Module 3: Semi-supervised learning (PheCAP)

Siyue Yang, Jianhui Gao, and Jesse Gronsbell

```
# Load the packages.
packages <- c("tidyverse", "PheCAP", "glmnet", "glmpath", "pROC")

# Check if the pacakages are missing or not.
# If missing, install automatically.
# If not missing, load the package.
package.check <- lapply(
      packages,
      FUN = function(x) {
        if (!require(x, character.only = TRUE)) {
          install.packages(x, dependencies = TRUE)
          library(x, character.only = TRUE)
        }
      }
    )
```

```
# Load environment.
load("environment.RData")
```

```
# Load helper functions.
source("../Rscripts/helper_function.R")
source("../Rscripts/ex2_helper_function.R")
```

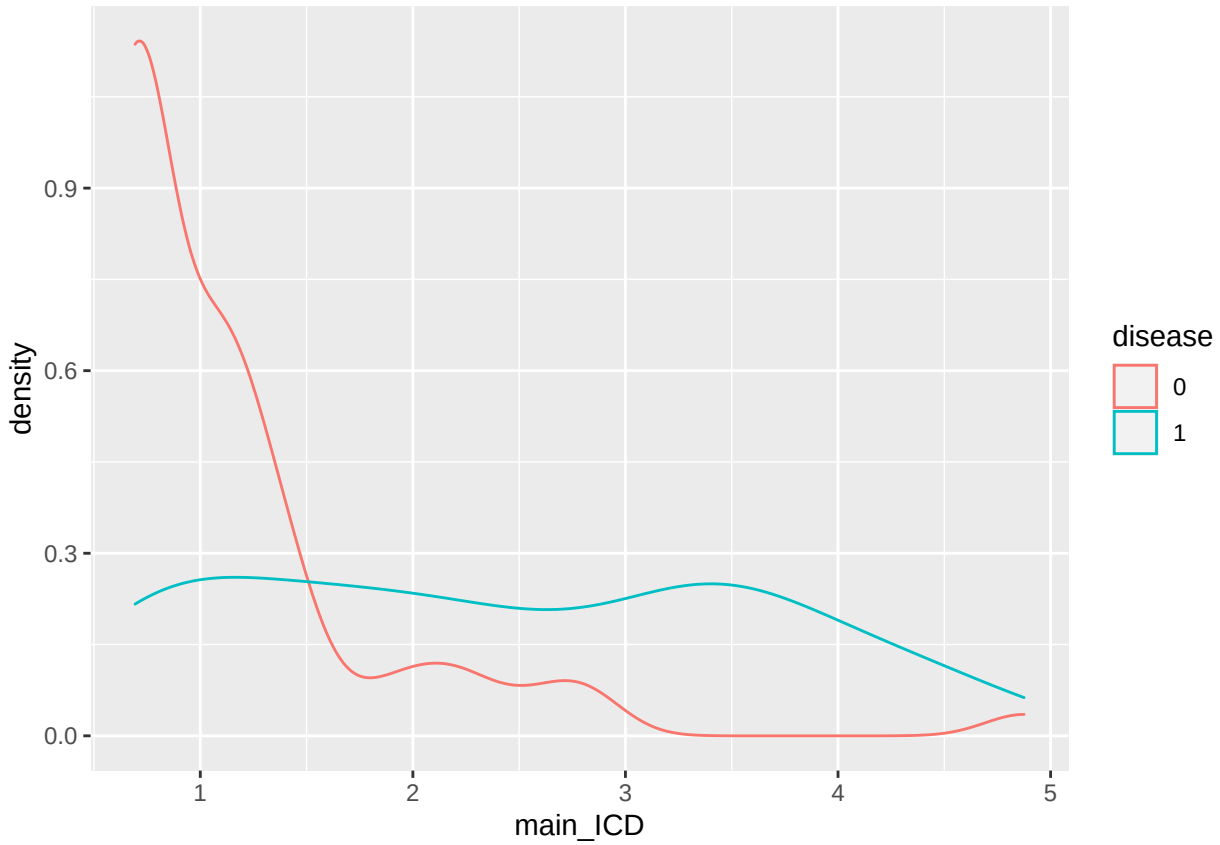
Feature selection

How to select features?

