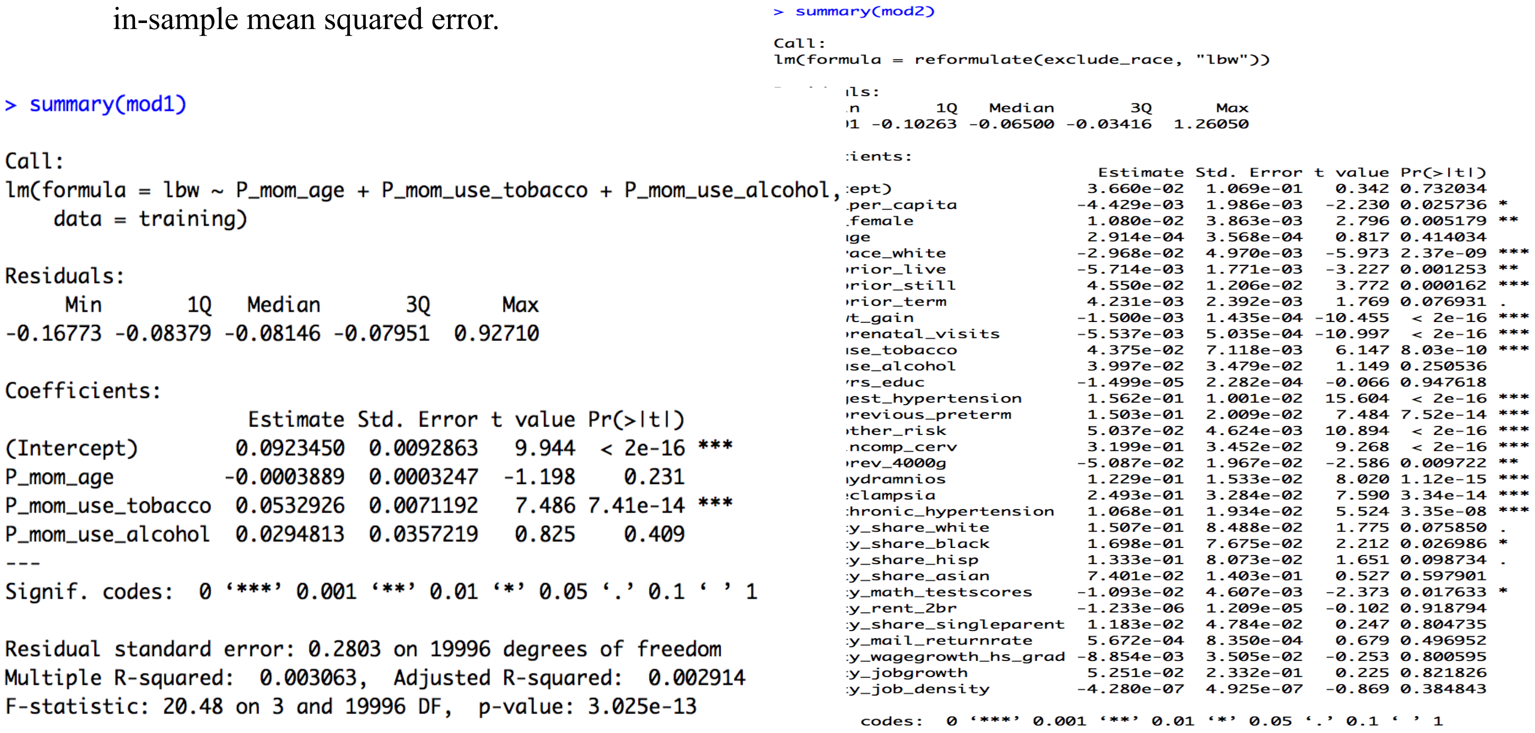
1. The first two models I explored in the coding exercise were multivariable linear regression models. Model 1 incorporated a few handpicked predictors that I initially thought would be important predictors of low birthweight. In particular, I felt that the mother's age, as well as substance use during pregnancy, would have strong relationships with low birthweight, so I chose to incorporate age, tobacco use, and alcohol use as predictors. In contrast, model 2 incorporates all variables of the data besides the mother's race. As models 1 and 2 are constructed by least-squares linear regression, they choose optimal coefficients that minimize in-sample mean squared error.

