BP			
Mean (SD)	129 (16.5)	134 (19.1)	0.012
Median [Min, Max]	130 [94.0, 180]	130 [100, 200]	
Cholesterol			
Mean (SD)	244 (54.0)	256 (48.0)	0.0497
Median [Min, Max]	236 [126, 564]	256 [149, 409]	
FBS.over.120			
no	127 (84.7%)	103 (85.8%)	0.924
yes	23 (15.3%)	17 (14.2%)	

From this table, it is clear that *Age*, *Sex* are associated with p values less than 0.001. *BP* and *Cholesterol* are associated with p values less than 0.01 and 0.05, whereas *fasting blood sugar over 120* is associated with a p value of 0.924 (ns)

Finally, we could change labels and add units of measurements to the table to give it a professional look, as presented in medical journals.

```
label(HD$BP)     <- "Blood Pressure"
label(HD$Cholesterol)  <- "Cholesterol"
label(HD$FBS.over.120) <- "FBS > 120 mg/dl"

units(HD$Age)      <- "years"
units(HD$BP)      <- "mm Hg"
units(HD$Cholesterol) <- "mg/dl"
units(HD$FBS.over.120) <- "mg/dl"
```

```
table1(~ Age + Sex + BP + Cholesterol +
         FBS.over.120 | status, data= HD,
       overall=F, extra.col=list(`P-value`=pvalue))
```

	no HD	HD	P-value
	(N=150)	(N=120)	
Age (years)			
Mean (SD)	52.7 (9.51)	56.6 (8.12)	<0.001
Median [Min, Max]	52.0 [29.0, 76.0]	58.0 [35.0, 77.0]	
Sex			
female	67 (44.7%)	20 (16.7%)	<0.001
male	83 (55.3%)	100 (83.3%)	
Blood Pressure (mm Hg)			
Mean (SD)	129 (16.5)	134 (19.1)	0.012
Median [Min, Max]	130 [94.0, 180]	130 [100, 200]	
Cholesterol (mg/dl)			
Mean (SD)	244 (54.0)	256 (48.0)	0.0497
Median [Min, Max]	236 [126, 564]	256 [149, 409]	
FBS > 120 mg/dl (mg/dl)			
no	127 (84.7%)	103 (85.8%)	0.924
yes	23 (15.3%)	17 (14.2%)	

The table is complete

Labels are understandable and the units of measurements were added to age, gender, blood pressure, cholesterol and blood sugar.

Graphical representation of the Risk Factor data

Age

Let start with the *Age* variable. Age distribution in the table looks like this:

```
table1(~ Age | status, data= HD,
       overall=F, extra.col=list(`P-value`=pvalue))
```

	no HD	HD	P-value
	(N=150)	(N=120)	
Age (years)			
Mean (SD)	52.7 (9.51)	56.6 (8.12)	<0.001
Median [Min, Max]	52.0 [29.0, 76.0]	58.0 [35.0, 77.0]	

The median age for patients without heart disease equals 52 and the median age for patients with heart disease is 58 (black horizontal lines). The mean ages are 52.7 and 56.6 respectively with a standard deviation of 9.51 and 8.12. The white box contains the ages of patients without heart disease (25-75%) and the grey box (25-75%) all ages in patients with heart disease.

Sex

The gender variable looks in the table like this:

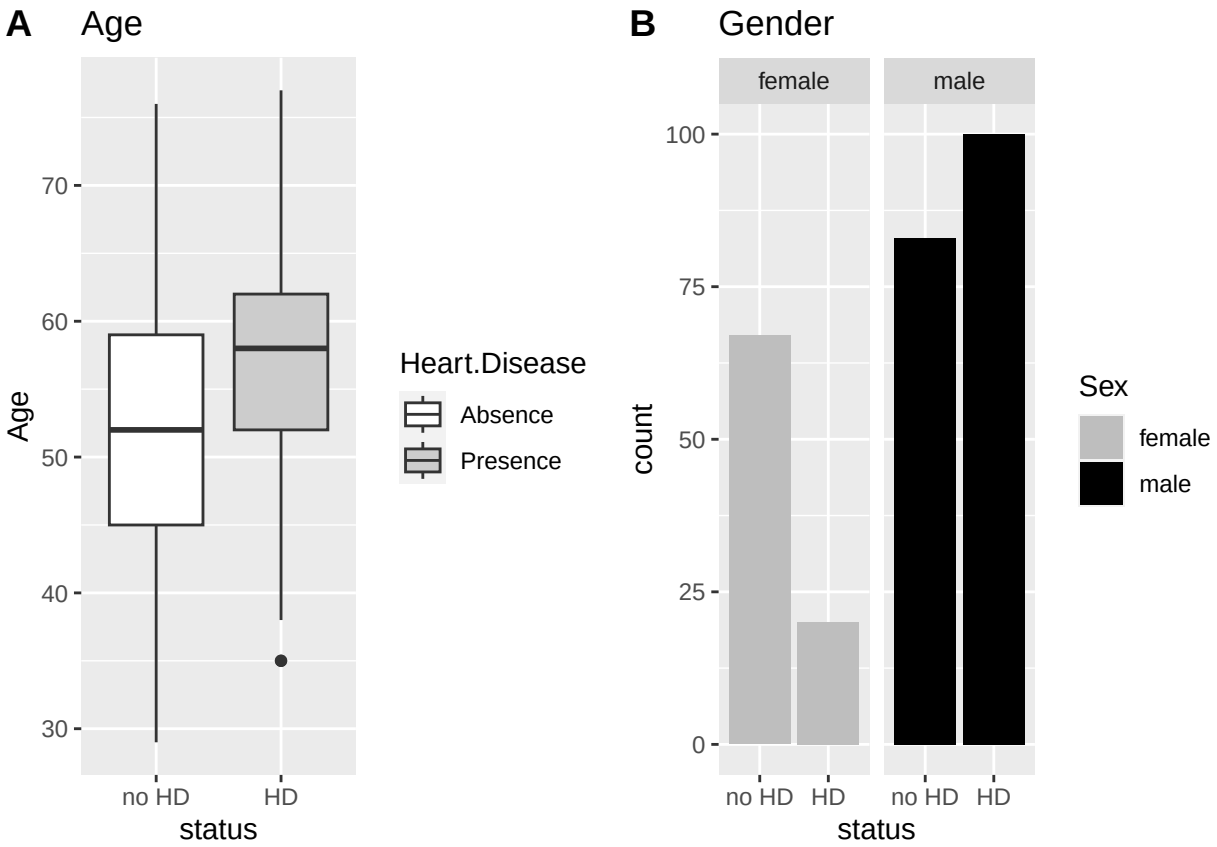
```
table1(~ Sex | status, data= HD,
       overall=F, extra.col=list(`P-value`=pvalue))
```

	no HD	HD	P-value
	(N=150)	(N=120)	
Sex			
female	67 (44.7%)	20 (16.7%)	<0.001
male	83 (55.3%)	100 (83.3%)	

There are 87 females and 183 male patient in the dataset. Of the females 67 have no heart disease and around one third 20 patients have heart disease. This is in strong contrasts with the 183 male patients of whom 83 do not have heart disease and 100 patients have. In a graphical representation this contrast is immediately obvious.

Graphical representation:

```
p1 <-ggplot(HD, aes(x = status, y = Age, fill = Heart.Disease)) +
  geom_boxplot() +
  scale_fill_manual(values=c("#FFFFFF",
                             "#CCCCCC")) +
  labs(title = "Age")



p2 <- ggplot(data = HD) +
  geom_bar(mapping = aes(x = status, fill = Sex)) +
  scale_fill_manual(values=c("grey",
                             "black")) +
  labs(title = "Gender") +
  facet_wrap(~Sex)

ggarrange(p1,p2,
          labels = c("A", "B"),
          ncol = 2, nrow = 1)
```



In this data set of 270 patients, there are 2 times more men than women included (183 versus 87). Of the 87 women around one third (20) has heart disease. Of the 183 men in this study more than 50 percent have heart disease (100/183).

