Project Check-In 2

Paulina Yao

2024-12-04

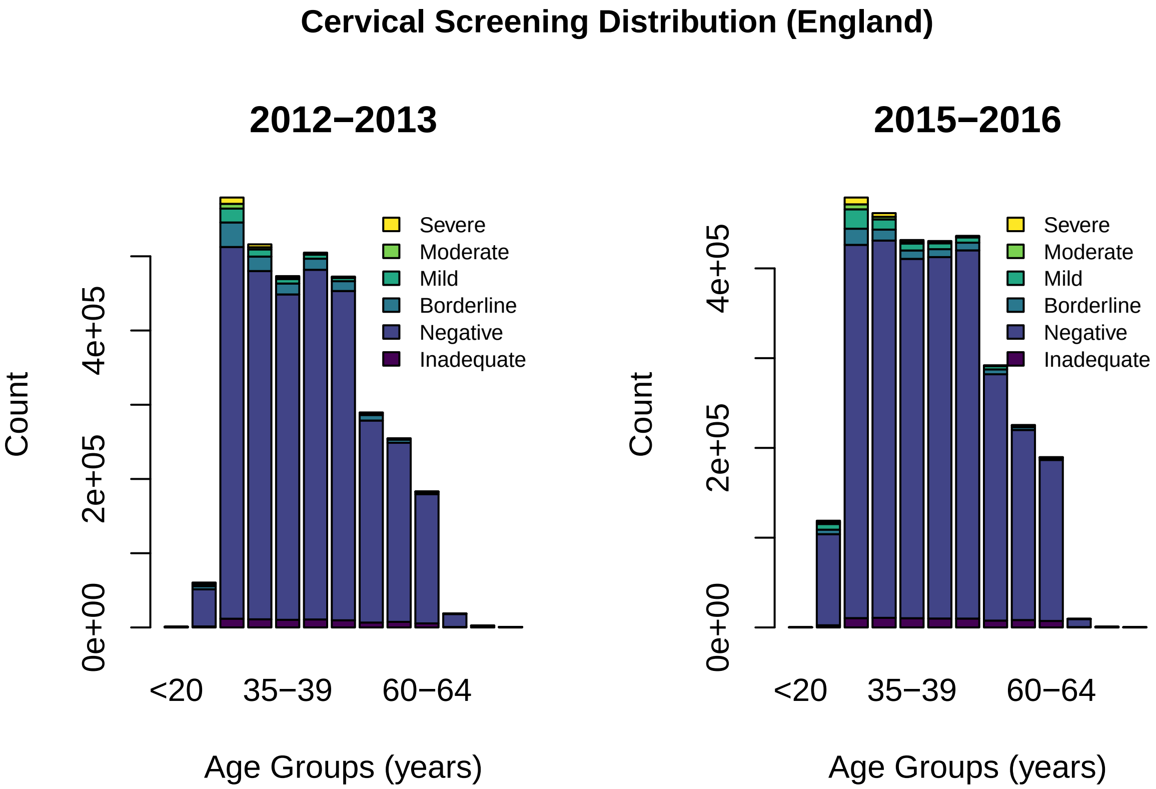
Stacked Barplots

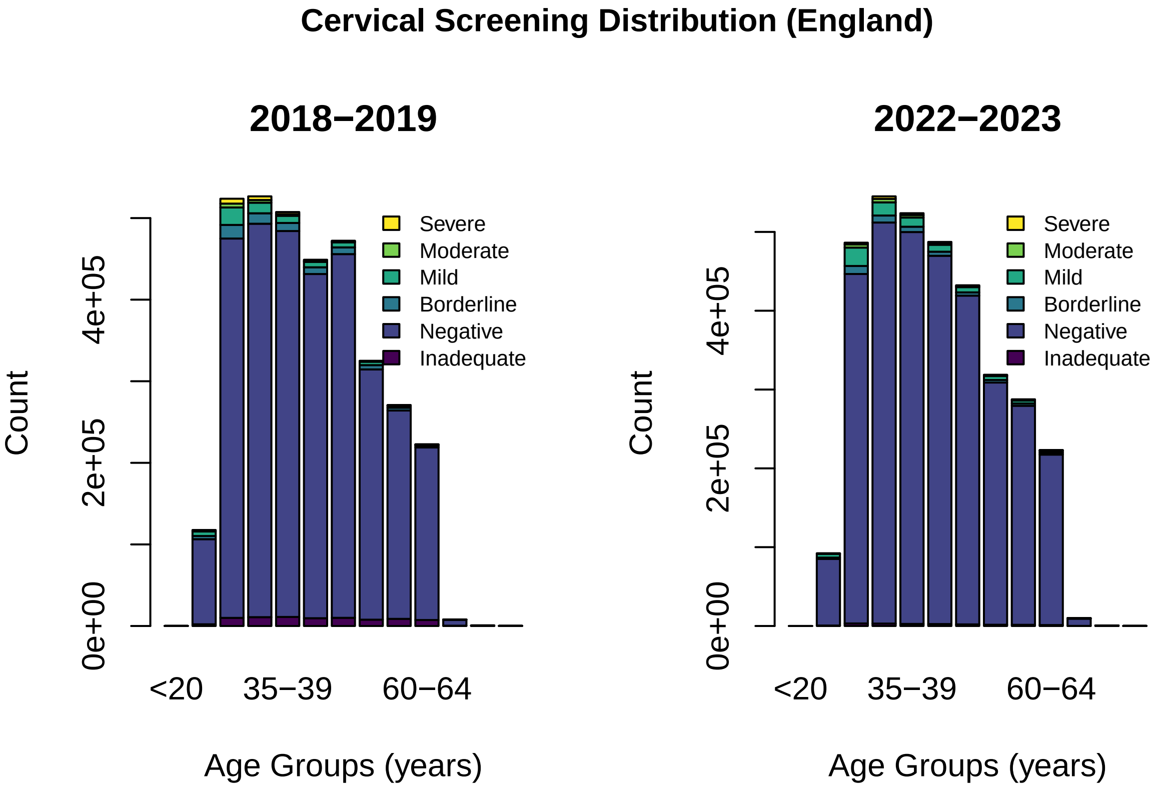
After shifting through the data and trying to understand the information better, I found that the data might be better interpreted through a different type of visualization. While I wasn't able to complete a regression analysis for this project check-in, I continued from my work cleaning the data to create barplots.

The barplots are able to better visualize the proportion of different diagnoses that the data is taken from.

After looking through the data, I found that for the England (national) data varied in the time frame that the data was taken; however, other data taken at the regional level are only taken from 2022-2023. The plots that were created may have been influenced from this discrepancy. I wanted to single out the data from each year frame.

