Seminar 4

Group A2

Yufeng Deng   **Yuanqing Wang**

Task 1

First, just have a brief look at the dataset:

```r
labels <- c('Positive','Negative')
diagnosis <- c(sum(data$Patient_Diagnosis[which(data$Patient_Diagnosis==1)]),
            nrow(data)-sum(data$Patient_Diagnosis[which(data$Patient_Diagnosis==1)]))
rater1 <- c(sum(data$Rater1[which(data$Rater1==1)]),
          nrow(data)-sum(data$Rater1[which(data$Rater1==1)]))
rater2 <- c(sum(data$Rater2[which(data$Rater2==1)]),
          nrow(data)-sum(data$Rater1[which(data$Rater1==1)]))
pie(diagnosis,labels)
pie(rater1,labels)
pie(rater2,labels)
```

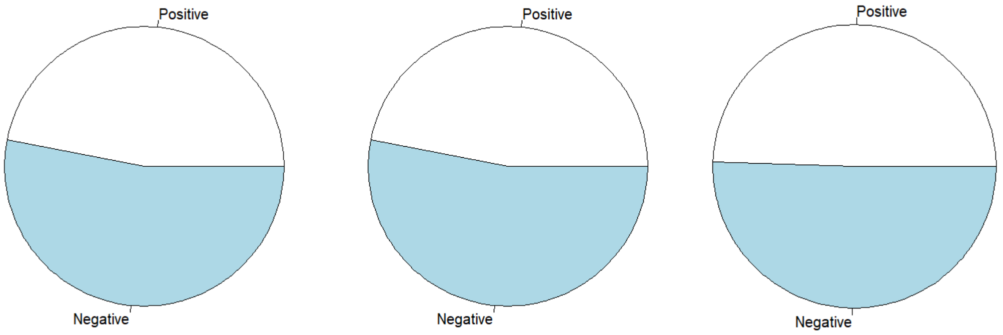

Figure 1.1 Distribution by the label. Diagnosis - Rater1 - Rater2

For Agreement analysis, confusion matrixes are needed to be provided.

```r
# between diagnosis and raters
test1 <- get_Metrics(data$Patient_Diagnosis,data$Rater1)
print(test1$confusion_matrix)
test2 <- get_Metrics(data$Patient_Diagnosis,data$Rater2)
print(test2$confusion_matrix)
# between two raters
test_agree <- get_Metrics(data$Rater1,data$Rater2)
agree_matrix <- test_agree$confusion_matrix
```

Table 1.1 Confusion Matrix between Diagnosis and Rater 1

	Pos R1	Neg R1	Total
Diseased	43	4	47
Healthy	4	49	53
Total	47	53	100

Table 1.2 Confusion Matrix between Diagnosis and Rater 2

	Pos R2	Neg R2	Total
Diseased	43	4	47
Healthy	4	49	53
Total	47	53	100

Table 1.3 Confusion Matrix between Rater 1 and Rater 2

	Pos R2	Neg R2	Total
Pos R1	39	8	47
Neg R1	13	40	53
Total	52	48	100

To test the agreement of Rater 1 and Rater 2, we calculated the kappa value:

```
# kappa value
data_2col <- data.frame(Rater1 = data$Rater1,Rater2 = data$Rater2)
result <- kappa2(data_2col)
print(result)
```

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

The kappa value is 0.581, which means that the two Raters have moderate agreement. It’s too low in medical field, which always requires kappa value higher than 0.7.

To test if there is a systematic difference between Rater 1 and Rater 2, we can apply the McNemar’s test:

```
# McNemar's test
agree_matrix_2dim <- get_2dim_matrix(agree_matrix)
result <- mcnemar.test(agree_matrix_2dim)
print(result)
```

```
                McNemar's Chi-squared test with continuity correction


data:  agree_matrix_2dim
McNemar's chi-squared = 0.7619, df = 1, p-value = 0.3827
```

Since the p-value is higher than 0.05, it can be considered as there is no systematic difference between Rater 1 and Rater 2.