Other risk factors

The other risk factors are : blood pressure measurements, blood cholesterol tests and blood sugar tests. Let us now focus on these 3 risk factors. In the table these variables look like this:

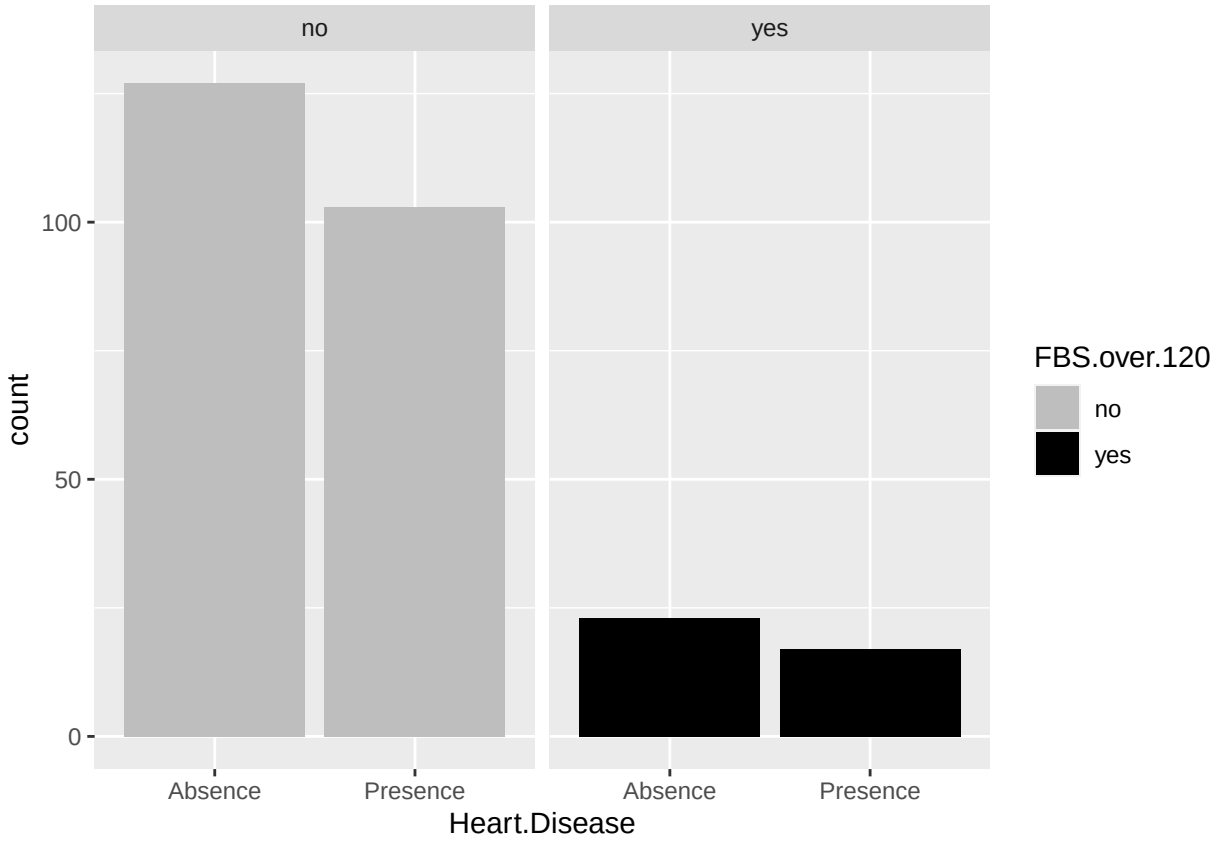
```
table1(~ BP + Cholesterol +
         FBS.over.120 | status, data= HD,
       overall=F, extra.col=list(`P-value`=pvalue))
```

	no HD	HD	P-value
	(N=150)	(N=120)	
Blood Pressure (mm Hg)			
Mean (SD)	129 (16.5)	134 (19.1)	0.012
Median [Min, Max]	130 [94.0, 180]	130 [100, 200]	
Cholesterol (mg/dl)			
Mean (SD)	244 (54.0)	256 (48.0)	0.0497
Median [Min, Max]	236 [126, 564]	256 [149, 409]	
FBS > 120 mg/dl (mg/dl)			
no	127 (84.7%)	103 (85.8%)	0.924
yes	23 (15.3%)	17 (14.2%)	

As one can see from the table, between patient with and without heart disease, there was no significant difference in number of patients with blood sugar levels over 120 mg/dl or not. It even looks like there were less patients with high levels in the heart disease group than in the group without heart disease.

How do these variables look in a graphical representation ?

```
ggplot(data = HD) +
  geom_bar(mapping = aes(x = Heart.Disease, fill = FBS.over.120)) +
  scale_fill_manual(values=c("grey",
                             "black")) +
  facet_wrap(~FBS.over.120)
```

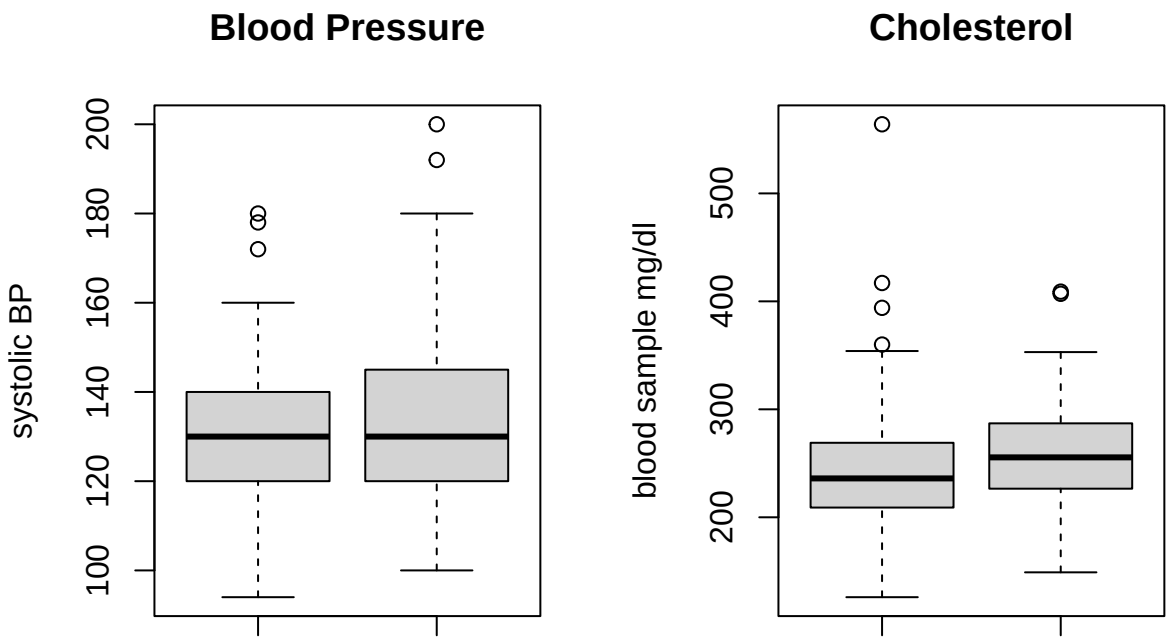


There were 23 patients with a fasting blood sugar over 120 mg/dl in the group without heart disease(n=150) and 17 patients in the group with heart disease (n=120). Therefore, there were less patients with high sugar levels in the heart disease group than in the group without heart disease (14.2% vs 15.3%).

Next, we look at the risk factors *blood pressure* and *blood cholesterol levels*. The graphs are presented as boxplots next to each other.

```
presHD <- HD %>%
  filter(Heart.Disease == "Presence")
noHD <- HD %>%
  filter(Heart.Disease == "Absence")

par(mfrow=c(1,2))
boxplot(noHD$BP,presHD$BP, main = "Blood Pressure",
        ylab = "systolic BP")
```

```
boxplot(noHD$Cholesterol, presHD$Cholesterol, main = "Cholesterol",
        ylab = "blood sample mg/dl")
```




```
par(mfrow=c(1,1))
```

It is obvious, from the boxplots that the median *BP* and median *Cholesterol* do not differ very much in patients without and with heart disease in this study. The differences between *Age* in patients with and without heart disease are much more outspoken. In addition, the differences in male *Sex*, was observed in the presence or absence of heart disease were also large.

In summary, in this cross-sectional study, risk factors *Age* and male *Sex* were much more often observed in heart disease, than in the absence of heart disease *Blood pressure* and *Cholesterol levels* showed statistically significant differences but were less outspoken and *FBG.over.120* as a surrogate for the presence of diabetes did not reach a sifnificant level.

