

Module 4: Alternative approaches

2-step Semi-supervised Approach

1. Regress the surrogate on the features with penalized least square to get the direction of β .

```
# COD + NLP + HU.
```

```
x <- log(ehr_data %>% select(starts_with("health") |  
  starts_with("COD") | starts_with("NLP")) + 1)  
S <- log(ehr_data$main_ICD + ehr_data$main_NLP + 1)
```

```
# Step 1.
```

```
beta_step1 <- adaptive_lasso_fit(  
  y = S[, # surrogate  
  x = x[, # all X  
  family = "gaussian",  
  tuning = "cv"  
)
```

2-step Semi-supervised Approach

1. Regress the surrogate on the features with penalized least square to get the direction of β .
2. Regress the outcome on the linear predictor to get the intercept and multiplier for the β .

```
# Linear predictor without intercept.
bhatx <- linear_model_predict(beta = beta_step1, x = as.matrix(x))

# Step 2.
step2 <- glm(
  train_y ~ bhatx[train_data$patient_id] + S[train_data$patient_id],
  family = "binomial"
)
beta_step2 <- coef(step2)
beta_step2
```

```
##                (Intercept) bhatx[train_data$patient_id]
##                -1.9461028                0.7057629
##      S[train_data$patient_id]
##                0.5988575
```

```
# Recover beta.
beta <- beta_step2[2] * beta_step1
```

Compare selected features

```
# LASSO.
```

```
names(beta_lasso[!beta_lasso == 0])[-1]
```

```
## [1] "COD2"      "COD10"     "NLP1"      "NLP17"     "NLP56"     "NLP82"
## [7] "NLP93"     "NLP104"    "NLP118"    "NLP130"    "NLP144"    "NLP164"
## [13] "NLP172"    "NLP193"    "NLP199"    "NLP222"    "NLP231"    "NLP265"
## [19] "NLP274"    "NLP280"    "NLP297"    "NLP299"    "NLP346"    "NLP362"
## [25] "NLP375"    "NLP382"    "NLP396"    "NLP401"    "NLP409"    "NLP435"
## [31] "NLP451"    "NLP462"    "NLP488"    "NLP533"    "NLP536"    "NLP552"
## [37] "NLP568"    "main_NLP"
```

```
# ALASSO.
```

```
names(beta_alasso[!beta_alasso == 0])[-1]
```

```
## [1] "NLP56"     "NLP93"     "NLP104"    "NLP118"    "NLP222"    "NLP231"
## [7] "NLP265"    "NLP280"    "NLP297"    "NLP299"    "NLP409"    "NLP536"
## [13] "main_NLP"
```

```
# PheCAP.
```

```
feature_selected
```

```
## Feature(s) selected by surrogate-assisted feature extraction (SAFE)
```

```
## [1] "main_ICD" "main_NLP" "NLP56"     "NLP93"     "NLP274"    "NLP306"
```

```
# Two Step.
```

```
names(beta[!beta == 0])[-1]
```

