Addition of Emphysema-CNN score improves upon prescreening risk calculated by LCRAT

Load packages

```
# List packages to be loaded (and installed if needed)
packages <-
    c(
        "readr",
        "here",
        "dplyr",
        "psych",
        "pROC"
    )

# List packages that are not installed
not_installed <-
    packages[!(packages %in% installed.packages()[, "Package"])]

# Install packages that are not installed
if (length(not_installed))
    install.packages(not_installed)

# Load all packages
lapply(packages, require, character.only = TRUE)
```

Merge in Wes' T0 data

```
nlst_emp <- read_csv(here('data/T0_data.csv'))
data_screen_abn_neg <- readRDS(here('data/data_screen_abn_neg.rds'))
data_screen_abn_neg_emp <-
    merge(data_screen_abn_neg, nlst_emp, by = "pid")
```

Fit models

Without physician-annotated features

```
# Prescreening risk only
glm_screen_neg <-
    glm(case ~ log1yrisk - 1,
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'))
summary(glm_screen_neg)
```

```
##
```

```
## Call:
## glm(formula = case ~ log1yrisk - 1, family = binomial(link = "log"),
##     data = data_screen_abn_neg_emp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7549  -0.4153  -0.3584  -0.3115   2.5678
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## log1yrisk  0.49427    0.01639   30.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:    Inf  on 1750  degrees of freedom
## Residual deviance: 879.18  on 1749  degrees of freedom
## AIC: 881.18
##
## Number of Fisher Scoring iterations: 6
```

```
predictions <- predict(glm_screen_neg, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6287
```

```
# Prescreening risk + p_emph
glm_screen_neg_pemph <-
    glm(case ~ log1yrisk + p_emph - 1,
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'),
        na.action = na.exclude)
summary(glm_screen_neg_pemph)
```

```
##
## Call:
## glm(formula = case ~ log1yrisk + p_emph - 1, family = binomial(link = "log"),
##     data = data_screen_abn_neg_emp, na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8022  -0.4161  -0.3543  -0.3010   2.6001
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## log1yrisk  0.51614    0.02082  24.790   <2e-16 ***
## p_emph     0.62740    0.32944   1.904   0.0569 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance:    Inf  on 1750  degrees of freedom
## Residual deviance: 875.93  on 1748  degrees of freedom
## AIC: 879.93
##
## Number of Fisher Scoring iterations: 6
```

```
predictions <- predict(glm_screen_neg_pemph, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6406
```

```
# Prescreening risk + logit p_emph
glm_screen_neg_logitpemph <-
    glm(case ~ log1yrisk + I(logit(p_emph)) - 1,
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'),
        na.action = na.exclude)
summary(glm_screen_neg_logitpemph)
```

```
##
## Call:
## glm(formula = case ~ log1yrisk + I(logit(p_emph)) - 1, family = binomial(link = "log"),
##     data = data_screen_abn_neg_emp, na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7680  -0.4150  -0.3548  -0.3048   2.5849
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## log1yrisk          0.45573    0.02319  19.652   <2e-16 ***
## I(logit(p_emph))   0.10048    0.04679   2.148   0.0317 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:    Inf  on 1750  degrees of freedom
## Residual deviance: 874.93  on 1748  degrees of freedom
## AIC: 878.93
##
## Number of Fisher Scoring iterations: 6
```

```
predictions <- predict(glm_screen_neg_logitpemph, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6463
```

```
# Prescreening risk + log p_emph
glm_screen_neg_logpemph <-
    glm(case ~ log1yrisk + I(log(p_emph)) - 1,
```

```
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'),
        na.action = na.exclude)
summary(glm_screen_neg_logpemph)
```

```
##
## Call:
## glm(formula = case ~ log1yrisk + I(log(p_emph)) - 1, family = binomial(link = "log"),
##     data = data_screen_abn_neg_emp, na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7482  -0.4162  -0.3571  -0.2987   2.6030
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## log1yrisk       0.41882    0.03509  11.936   <2e-16 ***
## I(log(p_emph))  0.17159    0.07477   2.295   0.0217 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:    Inf  on 1750  degrees of freedom
## Residual deviance: 873.91  on 1748  degrees of freedom
## AIC: 877.91
##
## Number of Fisher Scoring iterations: 6
```

```
predictions <- predict(glm_screen_neg_logpemph, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6477
```