Analysis of Chest Pain and EKG abnormalities

The variables *Chest.pain.type* and *EKG.results* are discussed in this section. First, we create a table with these variables.To do this the class of *Chest.pain.type* and *EKG.results* must be changed from **integer** to **factor**.

```
class(HD$Chest.pain.type)
```

```
## [1] "integer"
```

```
# retuns integer
table(HD$Chest.pain.type)
```

```
##
##   1   2   3   4
##  20  42  79 129
```

```
# returns the integers 1,2,3,4 change to factor variables
HD$Chest.pain.type <- factor(HD$Chest.pain.type,
                 levels=c(1,2,3,4))

class(HD$EKG.results)
```

```
## [1] "integer"
```

```
# returns integer
table(HD$EKG.results)
```

```
##
##   0   1   2
## 131   2 137
```

```
HD$EKG.results <- factor(HD$EKG.results,
                            levels=c(0,1,2))
```

Next, insert the table1 code:

```
table1(~ Chest.pain.type + EKG.results | status, data= HD,
       overall=F, extra.col=list(`P-value`=pvalue))
```

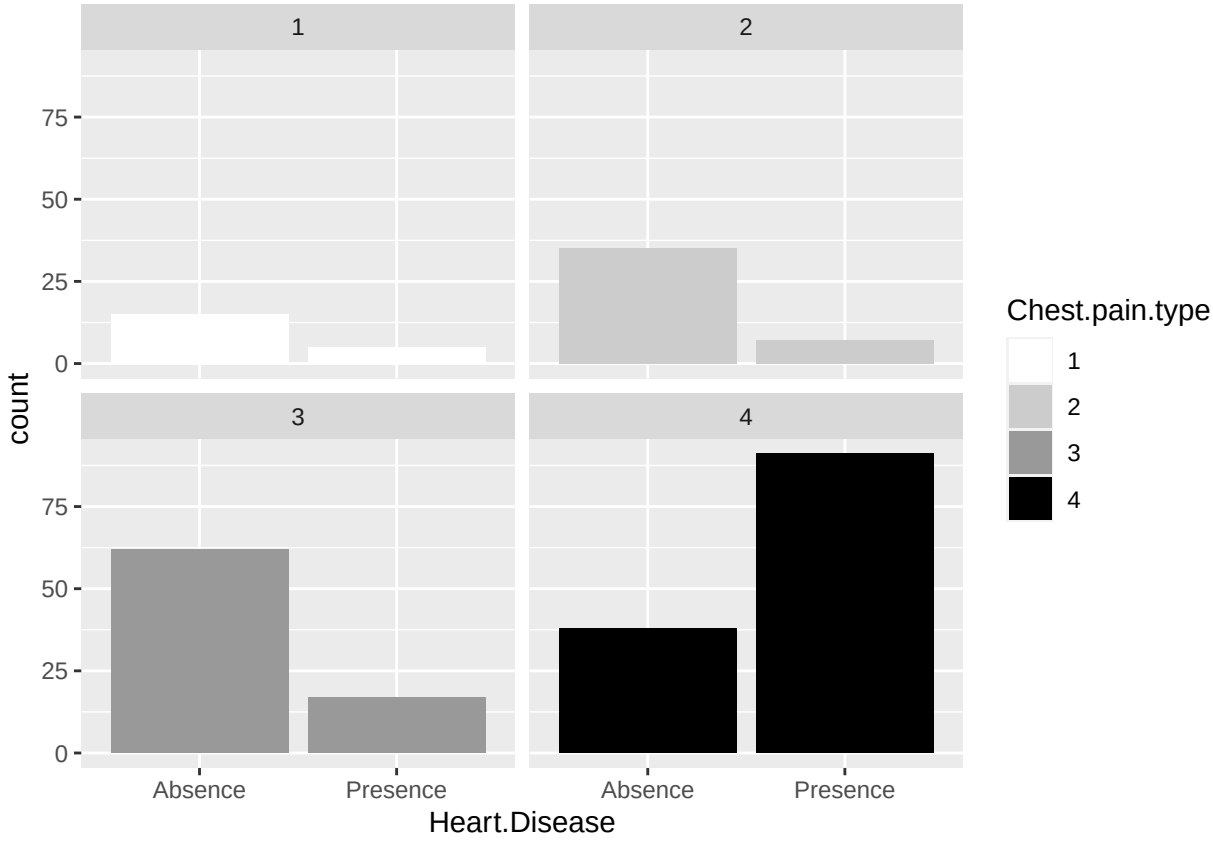
```
## Warning in chisq.test(table(y, g)): Chi-squared approximation may be incorrect
```

	no HD	HD	P-value
	(N=150)	(N=120)	
Chest.pain.type			
1	15 (10.0%)	5 (4.2%)	<0.001
2	35 (23.3%)	7 (5.8%)	
3	62 (41.3%)	17 (14.2%)	
4	38 (25.3%)	91 (75.8%)	
EKG.results			
0	85 (56.7%)	46 (38.3%)	0.0112
1	1 (0.7%)	1 (0.8%)	
2	64 (42.7%)	73 (60.8%)	

For *EKG.results* a warning message appears: Chi-squared approximation may be incorrect. This is correct because the numbers of EKG.results = 1 are to small to generate a reliable Chi-squared approximation. Therefore, we have to omit the EKG.results in the analysis.

The graphs of these findings look like this:

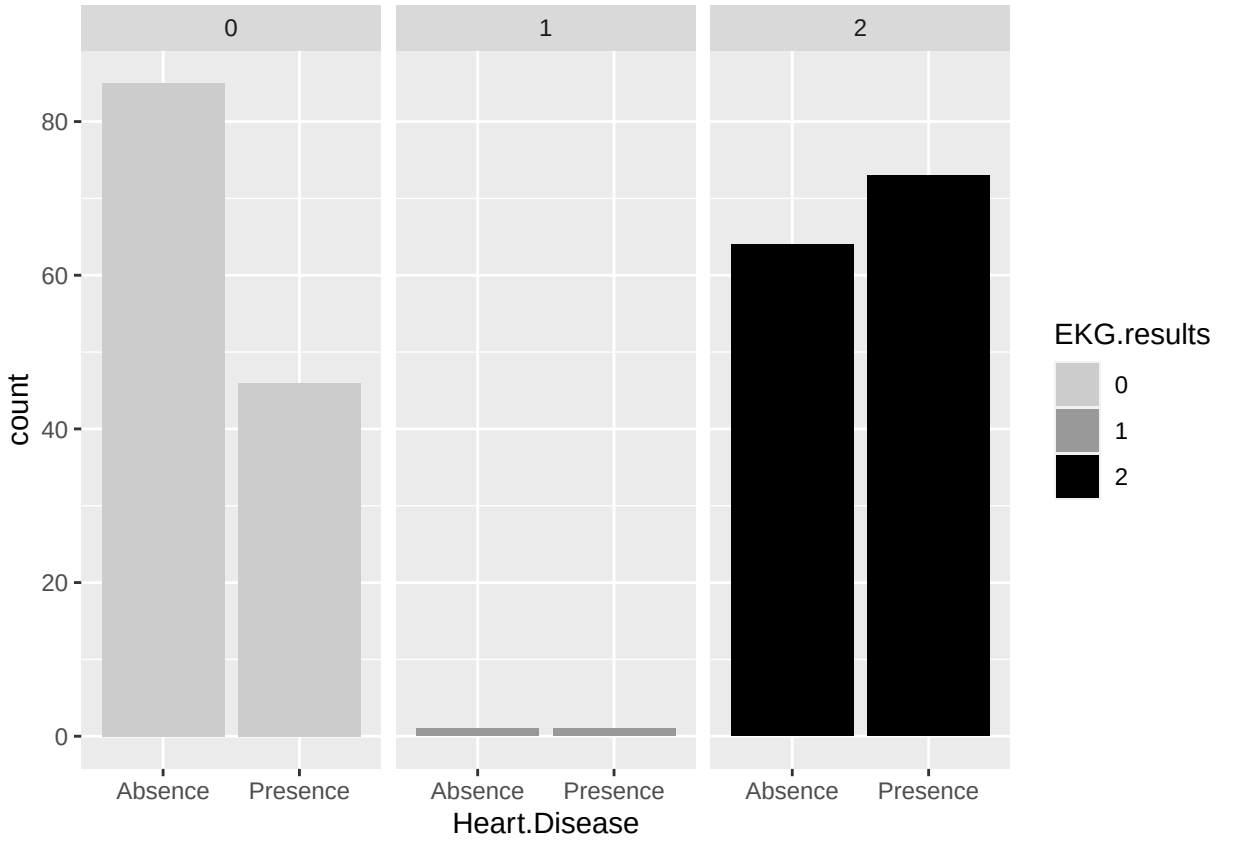
```
ggplot(data = HD) +
  geom_bar(mapping = aes(x = Heart.Disease, fill = Chest.pain.type)) +
  scale_fill_manual(values=c("#FFFFFF",
                             "#CCCCCC",
                             "#999999",
                             "#000000")) +
  facet_wrap(~Chest.pain.type)
```



Considering the chest pain scale going from 1 to 4, it comes as no surprise that severe typical chest pain is observed in a majority of patients with heart disease and chest pain type 1 and are more often observed in patients without the disease than in patients with heart disease.