To test differences in sensitivity & specificity between the two Raters, we still need to apply the McNemar's test:

```
# Test difference in sensitivity
subpop_diseased <- data.frame(data[which(data$Patient_Diagnosis == 0),])
test_subdis <- get_Metrics(subpop_diseased$Rater1,subpop_diseased$Rater2)
subdis_matrix_2dim <- get_2dim_matrix(test_subdis$confusion_matrix)
result <- mcnemar.test(subdis_matrix_2dim)
print(result)
```

```
	McNemar's Chi-squared test with continuity correction


data:  subdis_matrix_2dim
McNemar's chi-squared = 1.2308, df = 1, p-value = 0.2673
```

Table 1.4 subpopulation of diseased subjects

	Pos R2	Neg R2	Total
Pos R1	0	4	4
Neg R1	9	40	49
Total	9	44	53

```
# For specificity, look at the subpopulation of healthy subjects.
subpop_healthy <- data.frame(data[which(data$Patient_Diagnosis == 1),])
test_subheal <- get_Metrics(subpop_healthy$Rater1,subpop_healthy$Rater2)
subheal_matrix_2dim <- get_2dim_matrix(test_subheal$confusion_matrix)
result <- mcnemar.test(subheal_matrix_2dim)
print(result)
```

```
	McNemar's Chi-squared test


data:  subheal_matrix_2dim
McNemar's chi-squared = 0, df = 1, p-value = 1
```

Table 1.5 subpopulation of healthy subjects

	Pos R2	Neg R2	Total
Pos R1	39	4	43
Neg R1	4	0	4
Total	43	4	47

According to the results, we can conclude that there are no differences in Specificity and Sensitivity between Rater 1 and Rater 2.

Two functions are used:

```
get_Metrics <- function(diagnosis,rater){# generate the confusion matrix and metrics
  TP <- sum(diagnosis == 1 & rater == 1)
  FP <- sum(diagnosis == 0 & rater == 1)
  TN <- sum(diagnosis == 0 & rater == 0)
  FN <- sum(diagnosis == 1 & rater == 0)
  sumr1 <- sum(TP,FN)
  sumr2 <- sum(FP,TN)
  sumc1 <- sum(TP,FP)
  sumc2 <- sum(FN,TN)
  sum <- sum(TP,FP,FN,TN)
  table <- matrix(c(TP,FN,sumr1,FP,TN,sumr2,sumc1,sumc2,sum),
                  nrow=3,
                  byrow=TRUE)
  colnames(table) <- (c('PosTest','NegTest','Total'))
  rownames(table) <- (c('Diseased','Healthy','Total'))
  # df <- data.frame(table)
  output <- list(confusion_matrix=table,
                 Sensitivity=TP/sumr1,
                 Specificity=TN/sumr2,
                 PPV=TP/sumc1,
                 NPV=TN/sumc2,
                 Accuracy=(TP+TN)/sum)
  return (output)
}

get_2dim_matrix <- function(confmatrix){# turn a 3 dim matrix to a 2 dim one
  table <- matrix(c(confmatrix[1,1],confmatrix[1,2],confmatrix[2,1],confmatrix[2,2]),
                  nrow = 2,
                  byrow = TRUE,
                  dimnames = list("Rater1" = c("Positive", "Negative"),
                                  "Rater2" = c("Positive", "Negative")))
  return(table)
}
```

Task2

First, to show the data distribution, we draw a scatter plot between Kit1 and Kit2 as Figure 2.1.

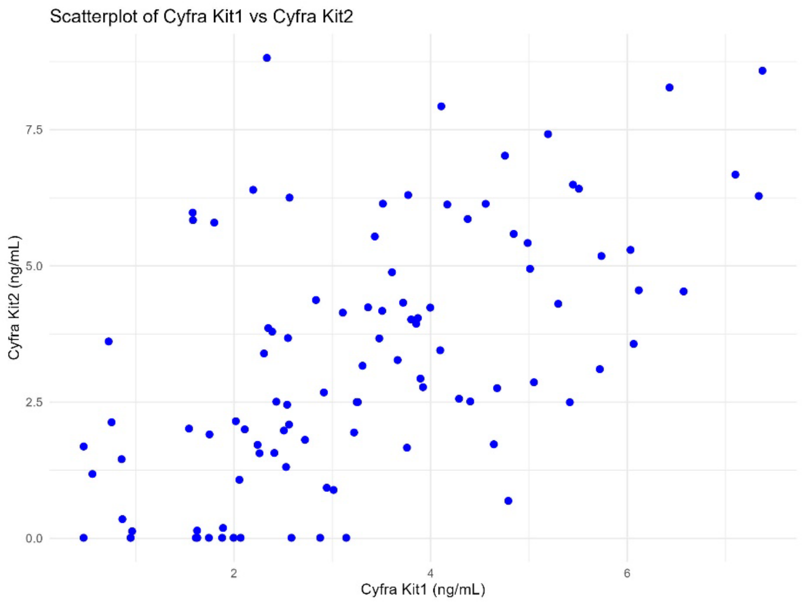

Figure 2.1 Scatter plot between Kit1 and Kit2

Then, we plotted a Bland-Altman plot to demonstrate the agreement between the two sets of data.

```
# Bland-Altman plot
blandr.plot <- blandr.draw(data$Cyfra_Kit1, data$Cyfra_Kit2, plotTitle = "Bland-Altman
Plot")
```