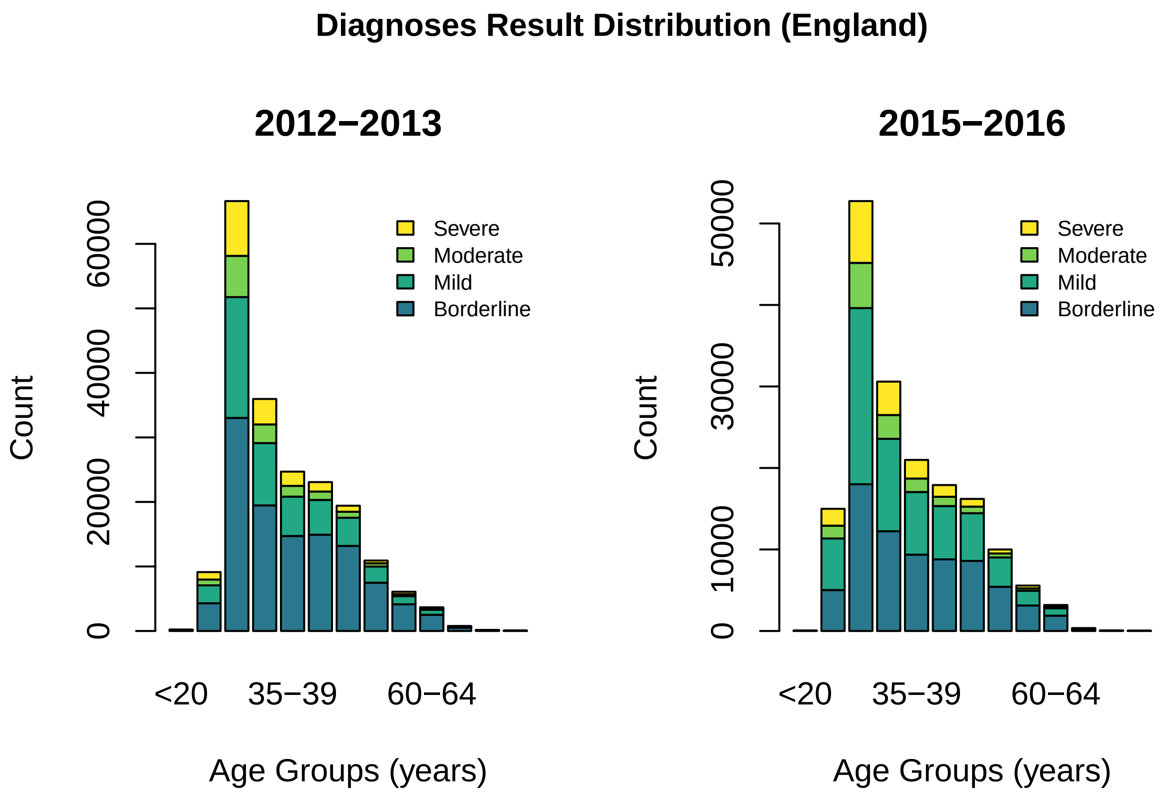
```
source <- read.csv("cervical-programme-annual-2022-23-csvs/kc61_sample_result_age_source.csv")

source$Severe_dyskaryosis <- gsub("not included", NA, source$Severe_dyskaryosis)

uniq_ind <- unique(source$Indicator)
for (i in 1:length(uniq_ind)){
  source$Indicator <- gsub(uniq_ind[i], i, source$Indicator)
}
source$Indicator <- as.numeric(source$Indicator)

for (i in 1:length(uniq_ind)){
  df <- data.frame(source[source$Indicator == i, ])
  df_name <- paste0("df", i)
  assign(df_name, df)
}

severe_NA <- source[is.na(source$Severe_dyskaryosis) == FALSE, ]
severe_NA$Severe_dyskaryosis <- as.numeric(severe_NA$Severe_dyskaryosis)
# all code taken above are taken from initial ProjectCheckIn document
# all code was used for cleaning data

suppressMessages(library(viridisLite))
suppressMessages(library(viridis))
suppressMessages(library(fields))

diagnosis_color <- color.scale(1:6, viridis(6))
diagnosis_col <- which(colnames(severe_NA) == "Inadequate" |
                         colnames(severe_NA) == "Negative" |
                         colnames(severe_NA) == "Borderline_changes" |
                         colnames(severe_NA) == "Mild_dyskaryosis" |
                         colnames(severe_NA) == "Moderate_dyskaryosis" |
                         colnames(severe_NA) == "Severe_dyskaryosis")

par(mfrow = c(1,2), oma = c(0, 0, 2, 0))
for (i in 1:2) {
  y <- c("2012-13", "2015-16", "2018-19", "2022-23")
  df <- which(severe_NA$CollectionYearRange == y[i] &
                severe_NA$Org_Name == "England" &
                suppressWarnings(severe_NA$Indicator == c(1:13)))
  plot <- severe_NA[df, diagnosis_col]
  plot <- t(plot)
  rownames(plot) <- c("Inadequate", "Negative",
                    "Borderline", "Mild",
                    "Moderate", "Severe")
  colnames(plot) <- c("<20", "20-24", "25-29", "30-34", "35-39",
                      "40-44", "45-49", "50-54", "55-59", "60-64",
                      "65-69", "70-74", ">=75")
  plot_title <- c("2012-2013", "2015-2016", "2018-2019", "2022-2023")
  plot <- as.matrix(plot)

  barplot(plot,
        col = diagnosis_color,
        legend.text = row.names(plot),
        args.legend = list(x = "topright", cex = 0.7, bty = "n"),
        main = plot_title[i],
        xlab = "Age Groups (years)",
        ylab = "Count")
}
mtext("Cervical Screening Distribution (England)", side = 3, outer = TRUE, font = 2)
```