Just for completeness, the EKG.results are also shown.

```
ggplot(data = HD) +
  geom_bar(mapping = aes(x = Heart.Disease, fill = EKG.results)) +

  scale_fill_manual(values=c("#CCCCCC",
                             "#999999",
                             "#000000")) +
  facet_wrap(~EKG.results)
```



The number of the EKG.results = 1 equals 2. These results are too small to generate a reliable Chi-squared approximation. As a solution, one could omit *EKG.results = 1*, and reduce the factor(HD$EKG.results) to 2 levels *0 =normal* and *1 = abnormal*. Of course, this should first be discussed with the authors or stakeholders.

Analysis of Exercise Tests

There are 4 variables related to exercise:

- Exercise angina The *Exercise.angina* variable is an integer and must be changed to a factor variable:

```
HD$Exercise.angina <- factor(HD$Exercise.angina,
                     levels=c(0,1),
                     labels=c("no angina",
                              "angina"))
```

- ST segment depression *ST.depression* is a numeric variable. A readable label and measurement units are obtained by the following code:

```
label(HD$ST.depression)     <- "ST depression"
units(HD$ST.depression)      <- "mm"
```

- Slope of the ST segment The *Slope.of.ST* variable is an integer and must be changed to a factor variable:

```
HD$Slope.of.ST <- factor(HD$Slope.of.ST,
                         levels=c(1,2,3),
                         labels=c("upsloping",
                                  "downsloping",
                                  "horizontal"))
```

Unfortunately, one does not know if these labels are correct. One has to consult the authors to find out where 1, 2 and 3 stand for. It might be possible that 1 stands for up sloping, 2 for horizontal and 3 for down sloping. We simple do not know from the dataset.

- Thallium test
  The *Thallium* variable is an integer and must be changed to a factor variable:

```
HD$Thallium <- factor(HD$Thallium,
                      levels=c(3,6,7),
                      labels=c("negative",
                               "inconclusive",
                               "positive"))
```

Again, one does not know if these labels are correct. One has to consult the authors to find out where 3, 6 and 7 stand for. So the label description might be incorrect.

In table format, these results look like this:

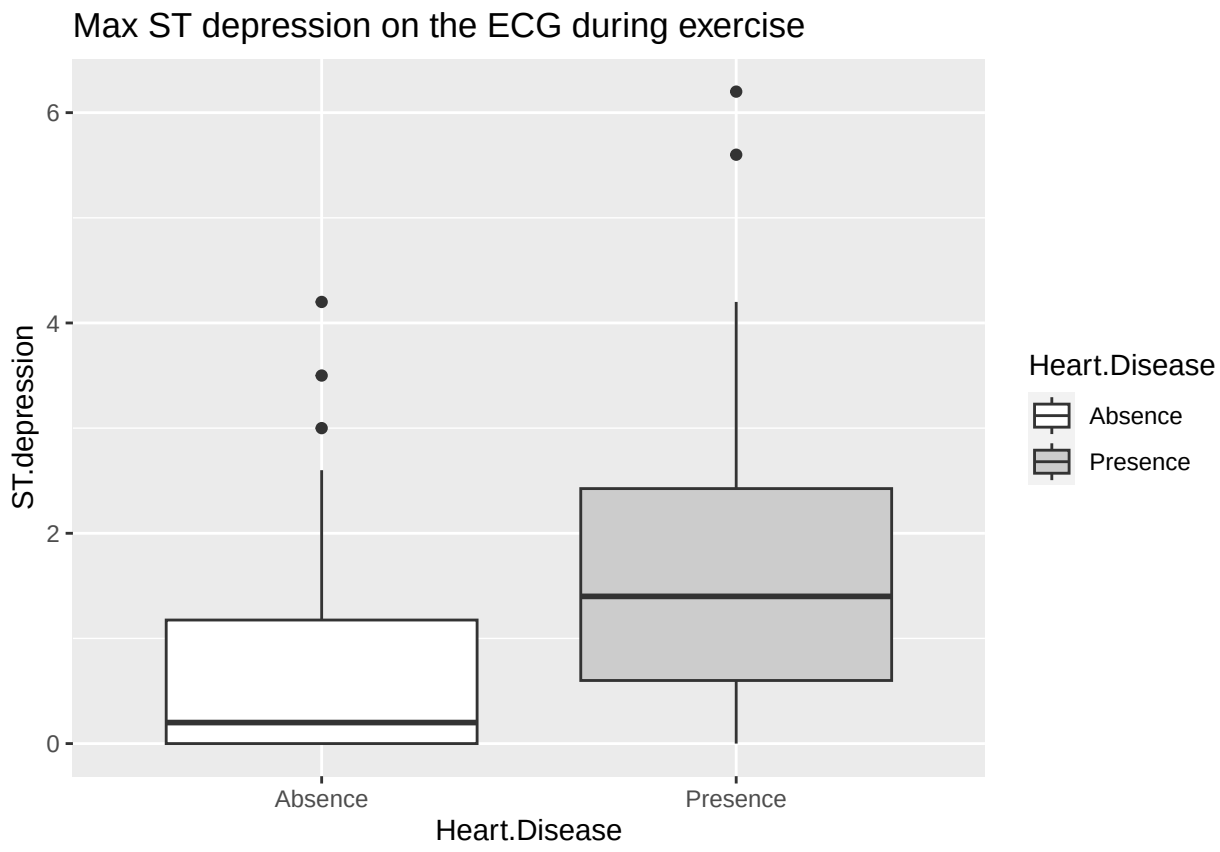
```
table1(~ Exercise.angina + ST.depression + Slope.of.ST +Thallium | status, data= HD,
       overall=F, extra.col=list(`P-value`=pvalue))
```

	no HD	HD	P-value
	(N=150)	(N=120)	
Exercise.angina			
no angina	127 (84.7%)	54 (45.0%)	<0.001
angina	23 (15.3%)	66 (55.0%)	
ST depression (mm)			
Mean (SD)	0.623 (0.801)	1.58 (1.28)	<0.001
Median [Min, Max]	0.200 [0, 4.20]	1.40 [0, 6.20]	
Slope.of.ST			
upsloping	98 (65.3%)	32 (26.7%)	<0.001
downsloping	44 (29.3%)	78 (65.0%)	
horizontal	8 (5.3%)	10 (8.3%)	
Thallium			
negative	119 (79.3%)	33 (27.5%)	<0.001
inconclusive	6 (4.0%)	8 (6.7%)	
positive	25 (16.7%)	79 (65.8%)	

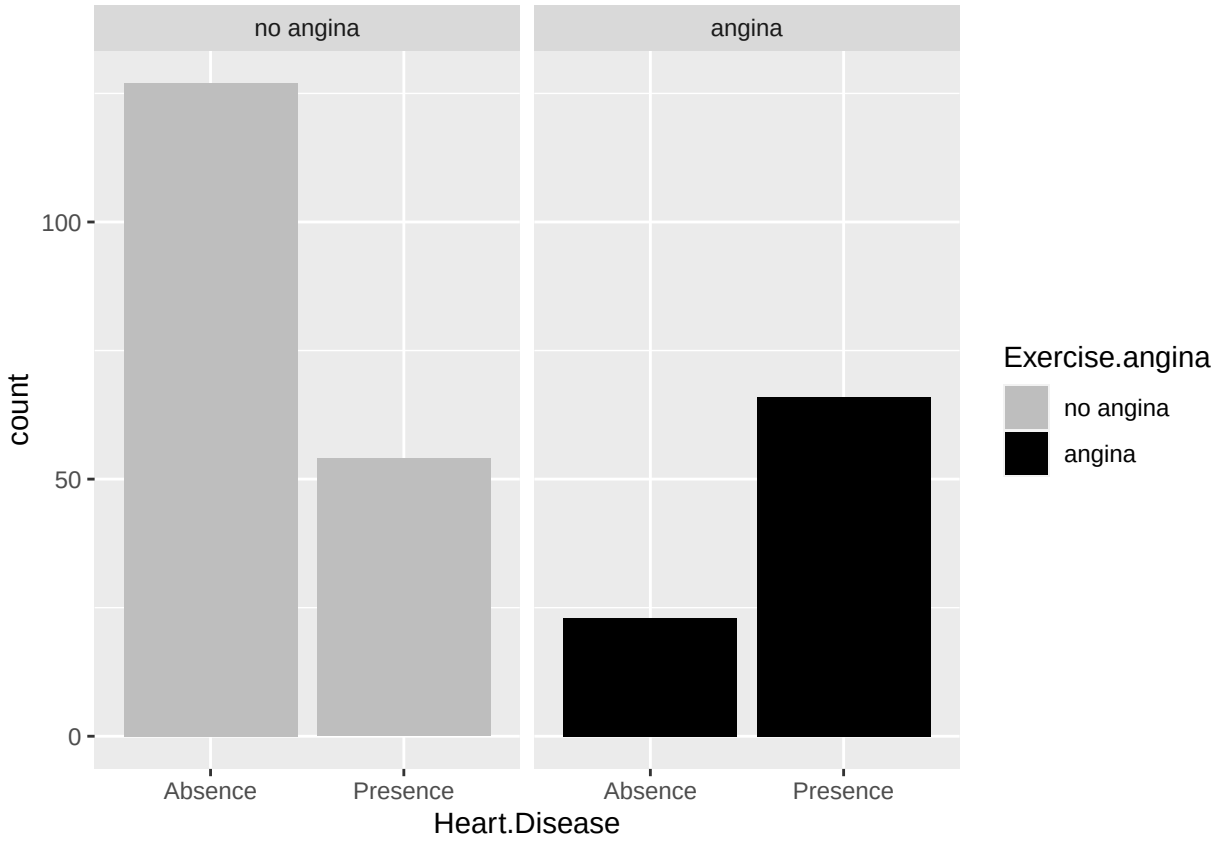
From this table, one thing is clear: The presence of angina during exercise, the level of ST depression, a down sloping ST segment or horizontal ST segment and an abnormal thallium test are show significant differences in patients with and without heart disease.

