Predicting TB Medication Adherence Using a Risk Score Model

February 2024

# Data Preprocessing

This study included 249 individual participants. Since some participants had multiple regimens, the raw data contained 266 regimens in total. For this analysis we considered only the first regimen for each participant. After removing the subsequent regimens, the data contained one row for each participant (n = 249).

The raw data included 64 potential covariates. We dropped PTID2, hunger\_freq, health\_ctr, and post\_tb\_pt\_work because they were either not listed in the data dictionary or were listed with a note that they should not be included in the model. We summarized the family support, evaluation of health services, motivation, and TB disinformation variables by taking the median value for each category.

The data dictionary noted that pills variable was coded as 1 for 0-3 pills, 2 for 4-6 pills, 3 for 7-9 pills, 4 for 10-11 pills, and 5 for 12+ pills. However, the values in the data were 0.25, 0.50, 0.75, and 1.00. We converted this variable to the scale in the data dictionary by multiplying each value by 4. Note that this results in no values of 5 in the cleaned data (indicating no subjects taking 12+ pills). The adr\_freq variable also needed to be multiplied by 4 for the same reason. In this variable’s case, all values listed in the data dictionary (0, 1, 2, 3, and 4) are present in the transformed data.

Distributions of the continuous covariates are visualized in the appendix. There were seven continuous variables that were converted to categorical using the following cutoffs (selected based on class balance or background knowledge):

* age\_BLchart: <16 years, 16-17 years, 18+ years
* audit\_cat: 0, >0
* ace\_cat: 0, 1, >1
* tx\_mos\_cat: months, >6 months
* self\_eff\_cat: score of , >12
* stig\_tot\_cat: score of , >30
* phq9\_tot\_cat: score of , >10

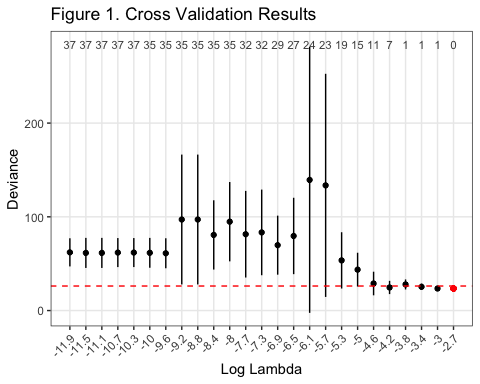
The categorical variables ram and regular\_drug were dropped because they each had only 5 patients responding with “yes”. The variables edu\_level\_mom and edu\_level\_dad could potentially be combined, but they contained many “I don’t know” responses so were also dropped.

After processing, the dataset contained 32 covariates (listed in the Appendix). We excluded 37 patients with missing covariate data, leaving us with 212 observations.

In order to create a risk score model, we must have a binary outcome with integer (or near-integer) covariates. Here, we are interested in medication adherence (PCTadherence), which is a continuous variable that measures the percentage of doses taken on time. To make this outcome binary, we define a cutoff of 90% above which we’ll consider the patient “adherent”. We observe a high level of class imbalance using this cutoff, with only 8.0% of participants classified as “non-adherent” (n = 17) and 92.0% of participants classified as “adherent” (n = 195).

# Model Fitting

Out model-fitting algorithm uses a parameter that penalizes nonzero coefficients. In other words, a higher value of will result in fewer covariates being included in the risk score model. To determine the best value for , we ran cross-validation. Figure 1 plots the mean model deviance for a range of potential values. The best fitting model will have the lowest deviance (shown in red). The numbers at the top of the plot show the number of nonzero coefficients in the model fit with each value of . We can see that the “best” model for these data has no nonzero coefficients, which suggests that there isn’t a very strong relationship between the features and the outcome, especially given the small sample size and class imbalance. However, Figure 1 shows that a model with seven nonzero coefficients had a deviance within one standard error of the “best” model’s deviance. Therefore, we will choose when fitting the full model.



