Task1Seminar4

Group A3

2025-12-5

Step 1: Data Preparation

We first load the data.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.1
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(irr)
```

```
## Warning: package 'irr' was built under R version 4.5.2
```

```
## Loading required package: lpSolve
```

```
## Warning: package 'lpSolve' was built under R version 4.5.2
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.5.2
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(pROC)
```

```
## Warning: package ’pROC’ was built under R version 4.5.2
```

```
## Type ’citation("pROC")’ for a citation.
##
## Attaching package: ’pROC’
##
## The following objects are masked from ’package:stats’:
##
##     cov, smooth, var
```

```r
library(lme4)
```

```
## Loading required package: Matrix
##
## Attaching package: ’Matrix’
##
## The following objects are masked from ’package:tidyr’:
##
##     expand, pack, unpack
```

```r
library(broom.mixed)
library(DescTools)
```

```
## Warning: package ’DescTools’ was built under R version 4.5.2
```

```
##
## Attaching package: ’DescTools’
##
## The following objects are masked from ’package:psych’:
##
##     AUC, ICC, SD
```

```r
library(knitr)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: ’caret’
##
## The following objects are masked from ’package:DescTools’:
##
##     MAE, RMSE
##
## The following object is masked from ’package:purrr’:
##
##     lift
```

```
df <- read_csv("data_t1.csv", col_types = cols())
```

```
## New names:
## * ‘‘ -> ‘...1‘
```

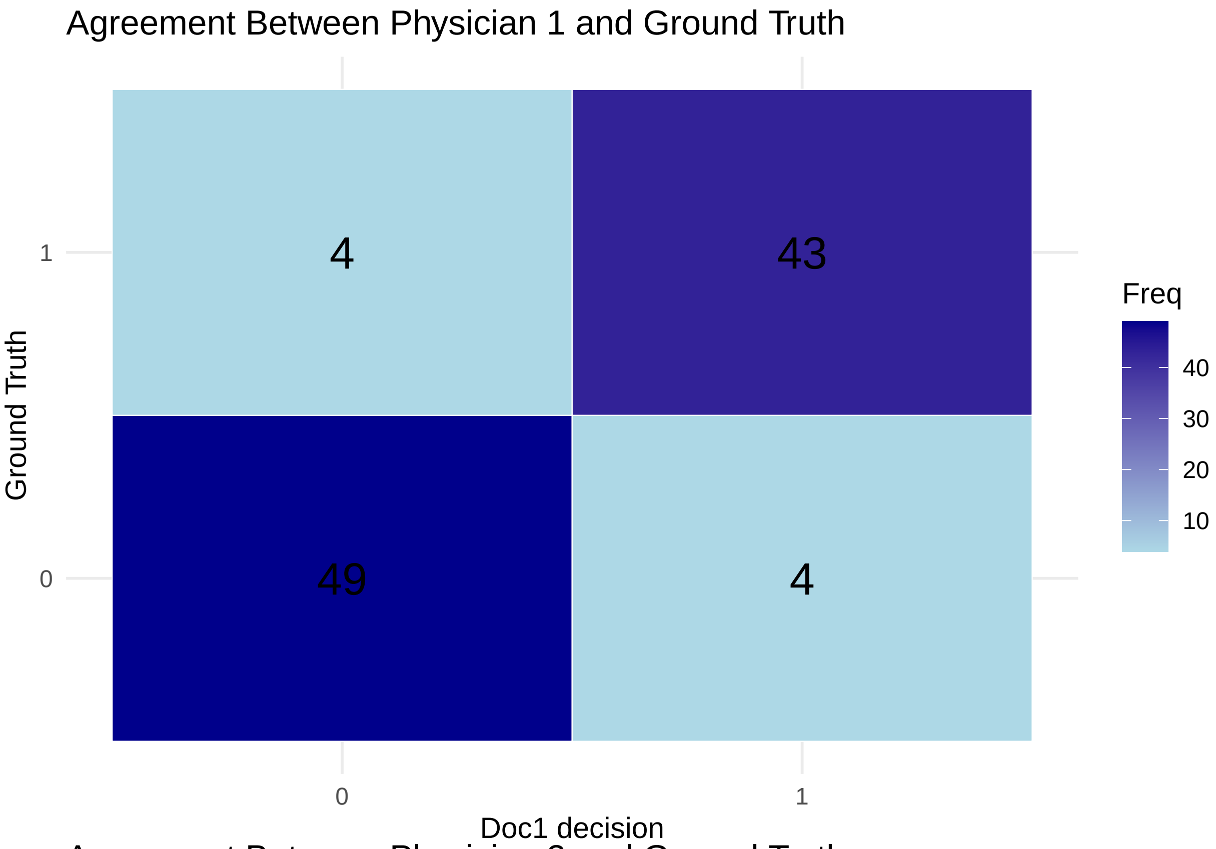
```
head(df)
```

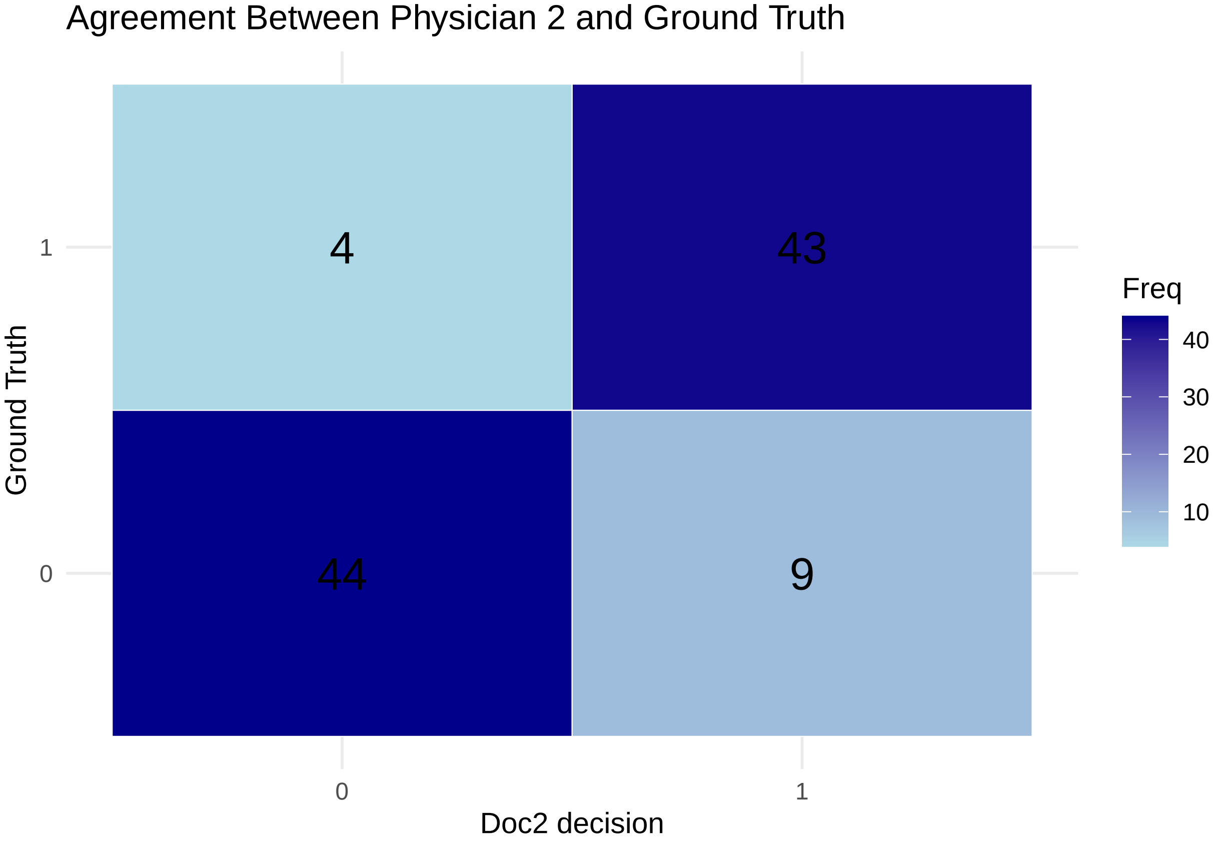
```
## # A tibble: 6 x 4
##    ...1 Patient_Diagnosis Rater1 Rater2
##   <dbl>             <dbl>  <dbl>  <dbl>
## 1     1                 0      0      0
## 2     2                 1      1      0
## 3     3                 0      0      0
## 4     4                 1      0      1
## 5     5                 1      1      1
## 6     6                 0      0      1
```