Graphical representation of these findings:

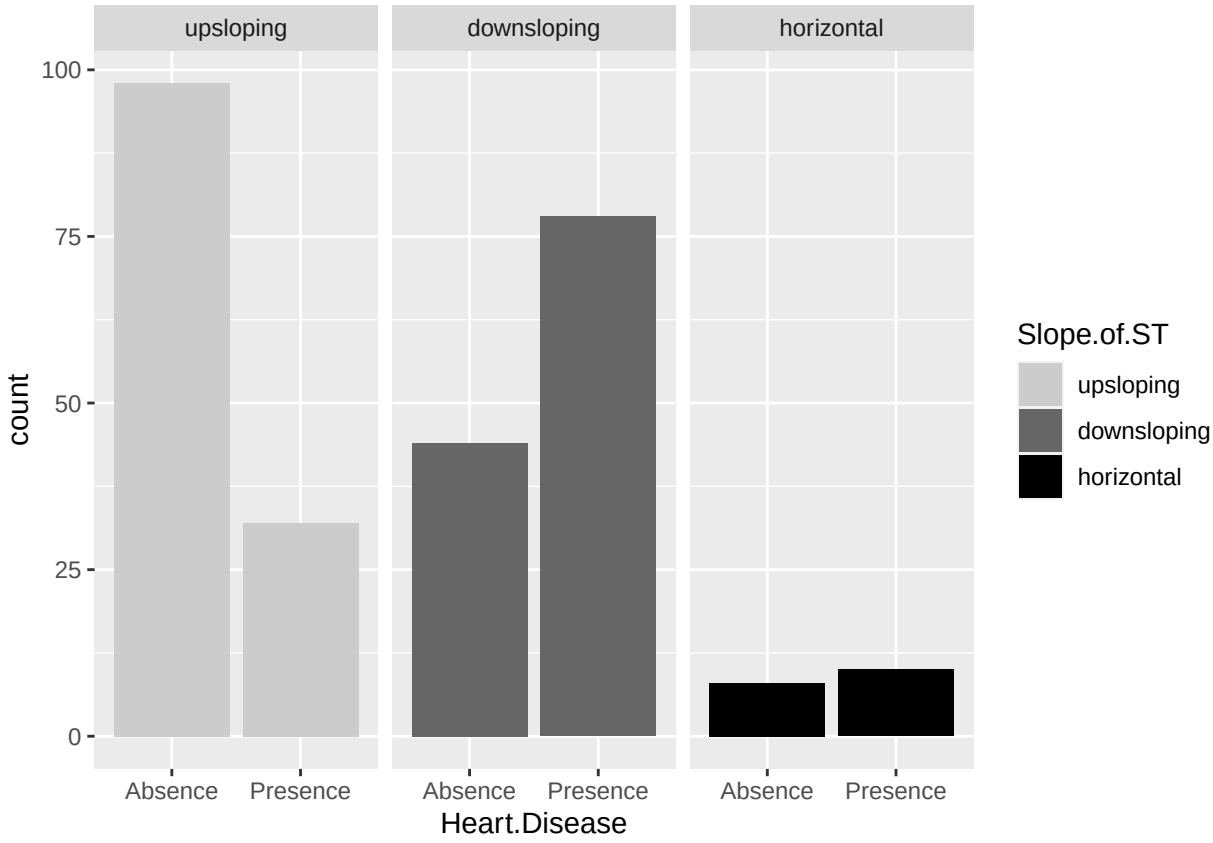
```
ggplot(HD, aes(x = Heart.Disease, y = ST.depression, fill = Heart.Disease)) +
  geom_boxplot() +
  scale_fill_manual(values=c("#FFFFFF",
                             "#CCCCCC")) +
  labs(title = "Max ST depression on the ECG during exercise")
```



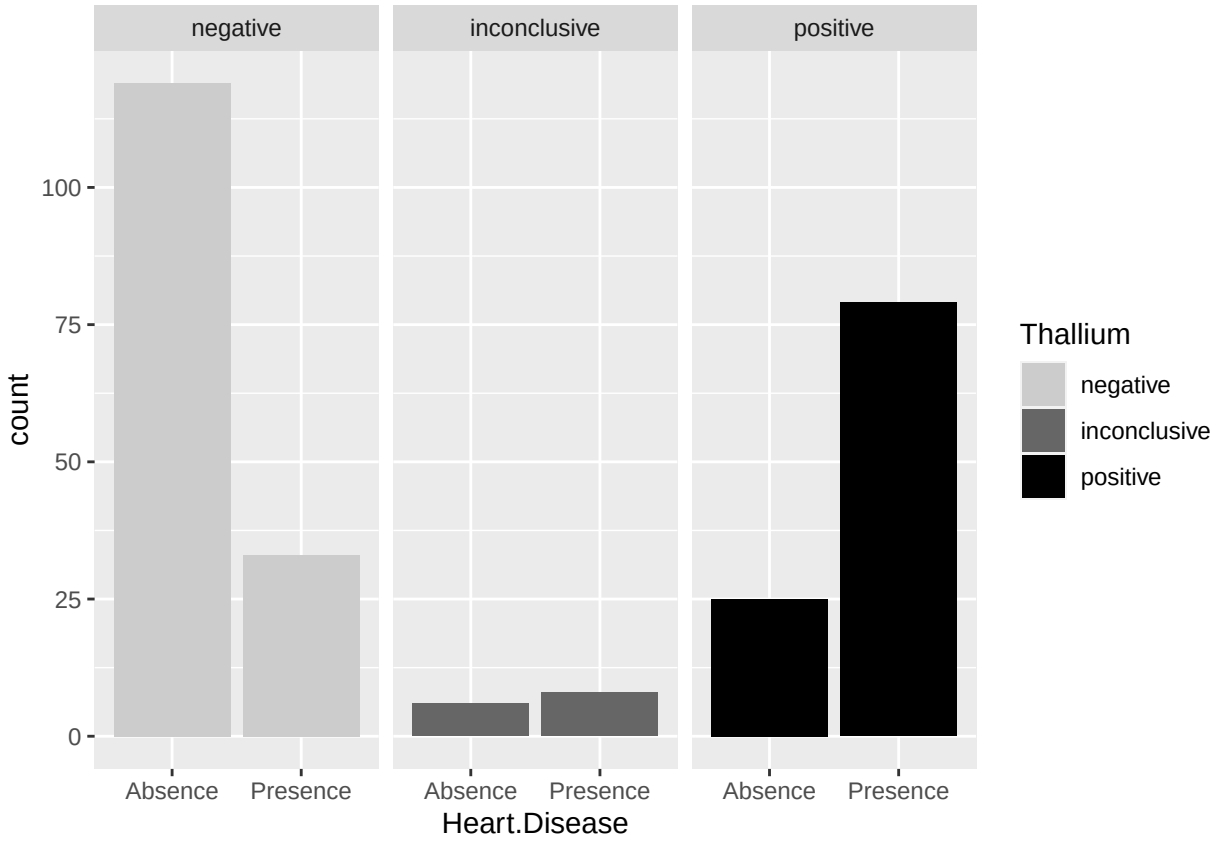

```
ggplot(data = HD) +
  geom_bar(mapping = aes(x = Heart.Disease, fill = Exercise.angina)) +
  scale_fill_manual(values=c("grey",
                             "black")) +
  facet_wrap(~Exercise.angina)
```




```
##########
ggplot(data = HD) +
  geom_bar(mapping = aes(x = Heart.Disease, fill = Slope.of.ST)) +
  scale_fill_manual(values=c("#CCCCCC",
                             "#666666",
                             "black")) +
  facet_wrap(~Slope.of.ST)
```




```
##############
ggplot(data = HD) +
  geom_bar(mapping = aes(x = Heart.Disease, fill = Thallium)) +
  scale_fill_manual(values=c("#CCCCCC",
                             "#666666",
                             "black")) +
  facet_wrap(~Thallium)
```



ST depression during exercise strongly predicts the presence of heart disease. In addition, *angina during exercise*, a *down sloping* ST segment or *horizontal* ST segment and an abnormal *thallium* test, all show significant differences in patients with and without heart disease.