The score card for the model with is presented in Table 1. A patient’s total score can be calculated by multiplying each variable’s response by the number of points shown on the right and then adding the points together. For example, if a patient’s responses to the variables listed in Table 1 were 1, 1, 1, 1, 5, 3, and 4, respectively, their total score would be . Table 2 can then be used to map this score to its associated risk of non-adherence. For this example patient, their risk of non-adherence would be 16%.

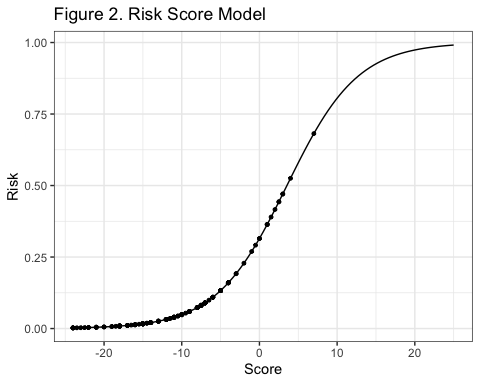
Score Card

|  | Points |
| --- | --- |
| Lives with Single Dad (0/1) | 10 |
| Pills Value (1-5) | 3 |
| Has Never Had Covid (0/1) | -4 |
| In-Person DOT Only (0/1) | 6 |
| Median Family Support Score (1-5) | -1 |
| Median Motivation Score (1-5) | -2 |
| Median TB Knowledge Score (1-5) | -2 |

Score-Risk Map

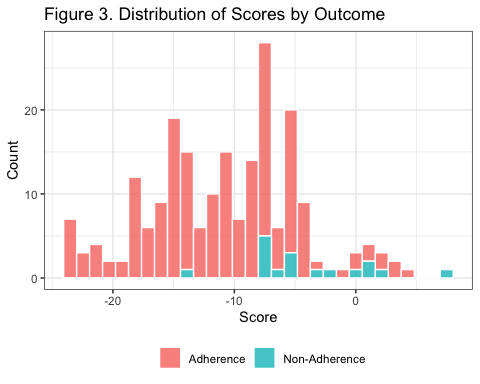
| Score | Risk |
| --- | --- |
| -16 | 0.01 |
| -14 | 0.02 |
| -12 | 0.03 |
| -10 | 0.05 |
| -8 | 0.07 |
| -6 | 0.11 |
| -4 | 0.16 |
| -2 | 0.23 |
| 0 | 0.31 |
| 2 | 0.42 |
| 4 | 0.53 |
| 6 | 0.63 |
| 8 | 0.73 |
| 10 | 0.81 |
| 12 | 0.87 |
| 14 | 0.91 |
| 16 | 0.94 |
| 18 | 0.96 |
| 20 | 0.97 |
| 22 | 0.98 |
| 24 | 0.99 |

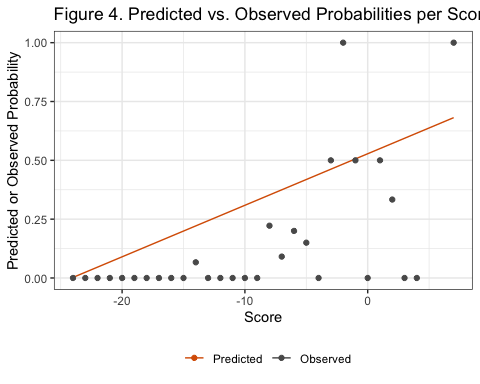
Figure 2 visualizes the logistic regression curve of this model. The observed scores from this dataset are plotted as points along the curve.



# Model Evaluation

Figure 3 plots the distribution of scores by adherence outcome. We can see that those with lower scores tended to be adherent while those with higher score tended to be non-adherent.

 We can also visualize the relationship between score and probability of adherence as in Figure 4. The grey points indicate the observed probability of adherence for each observed score. The orange line indicates the predicted probability of adherence associated with each score. We can see that this line tends to overestimate the probability of non-adherence.



We compared the performance of our risk score model to logistic regression, lasso regression, and rounded logistic regression. The coefficients for each of these models are reported in Table 3. As expected, the logistic regression model assigns a non-zero coefficient value to each covariate, while the lasso model shrinks many coefficients to zero.