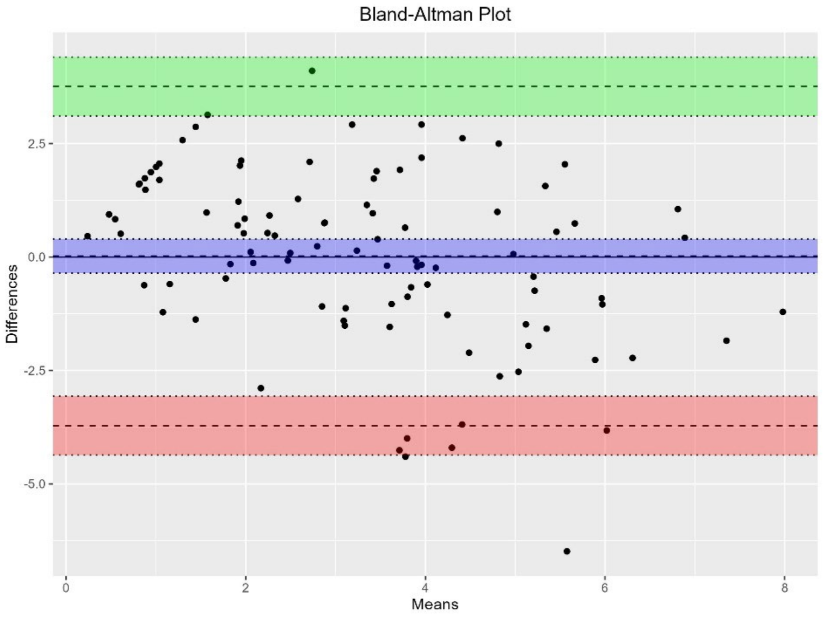

Figure 2.2 Bland-Altman Plot

The Bland-Altman plot is a graphical representation used to assess the agreement between two different measurement techniques. In this plot, the x-axis represents the mean of the two measurements, while the y-axis represents the difference between the two measurements. The central horizontal line represents the mean difference, and the two dashed lines represent the 95% limits of agreement (LoA). Figure 2.2 indicates that the two measurement methods are generally in good agreement, with a small bias and most data points falling within the 95% limits of agreement. However, there are a few outliers that suggest occasional discrepancies.

However, it should be noted that for a Bland-Altman analysis to be optimally reliable, it is advised that the distribution of differences between the two datasets follows a normal distribution. Upon examination of the difference distribution in our study, it was observed that this condition was not met. Additionally, when reviewing Figure 2.1 and closely inspecting the data, a notable presence of suspected erroneous values at 0.01 within Kit2 was identified. After excluding these potential outliers and reassessing the distribution of differences, the deviation from normality persisted. Consequently, given the lack of normal distribution even after addressing suspected bad values, we opted to investigate other analytical approaches.