Analysis of coronary angiography

The variable *Number.of.vessels.fluro* shows the results of the coronary vessels involved in the coronary disease of the patient. A score of 0 means normal coronary vessels, a score of 1 means there is one of the coronary vessels with a stenosis or obstruction. A score of 2 means 2 vessels are involved and a score of 3 means all vessels are involved in the process of coronary artery disease.

As a logical consequence, patients with normal coronary vessels do not have coronary artery disease but could have other forms of heart disease e.g. valvular disease, atrial or ventricular myocardial disease or cardiac arrhythmias. However, patients with 2 or 3 vessel disease all have coronary artery disease and would have heart disease because coronary disease is a subset under the total heart disease umbrella.

Now, let us look at the table *Number.of.vessels.fluro*:

```
table(HD$Number.of.vessels.fluro)
```

```
##
##   0   1   2   3
## 160  58  33  19
```

```
# change integer class to factor class
HD$Number.of.vessels.fluro <- factor(HD$Number.of.vessels.fluro,
                        levels=c(0,1,2,3),
                        labels=c("normal",
                                 "1-vessel",
                                 "2-vessel",
                                 "3-vessel"))
```

Fluorscopy

```
table1(~ Number.of.vessels.fluro | status, data= HD,
       overall=F, extra.col=list(`P-value`=pvalue))
```

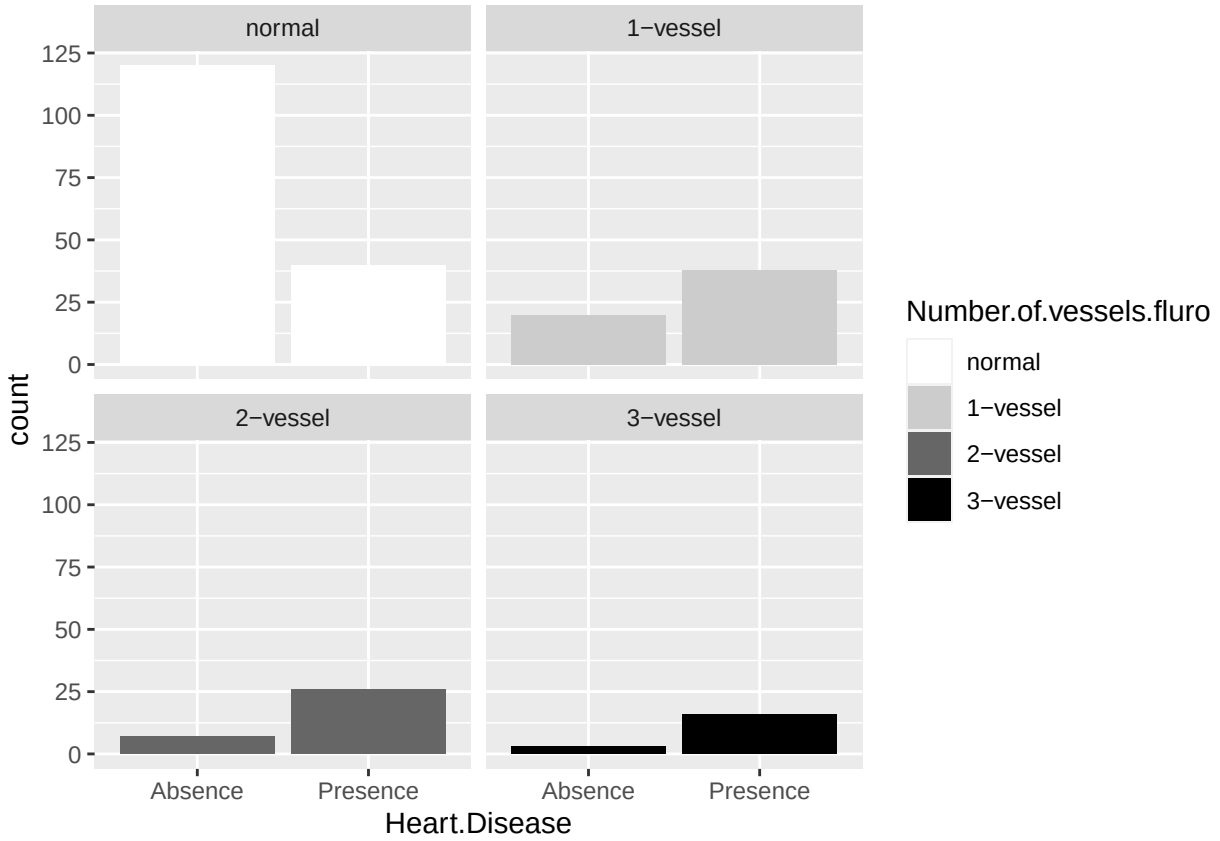
	no HD	HD	P-value
	(N=150)	(N=120)	
Number.of.vessels.fluro			
normal	120 (80.0%)	40 (33.3%)	<0.001
1-vessel	20 (13.3%)	38 (31.7%)	
2-vessel	7 (4.7%)	26 (21.7%)	
3-vessel	3 (2.0%)	16 (13.3%)	

There are 160 patients with **normal vessels** and 110 have **1-3 vessel** abnormalities.

The 110 with 1-3 vessel disease all have coronary artery disease and thus all have heart disease, because heart disease is the umbrella overlapping all variants of heart disease. From the table, we get a different impression, since in the 3 groups with 1, 2, or 3 vessel disease there are patients without heart disease. This is incomprehensible and we must ask the authors for an explanation.

Graphical representation looks like this:

```
ggplot(data = HD) +
  geom_bar(mapping = aes(x = Heart.Disease, fill = Number.of.vessels.fluro)) +
  scale_fill_manual(values=c("#FFFFFF",
                             "#CCCCCC",
                             "#666666",
                             "black")) +
  facet_wrap(~Number.of.vessels.fluro)
```



The graph illustrates this problem. How is it possible that patients with documented coronary artery disease have no heart disease, when coronary artery disease is a subset of patients with heart disease.

Conclusions

- This is a cross-sectional study, where 13 variables are tested against an outcome variable Heart.Disease. In a total of 270 patients, there were 150 patients without heart disease and 120 patients with heart disease.
- **Age** and **Gender** show significant differences in patients with and without heart disease. **Blood pressure**, **Cholesterol** also significant differences in patients with and without heart disease. **FBS.over.120** as a surrogate for diabetes did not reach the level of significance.
- Severe typical chest pain **type 4** was observed in the majority of patients with heart disease and in contrast chest pain pain **type 1** was more often observed in patients without the disease.
- There was a problem with the interpretation of the **EKG.results**, to be discussed with the authors.
- **Maximal heart rate during exercise**, **angina during exercise**, **ST depression**, the **down sloping or horizontal ST slope** and the **abnormal thallium** test also significant differences in patients with and without heart disease.
- The coronary vessel analysis was not problematic, but the results were **incomprehensible**. There must be a definition problem as we accept that patients with coronary vessel disease have coronary artery disease and thus have hearty disease, than it is **impossible** that they have no heart disease. We must ask the authors what their definition was of heart disease.

Logistic Regression analysis

In this analysis the question is answered: "Can we predict the **outcome** heart disease or no heart disease present based on the variables studied in this population?"