Performance metrics for these models are reported in Table 4. Overall, we observed that variable selection reduced the performance of the models, as the logistic and rounded regression models had AUC values above 0.97, while the risk score and lasso models had AUC values of 0.831 and 0.819. However, these metrics were evaluated using the training data, where the logistic and rounded logistic models were likely highly overfit. Rerunning these models on a validation data set would help estimate their true performance.

Model Coefficients

|  | Risk | Logistic | Lasso | Rounded |
| --- | --- | --- | --- | --- |
| gendermale | 0 | -5.365 | 0.000 | -3 |
| concomitant\_tbyes | 0 | 12.638 | 0.000 | 6 |
| lives\_w\_momyes | 0 | 13.073 | 0.000 | 6 |
| lives\_w\_parentsno parents | 0 | -27.889 | 0.000 | -14 |
| lives\_w\_parentssingle dad | 10 | 30.266 | 0.135 | 15 |
| lives\_w\_parentssingle mom | 0 | 9.077 | 0.000 | 4 |
| current\_sx\_none1 | 0 | 0.335 | 0.000 | 0 |
| pills | 3 | 16.100 | 0.014 | 8 |
| dosis\_fijasyes | 0 | 13.385 | 0.000 | 7 |
| daily\_contyes | 0 | -12.202 | 0.000 | -6 |
| monoRyes | 0 | -18.693 | 0.000 | -9 |
| adr\_freq | 0 | 5.245 | 0.000 | 3 |
| fam\_accompany\_dot | 0 | -5.438 | -0.002 | -3 |
| fam\_dislikefriends | 0 | 3.446 | 0.009 | 2 |
| autonomy\_obedient | 0 | -1.861 | -0.013 | -1 |
| tobacco\_freq | 0 | 12.185 | 0.000 | 6 |
| drug\_useyes | 0 | -3.950 | 0.022 | -2 |
| stig\_health\_ctr | 0 | -6.014 | 0.000 | -3 |
| tto\_anterior\_tb1 | 0 | 13.724 | 0.000 | 7 |
| prior\_covidno | -4 | -27.233 | 0.000 | -13 |
| prior\_covidsuspected (unconfirmed) | 0 | -20.150 | 0.000 | -10 |
| covid\_es | 0 | -11.090 | 0.000 | -5 |
| monitor1in-person DOT + family supervision | 0 | 15.671 | 0.000 | 8 |
| monitor1in-person DOT only | 6 | 27.149 | 0.000 | 13 |
| monitor1no supervision | 0 | 61.843 | 0.000 | 30 |
| monitor1VDOT + family supervision | 0 | 25.109 | 0.000 | 12 |
| psych\_interventionno intervention needed | 0 | -1.492 | 0.000 | -1 |
| psych\_interventionnot evaluated | 0 | -12.887 | 0.000 | -6 |
| psych\_interventionSAME | 0 | -1.593 | 0.000 | -1 |
| family\_median | -1 | -4.840 | -0.017 | -2 |
| health\_svc\_median | 0 | 6.369 | 0.000 | 3 |
| motiv\_median | -2 | -11.662 | 0.000 | -6 |
| knowledge\_median | -2 | -9.870 | 0.000 | -5 |
| age\_cat16-17 | 0 | 8.779 | 0.000 | 4 |
| age\_cat18+ | 0 | 2.848 | 0.000 | 1 |
| audit\_cat0 | 0 | 29.489 | 0.000 | 14 |
| ace\_cat0 | 0 | 2.059 | 0.000 | 1 |
| ace\_cat1 | 0 | 7.571 | 0.000 | 4 |
| self\_eff\_cat> 12 | 0 | -5.339 | 0.000 | -3 |
| stig\_tot\_cat> 30 | 0 | 14.379 | 0.000 | 7 |
| phq9\_tot\_cat> 10 | 0 | -21.760 | 0.000 | -11 |