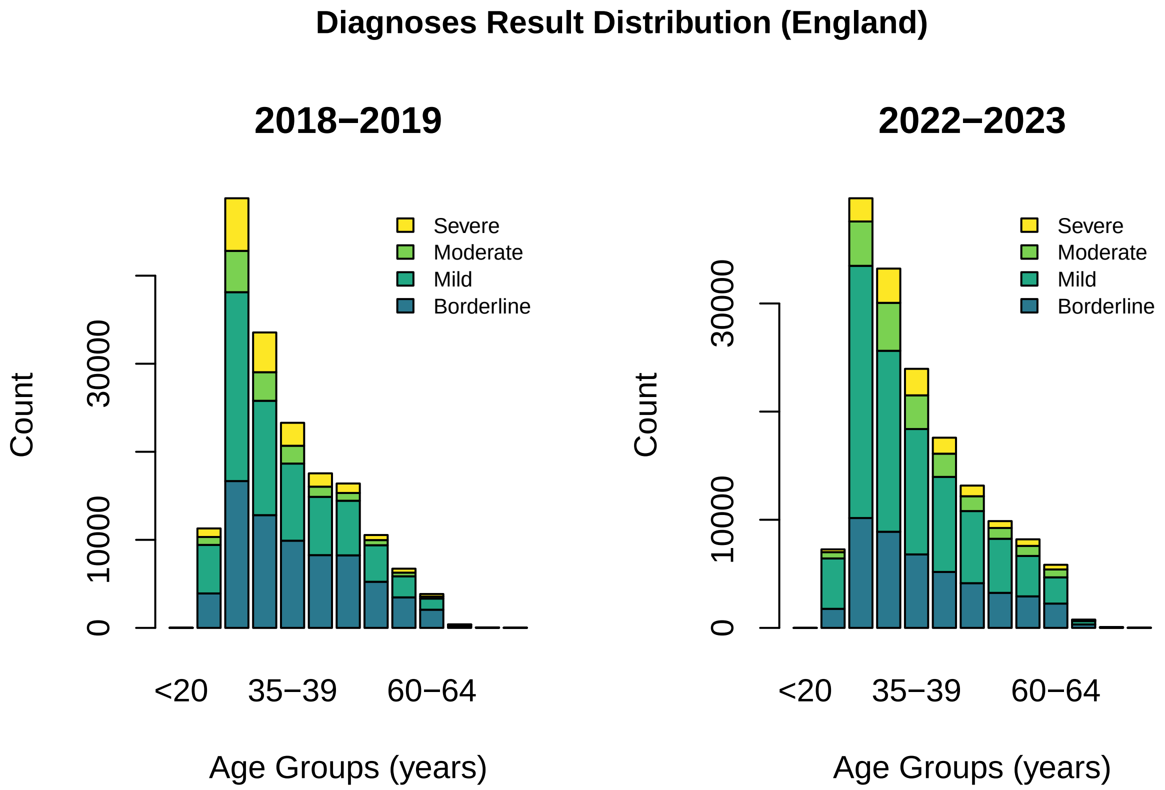




Each row of graphs were coded by the same `for` loop, running through different `i`'s. The code of the second row of graphs were excluded to minimize redundancy. The same is done for the next set of graphs.

```
par(mfrow = c(1,2), oma = c(0 , 0, 2, 0))
for (i in 1:2) {
  y <- c("2012-13", "2015-16", "2018-19", "2022-23")
  df <- which(severe_NA$CollectionYearRange == y[i] &
        severe_NA$Org_Name == "England" &
        suppressWarnings(severe_NA$Indicator == c(1:13)))
  df <- which(severe_NA$CollectionYearRange == y[i])
  plot <- severe_NA[df, diagnosis_col]
  plot <- t(plot)
  rownames(plot) <- c("Inadequate", "Negative",
          "Borderline", "Mild",
          "Moderate", "Severe")
  colnames(plot) <- c("<20", "20-24", "25-29", "30-34", "35-39",
           "40-44", "45-49", "50-54", "55-59", "60-64",
           "65-69", "70-74", ">=75")
  plot_title <- c("2012-2013", "2015-2016", "2018-2019", "2022-2023")
  plot <- as.matrix(plot)
  plot_2 <- plot[-(1:2), ]

  barplot(plot_2,
    col = diagnosis_color[c(3:6)],
    legend.text = row.names(plot_2),
    args.legend = list(x = "topright", cex = 0.7, bty = "n"),
    main = plot_title[i],
    xlab = "Age Groups (years)",
    ylab = "Count")
}
mtext("Diagnoses Result Distribution (England)", side = 3, outer = TRUE, font = 2)
```





From these bar graphs, we can observe that the general trend of diagnoses has decreased throughout the years. This means that there is an absolute decrease in the amount of patients that are diagnosed for any type of cervical tissue changes. We also see that the number of diagnoses for severe dyskaryosis has decreased as well. This change could be explained by the increasing administration of the HPV vaccine.

