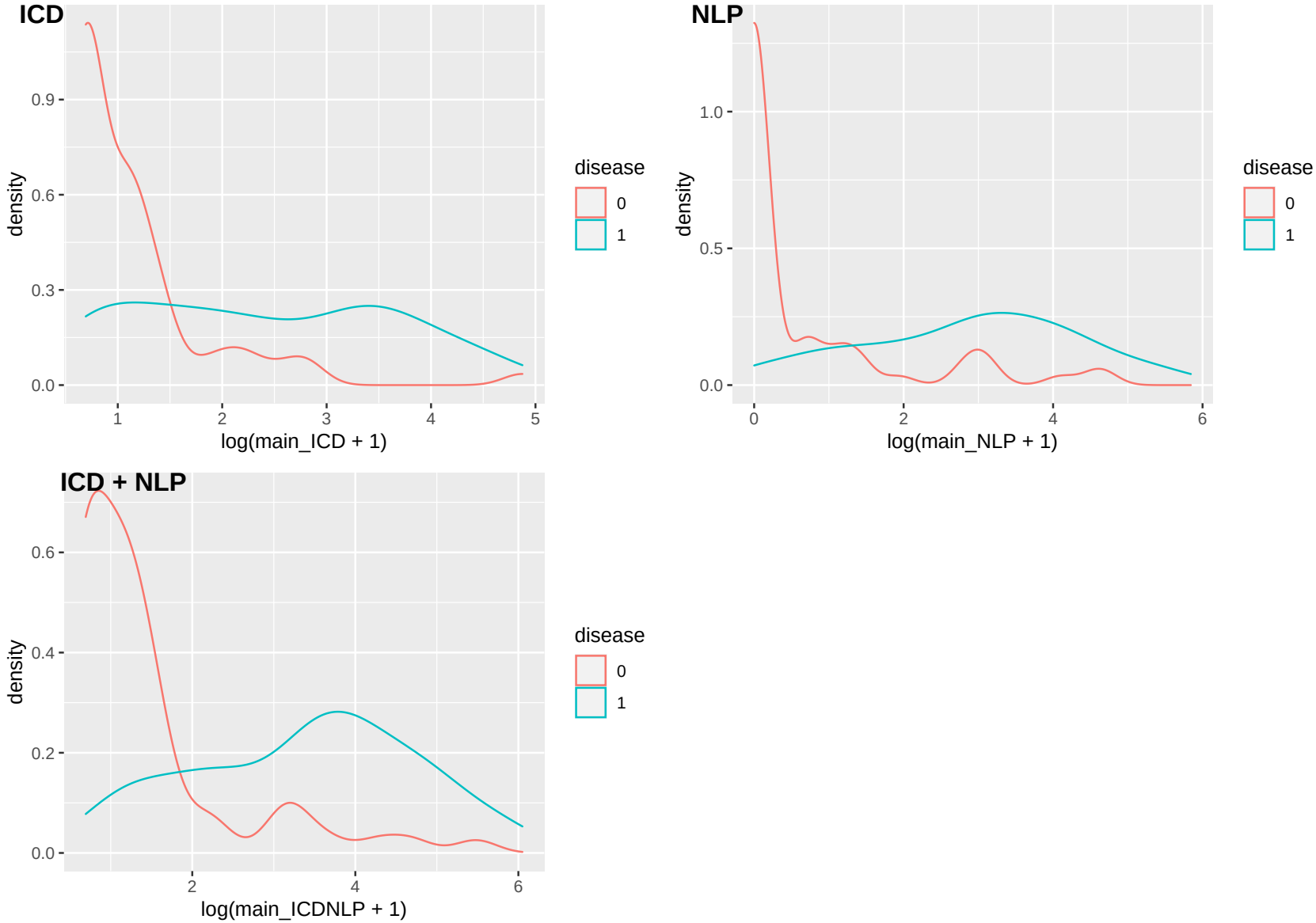
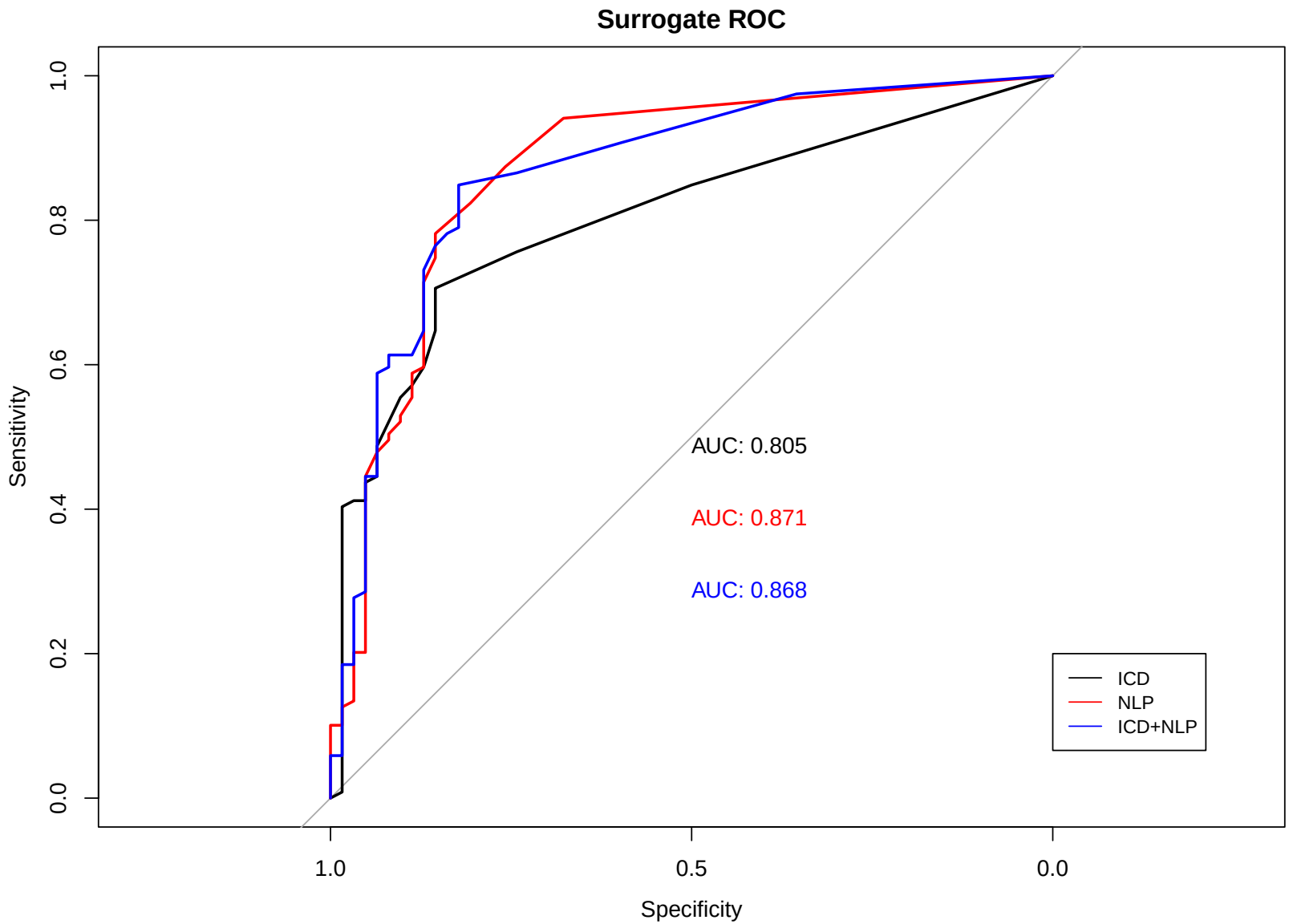
Module 3: Semi-supervised learning (PheCAP)