First, start with the original dataset:

Select the variables of interest.

```
HD <- read.csv("Heart_Disease_Prediction.csv")
colnames(HD)
```

```
##  [1] "Age"                      "Sex"
##  [3] "Chest.pain.type"          "BP"
##  [5] "Cholesterol"              "FBS.over.120"
##  [7] "EKG.results"              "Max.HR"
##  [9] "Exercise.angina"          "ST.depression"
## [11] "Slope.of.ST"              "Number.of.vessels.fluro"
## [13] "Thallium"                 "Heart.Disease"
```

```
newHD <- HD %>%
  select(Age, Sex, Chest.pain.type, EKG.results,
         Exercise.angina, ST.depression,
         Slope.of.ST, Thallium,Heart.Disease)


newHD$Sex <- factor(newHD$Sex,
                    levels=c(0,1),
                    labels=c("female",
                             "male"))

newHD$Chest.pain.type <- factor(newHD$Chest.pain.type,
                                levels=c(1,2,3,4))


status <- ifelse(HD$Heart.Disease == "Absence", 0,1)

newHD$status = status

newHD$status <- factor(newHD$status,
                       levels=c(0,1),
                       labels=c("no Heart Disease",
                                "Heart Disease"))
newHD$Exercise.angina <- factor(newHD$Exercise.angina,
                                levels=c(0,1),
                                labels=c("no angina",
                                         "angina"))
newHD$Slope.of.ST <- factor(newHD$Slope.of.ST,
                            levels=c(1,2,3),
                            labels=c("upsloping",
                                     "downsloping",
                                     "horizontal"))
```

```
newHD$Chest.pain.type <- factor(newHD$Chest.pain.type,
                                levels=c(1,2,3,4))

table(newHD$EKG.results)
```

```
##
##   0   1   2
## 131   2 137
```

```
newHD$EKG.results <- factor(newHD$EKG.results,
                            levels=c(0,1,2))

# skip the 2 results of factor 1
ks <- subset(newHD,EKG.results == 1)
ks  # gives the record numbers of these 2 patients
```

```
##     Age    Sex Chest.pain.type EKG.results Exercise.angina ST.depression
## 74   76 female               3           1       no angina           1.1
## 111  55 female               4           1          angina           3.4
##     Slope.of.ST Thallium Heart.Disease           status
## 74  downsloping        3       Absence no Heart Disease
## 111 downsloping        3      Presence    Heart Disease
```

```
newHD$EKG.results[74] <- 0
newHD$EKG.results[111] <- 2

newHD$EKG.results <- ifelse(newHD$EKG.results == 0, "normal","not normal")

newHD$EKG.results <- factor(newHD$EKG.results)
```

Next define the model:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. In this study, the variable **Heart.Disease** (Absence or Presence) was used to predict a binary outcome.

- status was tested against :
- Age
- Gender
- Chest.pain.type
- EKG.results
- Exercise angina
- ST.depression
- Slope.of.ST
- Thallium

```
mylogit <- glm(status ~ Age + Sex + Chest.pain.type + EKG.results +
                 Exercise.angina + ST.depression +
                 Slope.of.ST + Thallium, data = newHD,
               family = "binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = status ~ Age + Sex + Chest.pain.type + EKG.results +
##     Exercise.angina + ST.depression + Slope.of.ST + Thallium,
##     family = "binomial", data = newHD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3215  -0.5374  -0.1894   0.5069   2.6449
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -8.49687    1.55388  -5.468 4.55e-08 ***
## Age                     0.04118    0.02102   1.959  0.05012 .
## Sexmale                 1.23780    0.46346   2.671  0.00757 **
## Chest.pain.type2        1.34325    0.82139   1.635  0.10198
## Chest.pain.type3        0.67526    0.71761   0.941  0.34671
## Chest.pain.type4        2.54585    0.69889   3.643  0.00027 ***
## EKG.resultsnot normal   0.73747    0.36987   1.994  0.04617 *
## Exercise.anginaangina   0.66261    0.40531   1.635  0.10209
## ST.depression           0.66091    0.22466   2.942  0.00326 **
## Slope.of.STdownsloping  0.67093    0.43254   1.551  0.12087
## Slope.of.SThorizontal  -0.45426    0.93734  -0.485  0.62794
## Thallium                0.39372    0.10031   3.925 8.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.96  on 269  degrees of freedom
## Residual deviance: 200.87  on 258  degrees of freedom
## AIC: 224.87
##
## Number of Fisher Scoring iterations: 5
```

```
confint.default(mylogit)
```

```
##                                2.5 %      97.5 %
## (Intercept)            -1.154243e+01 -5.45131539
## Age                    -2.143451e-05  0.08238719
## Sexmale                 3.294384e-01  2.14617019
## Chest.pain.type2       -2.666415e-01  2.95313951
## Chest.pain.type3       -7.312261e-01  2.08175302
## Chest.pain.type4        1.176058e+00  3.91564948
## EKG.resultsnot normal   1.253720e-02  1.46240595
## Exercise.anginaangina  -1.317888e-01  1.45701313
```

```
## ST.depression           2.205943e-01  1.10122886
## Slope.of.STdownsloping -1.768317e-01  1.51868414
## Slope.of.SThorizontal  -2.291414e+00  1.38289505
## Thallium                1.971039e-01  0.59033030
```

```
exp(coef(mylogit))## odds ratios and 95% CI
```

```
##            (Intercept)                    Age                Sexmale
##           2.041058e-04           1.042043e+00           3.448034e+00
##       Chest.pain.type2       Chest.pain.type3       Chest.pain.type4
##           3.831472e+00           1.964551e+00           1.275411e+01
##  EKG.resultsnot normal  Exercise.anginaangina          ST.depression
##           2.090643e+00           1.939853e+00           1.936557e+00
## Slope.of.STdownsloping  Slope.of.SThorizontal               Thallium
##           1.956048e+00           6.349180e-01           1.482481e+00
```

```
exp(cbind(OR = coef(mylogit), confint(mylogit)))
```

```
## Waiting for profiling to be done...
```

```
##                                 OR        2.5 %       97.5 %
## (Intercept)            2.041058e-04 8.237830e-06  0.003738465
## Age                    1.042043e+00 1.000384e+00  1.086780671
## Sexmale                3.448034e+00 1.416004e+00  8.800859856
## Chest.pain.type2       3.831472e+00 7.876813e-01 20.451093095
## Chest.pain.type3       1.964551e+00 5.007280e-01  8.619492009
## Chest.pain.type4       1.275411e+01 3.452579e+00 55.152803299
## EKG.resultsnot normal  2.090643e+00 1.019681e+00  4.379625150
## Exercise.anginaangina  1.939853e+00 8.720006e-01  4.303093745
## ST.depression          1.936557e+00 1.270692e+00  3.079813029
## Slope.of.STdownsloping 1.956048e+00 8.359567e-01  4.597185541
## Slope.of.SThorizontal  6.349180e-01 9.239827e-02  3.699352595
## Thallium               1.482481e+00 1.221809e+00  1.814149797
```

The glm model was used analyzing the relationship between one or more existing independent variables against **Heart.Disease**, were status is either 0 or 1 (0 = HD not present and 1 = HD present).

Gender was a strong predictor of the presence of heart disease. In addition, **Chest.pain.type4**, **ST depression**, and a **positive thallium test** were all **strong predictors** of heart disease in this model.

The fact that all patients with **documented coronary abnormalities** (1 vessel, 2 vessel or 3 vessel disease) were strong predictors does not come as a surprise, since documented coronary artery disease is the same as presence of heart disease as we earlier mentioned.