Model Performance

|  | Threshold | Specificity | Sensitivity | AUC |
| --- | --- | --- | --- | --- |
| Risk | 0.067 | 0.667 | 0.941 | 0.831 |
| Logistic | 0.282 | 0.959 | 0.941 | 0.987 |
| Lasso | 0.101 | 0.846 | 0.765 | 0.819 |
| Rounded | 0.257 | 0.938 | 0.941 | 0.972 |

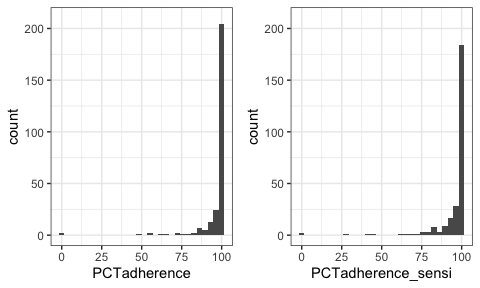
For our risk score model, a threshold of 6.7% risk resulted in the optimal performance. This probability corresponds to a score of -8.4.

# Conclusion

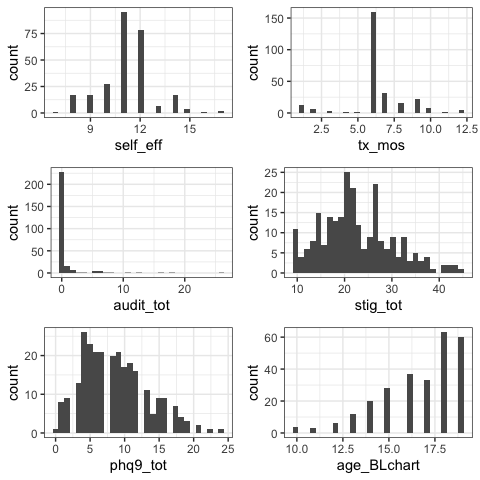
Despite a small sample size with only 17 non-adherent participants, our example model output made sense intuitively. Our model identified living with a single dad, having more pills to take, and having to take the medication in-person at a clinic as factors increasing the risk of non-adherence. Having never had COVID, higher family support, higher motivation, and higher knowledge of TB decreased the risk of non-adherence. Although the performance of this model should be validated with new data, it may still help with identifying which factors are the most important in predicting non-adherence.