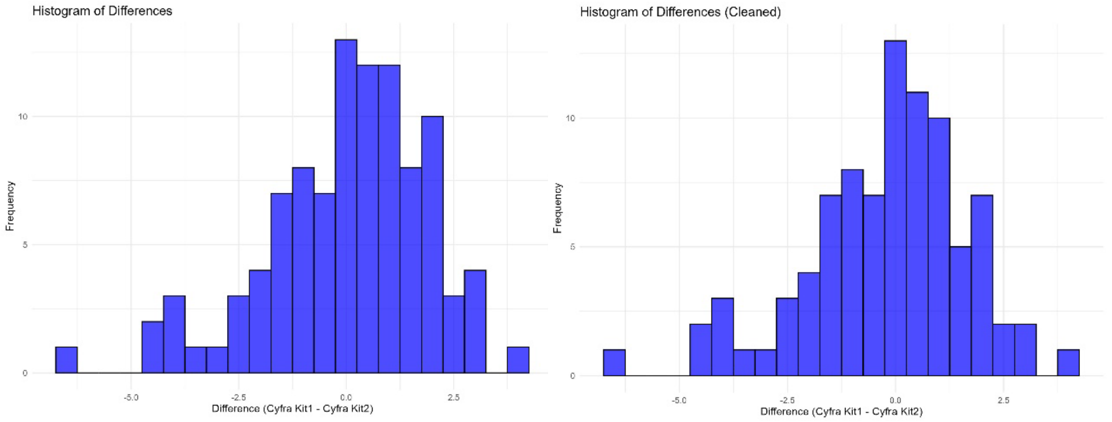

Figure 2.3 Difference Distribution

Then, we calculated the Intraclass Correlation Coefficient (ICC) for the two sets of data.

```
# calculate ICC
icc_result <- icc(data[, c("Cyfra_Kit1", "Cyfra_Kit2")], model = "twoway", type =
"agreement")
print(icc_result)
```

We obtain the result below,

```
Single Score Intraclass Correlation


   Model: twoway
   Type : agreement


   Subjects = 100
     Raters = 2
   ICC(A,1) = 0.554


 F-Test, H0: r0 = 0 ; H1: r0 > 0
   F(99,99) = 3.46 , p = 1.1e-09


 95%-Confidence Interval for ICC Population Values:
  0.402 < ICC < 0.677
```

The calculated Intraclass Correlation Coefficient (ICC) is 0.554, indicating moderate consistency between the two measurement methods. The ICC value ranges from 0 to 1, and a value close to 0.5 suggests a moderate level of agreement.

The F-test result with a p-value of 1.1e-09 is much smaller than the significance level (usually 0.05). Therefore, we reject the null hypothesis, indicating that there is significant consistency between the two measurement methods, i.e., the correlation between the methods is significantly greater than zero.

The 95% confidence interval for the ICC is 0.402 to 0.677, which indicates that with 95% probability, the true value of ICC falls within this range in the population. Since the confidence interval does not include zero, it further supports the conclusion of significant consistency between the two measurement methods.

Finally, we calculated the Concordance Correlation Coefficient (CCC) for the two sets of data.

```
# calculate CCC
ccc_result <- CCC(data$Cyfra_Kit1, data$Cyfra_Kit2)
print(ccc_result)
```

We obtain the result below,

```
$rho.c
        est    lwr.ci    upr.ci
1 0.5518589 0.4130038 0.6656401
$s.shift
[1] 1.408847
$l.shift
[1] -0.01024479
$C.b
[1] 0.9439519
```

This value indicates the level of agreement between the two measurement methods. A value of 0.5519 suggests a moderate correlation between the two methods. As with ICC, the closer the CCC value is to 1, the higher the consistency between the methods.

The 95% confidence interval for CCC is 0.4130 to 0.6656. This range indicates that, with 95% confidence, the true value of the CCC lies within this interval in the population. Since the entire interval is above zero, it further confirms that there is consistency between the two measurement methods.

A positive bias(s.shift=1.4088) suggests that there is a systematic deviation between the two methods, i.e., a difference in baseline measurements. The low-end bias(l.shift= -0.0102) is close to zero, indicating a high level of consistency between the methods at lower measurement values with minimal bias. The regression constant(C.b=0.9439) represents the overall level of consistency in the regression model. A high value indicates strong consistency in the regression.

Task3

First, we drew a box plot to examine the structure and distribution of the data. We then calculated the Intraclass Correlation Coefficient (ICC) for five different methods to assess their consistency.