```
df <- df[ , -1]
df <- df %>%
  rename(truth = Patient_Diagnosis,
         doc1 = Rater1,
         doc2 = Rater2) %>%
  mutate(across(c(truth, doc1, doc2), ~as.integer(.)))  # ensure 0/1 ints
```

From the loaded data, there are 2 raters in this study. The study include 100 individuals; each individual is diagnosed by two doctors separately, and has a reference diagnostic result also.

Step 2: Doctors' Decision vs. Ground Truth


Agreement Between Physician 1 and Ground Truth


Agreement Between Physician 2 and Ground Truth

```
##      Doc1
## Truth  0  1
##     0 49  4
##     1  4 43
```

```
##      Doc2
## Truth  0  1
##     0 44  9
##     1  4 43
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 49  4
##          1  4 43
##
##                Accuracy : 0.92
##                  95% CI : (0.8484, 0.9648)
##     No Information Rate : 0.53
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8394
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9149
##             Specificity : 0.9245
##          Pos Pred Value : 0.9149
##          Neg Pred Value : 0.9245
##              Prevalence : 0.4700
##          Detection Rate : 0.4300
##    Detection Prevalence : 0.4700
##       Balanced Accuracy : 0.9197
##
##        'Positive' Class : 1
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 44  4
##          1  9 43
##
##                Accuracy : 0.87
##                  95% CI : (0.788, 0.9289)
##     No Information Rate : 0.53
##     P-Value [Acc > NIR] : 4.774e-13
##
##                   Kappa : 0.7406
##
##  Mcnemar's Test P-Value : 0.2673
```