# EDA Appendix

### Figure A1. Outcome Distributions



### Figure A2. Continuous Covariate Distributions



### Table A1. Processed Variable Summary

| **Characteristic** | **N = 212** |
| --- | --- |
| gender |  |
| female | 78 (37%) |
| male | 134 (63%) |
| concomitant\_tb | 22 (10%) |
| lives\_w\_mom | 181 (85%) |
| lives\_w\_parents |  |
| 2 parents | 118 (56%) |
| no parents | 14 (6.6%) |
| single dad | 15 (7.1%) |
| single mom | 65 (31%) |
| current\_sx\_none |  |
| 0 | 165 (78%) |
| 1 | 47 (22%) |
| pills |  |
| 1 | 36 (17%) |
| 2 | 42 (20%) |
| 3 | 22 (10%) |
| 4 | 112 (53%) |
| dosis\_fijas | 71 (33%) |
| daily\_cont | 12 (5.7%) |
| monoR | 11 (5.2%) |
| adr\_freq |  |
| 0 | 88 (42%) |
| 1 | 68 (32%) |
| 2 | 38 (18%) |
| 3 | 11 (5.2%) |
| 4 | 7 (3.3%) |
| fam\_accompany\_dot |  |
| 1 | 33 (16%) |
| 2 | 32 (15%) |
| 3 | 40 (19%) |
| 4 | 16 (7.5%) |
| 5 | 91 (43%) |
| fam\_dislikefriends |  |
| 1 | 71 (33%) |
| 2 | 70 (33%) |
| 3 | 45 (21%) |
| 4 | 13 (6.1%) |
| 5 | 13 (6.1%) |
| autonomy\_obedient |  |
| 1 | 1 (0.5%) |
| 2 | 5 (2.4%) |
| 3 | 41 (19%) |
| 4 | 68 (32%) |
| 5 | 97 (46%) |
| tobacco\_freq |  |
| 0 | 182 (86%) |
| 1 | 24 (11%) |
| 2 | 5 (2.4%) |
| 3 | 1 (0.5%) |
| drug\_use | 25 (12%) |
| stig\_health\_ctr |  |
| 1 | 137 (65%) |
| 2 | 35 (17%) |
| 3 | 23 (11%) |
| 4 | 8 (3.8%) |
| 5 | 9 (4.2%) |
| tto\_anterior\_tb |  |
| 0 | 209 (99%) |
| 1 | 3 (1.4%) |
| prior\_covid |  |
| confirmed | 25 (12%) |
| no | 149 (70%) |
| suspected (unconfirmed) | 38 (18%) |
| covid\_es |  |
| 0 | 68 (32%) |
| 1 | 118 (56%) |
| 2 | 25 (12%) |
| 3 | 1 (0.5%) |
| monitor1 |  |
| family supervision only | 43 (20%) |
| in-person DOT + family supervision | 25 (12%) |
| in-person DOT only | 129 (61%) |
| no supervision | 1 (0.5%) |
| VDOT + family supervision | 14 (6.6%) |
| psych\_intervention |  |
| MINSA referral | 25 (12%) |
| no intervention needed | 47 (22%) |
| not evaluated | 90 (42%) |
| SAME | 50 (24%) |
| adherence\_outcome | 17 (8.0%) |
| family\_median |  |
| 1 | 2 (0.9%) |
| 1.5 | 1 (0.5%) |
| 2 | 3 (1.4%) |
| 2.5 | 2 (0.9%) |
| 3 | 28 (13%) |
| 3.5 | 13 (6.1%) |
| 4 | 50 (24%) |
| 4.5 | 8 (3.8%) |
| 5 | 105 (50%) |
| health\_svc\_median |  |
| 3 | 4 (1.9%) |
| 3.5 | 8 (3.8%) |
| 4 | 63 (30%) |
| 4.5 | 30 (14%) |
| 5 | 107 (50%) |
| motiv\_median |  |
| 2 | 1 (0.5%) |
| 2.5 | 1 (0.5%) |
| 3.5 | 1 (0.5%) |
| 4 | 32 (15%) |
| 4.5 | 24 (11%) |
| 5 | 153 (72%) |
| knowledge\_median |  |
| 2 | 2 (0.9%) |
| 3 | 53 (25%) |
| 3.5 | 3 (1.4%) |
| 4 | 154 (73%) |
| age\_cat |  |
| < 16 | 62 (29%) |
| 16-17 | 60 (28%) |
| 18+ | 90 (42%) |
| audit\_cat |  |
| >0 | 32 (15%) |
| 0 | 180 (85%) |
| ace\_cat |  |
| > 1 | 100 (47%) |
| 0 | 56 (26%) |
| 1 | 56 (26%) |
| self\_eff\_cat |  |
| <= 12 | 185 (87%) |
| > 12 | 27 (13%) |
| stig\_tot\_cat |  |
| <= 30 | 185 (87%) |
| > 30 | 27 (13%) |
| phq9\_tot\_cat |  |
| <= 10 | 145 (68%) |
| > 10 | 67 (32%) |

# Code Appendix

knitr::opts\_chunk$set(echo = FALSE)  
#devtools::install\_github("hjeglinton/riskscores", build\_vignettes = TRUE)  
  
library(knitr)  
library(xtable)  
library(glmnet)  
library(pROC)  
library(tableone)  
library(glmnet)  
library(caret)  
library(gridExtra)  
#library(kableExtra)  
library(gtsummary)  
library(riskscores)  
source('risk.R')  
  
raw\_data <- read.csv("../data/Peru\_TB\_data.csv")   
  
source('tb\_preprocessing.R')  
tb\_df <- tb\_preprocessing(raw\_data)   
tb\_mat <- tb\_as\_matrix(tb\_df)  
  