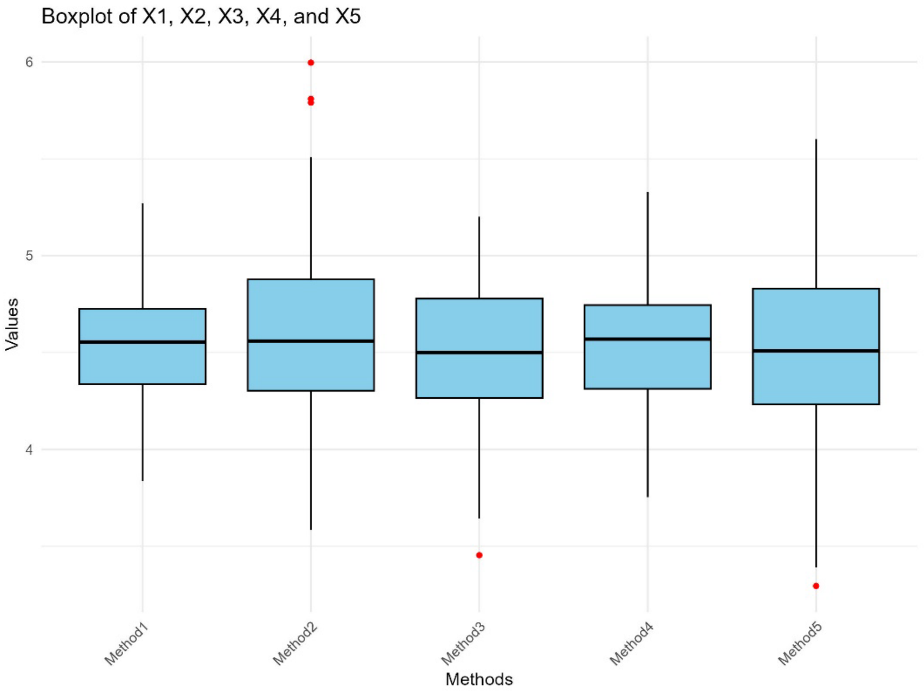

Figure 3.1 Boxplots of five methods

```
 Model: twoway
  Type : agreement

  Subjects = 100
    Raters = 5
  ICC(A,1) = 0.437

F-Test, H0: r0 = 0 ; H1: r0 > 0
 F(99,399) = 4.87 , p = 6.34e-30

95%-Confidence Interval for ICC Population Values:
 0.344 < ICC < 0.536
```

The ICC of 0.437 for the five evaluation methods indicates moderate agreement among the methods. While the methods do show a significant level of consistency, the moderate ICC value suggests that there is still room for improvement in aligning the methods.

The F-test result with a very small p-value (6.34e-30) confirms that the observed consistency is not due to chance, and we can confidently assert that the methods show a statistically significant level of agreement.

The 95% confidence interval of [0.344, 0.536] reinforces the conclusion of moderate consistency between the methods, and provides a range for the true ICC value in the population. While the agreement is statistically significant, further refinement or calibration of the methods could help increase consistency.

We created a correlation heatmap among the five methods to observe the correlations between these five sets of data as Figure 3.2

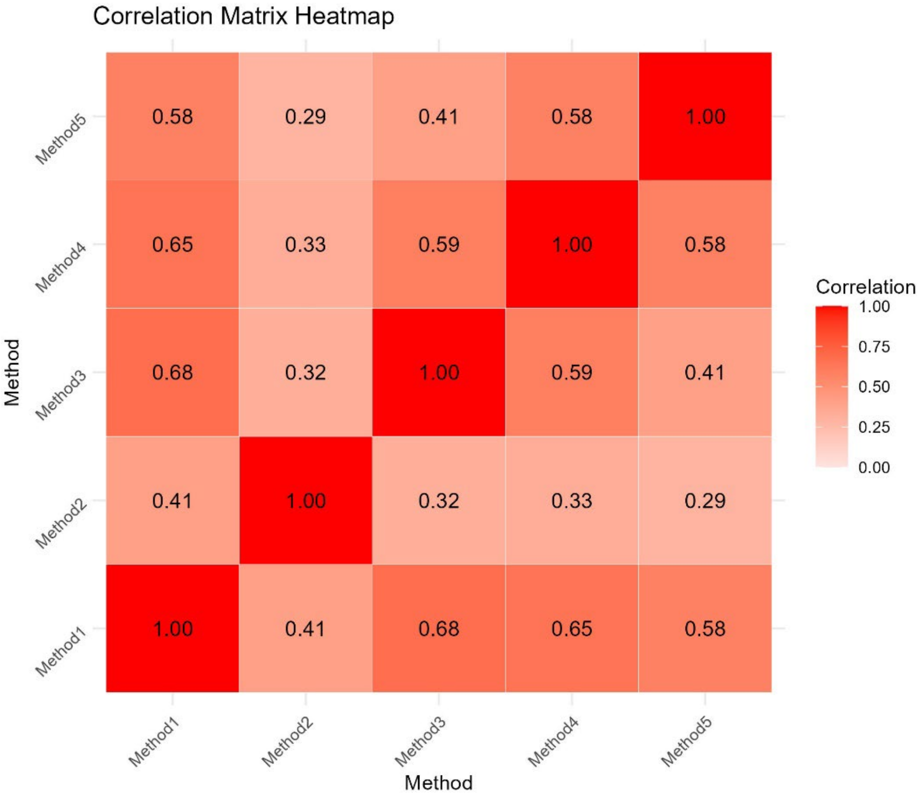

Figure 3.2 Correlation matrix heatmap

Upon inspection, we found that the correlation coefficients between X1 and X3, X1 and X4, X1 and X5, X3 and X4, as well as X4 and X5 were greater than 0.5. Consequently, we proceeded to construct Bland-Altman plots for these pairs to further analyze their agreement.