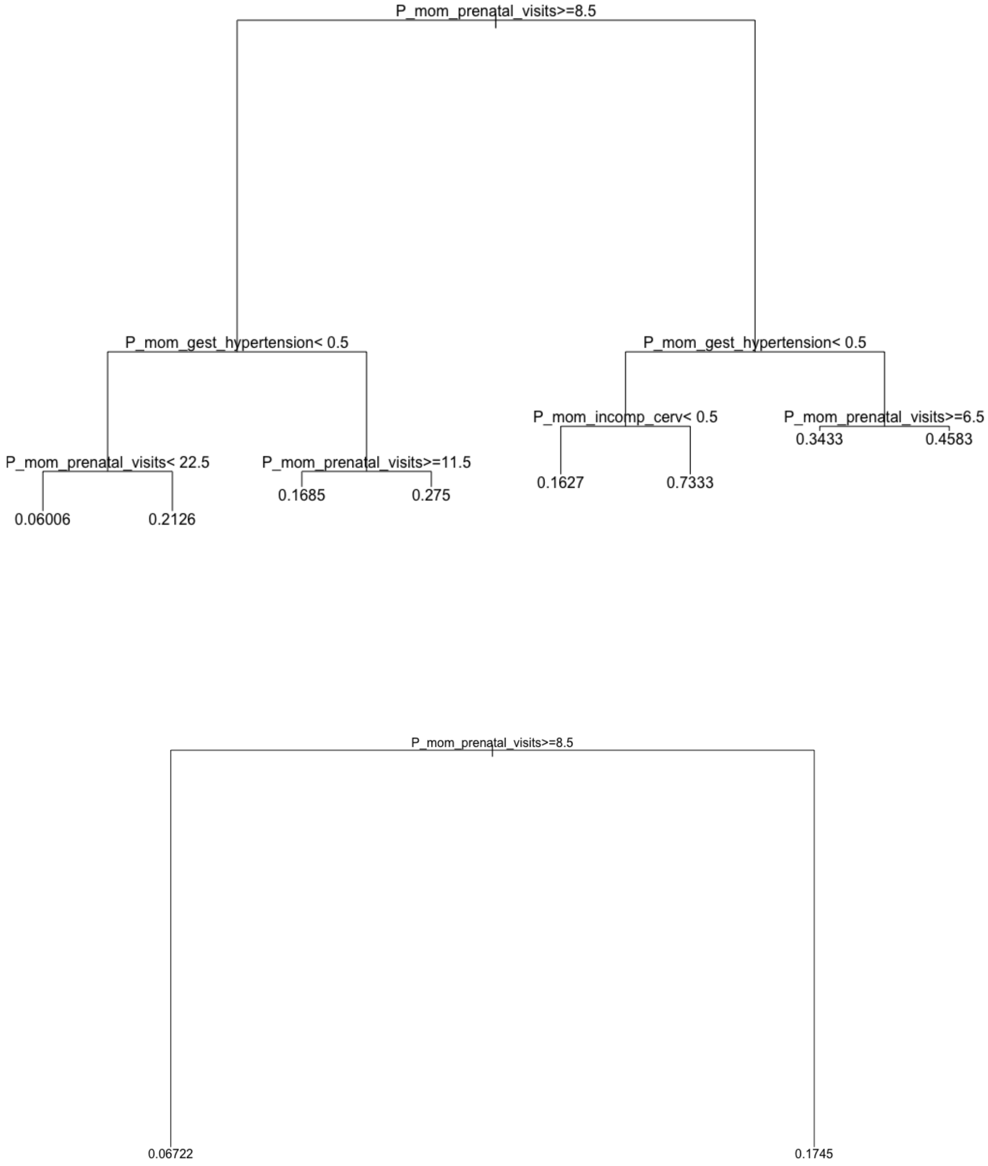


```
> summary(mod1)

Call:
lm(formula = lbw ~ P_mom_age + P_mom_use_tobacco + P_mom_use_alcohol,
    data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16773 -0.08379 -0.08146 -0.07951  0.92710 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.0923450  0.0092863   9.944  < 2e-16 ***
P_mom_age         -0.0003889  0.0003247  -1.198    0.231    
P_mom_use_tobacco  0.0532926  0.0071192   7.486 7.41e-14 ***
P_mom_use_alcohol  0.0294813  0.0357219   0.825    0.409    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2803 on 19996 degrees of freedom
Multiple R-squared:  0.003063,	Adjusted R-squared:  0.002914 
F-statistic: 20.48 on 3 and 19996 DF,  p-value: 3.025e-13
```

```
> summary(mod2)

Call:
lm(formula = reformulate(exclude_race, "lbw"))

ls:
n       1Q   Median       3Q      Max 
1 -0.10263 -0.06500 -0.03416  1.26050 

ients:
                            Estimate Std. Error t value Pr(>|t|)    
ept)                       3.660e-02  1.069e-01   0.342 0.732034    
per_capita                -4.429e-03  1.986e-03  -2.230 0.025736 *  
female                     1.080e-02  3.863e-03   2.796 0.005179 ** 
ge                         2.914e-04  3.568e-04   0.817 0.414034    
ace_white                 -2.968e-02  4.970e-03  -5.973 2.37e-09 ***
rior_live                 -5.714e-03  1.771e-03  -3.227 0.001253 ** 
rior_still                 4.550e-02  1.206e-02   3.772 0.000162 ***
rior_term                  4.231e-03  2.392e-03   1.769 0.076931 .  
t_gain                    -1.500e-03  1.435e-04 -10.455  < 2e-16 ***
renatal_visits            -5.537e-03  5.035e-04 -10.997  < 2e-16 ***
se_tobacco                 4.375e-02  7.118e-03   6.147 8.03e-10 ***
se_alcohol                 3.997e-02  3.479e-02   1.149 0.250536    
rs_educ                   -1.499e-05  2.282e-04  -0.066 0.947618    
est_hypertension           1.562e-01  1.001e-02  15.604  < 2e-16 ***
revious_preterm            1.503e-01  2.009e-02   7.484 7.52e-14 ***
ther_risk                  5.037e-02  4.624e-03  10.894  < 2e-16 ***