  
X <- as.matrix(tb\_mat[,-ncol(tb\_mat)])  
y <- tb\_mat[,ncol(tb\_mat)]  
  
# CV  
# get folds  
folds <- stratify\_folds(y, nfolds = 5, seed = 1)  
  
cv\_results <- cv\_risk\_mod(X, y, foldids = folds, a = -10, b = 10, nlambda = 25)  
  
plot(cv\_results, lambda\_text = FALSE) +   
 labs(title = "Figure 1. Cross Validation Results")  
  
  
lambda0 <- exp(-4.2)  
  
mod <- risk\_mod(X, y, lambda0 = lambda0, a = -10, b = 10)  
  
  
data.frame(mod$model\_card, row.names = c("Lives with Single Dad (0/1)",   
 "Pills Value (1-5)",  
 "Has Never Had Covid (0/1)",  
 "In-Person DOT Only (0/1)",  
 "Median Family Support Score (1-5)",  
 "Median Motivation Score (1-5)",  
 "Median TB Knowledge Score (1-5)")) %>%  
   
 kable(caption = "Score Card" ,   
 booktabs = TRUE, linesep = "") #%>%   
 #kableExtra::kable\_styling(latex\_options = c("HOLD\_position"), font\_size = 10)  
  
score\_range <- seq(-16, 25, 2)  
  
data.frame(Score = as.character(score\_range),   
 Risk = as.character(round(get\_risk(mod, score\_range), 2))) %>%  
 # t() %>%  
  
 kable(caption = "Score-Risk Map" ,   
 booktabs = TRUE)# %>%   
 #kableExtra::kable\_styling(latex\_options = c("HOLD\_position"), font\_size = 8) %>%  
 #row\_spec(row = 1, hline\_after = TRUE)  
  
  
  
predict\_df <- data.frame(score = predict(mod, type = "score"),  
 response = predict(mod, type = "response"))  
  
ggplot() +   
 geom\_point(data = predict\_df, aes(x = score, y = response), size = 1) +  
 geom\_function(data = data.frame(x = seq(-20,25)), aes(x),   
 fun = function(x) get\_risk(mod, x)) +   
 labs(x = "Score", y = "Risk", title = "Figure 2. Risk Score Model") +  
 #scale\_x\_continuous(breaks = seq(-60, 50, 10)) +   
 #scale\_y\_continuous(breaks = seq(0, 1, 0.10)) +   
 #geom\_point(aes(x = -14, y = get\_risk(mod, -14)), color = "blue",   
 # shape = 3, size = 3) +   
 #geom\_point(aes(x = 18, y = get\_risk(mod, 18)), color = "blue",   
 # shape = 3, size = 3) +   
  
 theme\_bw()  
# Figure 3  
ggplot() +   
 geom\_histogram(aes(x = predict\_df$score, fill = factor(tb\_df$adherence\_outcome)),  
 alpha = 0.8, color = "white") +   
 labs(x = "Score", y = "Count", fill = "",   
 title = "Figure 3. Distribution of Scores by Outcome") +   
 scale\_fill\_discrete(labels = c("Adherence", "Non-Adherence")) +   
 theme\_bw() +   
 theme(legend.position = "bottom")  
# Figure 4  
range\_scores <- range(predict\_df$score)  
all\_scores <- unique(predict\_df$score)  
vals <- mod$gamma\*(mod$beta[1]+range\_scores)  
probs <- exp(vals)/(1+exp(vals))  
props <- tb\_df %>%   
 mutate(rnd\_scores = floor(predict\_df$score)) %>%  
 group\_by(rnd\_scores) %>%   
 summarize(prop = sum(adherence\_outcome)/n())   
  
ggplot() +   
 geom\_line(aes(x=range\_scores, y=probs, color = "Predicted")) +   
 geom\_point(aes(x=props$rnd\_scores,y=props$prop, color = "Observed")) +   
 scale\_color\_manual(name = "",   
 breaks = c("Predicted", "Observed"),  
 values = c("Predicted" = "#d95f02", "Observed" = "#5A5A5A")) +   
 labs(x="Score", y="Predicted or Observed Probability",  
 title = "Figure 4. Predicted vs. Observed Probabilities per Score") +   
 theme\_bw() +   
 theme(legend.position = "bottom")  
  
  
coef\_vals <- matrix(0, ncol=4, nrow=ncol(X)-1)  
  
