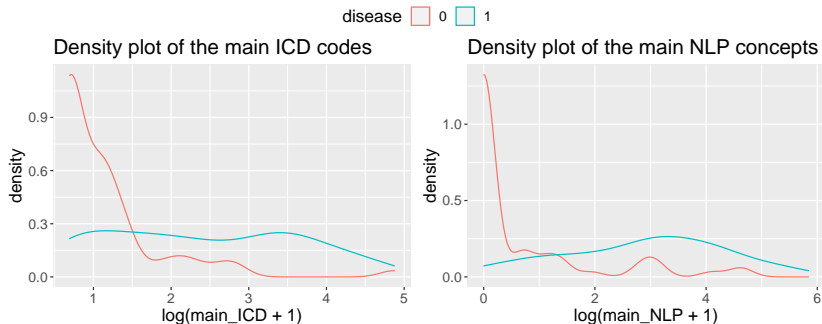# Module 3: Semi-supervised learning (PheCAP)
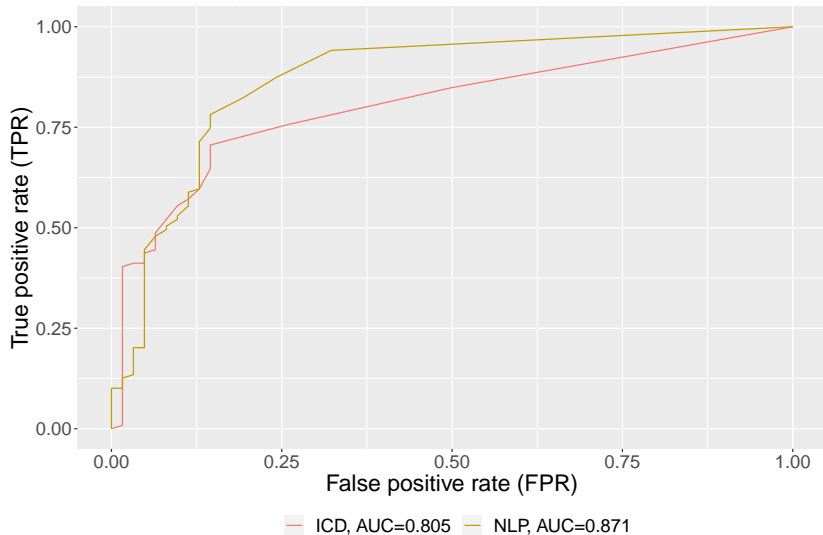
Surrogates for CAD

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

ROC Surrogates

The operating receiver characteristic (ROC) curve

ICD, AUC=0.805   NLP, AUC=0.871

Step 1: SAFE

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

feature_selected <- phecap_run_feature_extraction(data, surrogates)
feature_selected
```

```
## Feature(s) selected by surrogate-assisted feature extraction (SAFE)
## [1] "main_ICD" "main_NLP" "NLP56"    "NLP93"    "NLP274"    "NLP306"
```

Step 2: Orthogonalization + supervised learning

```
phecap_lasso <- phecap_train_phenotyping_model(
  data, surrogates, feature_selected,
  method = "lasso_cv"
)

phecap_lasso

## Phenotyping model:
## $lasso_cv
##             (Intercept)                 main_ICD                  main_NLP
##               1.9258667                0.2157399                 1.1666409
## healthcare_utilization                    NLP56                     NLP93
##              -0.9772753                0.0000000                -0.3242900
##                  NLP274                   NLP306
##               0.0000000                0.0000000
##
## AUC on training data: 0.93
## Average AUC on random splits: 0.889
```

Supervised learning (LASSO, ALASSO) vs. PheCAP

The operating receiver characteristic (ROC) curve

LASSO, AUC=0.914
ALASSO, AUC=0.905
PheCAP, AUC=0.937
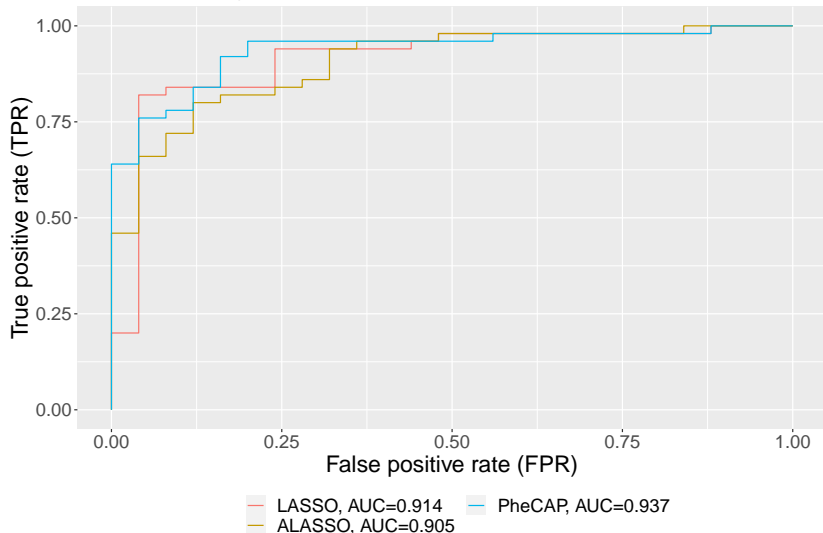
Supervised learning (LASSO, ALASSO) vs. PheCAP at FPR = 0.10

```
get_roc_parameter(0.1, roc_full_lasso)
```

```
##     cutoff  pos.rate FPR TPR       PPV       NPV       F1
## 1 0.8573308 0.5933333 0.1 0.84 0.9438202 0.7377049 0.8888889
```

```
get_roc_parameter(0.1, roc_full_alasso)
```

```
##     cutoff  pos.rate FPR TPR       PPV       NPV       F1
## 1 0.9228109 0.5133333 0.1 0.72 0.9350649 0.6164384 0.8135593
```

```
roc_full_phecap <- get_roc(y_true = test_y, y_score = y_hat_phecap) %>% data.frame()
get_roc_parameter(0.1, roc_full_phecap)
```

```
##     cutoff  pos.rate FPR TPR       PPV       NPV       F1
## 1 0.8342308 0.5533333 0.1 0.78 0.939759 0.6716418 0.852459
```

Supervised learning vs. PheCAP for different training size

▶ Randomly sample training size $= 50, 70, 90$
▶ Use the remaining data as the test set
▶ Repeat 600 times

```r
auc_phecap <- validate_phecap(
  dat = labeled_data,
  surrogates = surrogates,
  feature_selected = feature_selected,
  nsim = 600,
  ntrain = c(50, 70, 90)
)
```

Supervised learning vs. PheCAP for different training size

Area under the ROC curve (AUC) from 600 simulations