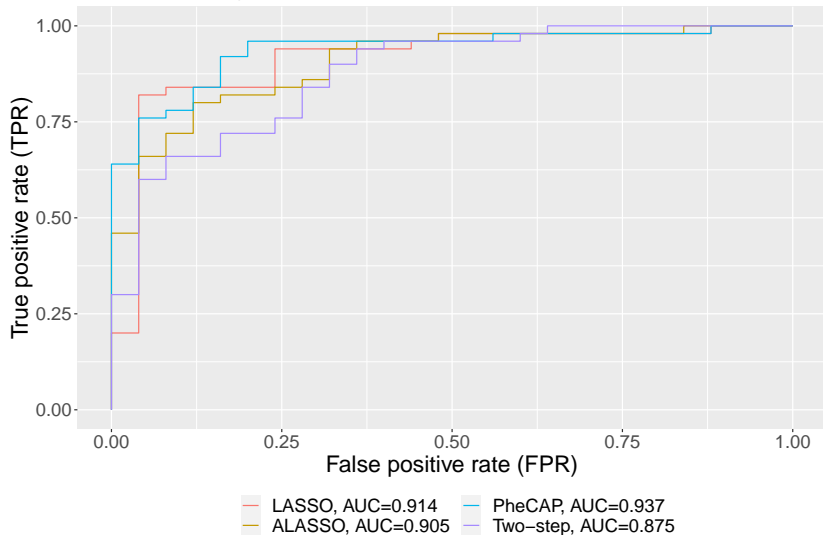
```
## [1] "COD6"      "COD8"      "COD10"     "NLP7"      "NLP14"     "NLP21"     "NLP24"     "NLP28"
## [9] "NLP31"     "NLP33"     "NLP44"     "NLP50"     "NLP56"     "NLP59"     "NLP61"     "NLP62"
## [17] "NLP66"     "NLP68"     "NLP70"     "NLP73"     "NLP74"     "NLP76"     "NLP81"     "NLP89"
## [25] "NLP92"     "NLP93"     "NLP95"     "NLP98"     "NLP102"    "NLP104"    "NLP108"    "NLP110"
## [33] "NLP116"    "NLP127"    "NLP130"    "NLP146"    "NLP160"    "NLP161"    "NLP172"    "NLP176"
## [41] "NLP178"    "NLP179"    "NLP183"    "NLP189"    "NLP190"    "NLP192"    "NLP199"    "NLP202"
## [49] "NLP208"    "NLP209"    "NLP215"    "NLP235"    "NLP236"    "NLP237"    "NLP242"    "NLP243"
## [57] "NLP244"    "NLP245"    "NLP246"    "NLP247"    "NLP248"    "NLP249"    "NLP250"    "NLP251"
## [65] "NLP252"    "NLP253"    "NLP254"    "NLP255"    "NLP256"    "NLP257"    "NLP258"    "NLP259"
## [73] "NLP260"    "NLP261"    "NLP262"    "NLP263"    "NLP264"    "NLP265"    "NLP266"    "NLP267"
## [81] "NLP268"    "NLP269"    "NLP270"    "NLP271"    "NLP272"    "NLP273"    "NLP274"    "NLP275"
## [89] "NLP276"    "NLP277"    "NLP278"    "NLP279"    "NLP280"    "NLP281"    "NLP282"    "NLP283"
## [97] "NLP284"    "NLP285"    "NLP286"    "NLP287"    "NLP288"    "NLP289"    "NLP290"    "NLP291"
## [105] "NLP292"    "NLP293"    "NLP294"    "NLP295"    "NLP296"    "NLP297"    "NLP298"    "NLP299"
## [113] "NLP300"    "NLP301"    "NLP302"    "NLP303"    "NLP304"    "NLP305"    "NLP306"    "NLP307"
## [121] "NLP308"    "NLP309"    "NLP310"    "NLP311"    "NLP312"    "NLP313"    "NLP314"    "NLP315"
## [129] "NLP316"    "NLP317"    "NLP318"    "NLP319"    "NLP320"    "NLP321"    "NLP322"    "NLP323"
## [137] "NLP324"    "NLP325"    "NLP326"    "NLP327"    "NLP328"    "NLP329"    "NLP330"    "NLP331"
## [145] "NLP332"    "NLP333"    "NLP334"    "NLP335"    "NLP336"    "NLP337"    "NLP338"    "NLP339"
## [153] "NLP340"    "NLP341"    "NLP342"    "NLP343"    "NLP344"    "NLP345"    "NLP346"    "NLP347"
## [161] "NLP348"    "NLP349"    "NLP350"    "NLP351"    "NLP352"    "NLP353"    "NLP354"    "NLP355"
## [169] "NLP356"    "NLP357"    "NLP358"    "NLP359"    "NLP360"    "NLP361"    "NLP362"    "NLP363"
## [177] "NLP364"    "NLP365"    "NLP366"    "NLP367"    "NLP368"    "NLP369"    "NLP370"    "NLP371"
## [185] "NLP372"    "NLP373"    "NLP374"    "NLP375"    "NLP376"    "NLP377"    "NLP378"    "NLP379"
## [193] "NLP380"    "NLP381"    "NLP382"    "NLP383"    "NLP384"    "NLP385"    "NLP386"    "NLP387"
## [201] "NLP388"    "NLP389"    "NLP390"    "NLP391"    "NLP392"    "NLP393"    "NLP394"    "NLP395"
## [209] "NLP396"    "NLP397"    "NLP398"    "NLP399"    "NLP400"    "NLP401"    "NLP402"    "NLP403"
## [217] "NLP404"    "NLP405"    "NLP406"    "NLP407"    "NLP408"    "NLP409"    "NLP410"    "NLP411"
## [225] "NLP412"    "NLP413"    "NLP414"    "NLP415"    "NLP416"    "NLP417"    "NLP418"    "NLP419"
## [233] "NLP420"    "NLP421"    "NLP422"    "NLP423"    "NLP424"    "NLP425"    "NLP426"    "NLP427"
## [241] "NLP428"    "NLP429"    "NLP430"    "NLP431"    "NLP432"    "NLP433"    "NLP434"    "NLP435"
## [249] "NLP436"    "NLP437"    "NLP438"    "NLP439"    "NLP440"    "NLP441"    "NLP442"    "NLP443"
## [257] "NLP444"    "NLP445"    "NLP446"    "NLP447"    "NLP448"    "NLP449"    "NLP450"    "NLP451"
## [265] "NLP452"    "NLP453"    "NLP454"    "NLP455"    "NLP456"    "NLP457"    "NLP458"    "NLP459"
## [273] "NLP460"    "NLP461"    "NLP462"    "NLP463"    "NLP464"    "NLP465"    "NLP466"    "NLP467"
## [281] "NLP468"    "NLP469"    "NLP470"    "NLP471"    "NLP472"    "NLP473"    "NLP474"    "NLP475"
## [289] "NLP476"    "NLP477"    "NLP478"    "NLP479"    "NLP480"    "NLP481"    "NLP482"    "NLP483"
## [297] "NLP484"    "NLP485"    "NLP486"    "NLP487"    "NLP488"    "NLP489"    "NLP490"    "NLP491"
## [305] "NLP492"    "NLP493"    "NLP494"    "NLP495"    "NLP496"    "NLP497"    "NLP498"    "NLP499"
## [313] "NLP500"    "NLP501"    "NLP502"    "NLP503"    "NLP504"    "NLP505"    "NLP506"    "NLP507"
## [321] "NLP508"    "NLP509"    "NLP510"    "NLP511"    "NLP512"    "NLP513"    "NLP514"    "NLP515"
## [329] "NLP516"    "NLP517"    "NLP518"    "NLP519"    "NLP520"    "NLP521"    "NLP522"    "NLP523"
## [337] "NLP524"    "NLP525"    "NLP526"    "NLP527"    "NLP528"    "NLP529"    "NLP530"    "NLP531"
## [345] "NLP532"    "NLP533"    "NLP534"    "NLP535"    "NLP536"    "NLP537"    "NLP538"    "NLP539"
## [353] "NLP540"    "NLP541"    "NLP542"    "NLP543"    "NLP544"    "NLP545"    "NLP546"    "NLP547"
## [361] "NLP548"    "NLP549"    "NLP550"    "NLP551"    "NLP552"    "NLP553"    "NLP554"    "NLP555"
## [369] "NLP556"    "NLP557"    "NLP558"    "NLP559"    "NLP560"    "NLP561"    "NLP562"    "NLP563"
## [377] "NLP564"    "NLP565"    "NLP566"    "NLP567"    "NLP568"    "NLP569"    "NLP570"    "NLP571"
## [385] "NLP572"    "NLP573"    "NLP574"    "NLP575"    "NLP576"    "NLP577"    "NLP578"    "NLP579"
## [393] "NLP580"    "NLP581"    "NLP582"    "NLP583"    "NLP584"    "NLP585"    "NLP586"    "NLP587"
## [401] "NLP588"    "NLP589"    "NLP590"    "NLP591"    "NLP592"    "NLP593"    "NLP594"    "NLP595"
## [409] "NLP596"    "NLP597"    "NLP598"    "NLP599"    "NLP600"    "NLP601"    "NLP602"    "NLP603"
## [417] "NLP604"    "NLP605"    "NLP606"    "NLP607"    "NLP608"    "NLP609"    "NLP610"    "NLP611"
## [425] "NLP612"    "NLP613"    "NLP614"    "NLP615"    "NLP616"    "NLP617"    "NLP618"    "NLP619"
## [433] "NLP620"    "NLP621"    "NLP622"    "NLP623"    "NLP624"    "NLP625"    "NLP626"    "NLP627"
## [441] "NLP628"    "NLP629"    "NLP630"    "NLP631"    "NLP632"    "NLP633"    "NLP634"    "NLP635"
## [449] "NLP636"    "NLP637"    "NLP638"    "NLP639"    "NLP640"    "NLP641"    "NLP642"    "NLP643"
## [457] "NLP644"    "NLP645"    "NLP646"    "NLP647"    "NLP648"    "NLP649"    "NLP650"    "NLP651"
## [465] "NLP652"    "NLP653"    "NLP654"    "NLP655"    "NLP656"    "NLP657"    "NLP658"    "NLP659"
## [473] "NLP660"    "NLP661"    "NLP662"    "NLP663"    "NLP664"    "NLP665"    "NLP666"    "NLP667"
## [481] "NLP668"    "NLP669"    "NLP670"    "NLP671"    "NLP672"    "NLP673"    "NLP674"    "NLP675"
## [489] "NLP676"    "NLP677"    "NLP678"    "NLP679"    "NLP680"    "NLP681"    "NLP682"    "NLP683"
## [497] "NLP684"    "NLP685"    "NLP686"    "NLP687"    "NLP688"    "NLP689"    "NLP690"    "NLP691"
## [505] "NLP692"    "NLP693"    "NLP694"    "NLP695"    "NLP696"    "NLP697"    "NLP698"    "NLP699"
## [513] "NLP700"    "NLP701"    "NLP702"    "NLP703"    "NLP704"    "NLP705"    "NLP706"    "NLP707"
## [521] "NLP708"    "NLP709"    "NLP710"    "NLP711"    "NLP712"    "NLP713"    "NLP714"    "NLP715"
## [529] "NLP716"    "NLP717"    "NLP718"    "NLP719"    "NLP720"    "NLP721"    "NLP722"    "NLP723"
## [537] "NLP724"    "NLP725"    "NLP726"    "NLP727"    "NLP728"    "NLP729"    "NLP730"    "NLP731"
## [545] "NLP732"    "NLP733"    "NLP734"    "NLP735"    "NLP736"    "NLP737"    "NLP738"    "NLP739"
## [553] "NLP740"    "NLP741"    "NLP742"    "NLP743"    "NLP744"    "NLP745"    "NLP746"    "NLP747"
## [561] "NLP748"    "NLP749"    "NLP750"    "NLP751"    "NLP752"    "NLP753"    "NLP754"    "NLP755"
## [569] "NLP756"    "NLP757"    "NLP758"    "NLP759"    "NLP760"    "NLP761"    "NLP762"    "NLP763"
## [577] "NLP764"    "NLP765"    "NLP766"    "NLP767"    "NLP768"    "NLP769"    "NLP770"    "NLP771"
## [585] "NLP772"    "NLP773"    "NLP774"    "NLP775"    "NLP776"    "NLP777"    "NLP778"    "NLP779"
## [593] "NLP780"    "NLP781"    "NLP782"    "NLP783"    "NLP784"    "NLP785"    "NLP786"    "NLP787"
## [601] "NLP788"    "NLP789"    "NLP790"    "NLP791"    "NLP792"    "NLP793"    "NLP794"    "NLP795"
## [609] "NLP796"    "NLP797"    "NLP798"    "NLP799"    "NLP800"    "NLP801"    "NLP802"    "NLP803"
## [617] "NLP804"    "NLP805"    "NLP806"    "NLP807"    "NLP808"    "NLP809"    "NLP810"    "NLP811"
## [625] "NLP812"    "NLP813"    "NLP814"    "NLP815"    "NLP816"    "NLP817"    "NLP818"    "NLP819"
## [633] "NLP820"    "NLP821"    "NLP822"    "NLP823"    "NLP824"    "NLP825"    "NLP826"    "NLP827"
## [641] "NLP828"    "NLP829"    "NLP830"    "NLP831"    "NLP832"    "NLP833"    "NLP834"    "NLP835"
## [649] "NLP836"    "NLP837"    "NLP838"    "NLP839"    "NLP840"    "NLP841"    "NLP842"    "NLP843"
## [657] "NLP844"    "NLP845"    "NLP846"    "NLP847"    "NLP848"    "NLP849"    "NLP850"    "NLP851"
## [665] "NLP852"    "NLP853"    "NLP854"    "NLP855"    "NLP856"    "NLP857"    "NLP858"    "NLP859"
## [673] "NLP860"    "NLP861"    "NLP862"    "NLP863"    "NLP864"    "NLP865"    "NLP866"    "NLP867"
## [681] "NLP868"    "NLP869"    "NLP870"    "NLP871"    "NLP872"    "NLP873"    "NLP874"    "NLP875"
## [689] "NLP876"    "NLP877"    "NLP878"    "NLP879"    "NLP880"    "NLP881"    "NLP882"    "NLP883"
## [697] "NLP884"    "NLP885"    "NLP886"    "NLP887"    "NLP888"    "NLP889"    "NLP890"    "NLP891"
## [705] "NLP892"    "NLP893"    "NLP894"    "NLP895"    "NLP896"    "NLP897"    "NLP898"    "NLP899"
## [713] "NLP900"    "NLP901"    "NLP902"    "NLP903"    "NLP904"    "NLP905"    "NLP906"    "NLP907"
## [721] "NLP908"    "NLP909"    "NLP910"    "NLP911"    "NLP912"    "NLP913"    "NLP914"    "NLP915"
## [729] "NLP916"    "NLP917"    "NLP918"    "NLP919"    "NLP920"    "NLP921"    "NLP922"    "NLP923"
## [737] "NLP924"    "NLP925"    "NLP926"    "NLP927"    "NLP928"    "NLP929"    "NLP930"    "NLP931"
## [745] "NLP932"    "NLP933"    "NLP934"    "NLP935"    "NLP936"    "NLP937"    "NLP938"    "NLP939"
## [753] "NLP940"    "NLP941"    "NLP942"    "NLP943"    "NLP944"    "NLP945"    "NLP946"    "NLP947"
## [761] "NLP948"    "NLP949"    "NLP950"    "NLP951"    "NLP952"    "NLP953"    "NLP954"    "NLP955"
## [769] "NLP956"    "NLP957"    "NLP958"    "NLP959"    "NLP960"    "NLP961"    "NLP962"    "NLP963"
## [777] "NLP964"    "NLP965"    "NLP966"    "NLP967"    "NLP968"    "NLP969"    "NLP970"    "NLP971"
## [785] "NLP972"    "NLP973"    "NLP974"    "NLP975"    "NLP976"    "NLP977"    "NLP978"    "NLP979"
## [793] "NLP980"    "NLP981"    "NLP982"    "NLP983"    "NLP984"    "NLP985"    "NLP986"    "NLP987"
## [801] "NLP988"    "NLP989"    "NLP990"    "NLP991"    "NLP992"    "NLP993"    "NLP994"    "NLP995"
## [809] "NLP996"    "NLP997"    "NLP998"    "NLP999"    "main_ICD"    "main_NLP"
```