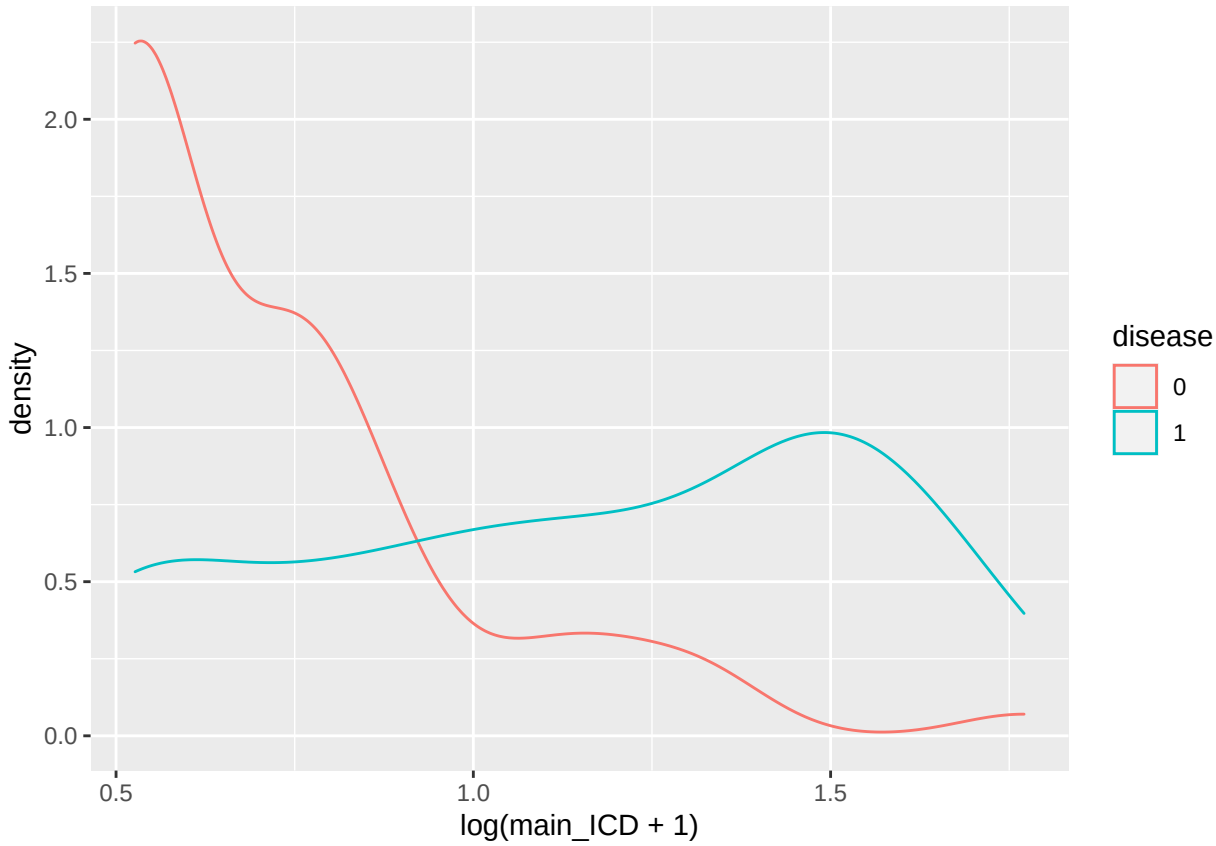
Can leverage some clinical-meaningful features that are related to Y.

e.g. Feature "main_ICD" = the total number of the disease-related billing codes.

```
ehr_data %>%
  filter(!is.na(label)) %>%
  mutate(disease = factor(label)) %>%
  ggplot(aes(x = main_ICD)) +
  geom_density(aes(color = disease))
```

