Sensitivity and Specificity

Derek Sonderegger

2018-11-19

# I haven't uploaded BurkPx to GitHub yet, but eventually we'll be  
# able to just download the BurkPx package.   
# library(devtools)   
# install\_github('dereksonderegger/BurkPx')   
library(BurkPx)  
library(ggplot2)

# Human ROC analysis on Human models

*Note to self: I need to ask Erik for a description of the human trials we used so as to document what data we used to fit the models.*

We first split the patients into test/training sets and then fit all the various models (IgG, IgM, IgGM, etc) using only patients from the training set. Because we have 100 Meliod patients, then 50 of those patients get assigned to the test group and 50 to the training group. Likewise of the 400 controls, 200 get assigned to the test set and 200 to the training.

Once the patients have been assigned to either the test or training set, all of the patients serologies are included in the set. This means that a single patient with many serologies might have an oversized effect. But we did this to try to keep our sample sizes as high as possible.

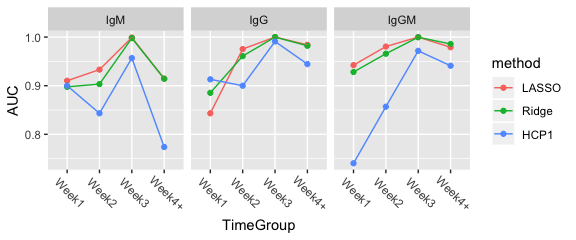
## Model Selection

To assess how well the models work, we will look at the Area Under the Curve (AUC) for the Reciever-Operator Curve (ROC). To generate this by week, We take the human data and split it into Healthy, Week 1, Week 2, etc. For each of the Weeks, we calculate probability of infection for the the Healthy and Infected groups using the six different human models.

ATPD, imps : we’ve run out of these lab supplies…

#################################################  
## ROC for each model ##  
#################################################  
time = 'Week1'  
method = 'LASSO'  
type = 'IgG'  
  
AUC\_results <- NULL  
for(time in c('Week1','Week2','Week3', 'Week4+')){  
 for(method in c('LASSO', 'Ridge', 'HCP1')){  
 for(type in c('IgM', 'IgG', 'IgGM')){  
 if( type %in% c('IgM', 'IgG') ){  
 df <- Human\_BurkPx\_test %>%  
 filter(Type == type) %>% filter(TimeGroup %in% c('Healthy', time)) %>%  
 spread(Antigen, Value) %>%  
 # group\_by(PatientID) %>% sample\_n(1) %>% # Use only 1 replicant from each patient  
 group\_by() %>% select(-PatientID, -SerumID, -TimeGroup, -Type, -Rep) %>% complete()  
 }else{  
 df <- Human\_BurkPx\_test %>%  
 group\_by(PatientID, Type, Rep) %>%  
 unite( 'Antigen', Type, Antigen ) %>%  
 filter(TimeGroup %in% c('Healthy', time)) %>%  
 spread(Antigen, Value) %>%  
 # group\_by(PatientID) %>% sample\_n(1) %>% # Use only 1 replicant from each patient  
 group\_by() %>% select(-PatientID, -TimeGroup, -Rep)  
 }  
  
 df$p <- predict(models[[str\_c('Human\_',type,'\_',method)]],   
 newdata=df, na.action = na.pass) %>%   
 as.vector()  
 temp <- pROC::roc(Status ~ p, data=df)  
 AUC\_results <- AUC\_results %>%  
 rbind( data.frame(Type = type, TimeGroup=time, method=method, AUC=pROC::auc(temp) ) )  
 }  
 }  
}

ggplot(AUC\_results, aes(x=TimeGroup, y=AUC, color=method)) +  
 facet\_grid(.~Type) +   
 geom\_point() +   
 geom\_line(aes(x=as.numeric(TimeGroup))) +  
 theme(axis.text.x = element\_text(angle=-45, vjust=0, hjust=.5) )



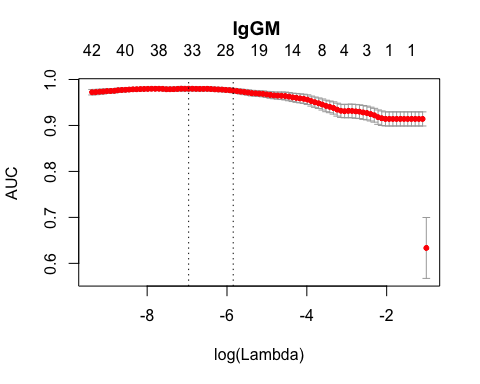
From this analysis, it is clear that for the first week, we need both the IgG and IgM sereologies. It also seems that the LASSO is working better than Ridge Regression and that IgG is working better than IgM. However the best performance is by the IgGM data which uses both IgG and IgM observation values.

