Sensitivity and Specificity

Derek Sonderegger

2019-06-28

# library(devtools)  
# install\_github('dereksonderegger/BurkPx')  
library(BurkPx)  
library(ggplot2)

# Methods

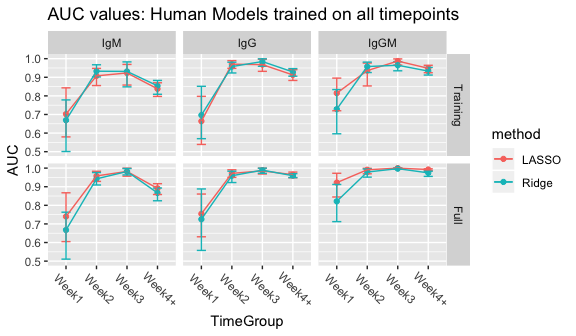
Our initial research identified 21 differnt antigens that could be predictive of Meliodosis. To create an assay with high sensitivity and specificity for future observations, we created a statistical model using our present data but protected ourselves from overfitting the data by using the *shrinkage methods* LASSO and ridge regression. These methods modify a a conventional linear model by continuously shrinking the model coeffients towards zero and then selecting a shrinkage amount via k-fold cross-validation. LASSO has the convenient property of pushing the regression coeffiencients to zero, effectively removing the antigen from the regression model, while ridge regression typcially will simply push coeffients to just be small. As a result, LASSO tends to produce a more parsimonious set of antigens to base the assay on.

# Human ROC analysis on Human models

We first split the human 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. We created statistical models using the training set using both LASSO and ridge regression. We will then challenge the models and calculate the Receiver Operator Characteristic (ROC) curves and its Area Under the Curve (AUC) using the testing set. Next we create LASSO and ridge regression models using the full set of data, which we will recommend for making predictions about future observations.

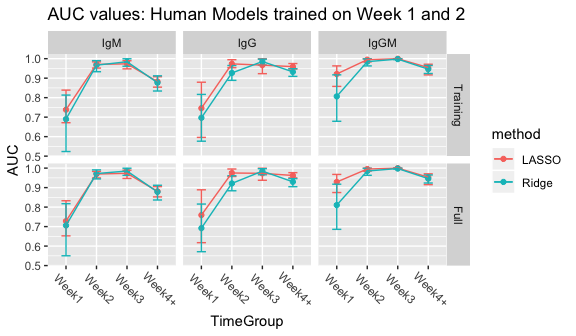
For the human data, we generated AUC values for each week after the patient was admitted to the hospital. To generate this by week, we split the testing data into Healthy, Week 1, Week 2, etc. For each, we calculate AUC using the three different human models (IgG, IgM, IgGM).



The graph columns denote the serum type and the graph rows represent which dataset the model was trained on. The training row was trained on the training set and tested on the test set. However the full row has models built using both the training and test sets and then challenged by the test set.

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 might be 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.

Next we look at the human models that were trained on only the Week 1 and 2 data. We compare the models against the same version 1 test set so that the comparisons



## How many covariates are used?

For each of the full LASSO models, it would be helpful to know which antigens where selected for inclusion into the model.

#### IgG

|  |  |
| --- | --- |
| Antigen | Coef |
| BPSS1498\_HCP1.B | 2.036e-05 |
| LPSA | 0.0001771 |
| LPSB | 5.506e-06 |

#### IgM

|  |  |
| --- | --- |
| Antigen | Coef |
| BPSL1201\_IMPS | -2.719e-05 |
| BPSL1404\_ClpX | -0.0007854 |
| BPSL2522\_OmpA | -0.0008399 |
| BPSL2827\_DNAK | 0.000272 |
| BPSS0476\_GroS | -0.0002748 |
| BPSS1498\_HCP1.B | 0.0002284 |
| BPSS1769\_NADH | -0.00016 |
| BPSS1850 | -6.548e-05 |
| CPS | 0.0001404 |
| LPSA | 0.0002212 |