```
##
##             Sensitivity : 0.9149
##             Specificity : 0.8302
##          Pos Pred Value : 0.8269
##          Neg Pred Value : 0.9167
##              Prevalence : 0.4700
##          Detection Rate : 0.4300
##    Detection Prevalence : 0.5200
##       Balanced Accuracy : 0.8725
##
##        'Positive' Class : 1
##
```

Doctor 1 has the accuracy of 92%, detects 91.5% of true cancer cases, correctly rules out cancer in 92.4% of non-cancer cases. Kappa value is 0.839, showing strong agreement.P-value for McNemar's test is 1, showing no significant difference between false positives and false negatives. Doc1 shows very strong diagnostic performance, does not show systematic bias toward over- or under-referral.

Doctor 2 has the accuracy of 87%, detects 91.5% of true cancer cases, correctly rules out cancer in 83.0% of non-cancer cases. Kappa value is 0.740, showing relatively strong agreement.P-value for McNemar's test is 0.267, showing no significant asymmetry in errors, but doc2 tends toward more false positives.

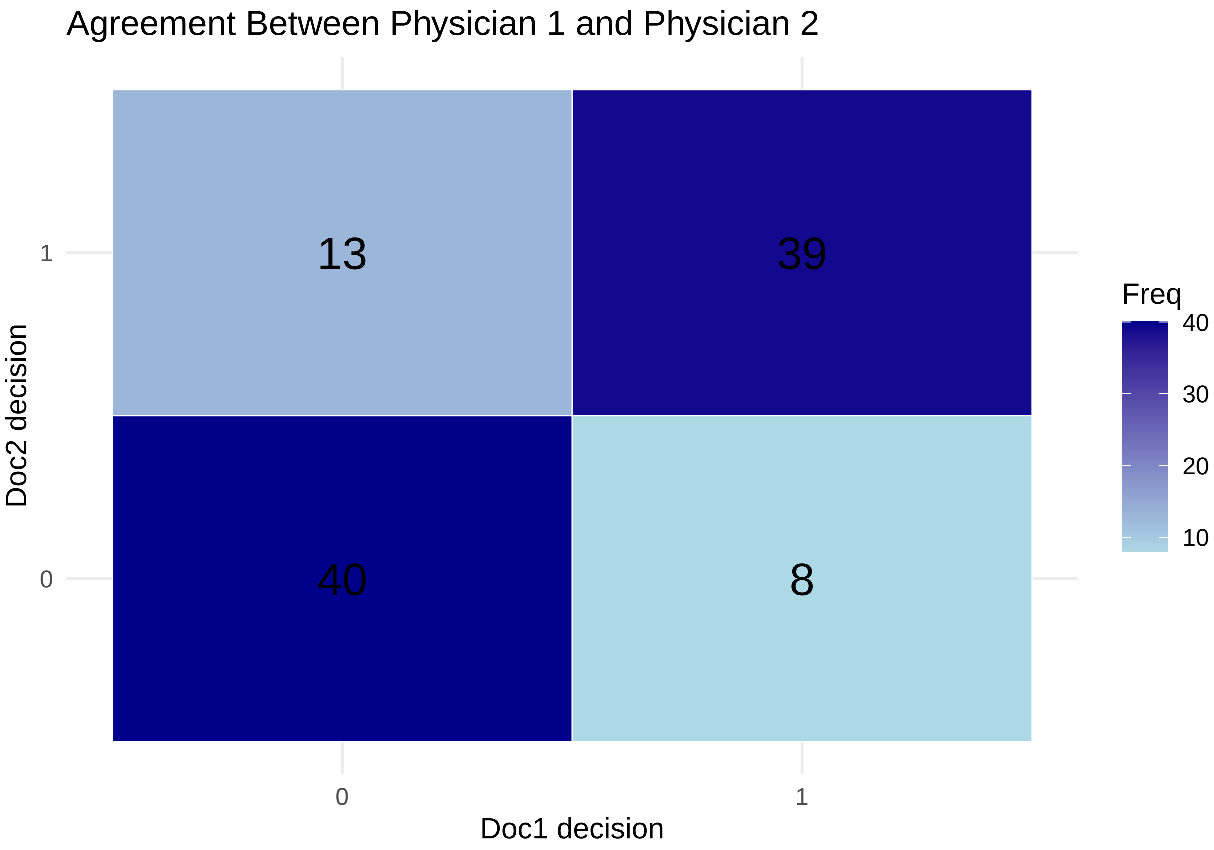
From the accuracy, specificity, PPV and NPV, diagnostic decision of doctor 1 is considered more precise than doctor 2. Doctor 2 tend to diagnose some healthy individuals as diseased.

Step 3: Comparing Between Doctors

We also did agreement tests for the two doctors.

```
##     Doc2
## Doc1  0  1
##    0 40 13
##    1  8 39
```

```
##     Doc2
## Doc1    0    1
##    0 0.40 0.13
##    1 0.08 0.39
```



```
##  Cohen's Kappa for 2 Raters (Weights: unweighted)
##
##  Subjects = 100
##    Raters = 2
##     Kappa = 0.581
##
##         z = 5.84
##   p-value = 5.25e-09
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tab_docs
## McNemar's chi-squared = 0.7619, df = 1, p-value = 0.3827
```

Out of 100 cases, two doctors agreed on most cases (79%). The Kappa value is 0.581, indicating moderate agreement. McNemar's test shows a p-value of 0.3827, neither doctor is systematically over- or under-referring compared to the other.

```
df$patient_id <- 1:nrow(df)
df_long <- df %>%
  pivot_longer(cols = c(doc1, doc2),
               names_to = "rater",
               values_to = "decision")

df_long$decision <- as.integer(df_long$decision)
```

```r
df_long$truth <- as.integer(df_long$truth)

model <- glmer(
  decision ~ truth + (1 | rater) + (1 | patient_id),
  data = df_long,
  family = binomial
)
```