# risk model prediction  
coef\_vals[, 1] <- coef(mod)[-1]  
risk\_probs <- predict\_df$response  
risk\_pred <- as.factor(ifelse(risk\_probs < 0.5, 0, 1))  
  
# glm prediction  
glm\_mod <- glm(y~X-1, family = "binomial")  
coef\_vals[, 2] <- coef(glm\_mod)[-1]  
glm\_probs<- predict(glm\_mod, type="response")  
glm\_pred <- as.factor(ifelse(glm\_probs < 0.5, 0, 1))  
  
# lasso prediction  
lasso\_res <- cv.glmnet(x=X[,-1], y=y, alpha=1)  
lasso\_mod <- glmnet(x=X[,-1], y=y, lambda=lasso\_res$lambda.min, alpha=1)  
coef\_vals[,3] <- coef(lasso\_mod)[-1]  
lasso\_probs <- as.vector(predict(lasso\_mod, newx=X[,-1]))  
lasso\_pred <- as.factor(ifelse(lasso\_probs < 0.5, 0, 1))  
  
# rounded logistic  
coef\_vals[,4] <- round(coef\_vals[,2] / abs(median(coef\_vals[,2])),0)  
x\_vars <- tb\_mat[,-c(1,ncol(tb\_mat))]  
round\_score <- x\_vars %\*% coef\_vals[,4]  
mod\_round <- glm(tb\_mat[,ncol(tb\_mat)] ~ round\_score, family = "binomial")  
round\_probs <- predict(mod\_round, type="response")  
round\_pred <- as.factor(ifelse(round\_probs < 0.5, 0, 1))  
  
  
# discrimination  
risk\_roc <- roc(factor(tb\_df$adherence\_outcome), risk\_probs)  
glm\_roc <- roc(factor(tb\_df$adherence\_outcome), glm\_probs)  
lasso\_roc <- roc(factor(tb\_df$adherence\_outcome), lasso\_probs)  
round\_roc <- roc(factor(tb\_df$adherence\_outcome), round\_probs)  
  
  
data.frame(coef\_vals,   
 row.names = dimnames(X)[[2]][-1]) %>%  
 kable(digits = 3, col.names = c("Risk", "Logistic", "Lasso", "Rounded"),  
 booktabs = T, caption = "Model Coefficients") #%>%  
  
 #kableExtra::kable\_styling(font\_size = 9,   
 # latex\_options = c("repeat\_header", "HOLD\_position"))  
rbind(coords(risk\_roc, "best"), coords(glm\_roc, "best"),   
 coords(lasso\_roc, "best"), coords(round\_roc, "best")) %>%  
 data.frame(auc = c(risk\_roc$auc[[1]], glm\_roc$auc[[1]], lasso\_roc$auc[[1]],  
 round\_roc$auc[[1]]),   
 row.names = c("Risk", "Logistic", "Lasso", "Rounded")) %>%  
 kable(digit = 3,   
 col.names = c("Threshold", "Specificity", "Sensitivity", "AUC"),  
 caption = "Model Performance",  
 booktabs = T)# %>%  
  
 # kableExtra::kable\_styling(font\_size = 9,   
 # latex\_options = c("repeat\_header", "HOLD\_position"))  
# outcome distributions  
p1 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = PCTadherence)) +   
 lims(y = c(0, 210)) +   
 theme\_bw()  
  
p2 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = PCTadherence\_sensi)) +   
 lims(y = c(0, 210)) +   
 theme\_bw()  
  
grid.arrange(p1, p2, ncol = 2)  
p3 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = self\_eff)) +   
 theme\_bw()  
  
p4 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = tx\_mos)) +   
 theme\_bw()  
  
