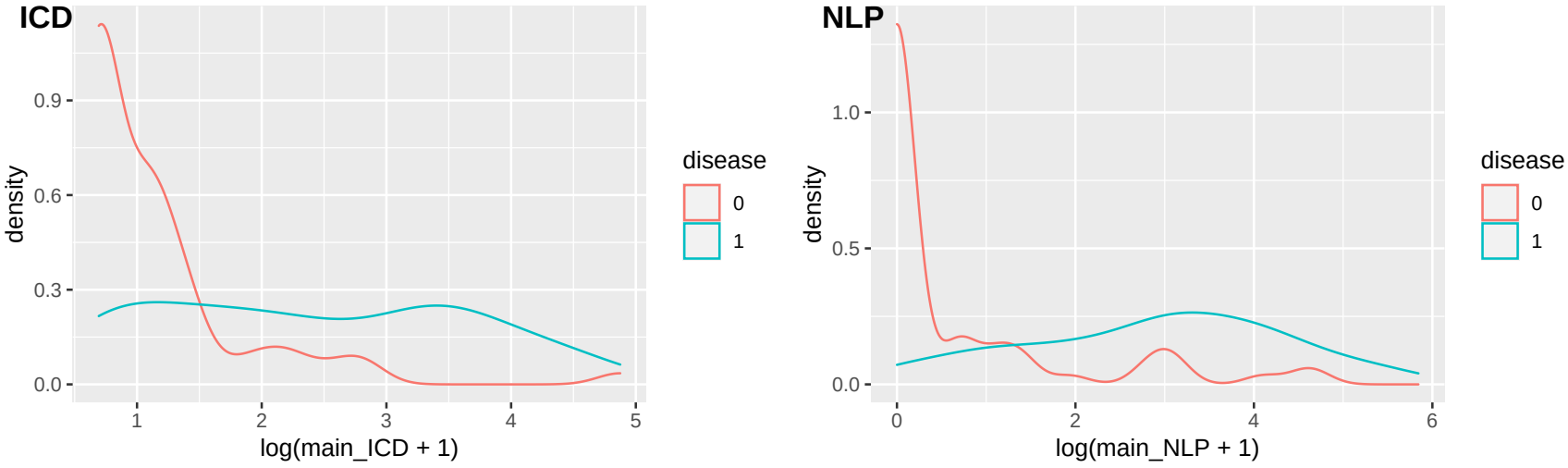
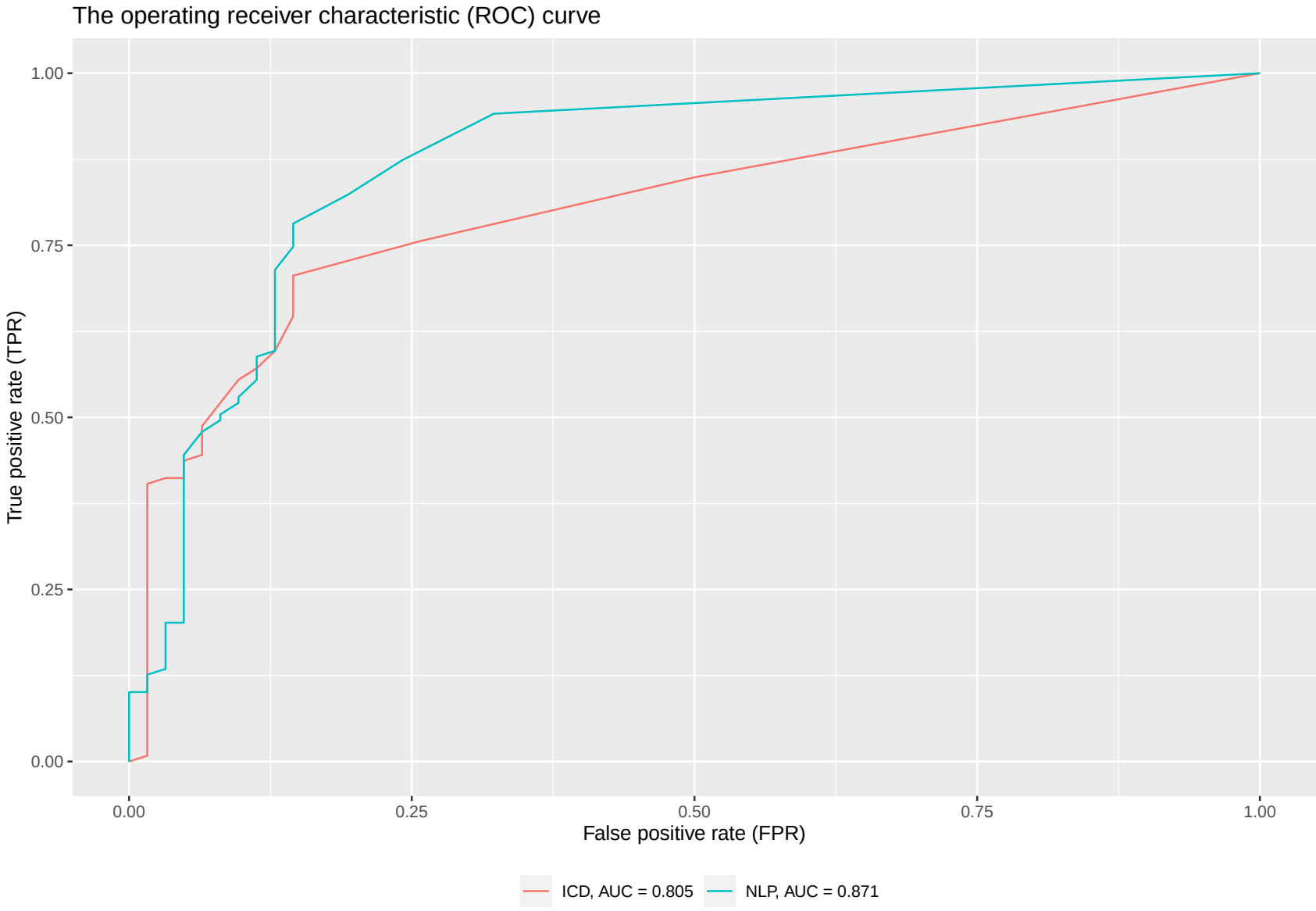
# Module 3: Semi-supervised learning (PheCAP)

