

## 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  $\beta$ .

```
# 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 = "ic"  
)
```

## 2-step Semi-supervised Approach

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

```
# 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$patient_id]
##                -1.9395295                0.6361248
##      S[train_data$patient_id]
##                0.6534730
```

```
# 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]
```

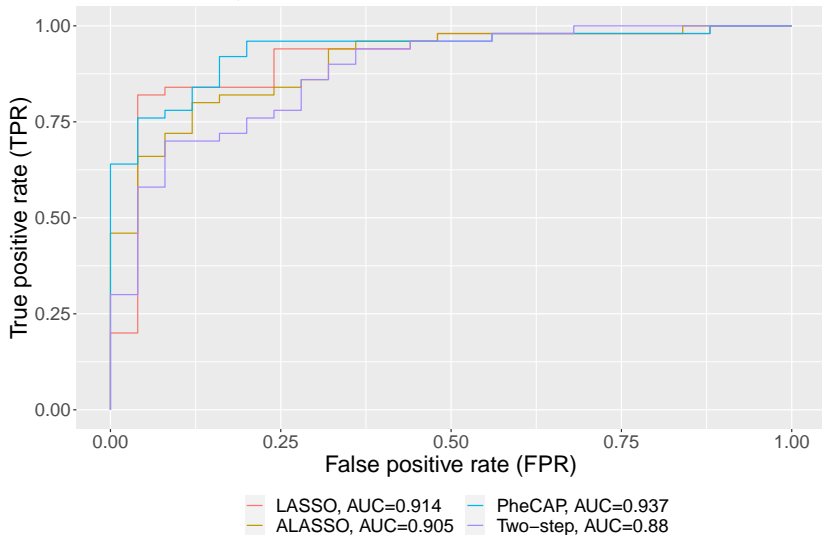
```
## [1] "COD10"     "NLP6"      "NLP14"     "NLP24"     "NLP31"     "NLP44"     "NLP56"     "NLP59"
## [9] "NLP61"     "NLP68"     "NLP73"     "NLP74"     "NLP93"     "NLP127"    "NLP130"    "NLP160"
## [17] "NLP161"    "NLP172"    "NLP176"    "NLP193"    "NLP199"    "NLP202"    "NLP215"    "NLP225"
## [25] "NLP231"    "NLP243"    "NLP294"    "NLP295"    "NLP302"    "NLP304"    "NLP306"    "NLP309"
## [33] "NLP321"    "NLP349"    "NLP350"    "NLP361"    "NLP403"    "NLP434"    "NLP446"    "NLP451"
## [41] "NLP456"    "NLP463"    "NLP465"    "NLP482"    "NLP495"    "NLP507"    "NLP536"    "NLP539"
## [49] "NLP544"    "NLP549"    "NLP558"    "NLP564"    "NLP568"    "NLP571"    "NLP584"    "NLP588"
## [57] "NLP591"    "NLP594"    "NLP597"    "NLP601"    "NLP604"    "NLP607"    "NLP610"    "NLP613"
## [65] "NLP616"    "NLP619"    "NLP622"    "NLP625"    "NLP628"    "NLP631"    "NLP634"    "NLP637"
## [73] "NLP640"    "NLP643"    "NLP646"    "NLP649"    "NLP652"    "NLP655"    "NLP658"    "NLP661"
## [81] "NLP664"    "NLP667"    "NLP670"    "NLP673"    "NLP676"    "NLP679"    "NLP682"    "NLP685"
## [89] "NLP688"    "NLP691"    "NLP694"    "NLP697"    "NLP700"    "NLP703"    "NLP706"    "NLP709"
## [97] "NLP712"    "NLP715"    "NLP718"    "NLP721"    "NLP724"    "NLP727"    "NLP730"    "NLP733"
## [105] "NLP736"    "NLP739"    "NLP742"    "NLP745"    "NLP748"    "NLP751"    "NLP754"    "NLP757"
## [113] "NLP760"    "NLP763"    "NLP766"    "NLP769"    "NLP772"    "NLP775"    "NLP778"    "NLP781"
## [121] "NLP784"    "NLP787"    "NLP790"    "NLP793"    "NLP796"    "NLP799"    "NLP802"    "NLP805"
## [129] "NLP808"    "NLP811"    "NLP814"    "NLP817"    "NLP820"    "NLP823"    "NLP826"    "NLP829"
## [137] "NLP832"    "NLP835"    "NLP838"    "NLP841"    "NLP844"    "NLP847"    "NLP850"    "NLP853"
## [145] "NLP856"    "NLP859"    "NLP862"    "NLP865"    "NLP868"    "NLP871"    "NLP874"    "NLP877"
## [153] "NLP880"    "NLP883"    "NLP886"    "NLP889"    "NLP892"    "NLP895"    "NLP898"    "NLP901"
## [161] "NLP904"    "NLP907"    "NLP910"    "NLP913"    "NLP916"    "NLP919"    "NLP922"    "NLP925"
## [169] "NLP928"    "NLP931"    "NLP934"    "NLP937"    "NLP940"    "NLP943"    "NLP946"    "NLP949"
## [177] "NLP952"    "NLP955"    "NLP958"    "NLP961"    "NLP964"    "NLP967"    "NLP970"    "NLP973"
## [185] "NLP976"    "NLP979"    "NLP982"    "NLP985"    "NLP988"    "NLP991"    "NLP994"    "NLP997"
## [193] "NLP1000"   "NLP1003"   "NLP1006"   "NLP1009"   "NLP1012"   "NLP1015"   "NLP1018"   "NLP1021"
## [201] "NLP1024"   "NLP1027"   "NLP1030"   "NLP1033"   "NLP1036"   "NLP1039"   "NLP1042"   "NLP1045"
## [209] "NLP1048"   "NLP1051"   "NLP1054"   "NLP1057"   "NLP1060"   "NLP1063"   "NLP1066"   "NLP1069"
## [217] "NLP1072"   "NLP1075"   "NLP1078"   "NLP1081"   "NLP1084"   "NLP1087"   "NLP1090"   "NLP1093"
## [225] "NLP1096"   "NLP1099"   "NLP1102"   "NLP1105"   "NLP1108"   "NLP1111"   "NLP1114"   "NLP1117"
## [233] "NLP1120"   "NLP1123"   "NLP1126"   "NLP1129"   "NLP1132"   "NLP1135"   "NLP1138"   "NLP1141"
## [241] "NLP1144"   "NLP1147"   "NLP1150"   "NLP1153"   "NLP1156"   "NLP1159"   "NLP1162"   "NLP1165"
## [249] "NLP1168"   "NLP1171"   "NLP1174"   "NLP1177"   "NLP1180"   "NLP1183"   "NLP1186"   "NLP1189"
## [257] "NLP1192"   "NLP1195"   "NLP1198"   "NLP1201"   "NLP1204"   "NLP1207"   "NLP1210"   "NLP1213"
## [265] "NLP1216"   "NLP1219"   "NLP1222"   "NLP1225"   "NLP1228"   "NLP1231"   "NLP1234"   "NLP1237"
## [273] "NLP1240"   "NLP1243"   "NLP1246"   "NLP1249"   "NLP1252"   "NLP1255"   "NLP1258"   "NLP1261"
## [281] "NLP1264"   "NLP1267"   "NLP1270"   "NLP1273"   "NLP1276"   "NLP1279"   "NLP1282"   "NLP1285"
## [289] "NLP1288"   "NLP1291"   "NLP1294"   "NLP1297"   "NLP1300"   "NLP1303"   "NLP1306"   "NLP1309"
## [297] "NLP1312"   "NLP1315"   "NLP1318"   "NLP1321"   "NLP1324"   "NLP1327"   "NLP1330"   "NLP1333"
## [305] "NLP1336"   "NLP1339"   "NLP1342"   "NLP1345"   "NLP1348"   "NLP1351"   "NLP1354"   "NLP1357"
## [313] "NLP1360"   "NLP1363"   "NLP1366"   "NLP1369"   "NLP1372"   "NLP1375"   "NLP1378"   "NLP1381"
## [321] "NLP1384"   "NLP1387"   "NLP1390"   "NLP1393"   "NLP1396"   "NLP1399"   "NLP1402"   "NLP1405"
## [329] "NLP1408"   "NLP1411"   "NLP1414"   "NLP1417"   "NLP1420"   "NLP1423"   "NLP1426"   "NLP1429"
## [337] "NLP1432"   "NLP1435"   "NLP1438"   "NLP1441"   "NLP1444"   "NLP1447"   "NLP1450"   "NLP1453"
## [345] "NLP1456"   "NLP1459"   "NLP1462"   "NLP1465"   "NLP1468"   "NLP1471"   "NLP1474"   "NLP1477"
## [353] "NLP1480"   "NLP1483"   "NLP1486"   "NLP1489"   "NLP1492"   "NLP1495"   "NLP1498"   "NLP1501"
## [361] "NLP1504"   "NLP1507"   "NLP1510"   "NLP1513"   "NLP1516"   "NLP1519"   "NLP1522"   "NLP1525"
## [369] "NLP1528"   "NLP1531"   "NLP1534"   "NLP1537"   "NLP1540"   "NLP1543"   "NLP1546"   "NLP1549"
## [377] "NLP1552"   "NLP1555"   "NLP1558"   "NLP1561"   "NLP1564"   "NLP1567"   "NLP1570"   "NLP1573"
## [385] "NLP1576"   "NLP1579"   "NLP1582"   "NLP1585"   "NLP1588"   "NLP1591"   "NLP1594"   "NLP1597"
## [393] "NLP1600"   "NLP1603"   "NLP1606"   "NLP1609"   "NLP1612"   "NLP1615"   "NLP1618"   "NLP1621"
## [401] "NLP1624"   "NLP1627"   "NLP1630"   "NLP1633"   "NLP1636"   "NLP1639"   "NLP1642"   "NLP1645"
## [409] "NLP1648"   "NLP1651"   "NLP1654"   "NLP1657"   "NLP1660"   "NLP1663"   "NLP1666"   "NLP1669"
## [417] "NLP1672"   "NLP1675"   "NLP1678"   "NLP1681"   "NLP1684"   "NLP1687"   "NLP1690"   "NLP1693"
## [425] "NLP1696"   "NLP1699"   "NLP1702"   "NLP1705"   "NLP1708"   "NLP1711"   "NLP1714"   "NLP1717"
## [433] "NLP1720"   "NLP1723"   "NLP1726"   "NLP1729"   "NLP1732"   "NLP1735"   "NLP1738"   "NLP1741"
## [441] "NLP1744"   "NLP1747"   "NLP1750"   "NLP1753"   "NLP1756"   "NLP1759"   "NLP1762"   "NLP1765"
## [449] "NLP1768"   "NLP1771"   "NLP1774"   "NLP1777"   "NLP1780"   "NLP1783"   "NLP1786"   "NLP1789"
## [457] "NLP1792"   "NLP1795"   "NLP1798"   "NLP1801"   "NLP1804"   "NLP1807"   "NLP1810"   "NLP1813"
## [465] "NLP1816"   "NLP1819"   "NLP1822"   "NLP1825"   "NLP1828"   "NLP1831"   "NLP1834"   "NLP1837"
## [473] "NLP1840"   "NLP1843"   "NLP1846"   "NLP1849"   "NLP1852"   "NLP1855"   "NLP1858"   "NLP1861"
## [481] "NLP1864"   "NLP1867"   "NLP1870"   "NLP1873"   "NLP1876"   "NLP1879"   "NLP1882"   "NLP1885"
## [489] "NLP1888"   "NLP1891"   "NLP1894"   "NLP1897"   "NLP1900"   "NLP1903"   "NLP1906"   "NLP1909"
## [497] "NLP1912"   "NLP1915"   "NLP1918"   "NLP1921"   "NLP1924"   "NLP1927"   "NLP1930"   "NLP1933"
## [505] "NLP1936"   "NLP1939"   "NLP1942"   "NLP1945"   "NLP1948"   "NLP1951"   "NLP1954"   "NLP1957"
## [513] "NLP1960"   "NLP1963"   "NLP1966"   "NLP1969"   "NLP1972"   "NLP1975"   "NLP1978"   "NLP1981"
## [521] "NLP1984"   "NLP1987"   "NLP1990"   "NLP1993"   "NLP1996"   "NLP1999"   "NLP2002"   "NLP2005"
## [529] "NLP2008"   "NLP2011"   "NLP2014"   "NLP2017"   "NLP2020"   "NLP2023"   "NLP2026"   "NLP2029"
## [537] "NLP2032"   "NLP2035"   "NLP2038"   "NLP2041"   "NLP2044"   "NLP2047"   "NLP2050"   "NLP2053"
## [545] "NLP2056"   "NLP2059"   "NLP2062"   "NLP2065"   "NLP2068"   "NLP2071"   "NLP2074"   "NLP2077"
## [553] "NLP2080"   "NLP2083"   "NLP2086"   "NLP2089"   "NLP2092"   "NLP2095"   "NLP2098"   "NLP2101"
## [561] "NLP2104"   "NLP2107"   "NLP2110"   "NLP2113"   "NLP2116"   "NLP2119"   "NLP2122"   "NLP2125"
## [569] "NLP2128"   "NLP2131"   "NLP2134"   "NLP2137"   "NLP2140"   "NLP2143"   "NLP2146"   "NLP2149"
## [577] "NLP2152"   "NLP2155"   "NLP2158"   "NLP2161"   "NLP2164"   "NLP2167"   "NLP2170"   "NLP2173"
## [585] "NLP2176"   "NLP2179"   "NLP2182"   "NLP2185"   "NLP2188"   "NLP2191"   "NLP2194"   "NLP2197"
## [593] "NLP2200"   "NLP2203"   "NLP2206"   "NLP2209"   "NLP2212"   "NLP2215"   "NLP2218"   "NLP2221"
## [601] "NLP2224"   "NLP2227"   "NLP2230"   "NLP2233"   "NLP2236"   "NLP2239"   "NLP2242"   "NLP2245"
## [609] "NLP2248"   "NLP2251"   "NLP2254"   "NLP2257"   "NLP2260"   "NLP2263"   "NLP2266"   "NLP2269"
## [617] "NLP2272"   "NLP2275"   "NLP2278"   "NLP2281"   "NLP2284"   "NLP2287"   "NLP2290"   "NLP2293"
## [625] "NLP2296"   "NLP2299"   "NLP2302"   "NLP2305"   "NLP2308"   "NLP2311"   "NLP2314"   "NLP2317"
## [633] "NLP2320"   "NLP2323"   "NLP2326"   "NLP2329"   "NLP2332"   "NLP2335"   "NLP2338"   "NLP2341"
## [641] "NLP2344"   "NLP2347"   "NLP2350"   "NLP2353"   "NLP2356"   "NLP2359"   "NLP2362"   "NLP2365"
## [649] "NLP2368"   "NLP2371"   "NLP2374"   "NLP2377"   "NLP2380"   "NLP2383"   "NLP2386"   "NLP2389"
## [657] "NLP2392"   "NLP2395"   "NLP2398"   "NLP2401"   "NLP2404"   "NLP2407"   "NLP2410"   "NLP2413"
## [665] "NLP2416"   "NLP2419"   "NLP2422"   "NLP2425"   "NLP2428"   "NLP2431"   "NLP2434"   "NLP2437"
## [673] "NLP2440"   "NLP2443"   "NLP2446"   "NLP2449"   "NLP2452"   "NLP2455"   "NLP2458"   "NLP2461"
## [681] "NLP2464"   "NLP2467"   "NLP2470"   "NLP2473"   "NLP2476"   "NLP2479"   "NLP2482"   "NLP2485"
## [689] "NLP2488"   "NLP2491"   "NLP2494"   "NLP2497"   "NLP2500"   "NLP2503"   "NLP2506"   "NLP2509"
## [697] "NLP2512"   "NLP2515"   "NLP2518"   "NLP2521"   "NLP2524"   "NLP2527"   "NLP2530"   "NLP2533"
## [705] "NLP2536"   "NLP2539"   "NLP2542"   "NLP2545"   "NLP2548"   "NLP2551"   "NLP2554"   "NLP2557"
## [713] "NLP2560"   "NLP2563"   "NLP2566"   "NLP2569"   "NLP2572"   "NLP2575"   "NLP2578"   "NLP2581"
## [721] "NLP2584"   "NLP2587"   "NLP2590"   "NLP2593"   "NLP2596"   "NLP2599"   "NLP2602"   "NLP2605"
## [729] "NLP2608"   "NLP2611"   "NLP2614"   "NLP2617"   "NLP2620"   "NLP2623"   "NLP2626"   "NLP2629"
## [737] "NLP2632"   "NLP2635"   "NLP2638"   "NLP2641"   "NLP2644"   "NLP2647"   "NLP2650"   "NLP2653"
## [745] "NLP2656"   "NLP2659"   "NLP2662"   "NLP2665"   "NLP2668"   "NLP2671"   "NLP2674"   "NLP2677"
## [753] "NLP2680"   "NLP2683"   "NLP2686"   "NLP2689"   "NLP2692"   "NLP2695"   "NLP2698"   "NLP2701"
## [761] "NLP2704"   "NLP2707"   "NLP2710"   "NLP2713"   "NLP2716"   "NLP2719"   "NLP2722"   "NLP2725"
## [769] "NLP2728"   "NLP2731"   "NLP2734"   "NLP2737"   "NLP2740"   "NLP2743"   "NLP2746"   "NLP2749"
## [777] "NLP2752"   "NLP2755"   "NLP2758"   "NLP2761"   "NLP2764"   "NLP2767"   "NLP2770"   "NLP2773"
## [785] "NLP2776"   "NLP2779"   "NLP2782"   "NLP2785"   "NLP2788"   "NLP2791"   "NLP2794"   "NLP2797"
## [793] "NLP2800"   "NLP2803"   "NLP2806"   "NLP2809"   "NLP2812"   "NLP2815"   "NLP2818"   "NLP2821"
## [801] "NLP2824"   "NLP2827"   "NLP2830"   "NLP2833"   "NLP2836"   "NLP2839"   "NLP2842"   "NLP2845"
## [809] "NLP2848"   "NLP2851"   "NLP2854"   "NLP2857"   "NLP2860"   "NLP2863"   "NLP2866"   "NLP2869"
## [817] "NLP2872"   "NLP2875"   "NLP2878"   "NLP2881"   "NLP2884"   "NLP2887"   "NLP2890"   "NLP2893"
## [825] "NLP2896"   "NLP2899"   "NLP2902"   "NLP2905"   "NLP2908"   "NLP2911"   "NLP2914"   "NLP2917"
## [833] "NLP2920"   "NLP2923"   "NLP2926"   "NLP2929"   "NLP2932"   "NLP2935"   "NLP2938"   "NLP2941"
## [841] "NLP2944"   "NLP2947"   "NLP2950"   "NLP2953"   "NLP2956"   "NLP2959"   "NLP2962"   "NLP2965"
## [849] "NLP2968"   "NLP2971"   "NLP2974"   "NLP2977"   "NLP2980"   "NLP2983"   "NLP2986"   "NLP2989"
## [857] "NLP2992"   "NLP2995"   "NLP2998"   "NLP3001"   "NLP3004"   "NLP3007"   "NLP3010"   "NLP3013"
## [865] "NLP3016"   "NLP3019"   "NLP3022"   "NLP3025"   "NLP3028"   "NLP3031"   "NLP3034"   "NLP3037"
## [873] "NLP3040"   "NLP3043"   "NLP3046"   "NLP3049"   "NLP3052"   "NLP3055"   "NLP3058"   "NLP3061"
## [881] "NLP3064"   "NLP3067"   "NLP3070"   "NLP3073"   "NLP3076"   "NLP3079"   "NLP3082"   "NLP3085"
## [889] "NLP3088"   "NLP3091"   "NLP3094"   "NLP3097"   "NLP3100"   "NLP3103"   "NLP3106"   "NLP3109"
## [897] "NLP3112"   "NLP3115"   "NLP3118"   "NLP3121"   "NLP3124"   "NLP3127"   "NLP3130"   "NLP3133"
## [905] "NLP3136"   "NLP3139"   "NLP3142"   "NLP3145"   "NLP3148"   "NLP3151"   "NLP3154"   "NLP3157"
## [913] "NLP3160"   "NLP3163"   "NLP3166"   "NLP3169"   "NLP3172"   "NLP3175"   "NLP3178"   "NLP3181"
## [921] "NLP3184"   "NLP3187"   "NLP3190"   "NLP3193"   "NLP3196"   "NLP3199"   "NLP3202"   "NLP3205"
## [929] "NLP3208"   "NLP3211"   "NLP3214"   "NLP3217"   "NLP3220"   "NLP3223"   "NLP3226"   "NLP3229"
## [937] "NLP3232"   "NLP3235"   "NLP3238"   "NLP3241"   "NLP3244"   "NLP3247"   "NLP3250"   "NLP3253"
## [945] "NLP3256"   "NLP3259"   "NLP3262"   "NLP3265"   "NLP3268"   "NLP3271"   "NLP3274"   "NLP3277"
## [953] "NLP3280"   "NLP3283"   "NLP3286"   "NLP3289"   "NLP3292"   "NLP3295"   "NLP3298"   "NLP3301"
## [961] "NLP3304"   "NLP3307"   "NLP3310"   "NLP3313"   "NLP3316"   "NLP3319"   "NLP3322"   "NLP3325"
## [969] "NLP3328"   "NLP3331"   "NLP3334"   "NLP3337"   "NLP3340"   "NLP3343"   "NLP3346"   "NLP3349"
## [977] "NLP3352"   "NLP3355"   "NLP3358"   "NLP3361"   "NLP3364"   "NLP3367"   "NLP3370"   "NLP3373"
## [985] "NLP3376"   "NLP3379"   "NLP3382"   "NLP3385"   "NLP3388"   "NLP3391"   "NLP3394"   "NLP3397"
## [993] "NLP3400"   "NLP3403"   "NLP3406"   "NLP3409"   "NLP3412"   "NLP3415"   "NLP3418"   "NLP3421"
## [997] "NLP3424"   "NLP3427"   "NLP3430"   "NLP3433"   "NLP3436"   "NLP3439"   "NLP3442"   "NLP3445"
## [1000] "NLP3448"   "NLP3451"   "NLP3454"   "NLP3457"   "NLP3460"   "NLP3463"   "NLP3466"   "NLP3469"
## [1000] "NLP3472"   "NLP3475"   "NLP3478"   "NLP3481"   "NLP3484"   "NLP3487"   "NLP3490"   "NLP3493"
## [1000] "NLP3496"   "NLP3499"   "NLP3502"   "NLP3505"   "NLP3508"   "NLP3511"   "NLP3514"   "NLP3517"
## [1000] "NLP3520"   "NLP3523"   "NLP3526"   "NLP3529"   "NLP3532"   "NLP3535"   "NLP3538"   "NLP3541"
## [1000] "NLP3544"   "NLP3547"   "NLP3550"   "NLP3553"   "NLP3556"   "NLP3559"   "NLP3562"   "NLP3565"
## [1000] "NLP3568"   "NLP3571"   "NLP3574"   "NLP3577"   "NLP3580"   "NLP3583"   "NLP3586"   "NLP3589"
## [1000] "NLP3592"   "NLP3595"   "NLP3598"   "NLP3601"   "NLP3604"   "NLP3607"   "NLP3610"   "NLP3613"
## [1000] "NLP3616"   "NLP3619"   "NLP3622"   "NLP3625"   "NLP3628"   "NLP3631"   "NLP3634"   "NLP3637"
## [1000] "NLP3640"   "NLP3643"   "NLP3646"   "NLP3649"   "NLP3652"   "NLP3655"   "NLP3658"   "NLP3661"
## [1000] "NLP3664"   "NLP3667"   "NLP3670"   "NLP3673"   "NLP3676"   "NLP3679"   "NLP3682"   "NLP3685"
## [1000] "NLP3688"   "NLP3691"   "NLP3694"   "NLP3697"   "NLP3700"   "NLP3703"   "NLP3706"   "NLP3709"
## [1000] "NLP3712"   "NLP3715"   "NLP3718"   "NLP3721"   "NLP3724"   "NLP3727"   "NLP3730"   "NLP3733"
## [1000] "NLP3736"   "NLP3739"   "NLP3742"   "NLP3745"   "NLP3748"   "NLP3751"   "NLP3754"   "NLP3757"
## [1000] "NLP3760"   "NLP3763"   "NLP3766"   "NLP3769"   "NLP3772"   "NLP3775"   "NLP3778"   "NLP3781"
## [1000] "NLP3784"   "NLP3787"   "NLP3790"   "NLP3793"   "NLP3796"   "NLP3799"   "NLP3802"   "NLP3805"
## [1000] "NLP3808"   "NLP3811"   "NLP3814"   "NLP3817"   "NLP3820"   "NLP3823"   "NLP3826"   "NLP3829"
## [1000] "NLP3832"   "NLP3835"   "NLP3838"   "NLP3841"   "NLP3844"   "NLP3847"   "NLP3850"   "NLP3853"
## [1000] "NLP3856"   "NLP3859"   "NLP3862"   "NLP3865"   "NLP3868"   "NLP3871"   "NLP3874"   "NLP3877"
## [1000] "NLP3880"   "NLP3883"   "NLP3886"   "NLP3889"   "NLP3892"   "NLP3895"   "NLP3898"   "NLP3901"
## [1000] "NLP3904"   "NLP3907"   "NLP3910"   "NLP3913"   "NLP3916"   "NLP3919"   "NLP3922"   "NLP3925"
## [1000] "NLP3928"   "NLP3931"   "NLP3934"   "NLP3937"   "NLP3940"   "NLP3943"   "NLP3946"   "NLP3949"
## [1000] "NLP3952"   "NLP3955"   "NLP3958"   "NLP3961"   "NLP3964"   "NLP3967"   "NLP3970"   "NLP3973"
## [1000] "NLP3976"   "NLP3979"   "NLP3982"   "NLP3985"   "NLP398
```

# 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)
```

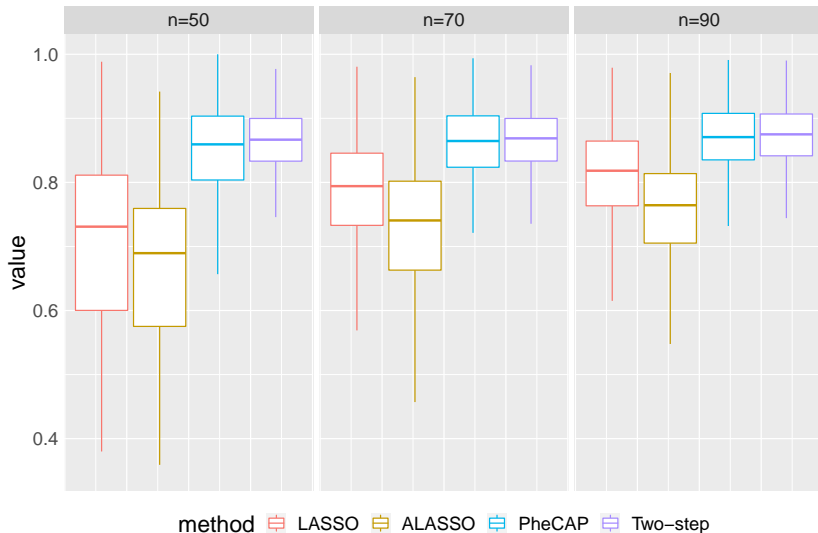
# ROC

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