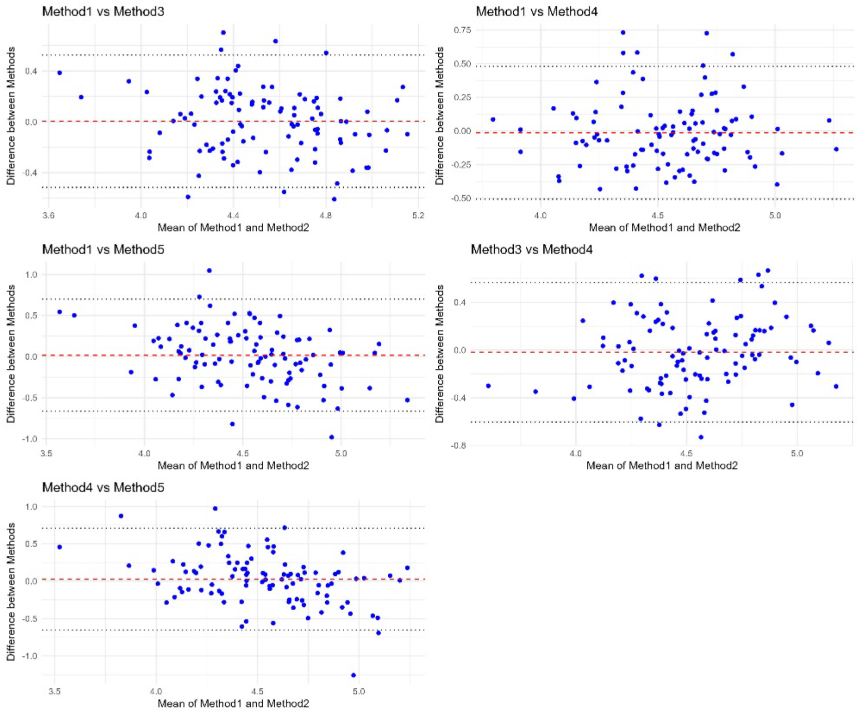

Figure 3.3 Bland-Altman plots

The Bland-Altman plots for the five pairs of methods (Method1 vs Method3, Method1 vs Method4, Method1 vs Method5, Method3 vs Method4, and Method4 vs Method5) provide a comprehensive visual assessment of the agreement between these methods. Specifically, the mean differences (bias) and the 95% limits of agreement (LoA) are key metrics in these plots. The mean differences for all pairs are close to zero, indicating that on average, the methods do not differ significantly. However, the spread of the data points around the mean difference line varies across the pairs, with some showing a tighter clustering and others a wider dispersion.

For Method1 vs Method3, Method1 vs Method4, and Method1 vs Method5, the data points are relatively evenly distributed around the mean difference line, with a few outliers. This suggests a generally good agreement, although the presence of outliers indicates occasional discrepancies. The 95% LoA for these pairs are relatively narrow, further supporting the consistency of these methods. For Method3 vs Method4 and Method4 vs Method5, the data points also show a relatively even distribution around the mean difference line, with a similar pattern of occasional outliers. The 95% LoA for these pairs are slightly wider compared to the previous pairs, indicating a slightly higher variability in the differences between these methods.

Overall, the Bland-Altman plots indicate that the five methods exhibit a reasonable level of agreement, with most data points falling within the 95% limits of agreement. However, the presence of outliers and the variability in the spread of data points suggest that while the methods are generally consistent, there are occasional discrepancies that may need further investigation. The moderate spread of data points around the mean difference line indicates that the methods are generally reliable, but the occasional outliers highlight the need for caution in interpreting the results and potentially refining the measurement techniques to reduce these discrepancies.