Surrogates for CAD



The more the disease-related codes and NLP mentions, the more **likely** the patient has the disease

ROC Surrogates



Step 1: SAFE

```
surrogates <- list(
  PhecapSurrogate(
    variable_names = "main_ICD",
    lower_cutoff = 1, upper_cutoff = 10),
  PhecapSurrogate(
    variable_names = "main_NLP",
    lower_cutoff = 1, upper_cutoff = 10),
  PhecapSurrogate(
    variable_names = c("main_ICD", "main_NLP"),
    lower_cutoff = 1, upper_cutoff = 10))

feature_selected <- phecap_run_feature_extraction(data, su
```

Step 2: Orthogonalization + supervised learning

```
phecap_lasso <- phecap_train_phenotyping_model(data, surrog
                                               method = "lasso_cv"
```

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

```
plot(roc.lasso,
  print.auc = TRUE, main = "n_training = 90 (50%)"
)
plot(roc.alasso,
  print.auc = TRUE, col = 'red', add = TRUE, print.auc.y =
)
plot(roc.phecap,
  print.auc = TRUE, col = 'blue', add = TRUE, print.auc.y =
)
legend(0, 0.2, legend = c("LASSO", "ALASSO", "PheCAP"), col
       lty = 1, cex = 0.8)
```

Supervised learning vs. PheCAP for different training size

```
# selected_index <- which(colnames(ehr_data) %in% vars)
# start<- Sys.time()
# auc_phecap <- validate_phecap(dat = labeled_data, nsim =
#                                 n.train = c(50, 70, 90),
#                                 selected_features = select
# end <- Sys.time()
# end - start
#
# auc_all <- cbind(auc_supervised, auc_phecap)
```

```
# par(mfrow = c(1, 3))
# boxplot(auc_all %>% select(starts_with("n=50")),
#   ylim = c(0.5, 1),
#   names = c("LASSO", "ALASSO", "PheCAP"), main = "n=50"
# )
# boxplot(auc_all %>% select(starts_with("n=70")),
#   ylim = c(0.5, 1),
#   names =  c("LASSO", "ALASSO", "PheCAP"), main = "n=70"
```