ROC

```
mu <- beta_step2[1] +  
  as.numeric(as.matrix(x[test_data$patient_id, ]  
    %*% beta[-1]) +  
  as.numeric(beta_step2[3] %*% S[test_data$patient_id])  
  
# Expit.  
y_hat_twostep <- plogis(mu)  
  
roc_twostep <- roc(test_y, y_hat_twostep)
```

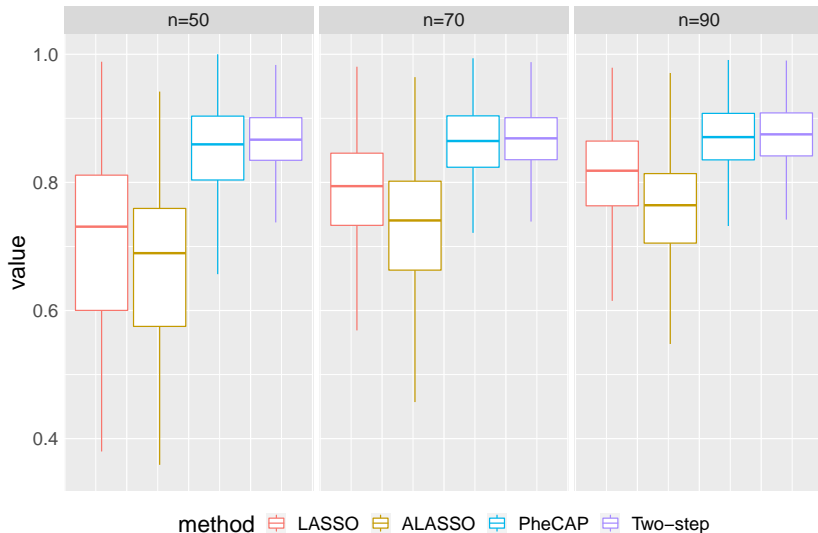
ROC

The operating receiver characteristic (ROC) curve



Model Evaluation

Area under the ROC curve (AUC) from 600 simulations



MAP

```
# Use un-transformed data; MAP requires sparse matrix.
# Create sparse matrix for surrogates.
data_fit <- sparsify(
  PheCAP::ehr_data %>%
    select(main_ICD, main_NLP) %>%
    rename(ICD = main_ICD) %>% data.table()
)

# Create sparse matrix for HU.
note <- Matrix(
  PheCAP::ehr_data$healthcare_utilization,
  ncol = 1, sparse = TRUE
)
model_map <- MAP(mat = data_fit, note = note, full.output = TRUE)
```

```
## #####
## MAP only considers patients who have note count data and
##      at least one nonmissing variable!
## ####
## Here is a summary of the input data:
## Total number of patients: 10000
##   ICD main_NLP note   Freq
## 1 YES      YES   YES 10000
## ####
```

```
y_hat_map <- model_map$scores[data$validation_set]
roc_map <- roc(test_y, y_hat_map)
```


ROC

The operating receiver characteristic (ROC) curve