ncomp_cerv                 3.199e-01  3.452e-02   9.268  < 2e-16 ***
rev_4000g                 -5.087e-02  1.967e-02  -2.586 0.009722 ** 
ydramnios                  1.229e-01  1.533e-02   8.020 1.12e-15 ***
clampsia                   2.493e-01  3.284e-02   7.590 3.34e-14 ***
hronic_hypertension        1.068e-01  1.934e-02   5.524 3.35e-08 ***
y_share_white              1.507e-01  8.488e-02   1.775 0.075850 .  
y_share_black              1.698e-01  7.675e-02   2.212 0.026986 *  
y_share_hisp               1.333e-01  8.073e-02   1.651 0.098734 .  
y_share_asian              7.401e-02  1.403e-01   0.527 0.597901    
y_math_testscores         -1.093e-02  4.607e-03  -2.373 0.017633 *  
y_rent_2br                -1.233e-06  1.209e-05  -0.102 0.918794    
y_share_singleparent       1.183e-02  4.784e-02   0.247 0.804735    
y_mail_returnrate          5.672e-04  8.350e-04   0.679 0.496952    
y_wagegrowth_hs_grad      -8.854e-03  3.505e-02  -0.253 0.800595    
y_jobgrowth                5.251e-02  2.332e-01   0.225 0.821826    
y_job_density             -4.280e-07  4.925e-07  -0.869 0.384843    

 codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figures 1 & 2: Summaries of linear regression models.

The next two models I explored were both decision trees. From the significance levels of all predictors (Figure 2), I chose hypertension, frequency of prenatal visits, and cervix incompetency as significant predictors to incorporate in model 3. I felt these variables could plausibly have causal relationships with low birthweights, so I chose to explore these in a decision tree. In contrast, model 4 is a decision tree that incorporated all predictors other than race. As opposed to linear regression, decision trees work by considering a series of splits. At each tree split, the model selects the best predictor, or the predictor with the greatest predictive power of the outcome variable, and splits the data based on its values of the predictor. Then, at each successive split, this process is repeated until maximum tree depth is reached.