Task4

To analyze the performance of the model, we can calculate the AUC of its ROC curve.

```
roc_obj <- roc(data$labels_obs,data$prob_pred)
AUC <- roc_obj$auc
print(AUC)
```

```
Area under the curve: 0.8533
```

The AUC of ROC curve is 0.8533, which can be considered clinically useful[1].

To determine the threshold, we tried to use the features of ROC curve:

```
# Using ROC to determine the threshold
roc_data <- data.frame(
  Spec = roc_obj$specificities,      # Specificity
  Sens = roc_obj$sensitivities,      # Sensitivities
  Thresholds = roc_obj$thresholds
)
# draw ROC
ggplot(roc_data, aes(x = Spec, y = Sens)) +
  geom_line(color = "blue", size = 1) +
  geom_abline(slope = 1, intercept = 1, linetype = "dashed", color = "red") +  # 参考线
  labs(
    title = "ROC Curve",
    x = "Specificity",
    y = "Sensitivity"
  ) +
  theme_minimal() +
  scale_x_reverse()
```

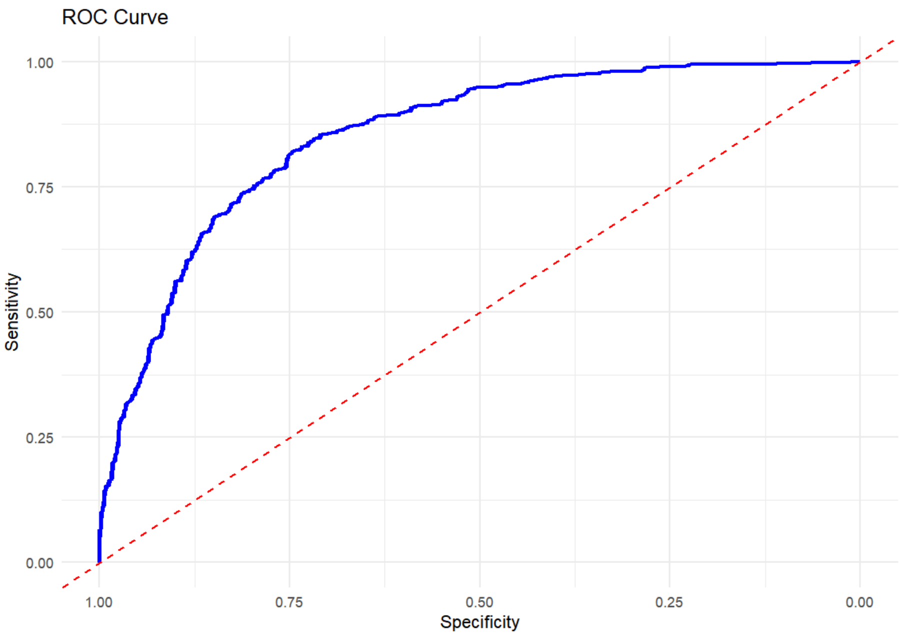

Figure 4.1 ROC curve

To find the optimal threshold, we assume that the threshold which can obtain the maximum of (Specificity + Sensitivity) is the optimal threshold.

```
optimal_idx <- which.max(roc_obj$sensitivities + roc_obj$specificities)
optimal_threshold <- roc_obj$thresholds[optimal_idx]
print(paste("Optimal Threshold:", optimal_threshold))
```

```
[1] "Optimal Threshold: 0.602483458602781"
```