Lasso Model

```r
row_select <- sample(1:nrow(severe_NA), nrow(severe_NA) / 2)
col_exclude <- c(1:8, 17)

source_train <- severe_NA[row_select, -col_exclude]
source_test <- severe_NA[-row_select, -col_exclude]

suppressMessages(library(glmnet))

cn <- colnames(source_train)
exclude <- which(cn == "Severe_dyskaryosis" | cn == "CollectionYearRange")
X <- as.matrix(source_train[ , -exclude])

suppressWarnings(source_lasso <- cv.glmnet(X, source_train$Severe_dyskaryosis))
coef(source_lasso)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                                   s1
## (Intercept)            30.544009129
## Inadequate              0.008278981
## Negative               -0.004412804
## Borderline_changes      0.057829491
## Mild_dyskaryosis        .
## Moderate_dyskaryosis    0.698899595
## Severe_dyskaryosis_Inv 22.356820919
## Glandular_neoplasia     1.402913280
```

```
Y <- as.matrix(source_test[ , -exclude])
suppressWarnings(source_predict <- predict(source_lasso, newx = Y))

cor(source_predict, source_test$Severe_dyskaryosis) ^ 2
```

```
##                 [,1]
## lambda.1se 0.9134412
```

```
source_errors <- source_test$Severe_dyskaryosis - source_predict
sqrt(mean(source_errors ^ 2))
```

```
## [1] 421.4922
```

```
mean(abs(source_errors))
```

```
## [1] 153.7299
```

For the lasso model, the independent variables acting as name tags for the data were removed so the model could be simplified and fitted to the best predictors. The goal of this model is to see if any variables within this data would be a good predictor of a severe dyskaryosis diagnosis. Looking at the R-squared of the lasso model, we see that the model explains 95.18% of the variability within the data which means the model is fitted very well to the data. However, looking at the RMSE and MAE, we see that the model is very lacking in its predictions. This can be because of the odd way the data is organized and collected.

Some data within this dataset is taken across multiple year frames, while others are taken within the same year. The data also varies widely between the number of observations that are taken at each "location". The largely variability and inconsistency of the data could contribute to a high RMSE and MAE seen within the lasso model.

Because of the data, a portion of the data is singled out to be evaluated again. The data taken from 2012-2023 in England for all ages were subsetted into a new data frame, and the lasso model was ran again under these restrictions.

```
lasso_data <- severe_NA[severe_NA$CollectionYearRange == "2022-23" &
                          suppressWarnings(severe_NA$Indicator == c(1:13)), ]
row_select <- sample(1:nrow(lasso_data), nrow(lasso_data) / 2)
col_exclude <- c(1:8, 17)

source_train <- lasso_data[row_select, -col_exclude]
source_test <- lasso_data[-row_select, -col_exclude]

cn <- colnames(source_train)
exclude <- which(cn == "Severe_dyskaryosis")
```

```r
X <- as.matrix(source_train[ , -exclude])

suppressWarnings(source_lasso <- cv.glmnet(X, source_train$Severe_dyskaryosis))
coef(source_lasso)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                                  s1
## (Intercept)             4.990689306
## Inadequate              .
## Negative               -0.001441650
## Borderline_changes      0.180269162
## Mild_dyskaryosis        0.007158867
## Moderate_dyskaryosis    .
## Severe_dyskaryosis_Inv  .
## Glandular_neoplasia     6.620355470
```

```r
Y <- as.matrix(source_test[ , -exclude])
suppressWarnings(source_predict <- predict(source_lasso, newx = Y))

cor(source_predict, source_test$Severe_dyskaryosis) ^ 2
```

```
##                 [,1]
## lambda.1se 0.9871171
```

```r
source_errors <- source_test$Severe_dyskaryosis - source_predict
sqrt(mean(source_errors ^ 2))
```

```
## [1] 124.1981
```

```r
mean(abs(source_errors))
```

```
## [1] 59.77411
```

The R-squared, RMSE, and MAE have all improved from the previous model. However, since the data does have the quality of time attached to it, it may still be an issue in terms of predictions. This may explain the large RMSE and MAE values.