p5 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = audit\_tot)) +   
 theme\_bw()  
  
p6 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = stig\_tot)) +   
 theme\_bw()  
  
p7 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = phq9\_tot)) +   
 theme\_bw()  
  
p8 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = age\_BLchart)) +   
 theme\_bw()  
  
p9 <- ggplot(raw\_data) +  
 geom\_histogram(aes(x = ace\_score)) +   
 theme\_bw()  
  
grid.arrange(p3, p4, p5, p6, p7, p8, ncol = 2)  
tbl\_summary(tb\_df)#%>%  
 #as\_kable\_extra(booktabs = TRUE,   
 # longtable = TRUE) #%>%  
 # kableExtra::kable\_styling(font\_size = 9,   
 # latex\_options = c("repeat\_header", "HOLD\_position"))  
  
  
  
# load risk model files  
X <- tb\_matrix[, -ncol(tb\_matrix)]  
y <- tb\_matrix[, ncol(tb\_matrix)]  
  
coef\_vals <- matrix(0, ncol=3, nrow=ncol(X)-1)  
  
# risk model prediction   
risk\_output\_cv <- cv\_risk\_mod(X, y, a=-5, b=5)  
risk\_output <- risk\_mod(X, y, a=-5, b=5, lambda0=risk\_output\_cv$lambda\_min)  
coef\_vals[,1] <- risk\_output$beta[-1]  
risk\_probs <- predict(risk\_output$glm\_mod, type="response")  
risk\_pred <- as.factor(ifelse(risk\_probs < 0.5, FALSE, TRUE))  
  
# glm prediction  
glm\_mod <- glm(y~X-1, family = "binomial")  
coef\_vals[, 2] <- coef(glm\_mod)[-1]  
glm\_probs<- predict(glm\_mod, type="response")  
glm\_pred <- as.factor(ifelse(glm\_probs < 0.5, FALSE, TRUE))  
  
# lasso prediction  
  
lasso\_res <- cv.glmnet(x=X[,-1], y=y, alpha=1)  
lasso\_mod <- glmnet(x=X[,-1], y=y, lambda=lasso\_res$lambda.min, alpha=1)  
coef\_vals[,3] <- coef(lasso\_mod)[-1]  
lasso\_probs <- as.vector(predict(lasso\_mod, newx=X[,-1]))  
lasso\_pred <- as.factor(ifelse(lasso\_probs < 0.5, FALSE, TRUE))  
  
print(coef\_vals)  
  
# confusion matrices  
confusionMatrix(as.factor(tb\_df$adherence\_outcome), risk\_pred)  
confusionMatrix(as.factor(tb\_df$adherence\_outcome), glm\_pred)  
confusionMatrix(as.factor(tb\_df$adherence\_outcome), lasso\_pred)  
  
# discrimination  
roc(as.factor(tb\_df$adherence\_outcome), risk\_probs)  
roc(as.factor(tb\_df$adherence\_outcome), glm\_probs)  
roc(as.factor(tb\_df$adherence\_outcome), lasso\_probs)  
  
# find risk score probs to summarize  
tb\_df$scores <- (X[,-1] %\*% risk\_output$beta[-1])  
ggplot(tb\_df)+geom\_histogram(aes(x=scores, fill=adherence\_outcome), alpha=0.5)  
  
# for each score find predicted probability and also find percent class 1  
range\_scores <- range(tb\_df$scores)  
all\_scores <- seq(range\_scores[1], range\_scores[2])  
vals <- risk\_output$gamma\*(risk\_output$beta[1]+range\_scores)  
probs <- exp(vals)/(1+exp(vals))  
props <- tb\_df %>%   
 mutate(rnd\_scores = floor(scores)) %>%  
 group\_by(rnd\_scores) %>%   
 summarize(prop = sum(adherence\_outcome)/n())   
  
ggplot()+geom\_line(aes(x=range\_scores, y=probs)) +   
 geom\_point(aes(x=props$rnd\_scores,y=props$prop)) +   
 labs(x="Score", y="Predicted or Observed Probability")