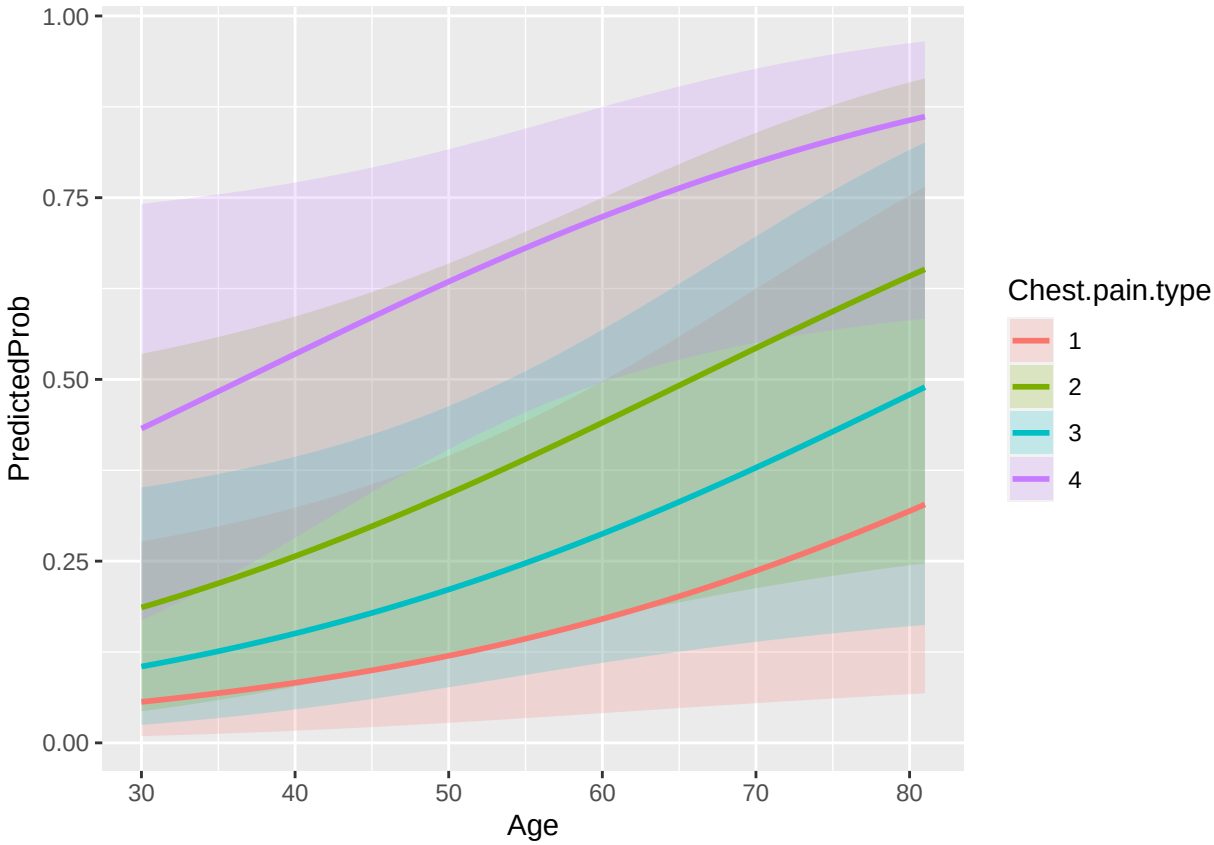
Example of Age and Chest-pain type against outcome

Logistic regression analysis for **men**, where Age and Chest pain type (1:4) as independent variables were tested against the dependant variable outcome(HD present or HD not present);

```
newdata1 <- with(newHD, data.frame(Age = mean(Age),
                                   Sex = "male",
                                   EKG.results = "normal",
```

```
                                    Exercise.angina ="angina",
                                    ST.depression = mean(ST.depression),
                                    Slope.of.ST= "upsloping",
                                    Thallium = mean(Thallium),
                                    Chest.pain.type = factor(1:4)))


# mylogit is the glm result of the original data
newdata1$rankP <- predict(mylogit, newdata = newdata1, type = "response")


newdata2 <- with(newHD, data.frame(Age = rep(seq(from=30, to = 81, by =1),8),
                                   Sex = "male",
                                   EKG.results = "normal",
                                   Exercise.angina ="angina",
                                   ST.depression = mean(ST.depression),
                                   Slope.of.ST= "upsloping",
                                   Thallium = mean(Thallium),
                                   Chest.pain.type = factor(rep(1:4, each = 104))))



newdata3 <- cbind(newdata2, predict(mylogit, newdata = newdata2,
                                    type = "link", se = TRUE))

newdata3 <- within(newdata3, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})

ggplot(newdata3, aes(x = Age, y = PredictedProb)) +
  geom_ribbon(aes(ymin = LL, ymax = UL, fill = Chest.pain.type), alpha = 0.2) +
  geom_line(aes(colour = Chest.pain.type),size = 1)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```

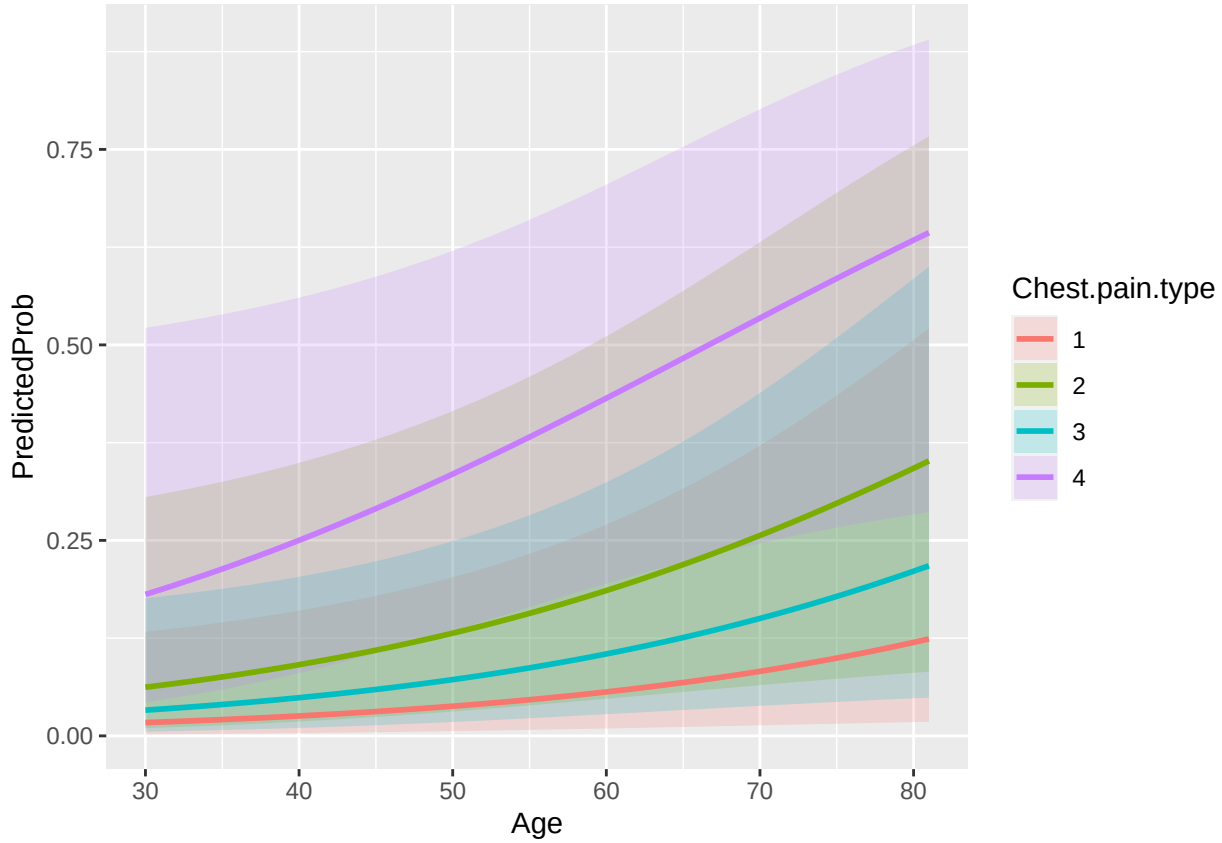


From this illustration, typical chest pain or type-4 is a strong predictor of the presence of heart disease at any age. We can do the same for other strong predictors, like ST depression and an abnormal Thallium test. We can **not** do this for documented coronary disease (one-vessel, two-vessel or three-vessel) because the presence of coronary artery disease implies the presence of heart disease by definition, as mentioned before. This item has to be discussed with the authors of the data set.

The same was done for the **female** population :

```
# female
newdata1 <- with(newHD, data.frame(Age = mean(Age),
                                   Sex = "female",
                                   EKG.results = "normal",
                                   Exercise.angina ="angina",
                                   ST.depression = mean(ST.depression),
                                   Slope.of.ST= "upsloping",
                                   Thallium = mean(Thallium),
                                   Chest.pain.type = factor(1:4)))



# mylogit is the glm result of the original data
newdata1$rankP <- predict(mylogit, newdata = newdata1, type = "response")


newdata2 <- with(newHD, data.frame(Age = rep(seq(from=30, to = 81, by =1),8),
                                   Sex = "female",
                                   EKG.results = "normal",
```

```
                                    Exercise.angina ="angina",
                                    ST.depression = mean(ST.depression),
                                    Slope.of.ST= "upsloping",
                                    Thallium = mean(Thallium),
                                    Chest.pain.type = factor(rep(1:4, each = 104))))

newdata3 <- cbind(newdata2, predict(mylogit, newdata = newdata2,
                                     type = "link", se = TRUE))

newdata3 <- within(newdata3, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})

ggplot(newdata3, aes(x = Age, y = PredictedProb)) +
  geom_ribbon(aes(ymin = LL, ymax = UL, fill = Chest.pain.type), alpha = 0.2) +
  geom_line(aes(colour = Chest.pain.type),size = 1)
```



From this illustration, it is obvious that typical chest pain or chest pain type-4, is at any age is the best predictor of heart disease in women, which underscores the importance of the clinical history of a patient.