```
## boundary (singular) fit: see help(’isSingular’)
```

```r
summary(model)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: decision ~ truth + (1 | rater) + (1 | patient_id)
##    Data: df_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##     141.6     154.8     -66.8     133.6       196
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2787 -0.3739 -0.3739  0.3050  2.6747
##
## Random effects:
##  Groups     Name        Variance  Std.Dev.
##  patient_id (Intercept) 7.272e-16 2.697e-08
##  rater      (Intercept) 0.000e+00 0.000e+00
## Number of obs: 200, groups:  patient_id, 100; rater, 2
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9677     0.2961  -6.645 3.03e-11 ***
## truth         4.3426     0.4736   9.169  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##       (Intr)
## truth -0.625
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help(’isSingular’)
```

```r
library(DHARMa)
```

```
## This is DHARMa 0.4.7. For overview type ’?DHARMa’. For recent changes, type news(package = ’DHARMa’)
```

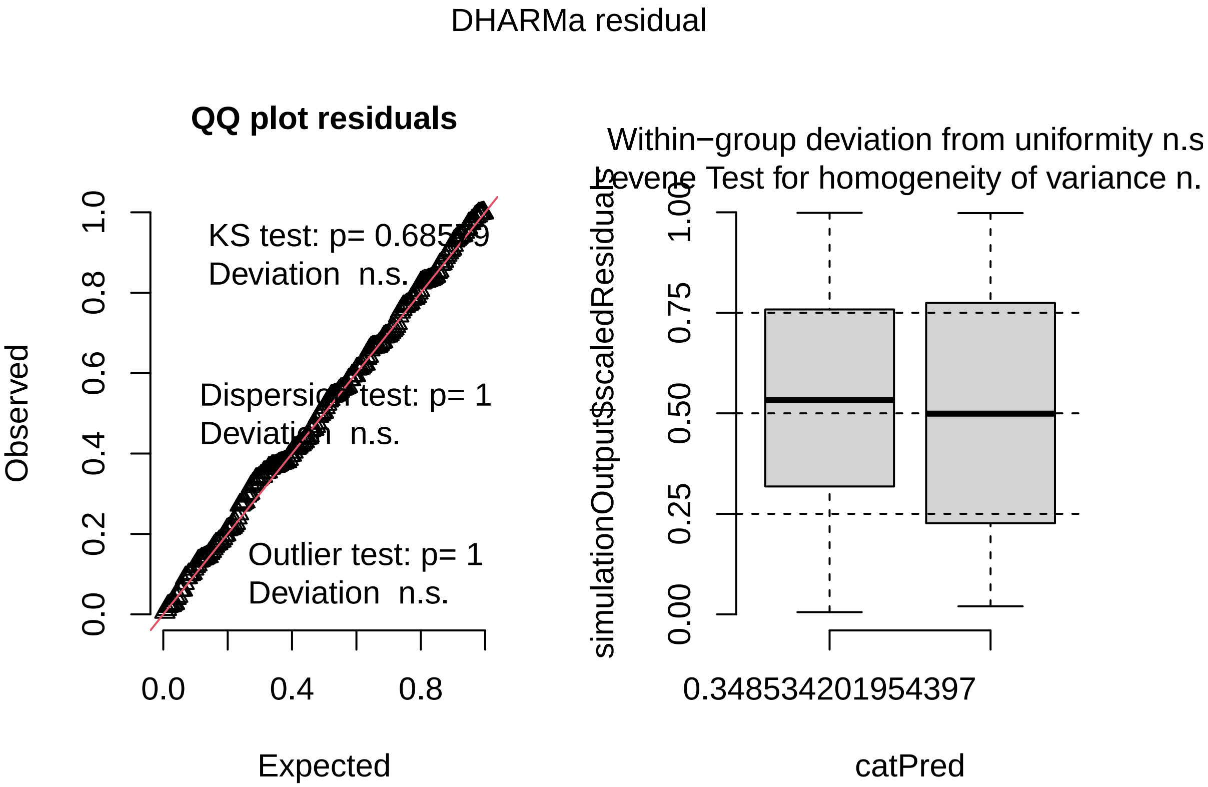
```r
library(jtools)
```

```
## Warning: package ’jtools’ was built under R version 4.5.2
```

```
##
## Attaching package: 'jtools'

## The following object is masked from 'package:DescTools':
##
##     %nin%
```

```
m1_res <- simulateResiduals(model)
plot(m1_res)
```



From the mixed effect model, we can se that random effect is insignificant in this case, showing patient individuals and doctors' decisions behave almost identically in this dataset. Decisions are almost entirely explained by the truth, not by which doctor or which patient.

Residuals are mostly small meaning model fits well.