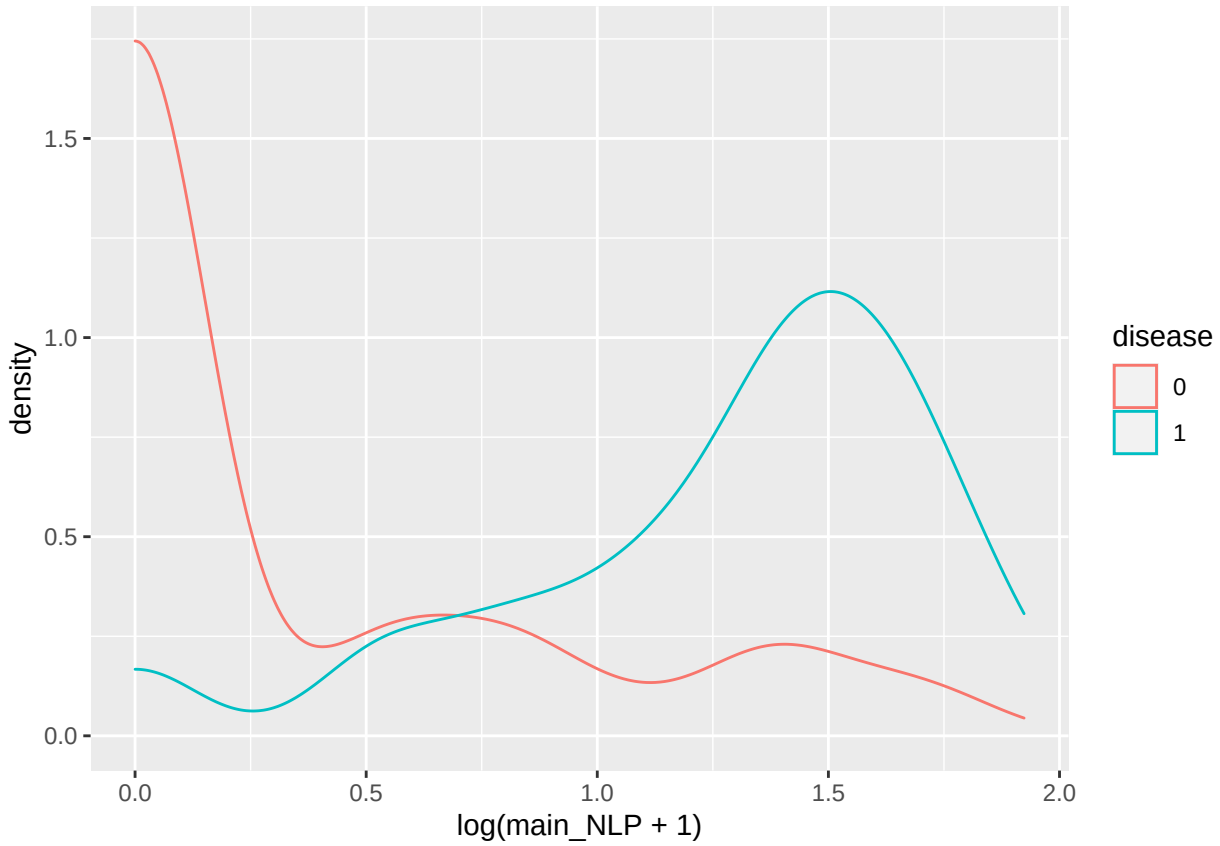



```
# With log transformation.
ehr_data %>%
  filter(!is.na(label)) %>%
  mutate(disease = factor(label)) %>%
  ggplot(aes(x = log(main_ICD + 1))) +
  geom_density(aes(color = disease))
```



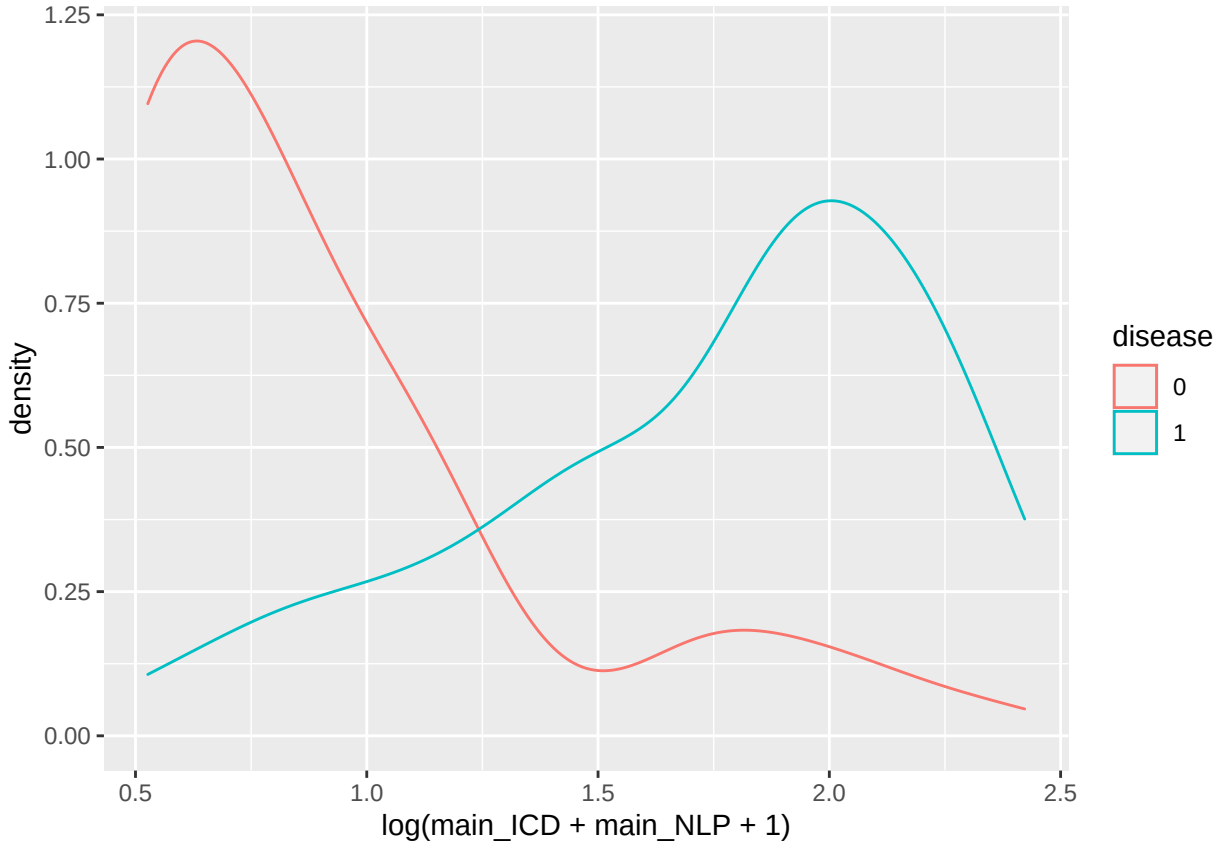
The more the disease-related codes, the more **likely** the patient has the disease.

```
ehr_data %>%
  filter(!is.na(label)) %>%
  mutate(disease = factor(label)) %>%
  ggplot(aes(x = log(main_NLP + 1))) +
  geom_density(aes(color = disease))
```