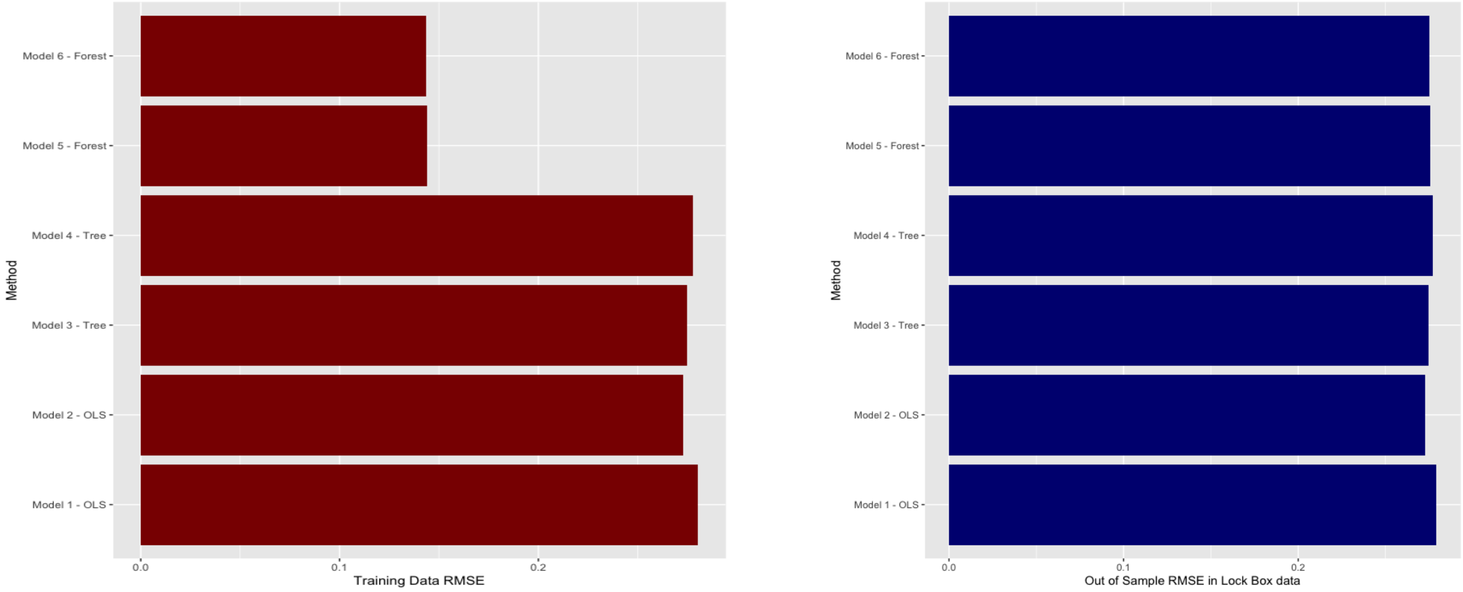

P_mom_prenatal_visits>=8.5
P_mom_gest_hypertension< 0.5
P_mom_gest_hypertension< 0.5
P_mom_incomp_cerv< 0.5
P_mom_prenatal_visits>=6.5
0.3433
0.4583
P_mom_prenatal_visits< 22.5
P_mom_prenatal_visits>=11.5
0.1685
0.275
0.1627
0.7333
0.06006
0.2126
P_mom_prenatal_visits>=8.5
0.06722
0.1745

Figures 3 & 4: Plots of decision tree models 3 (top) and 4 (bottom).

For example, the decision tree displayed in figure 4 splits the data based on whether or not the number of prenatal visits is at least 8.5. The model then predicts the probability of low birth weight to be about 0.067 for observations with at least 8.5 prenatal visits and about 0.175 for observations with less than 8.5 prenatal visits.

My last two models were random forests, which are essentially large groups of decision trees that each consider a random set of predictors. Model 5 excludes race as a potential predictor, while model 6 incorporates every potential predictor of the dataset. I obtained the optimal tuning parameters by minimizing lockbox RMSE through trial-and-error.