```
roc1 <- roc(df$truth, df$doc1)
```

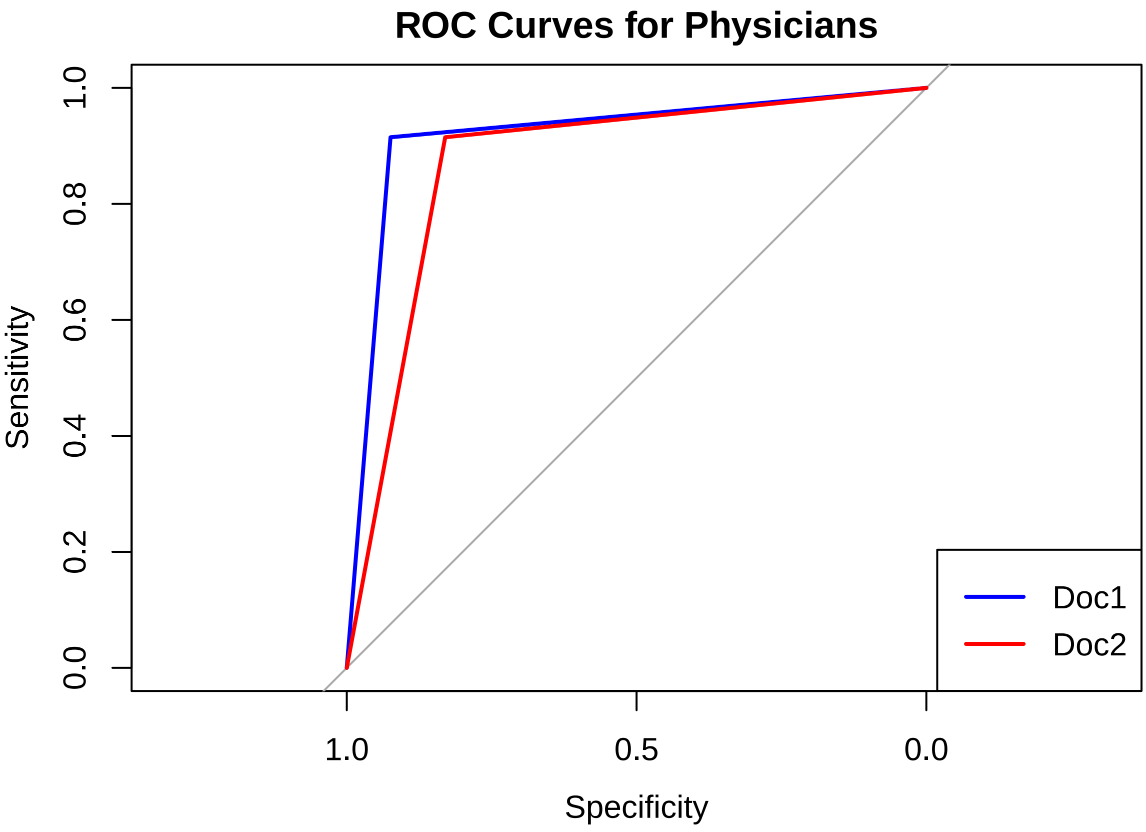
```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
roc2 <- roc(df$truth, df$doc2)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot(roc1, col="blue", main="ROC Curves for Physicians")
lines(roc2, col="red")
legend("bottomright", legend=c("Doc1","Doc2"), col=c("blue","red"), lwd=2)
```



```
roc.test(roc1, roc2)
```

```
##
##  DeLong's test for two correlated ROC curves
##
## data:  roc1 and roc2
## Z = 1.0389, p-value = 0.2989
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  -0.04182046  0.13616008
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.9197110   0.8725411
```

AUC of both doctor is high; and there is no statistically significant difference in AUC between Doc1 and Doc2.