Other options? - Combine them.

```
ehr_data %>%
  filter(!is.na(label)) %>%
  mutate(disease = factor(label)) %>%
  ggplot(aes(x = log(main_ICD+main_NLP+1))) +
  geom_density(aes(color = disease))
```




```
nonmissing_index <- which(!is.na(ehr_data$label))
y <- ehr_data$label[nonmissing_index]
sicd <- ehr_data$main_ICD
snlp <- ehr_data$main_NLP
sicdnlp <- ehr_data$main_ICD + ehr_data$main_NLP

get_auc(y, sicd[nonmissing_index])
```

```
## [1] 0.8046218
```

```
get_auc(y, snlp[nonmissing_index])
```

```
## [1] 0.8712388
```

```
get_auc(y, sicdnlp[nonmissing_index])
```

```
## [1] 0.8774058
```

We call these highly predictive features of the true disease status "surrogates".

Opportunities of using surrogate features

1. Feature selection to reduce p

2. Algorithm development with limited Y
3. Algorithm validation with limited Y

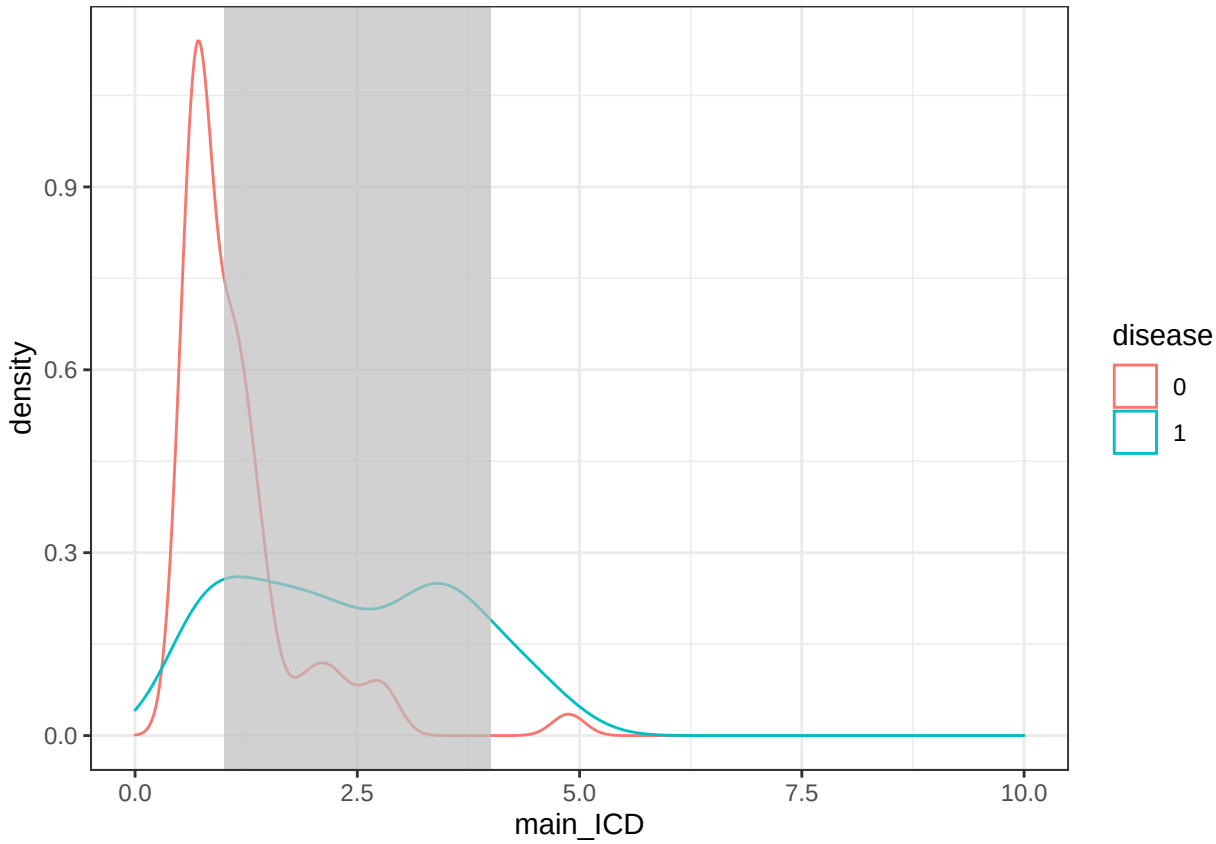
Opportunity 2 and 3 will be covered in the next module!