With physician-annotated features

```
glm_screen_neg_cons <-
    glm(
        case ~ log1yrisk + log1yrisk:consolidation - 1,
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'),
        na.action = na.exclude
    )
summary(glm_screen_neg_cons)
```

```
##
## Call:
## glm(formula = case ~ log1yrisk + log1yrisk:consolidation - 1,
##     family = binomial(link = "log"), data = data_screen_abn_neg_emp,
##     na.action = na.exclude)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7233  -0.4049  -0.3462  -0.3003   2.6045
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## log1yrisk                 0.50851    0.01842  27.602   <2e-16 ***
## log1yrisk:consolidation -0.10453    0.09195  -1.137    0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:   Inf  on 1536  degrees of freedom
## Residual deviance: 740.4  on 1534  degrees of freedom
##   (214 observations deleted due to missingness)
## AIC: 744.4
##
## Number of Fisher Scoring iterations: 6
```

```
predictions <- predict(glm_screen_neg_cons, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6411
```

```
glm_screen_neg_emph <-
    glm(
        case ~ log1yrisk + log1yrisk:emphysema - 1,
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'),
        na.action = na.exclude
    )
summary(glm_screen_neg_emph)
```

```
##
## Call:
## glm(formula = case ~ log1yrisk + log1yrisk:emphysema - 1, family = binomial(link = "log"),
##     data = data_screen_abn_neg_emp, na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7806  -0.4157  -0.3479  -0.2724   2.6343
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## log1yrisk              0.55480    0.02842  19.521   <2e-16 ***
## log1yrisk:emphysema -0.09376    0.03682  -2.547   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance:    Inf  on 1536  degrees of freedom
## Residual deviance: 734.85  on 1534  degrees of freedom
##   (214 observations deleted due to missingness)
## AIC: 738.85
##
## Number of Fisher Scoring iterations: 6
```

```
predictions <- predict(glm_screen_neg_emph, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6555
```

```
glm_screen_neg_cons_emph <-
    glm(
        case ~ log1yrisk + log1yrisk:consolidation + log1yrisk:emphysema - 1,
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'),
        na.action = na.exclude
    )
summary(glm_screen_neg_cons_emph)
```

```
##
## Call:
## glm(formula = case ~ log1yrisk + log1yrisk:consolidation + log1yrisk:emphysema -
##     1, family = binomial(link = "log"), data = data_screen_abn_neg_emp,
##     na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8406  -0.4164  -0.3482  -0.2679   2.6487
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## log1yrisk                 0.56087    0.02897  19.360  < 2e-16 ***
## log1yrisk:consolidation -0.12984    0.08974  -1.447  0.14792
## log1yrisk:emphysema     -0.09777    0.03691  -2.649  0.00808 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:    Inf  on 1536  degrees of freedom
## Residual deviance: 733.27  on 1533  degrees of freedom
##   (214 observations deleted due to missingness)
## AIC: 739.27
##
## Number of Fisher Scoring iterations: 7
```

```
predictions <- predict(glm_screen_neg_cons_emph, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6568
```

```r
glm_screen_neg_cons_emph_pemph <-
    glm(
        case ~ log1yrisk + log1yrisk:consolidation + log1yrisk:emphysema + I(log(p_emph)) - 1,
        data = data_screen_abn_neg_emp,
        family = binomial(link = 'log'),
        na.action = na.exclude
    )
summary(glm_screen_neg_cons_emph_pemph)
```

```
##
## Call:
## glm(formula = case ~ log1yrisk + log1yrisk:consolidation + log1yrisk:emphysema +
##     I(log(p_emph)) - 1, family = binomial(link = "log"), data = data_screen_abn_neg_emp,
##     na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7769  -0.4153  -0.3389  -0.2597   2.6374
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## log1yrisk                 0.45693    0.04953   9.225   <2e-16 ***
## I(log(p_emph))            0.21134    0.08587   2.461   0.0139 *
## log1yrisk:consolidation -0.14210    0.08549  -1.662   0.0965 .
## log1yrisk:emphysema     -0.07085    0.03846  -1.842   0.0655 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:    Inf  on 1536  degrees of freedom
## Residual deviance: 727.13  on 1532  degrees of freedom
##   (214 observations deleted due to missingness)
## AIC: 735.13
##
## Number of Fisher Scoring iterations: 7
```

```r
predictions <- predict(glm_screen_neg_cons_emph_pemph, type = "response")
pROC::auc(data_screen_abn_neg_emp$case, predictions)
```

```
## Area under the curve: 0.6706
```