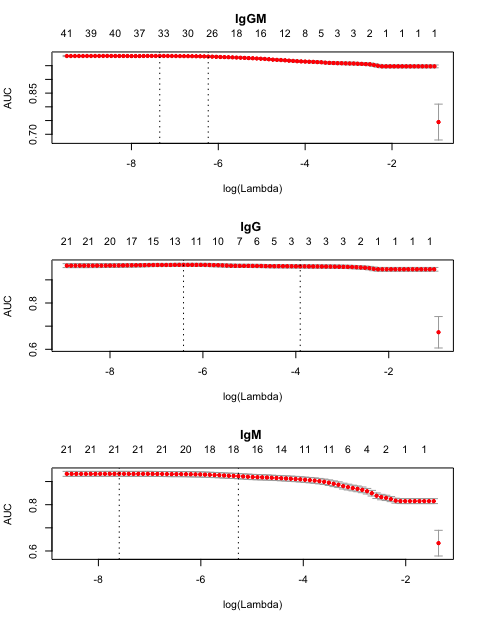
#### IgGM

|  |  |
| --- | --- |
| Antigen | Coef |
| IgG\_BPSL1743\_Arg | 0.000354 |
| IgG\_BPSL2522\_OmpA | -0.0001495 |
| IgG\_BPSL2827\_DNAK | -0.0004162 |
| IgG\_BPSS0135 | 0.0003345 |
| IgG\_BPSS0530 | 0.0002753 |
| IgG\_BPSS1498\_HCP1.B | 0.0002604 |
| IgG\_BPSS1652 | 5.71e-05 |
| IgG\_BPSS1850 | -0.000105 |
| IgG\_LPSA | 0.0002906 |
| IgG\_LPSB | 2.319e-05 |
| IgM\_BPSL1201\_IMPS | -0.0006715 |
| IgM\_BPSL1743\_Arg | -0.0001083 |
| IgM\_BPSL2522\_OmpA | -0.001747 |
| IgM\_BPSL2827\_DNAK | 0.0007385 |
| IgM\_BPSL3222\_rpIL | -0.0003492 |
| IgM\_BPSL3396\_AtpD | -0.0005359 |
| IgM\_BPSS0135 | -0.000111 |
| IgM\_BPSS0476\_GroS | -0.000494 |
| IgM\_BPSS0477\_GroEL2 | 0.0007597 |
| IgM\_BPSS0530 | 0.001095 |
| IgM\_BPSS1498\_HCP1.B | 0.0001904 |
| IgM\_BPSS1652 | 0.0008359 |
| IgM\_BPSS1769\_NADH | 0.0006622 |
| IgM\_BPSS1850 | -7.593e-05 |
| IgM\_CPS | 0.0001627 |
| IgM\_LPSA | 4.834e-05 |
| IgM\_LPSB | -0.0002087 |
| IgM\_MSHR5855.WCL | -0.001502 |

#### Both IgG and IgM

|  |  |
| --- | --- |
| Antigen | n |
| BPSL1201\_IMPS | 1 |
| BPSL1743\_Arg | 2 |
| BPSL2522\_OmpA | 2 |
| BPSL2827\_DNAK | 2 |
| BPSL3222\_rpIL | 1 |
| BPSL3396\_AtpD | 1 |
| BPSS0135 | 2 |
| BPSS0476\_GroS | 1 |
| BPSS0477\_GroEL2 | 1 |
| BPSS0530 | 2 |
| BPSS1498\_HCP1.B | 2 |
| BPSS1652 | 2 |
| BPSS1769\_NADH | 1 |
| BPSS1850 | 2 |
| CPS | 1 |
| LPSA | 2 |
| LPSB | 2 |
| MSHR5855.WCL | 1 |

## What happens as we decrease the number of covariates?

For the LASSO models, as the shrinkage parameter increases, the number of non-zero model coefficients decreases. We can examine how the AUC decreases as a function of , and therefore indirectly as a function of number of non-zero model coefficients. In the following graphs, the numbers along the top of the graph represent the number of non-zero coefficients. The dotted line on the right represents the selected shrinkage value we used for the Full LASSO models. 

# Nonhuman Primate

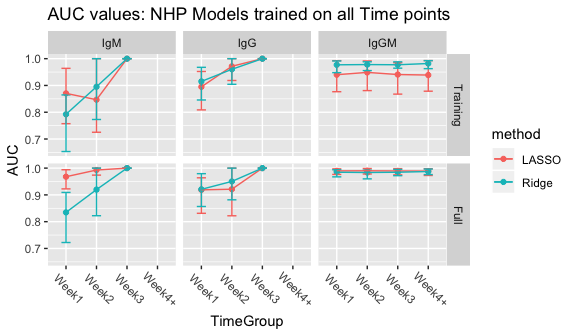
The nonhuman primate data was collected from both the Batelle and Tulane information. When I created the testing and training sets, we split the Batelle individual subjects equally into the test and training sets and likewise split the Tulane negative individuals equally into the test and training sets. We created two NHP train/test sets. The first version Batelle animals were selected randomly from Batelle animals that survived for the full experiment. The second training/testing set were created from the complete dataset.