Feature selection method

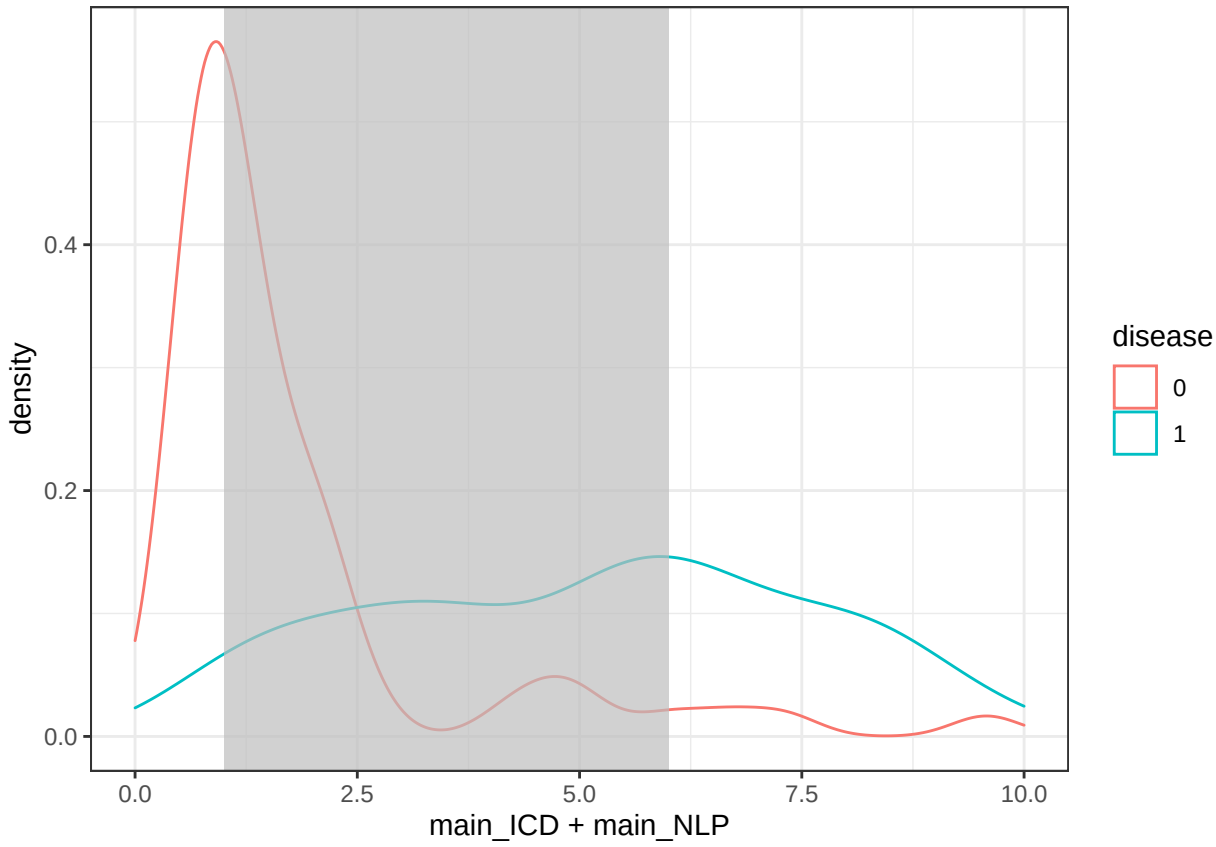
Motivation (Extreme assumption):