## How many covariates are used?

# the models object is a list of all the models I created. The naming convention  
# follows the following convention: Human\_IgGM\_LASSO. Species\_AntigenList\_Method  
  
# Which covariates are used in the IgG LASSO model?  
temp <- coef( models$Human\_IgG\_LASSO )   
data.frame(Antigen=rownames(temp), Coef=as.vector(temp)) %>%  
 filter(Antigen != '(Intercept)', Coef != 0 )   
## Antigen Coef  
## 1 AtpD -1.813250e-05  
## 2 ClpX 4.510126e-04  
## 3 DNAK -3.195289e-04  
## 4 HCP1 1.256245e-04  
## 5 IMPS -1.344398e-04  
## 6 LPSA 3.155172e-04  
## 7 LPSB 7.675884e-05  
## 8 NADH -1.056039e-04  
## 9 rpIL 3.318410e-06  
## 10 S1652 5.753299e-05  
## 11 S1850 1.032167e-05  
  
# Which covariates are used in the IgGM LASSO model?  
temp <- coef( models$Human\_IgGM\_LASSO )   
data.frame(Antigen=rownames(temp), Coef=as.vector(temp)) %>%  
 filter(Antigen != '(Intercept)', Coef != 0 )   
## Antigen Coef  
## 1 IgG\_Arg 2.075479e-05  
## 2 IgG\_ClpX 9.754925e-04  
## 3 IgG\_DNAK -7.026007e-04  
## 4 IgG\_GroS 2.829079e-05  
## 5 IgG\_HCP1 3.777466e-04  
## 6 IgG\_LPSA 4.057119e-04  
## 7 IgG\_LPSB 9.624657e-05  
## 8 IgG\_NADH -6.125564e-04  
## 9 IgG\_rpIL 2.759273e-05  
## 10 IgG\_S0135 3.164166e-04  
## 11 IgG\_S1652 6.600877e-05  
## 12 IgG\_S1850 6.570940e-05  
## 13 IgM\_AtpD -7.401217e-04  
## 14 IgM\_ClpX -1.067747e-04  
## 15 IgM\_CPS 7.428564e-05  
## 16 IgM\_DNAK 2.971660e-03  
## 17 IgM\_GroS -1.803437e-04  
## 18 IgM\_HCP1 4.083163e-08  
## 19 IgM\_IMPS 1.296829e-04  
## 20 IgM\_LPSB -5.187237e-04  
## 21 IgM\_NADH 1.395115e-03  
## 22 IgM\_OmpA -1.798301e-03  
## 23 IgM\_rpIL -3.902404e-03  
## 24 IgM\_S0135 -5.788038e-04  
## 25 IgM\_S0530 3.258072e-03  
## 26 IgM\_S1652 5.625445e-04  
## 27 IgM\_S1850 -5.612173e-04  
## 28 IgM\_WCL -1.802317e-04  
  
# Of those, which are used in both IgG and IgM?  
data.frame(Antigen=rownames(temp), Coef=as.vector(temp)) %>%  
 filter(Antigen != '(Intercept)', Coef != 0 ) %>%  
 separate(Antigen, into=c('Type','Antigen'), extra='merge') %>%  
 group\_by(Antigen) %>% arrange(Antigen) %>% count()  
## # A tibble: 18 x 2  
## # Groups: Antigen [18]  
## Antigen n  
## <chr> <int>  
## 1 Arg 1  
## 2 AtpD 1  
## 3 ClpX 2  
## 4 CPS 1  
## 5 DNAK 2  
## 6 GroS 2  
## 7 HCP1 2  
## 8 IMPS 1  
## 9 LPSA 1  
## 10 LPSB 2  
## 11 NADH 2  
## 12 OmpA 1  
## 13 rpIL 2  
## 14 S0135 2  
## 15 S0530 1  
## 16 S1652 2  
## 17 S1850 2  
## 18 WCL 1

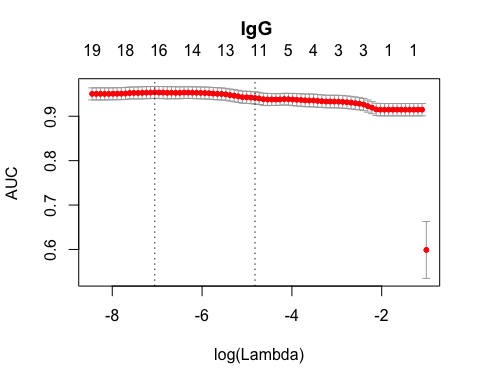
## What happens to the IgGM model as we decrease the number of covariates.

plot(models$Human\_IgGM\_LASSO, main='IgGM \n')



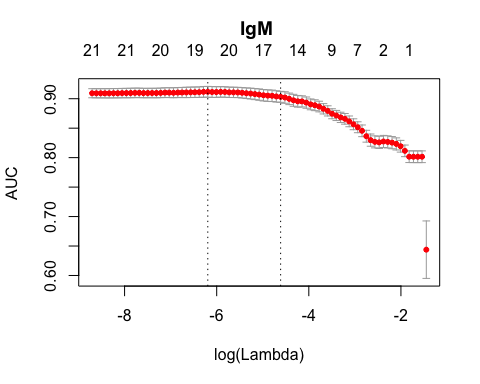
## What happens to the IgG model as we decrease the number of covariates.

plot(models$Human\_IgG\_LASSO, main='IgG \n')



## What happens to the IgM model as we decrease the number of covariates.

plot(models$Human\_IgM\_LASSO, main='IgM \n')



I am a little concerned with how large the IgGM model is, but considering our sample sizes..

Human\_BurkPx %>%  
 group\_by(PatientID, Status) %>%  
 count() %>%   
 group\_by(Status) %>% count() %>%  
 group\_by() %>% mutate(perc = nn/sum(nn) )  
## # A tibble: 2 x 3  
## Status nn perc  
## <fct> <int> <dbl>  
## 1 Negative 399 0.808  
## 2 Melioid 95 0.192

maybe this isn’t a big deal. The data consists of nearly 500 subjects split 80% / 20% between Healthy and Melioid.