# 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))

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"
```

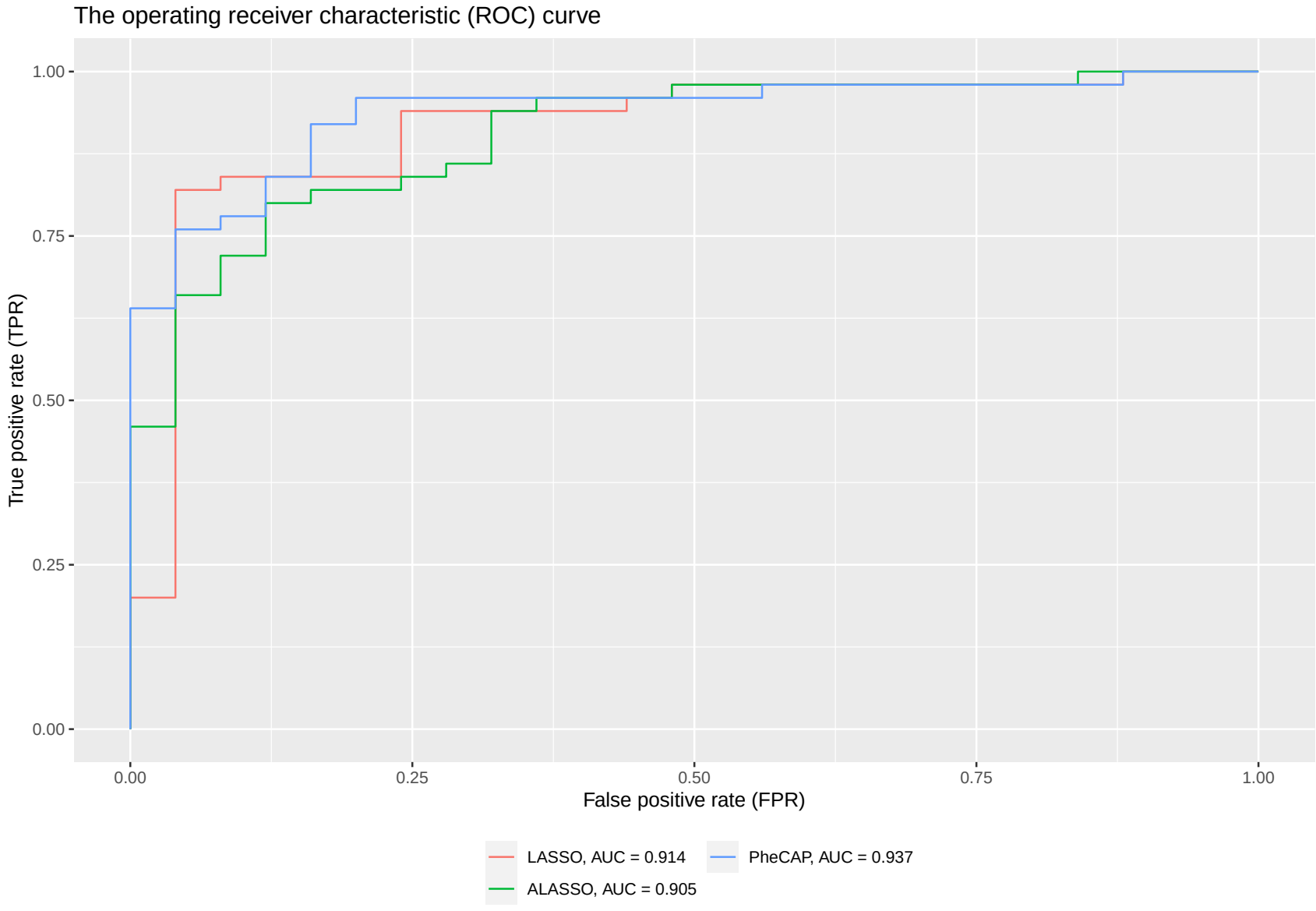
# Step 2: Orthogonalization + supervised learning

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

```
phecap_lasso
```

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

# Supervised learning (LASSO, ALASSO) vs. PheCAP



# 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

```
auc_phecap <- validate_phecap(dat = labeled_data,
                              orig_data = ehr_data,
                              surrogates = surrogates,
                              feature_selected = feature_selected,
                              nsim = 600,
                              n.train = c(50, 70, 90))
```

# Supervised learning vs. PheCAP for different training size