- Patients with **high** main ICD or NLP mentions generally have the phenotype.
- Patients with **extremely** low counts are unlikely to have the phenotype.

```
ehr_data %>%
  filter(!is.na(label)) %>%
  mutate(disease = factor(label)) %>%
  ggplot(aes(x = main_ICD)) +
  geom_density(aes(color = disease)) +
  theme_bw() +
  annotate("rect", fill = "grey", alpha = 0.7, xmin = 1, xmax = 4,
           ymin = -Inf, ymax = Inf) +
  xlim(c(0,10))
```



```
ehr_data %>%
  filter(!is.na(label)) %>%
  mutate(disease = factor(label)) %>%
  ggplot(aes(x = main_ICD + main_NLP)) +
  geom_density(aes(color = disease)) +
  theme_bw() +
  annotate("rect", fill = "grey", alpha = 0.7, xmin = 1, xmax = 6,
           ymin = -Inf, ymax = Inf) +
  xlim(c(0,10))
```



- Left white rect: patients not having the disease.
- Right white rect: patients having the disease.

Prepare data for feature selection

Prepare surrogates

Surrogates are available for all the patients!

```
# Prepare 3 surrogates.
sicd <- ehr_data$main_ICD
snlp <- ehr_data$main_NLP
sicdnlp <- ehr_data$main_ICD + ehr_data$main_NLP
```

```r
# Prepare features to be selected.
x <- data.matrix(ehr_data %>% select(starts_with("COD") | starts_with("NLP")))
```

Run surrogate-assisted feature extraction (SAFE) and show result.

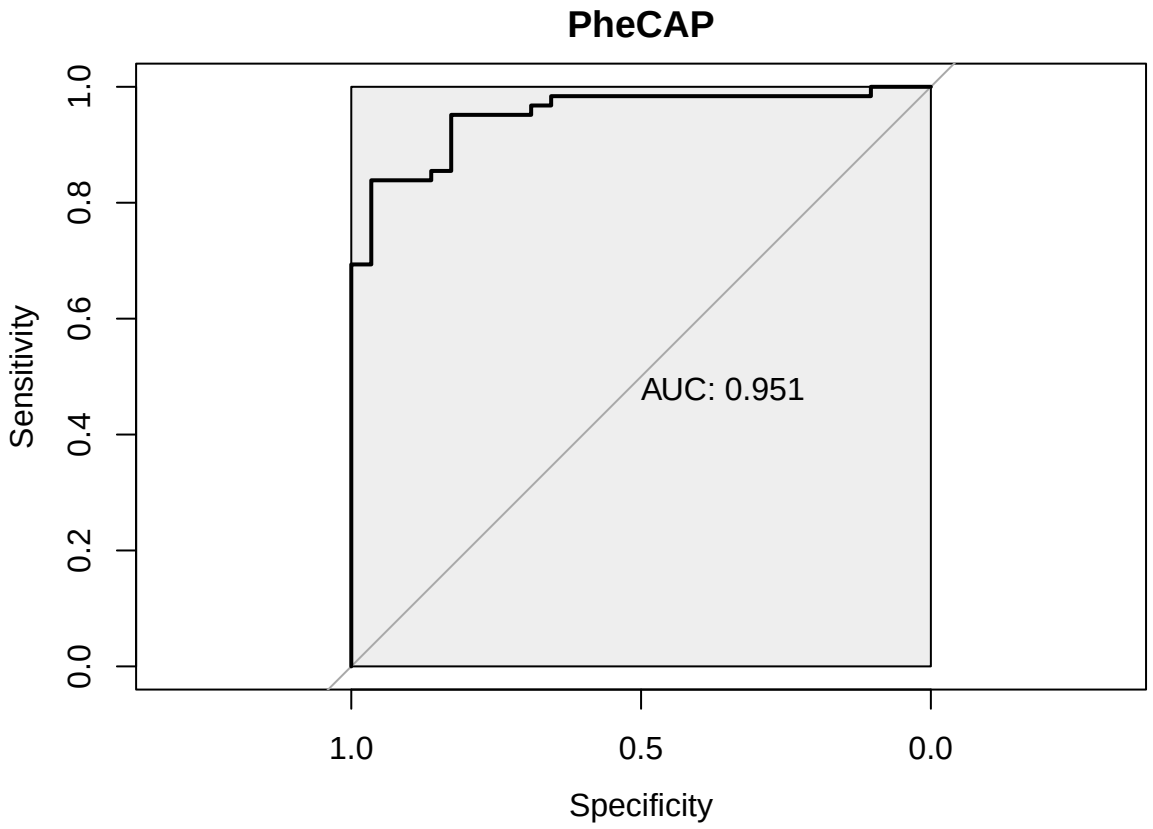
```r
# Truncated using 3 and 1.
SAFE_icd <- extreme_method(sicd, x, u_bound = 4, l_bound = 1)
SAFE_nlp <- extreme_method(snlp, x, u_bound = 4, l_bound = 1)
SAFE_both <- extreme_method(sicd+snlp, x, u_bound = 6, l_bound = 1)

# Majority voting.
beta <- rbind(SAFE_icd$beta_all, SAFE_nlp$beta_all, SAFE_both$beta_all)
SAFE_select <- which(colMeans(beta, na.rm = T) >= 0.5)
SAFE_feature <- colnames(x)[SAFE_select]
SAFE_feature
```

```
##  [1] "NLP56"  "NLP93"  "NLP160" "NLP161" "NLP231" "NLP306" "NLP321" "NLP349"
##  [9] "NLP403" "NLP434" "NLP446"
```