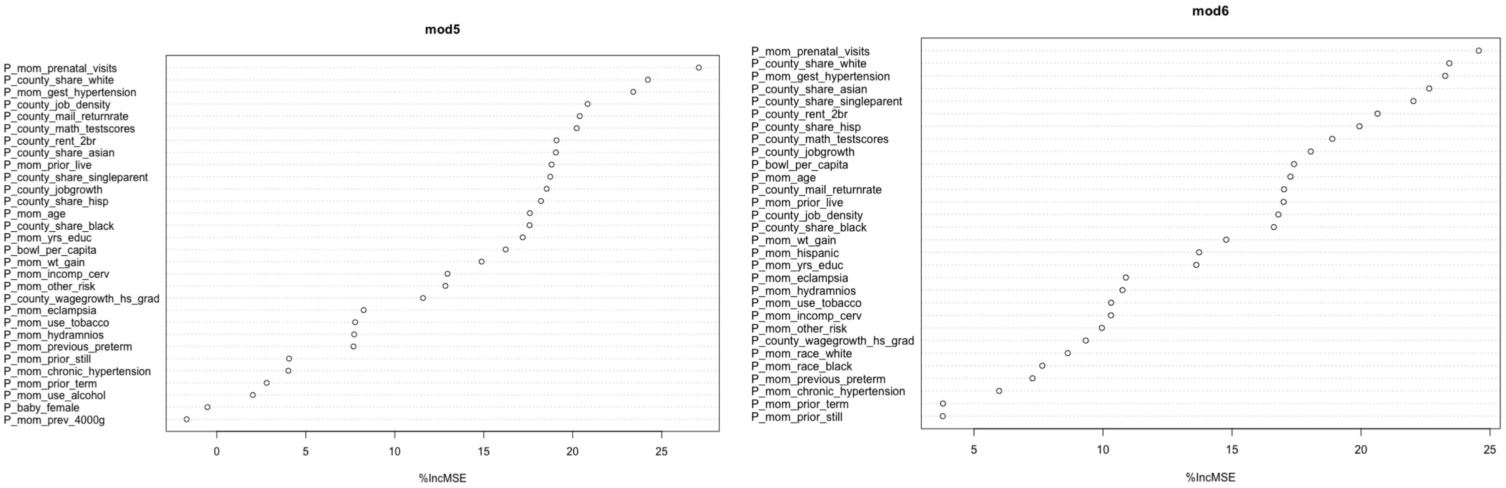
2. The criterion I used to select my ideal model was out-of-sample RMSE. I chose to look at out-of-sample, rather than in-sample, RMSE because I would want to know how a model would perform on data it hasn't seen before. Model 2, a linear regression including all non-race predictors, resulted in the lowest test/lockbox sample RMSE.



Figures 5 & 6: Plots of training RMSE (left) and testing RMSE (right) across all models.

3. Across both random forest models, the three most important predictors were the number of prenatal visits, the proportion of white residents in the county, and whether or not the mother had gestational hypertension.

It is important to recognize that a predictor with "importance" does not necessarily have a larger causal relationship with birth weight. Instead, a feature's importance is measured by the increase in mean squared error when that feature is not considered when fitting the random forest model. The variables are important in the sense that they reduce the model's prediction error when included as potential predictors.