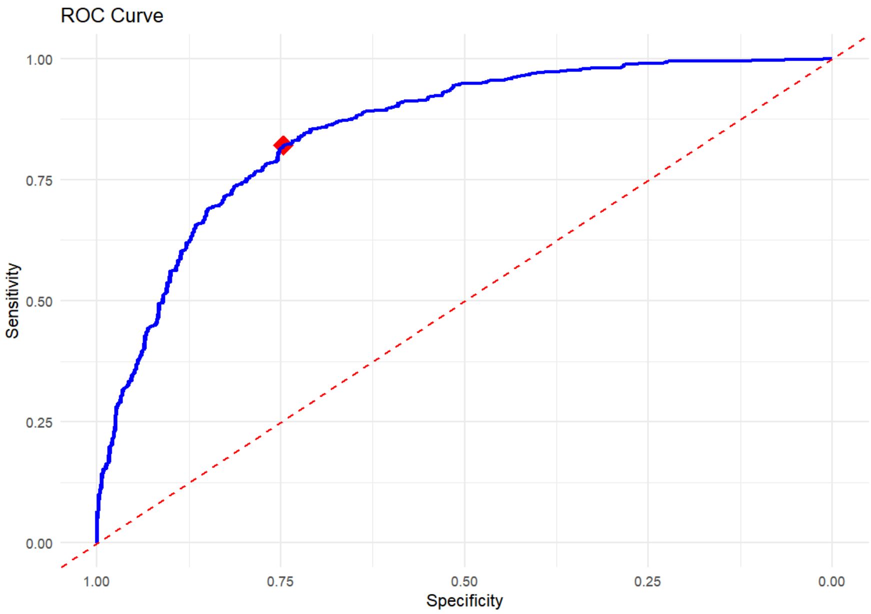

Figure 4.2 ROC curve with the optimal point

We can evaluate this way of choosing threshold by calculating several metrics:

```
binary_pred = numeric(nrow(data))
for (i in data$X){
  if (data$prob_pred[i] >= optimal_threshold){
    binary_pred[i] = 1
  }
  else{
    binary_pred[i] = 0
  }
}
CM <- confusionMatrix(factor(binary_pred),factor(data$labels_obs))
print(CM$overall)
print(CM$byclass)
```

Table 4.1 Metrics of the optimal threshold

Metrics	Value
Accuracy	0.784
Kappa	0.568
Sensitivity	0.821
Specificity	0.746

Since we know that lowering FPR (1-Specificity) is the most important thing in medical diagnosis, we need to try some methods to increase Specificity. Here we force the Specificity to be 0.9 and 0.95.

```
# determine target
target_specificity <- 0.9
# filter
filtered <- roc_data[roc_data$Spec > target_specificity, ]
# find the max Sensitivity in this situation
```

```
threshold_90 <- filtered$Threshold[which.max(filtered$Sens)]
print(paste("Threshold 0.9 Specificity:", threshold_90))
```

```
[1] "Threshold 0.9 Specificity: 0.709243576435695"
[1] "Threshold 0.95 Specificity: 0.781006462207699"
```

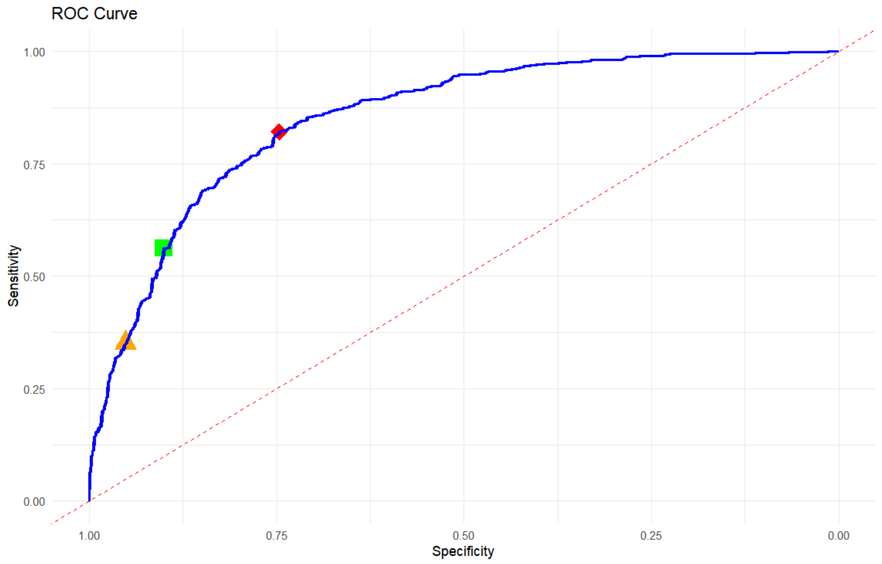

Figure 4.3 Three different thresholds (Specificity of orange, green: 0.95, 0.90)

We can evaluate them by metrics:

Table 4.2 Metrics of the three different thresholds

Metrics	Optimal	0.90 Specificity	0.95 Specificity
Accuracy	0.784	0.729	0.647
Kappa	0.568	0.461	0.300
Sensitivity	0.821	0.562	0.351
Specificity	0.746	0.901	0.951

It is obvious that when Specificity increases, the accuracy will decrease. It is hard to say which threshold is the best one.

References

1. Çorbacıoğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turk J Emerg Med*. 2023;23(4):195-198. doi:10.4103/tjem.tjem_182_23