We select features that occur 50% among the three different surrogate-selected feature sets. This is the idea of *majority voting*.

Train phenotyping model and show the AUC on the testing set.

```r
selected_features <- c("main_ICD", "main_NLP", "healthcare_utilization", SAFE_feature)

modelssl_lr <- fit_alasso_bic(
  y = train_y,
  x = train_x[, selected_features]
)

newx <- cbind(1, test_x[, selected_features])

# Prediction on testing set.
y_hat <- expit(newx %*% modelssl_lr$beta_hat) # Inverse Logit

plot(roc(test_y, y_hat),
  print.auc = TRUE,
  max.auc.polygon = TRUE, main = "PheCAP"
)
```

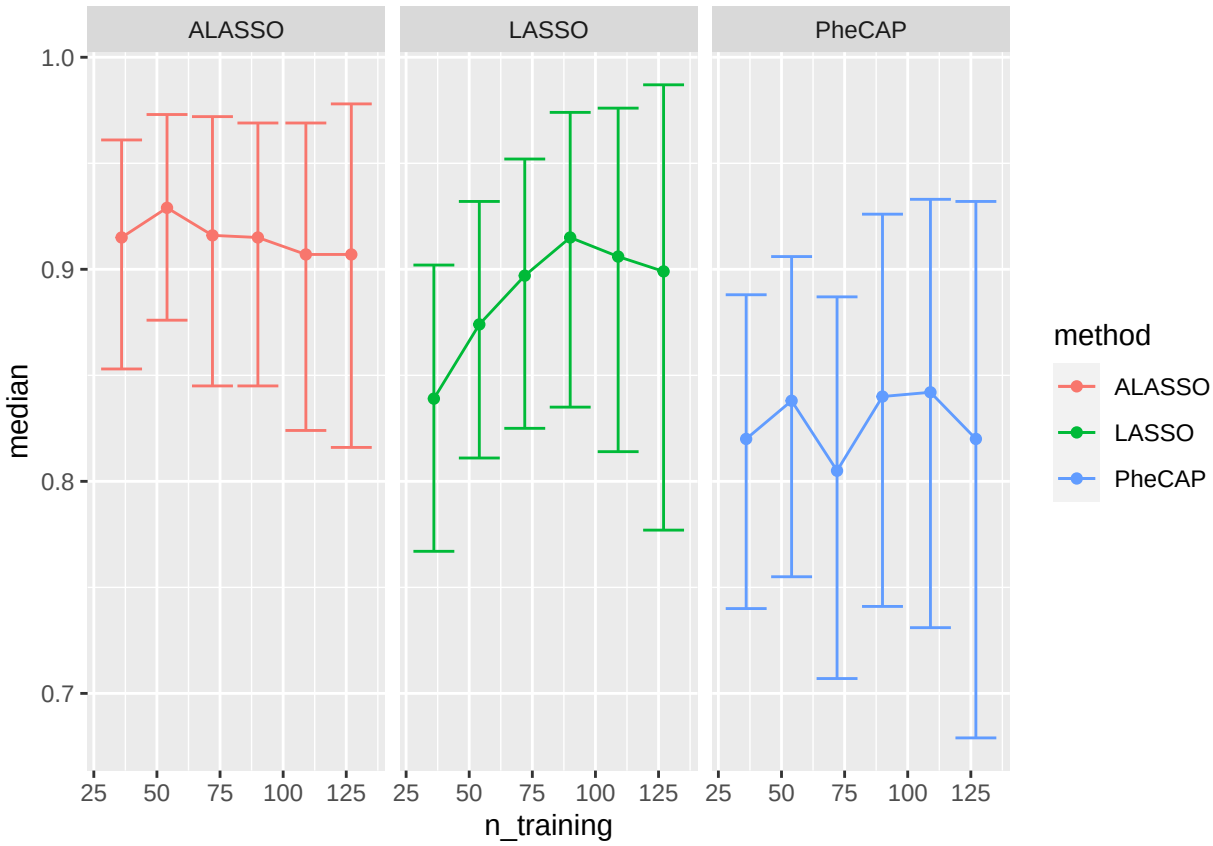



```
head(get_roc(test_y, y_hat))
```

```
##         cutoff    pos.rate        FPR       TPR       PPV       NPV        F1
## [1,] 0.9718918 0.005494505 0.00000000 0.3388930 1.0000000 0.4143530 0.5062286
## [2,] 0.9004385 0.395604396 0.00000000 0.5445931 1.0000000 0.5066810 0.7051606
## [3,] 0.8471048 0.483516484 0.03448276 0.7051613 0.9776386 0.6050130 0.8193403
## [4,] 0.8413351 0.483516484 0.03448276 0.7472581 0.9788718 0.6411724 0.8475258
## [5,] 0.8144090 0.549450549 0.03448276 0.7893548 0.9799760 0.6819289 0.8743970
## [6,] 0.7702933 0.571428571 0.03448276 0.8314516 0.9809705 0.7282185 0.9000436
```

```
labeled_data <- ehr_data %>% dplyr::filter(!is.na(label))

test_auc <- c()
for (i in round(seq(0.2, 0.7, 0.1) * nrow(labeled_data))) {
  set.seed(123456)
  idx <- sample(labeled_data$patient_id, i)
  train_data <- labeled_data %>% filter(patient_id %in% idx)
  test_data <- labeled_data %>% filter(!(patient_id %in% idx))
  idy <- test_data$patient_id

  # LASSO
  metric <- validate_model(
    train_y = train_data$label,