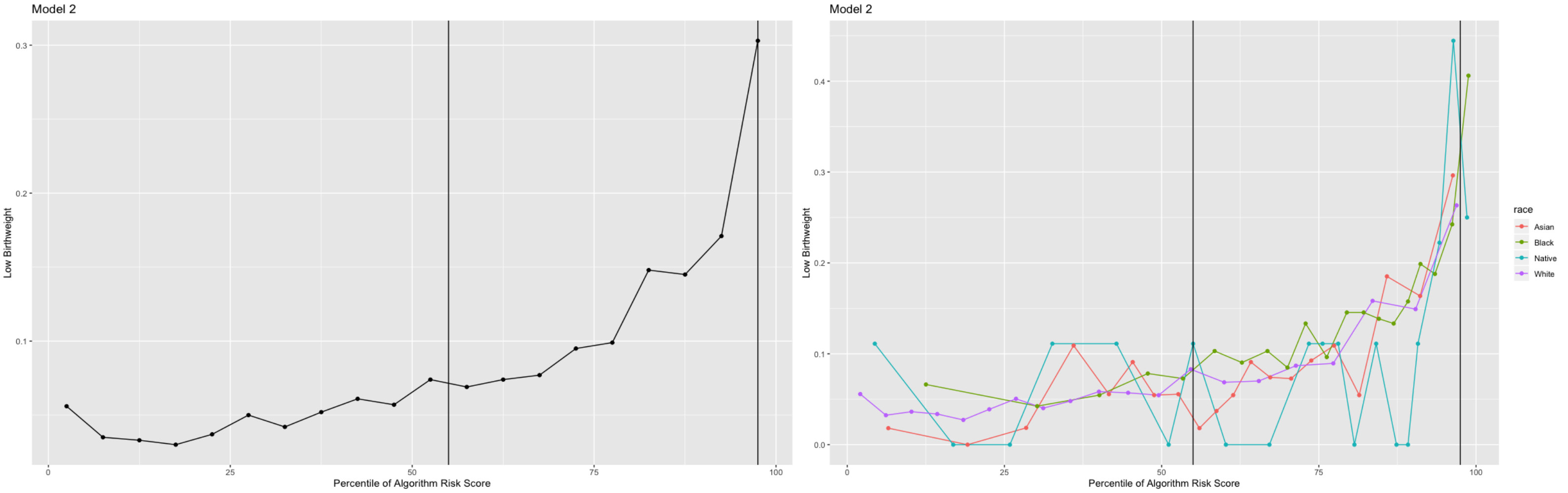
Figures 7 & 8: Predictor importance plots for random forest models 5 (left) and 6 (right).

4. Model 6, which incorporates race as a potential predictor, does not provide a significant increase in performance from model 5, which excludes race, based on test RMSE. Furthermore, model 6's most important predictors do not include race. Including race does not necessarily offer a predictive advantage, but it may actually add to existing algorithmic bias. As race is heavily correlated with many other data features, it is easily "reconstructable" through other variables, which leads to racial bias in algorithms even when it is not explicitly considered as a predictor (Obermeyer et al., 2019). I believe that including race/ethnicity as a model predictor would serve only to affirm this bias.

5. Deciding how to allocate resources is fundamentally a causal inference policy problem. However, health systems operate under the assumption that "those with the greatest care needs will benefit the most from the program", and this key assumption makes the allocation problem a prediction policy problem (Ibid).

6. A model's "label" is what the model predicts based on observed data. Following the greatest care needs assumption for resource allocation, an ideal label for this scenario would be need for medical care. However, this is not a quantifiable variable, so need is an infeasible label. Instead of directly using need as a label, health plans could feasibly use low birthweight as a proxy for need. Because birth weight is observable and correlates strongly with future development and life expectancy, it serves as a useful label. In addition, health plans must consider whether or not their selected labels add bias to the algorithm, and I would not expect birthweight to be biased by race.

7. At the same predicted risk score, a black mother is more likely than a white mother is to have a baby with low birth weight on average. This result confirms the trends shown in graphs 1a and 3a: despite algorithms predicting the same risk across races, black patients are much more likely to have a higher number of chronic illnesses (1a) as well as higher medical costs (3a) than their white counterparts (Ibid). In other words, current algorithms tend to be biased by race.



Figures 9 & 10: Binned plots of predicted risk against low birth weight prevalence for all groups (left) and stratified by race (right).

8. The model's eligible population is composed of about 60% white patients, 33% black patients, 6% Asian patients, and 2% Native American patients. Out of the eligible white patients, 506 had low-weight births, compared to 304 for black patients, 45 for Asian patients, and 11 for Native American patients.

9. Involving humans in the decision making process can be helpful to a certain extent. As machines cannot recognize the biases of their own models, it is essential that the humans designing these models can recognize and appropriately adjust for algorithmic bias. However, humans themselves frequently have preconceived notions especially when it comes to race. Studies like the Implicit Association Test suggest that improper human involvement could actually worsen the racial bias dilemma. In order for an algorithm to maintain a high level of predictive power without significantly favoring one group over another, there must be a happy medium between blind machine computation and complete human involvement.

10. I would suggest a regression discontinuity research design to assess whether or not there is a significant difference in outcomes between patients who are eligible and patients who are not eligible for the program. Regression discontinuity design is useful for testing differences in subjects on either side of a certain treatment threshold - in this case, the top 25% risk score cutoff.