## NHP Version 1

|  |  |  |  |
| --- | --- | --- | --- |
| Origin | Status | n | Type |
| Battelle | Negative | 6 | Training |
| Battelle | Melioid | 10 | Training |
| Tulane | Negative | 24 | Training |
| Battelle | Negative | 6 | Test |
| Battelle | Melioid | 12 | Test |
| Tulane | Negative | 23 | Test |

We next look at the AUC values for the Testing set based on either the model trained on just the training set vs the model trained on the full Version 1 dataset. Because we would expect the model trained on the test data would perform well on the test set, we should only draw conclusions from the model built from the training set. The only reason the model built from the full dataset exists is to be able to use it on future data. I show these results here just to verify that the full model doesn’t do worse than the training model.

## Warning: Removed 8 rows containing missing values (geom\_point).  
## Warning: Removed 8 rows containing missing values (geom\_errorbar).

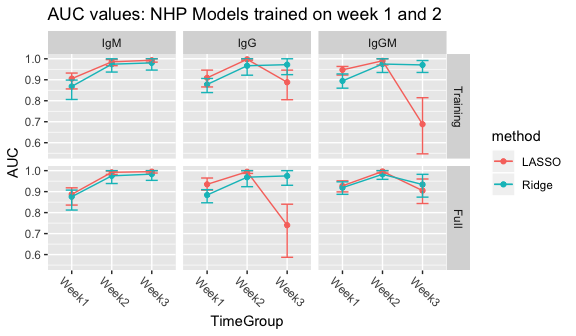


For the Non-human primates, we see that the models that were trained on only the training data still did quite well in predicting the testing set with AUC values in the 90s. However the models trained with both the test and training sets (aka “Full” data), did a better job predicting the testing set, as we would expect. Notably, there isn’t a huge difference in prediction capabilities and, unlike the human data, we see that the IgM serum is better at prediction than the IgG.

## NHP Version 3

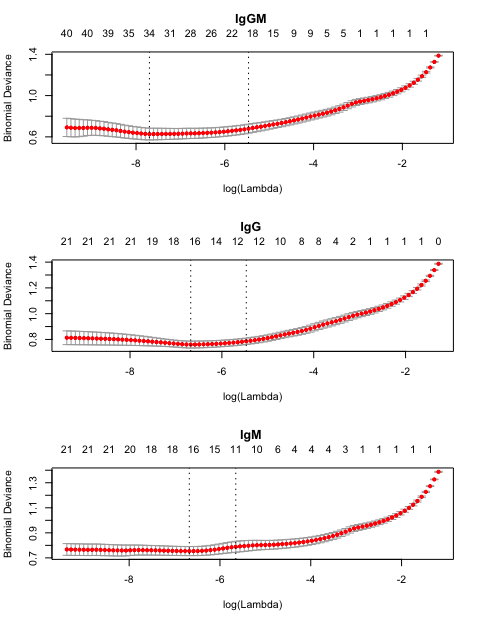
|  |  |  |  |
| --- | --- | --- | --- |
| Origin | Status | n | Type |
| Battelle | Negative | 40 | Training |
| Battelle | Melioid | 39 | Training |
| Tulane | Negative | 24 | Training |
| Battelle | Negative | 39 | Test |
| Battelle | Melioid | 43 | Test |
| Tulane | Negative | 23 | Test |

We next look at the AUC values for the Testing set based on either the model trained on just the training set vs the model trained on the full Version 1 dataset. Because we would expect the model trained on the test data would perform well on the test set, we should only draw conclusions from the model built from the training set. The only reason the model built from the full dataset exists is to be able to use it on future data. I show these results here just to verify that the full model doesn’t do worse than the training model.



For the Non-human primates, we see that the models that were trained on only the training data still did quite well in predicting the testing set with AUC values in the 90s. However the models trained with both the test and training sets (aka “Full” data), did a better job predicting the testing set, as we would expect. Notably, there isn’t a huge difference in prediction capabilities and, unlike the human data, we see that the IgGM combination of serums is better at prediction than the IgG serum.

## What happens as we decrease the number of covariates.

 Because of the reduced sample size compared to the Human models, the NHP models were fitted by minimizing the binomial deviance instead of AUC.