    test_y = test_data$label,
    test_y_hat = lasso_pred(train_data, test_data),
    train_y_hat = lasso_pred(train_data, train_data)
```

```
  )

  # Formatting
  test_auc <- rbind(test_auc, data.frame(
    n_training = i,
    method = "LASSO",
    median = metric$test_AUC,
    L = as.numeric(sub(".*?(\\d+\\.\\d+).*", "\\1", metric$test_CI)),
    U = as.numeric(sub(".*\\b(\\d+\\.\\d+).*", "\\1", metric$test_CI))
  ))

  # ALASSO
  metric <- validate_model(
    train_y = train_data$label,
    test_y = test_data$label,
    test_y_hat = alasso_pred(train_data, test_data),
    train_y_hat = alasso_pred(train_data, train_data)
  )

  test_auc <- rbind(test_auc, data.frame(
    n_training = i,
    method = "ALASSO",
    median = metric$test_AUC,
    L = as.numeric(sub(".*?(\\d+\\.\\d+).*", "\\1", metric$test_CI)),
    U = as.numeric(sub(".*\\b(\\d+\\.\\d+).*", "\\1", metric$test_CI))
  ))

  # PheCAP
  metric <- validate_model(
    train_y = train_data$label,
    test_y = test_data$label,
    test_y_hat = phe_pred(train_data, test_data),
    train_y_hat = phe_pred(train_data, train_data)
  )

  test_auc <- rbind(test_auc, data.frame(
    n_training = i,
    method = "PheCAP",
    median = metric$test_AUC,
    L = as.numeric(sub(".*?(\\d+\\.\\d+).*", "\\1", metric$test_CI)),
    U = as.numeric(sub(".*\\b(\\d+\\.\\d+).*", "\\1", metric$test_CI))
  ))
}

# Facet Plot
test_auc %>% ggplot(aes(
  x = n_training, y = median,
  group = method, color = method
)) +
  geom_point() +
  geom_line() +
  geom_errorbar(aes(ymin = L, ymax = U)) +
  facet_grid(. ~ method)
```



Save the data and feature selected for module 4 and model fitting.

```
save(list = ls(), file = "../module4/environment.RData")
```