**Technical Report**: Qualification of a set of novel biomarkers when limited by small sample size for prediction of response to treatment in lupus nephritis.

Wolf BJ, Spainhour JCG, Tom Fleury, Oates JC.

A biomarker, as defined by the National Institutes of Health biomarkers definition working group, is “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (14)”. Biomarkers for disease have potential uses for early detection, individual risk assessment, prognostic capability, and determination of therapeutic response (9). Historically biomarkers have included physical traits or physiologic metrics such as blood pressure or body mass index and are designed to provide evidence of an individual’s health status (73,79). However, rapid advances in high-throughput technologies and increasing OMICs data provide the potential for discovery of molecular biomarkers directly relevant to patient health and disease status (101, Deyati et al. 2012). While discovery of molecular biomarkers has the potential to improve both disease diagnosis and treatment and reduce therapeutic cost through personalized medicine, few biomarkers are approved for use in clinical applications (Drucker and Krapfenbauer, 66,73,79, Poste- bring on the biomarkers 2011).

One reason few biomarkers are used in clinical practice is failure to move from the discovery phase to into clinical validation (79). There are many potential difficulties that may lead to failure to qualify and validate potentially relevant biomarkers. One difficulty in identifying relevant predictive biomarkers arises from low prevalence of disease prevalence in the population. When disease prevalence or response rate to therapy is low, it is difficult to gather sufficient information in initial studies to evaluate improvements in prediction that can be made using novel biomarkers. A second reason many biomarkers fail to progress to clinical application is inadequate sensitivity and specificity (59, 65, 76, 87). Poor sensitivity and/or specificity of individual markers results in part from the fact that many disease stem from complex interactions among genetic and environmental factors (5, 57). Failure to recognize interactions among biomarkers may lead to an inability to replicate study results in a validation study (24). Additionally, difficulties can also arise if with the data including differing types of biomarkers, for example if a data set includes combinations of categorical, ordinal, and continuous markers. Similarly, many biomarkers fail to meet statistical assumptions such as normality, which is problematic when using traditional statistical approaches to evaluate novel markers.

When evaluating a set of novel biomarkers to determine their prediction capability and which markers are the best candidates for a further study, all of these issues should be considered. The choice of statistical method can impact which markers are selected. The chosen method should be able to handle multiple types of data, be able to evaluate multiple markers simultaneously, handle potentially limited prevalence of the response, and ideally provide some measure of importance for each biomarker under consideration to aid in selection of additional markers when moving forward. Logistic regression (LR) is a traditional parametric statistical approach for building prediction models for classification and can evaluate the association between disease status and individual markers simultaneously. Additionally LR has the ability to evaluate different types of predictors simultaneously. However, LR suffers from loss of precision of regression estimates and poor predictive performance in new data as the number of predictor variables increases relative the number of observed events (ref). Many machine learning methods do not require several of the assumptions of traditional statistical models and are able to handle large numbers of predictors. However, machine learning models are often less interpretable than logistic regression models and may also be subject to over-fitting resulting in poor prediction performance.

There is not one statistical method that can provide a “best” model for all types of data, particularly when data sets are small. Thus it may be beneficial to evaluate predictive performance of multiple statistical models when examining novel biomarkers. Popular approaches for evaluating the predictive performance of statistical models include the split sample method in which the data randomly splitting the data into a training and a test set, *k*-fold cross validation which splits the data into *k* training sets and evaluates performance on the associated *k* test sets, and bootstrap validation in which samples are drawn from the data with replacement from the original data and evaluates predictive performance on observations not included in the bootstrap sample (Picard and Berk 1990, Streyerberg et al. 2001, Streyerberg 2003, Austin and Tu 2004). Use of training and test data provides an unbiased examination of how the model will predict new observations (Streyerberg et al. 2001). Additionally both cross-validation and bootstrapping provide better estimates of the expected prediction error of a model. However, when the original data set is small or the number of “responders” in the data set is small, both bootstrapping and cross-validation may select training or test sets with an insufficient number of responders to determine prediction performance. Additionally, in small data sets it may be desirable to ensure that categorical predictors are also well represented in both training and test data sets.

In this paper we present a strategy for evaluating the prediction capability of a set of novel biomarkers using different statistical models in data with different types of predictors and small sample size. We apply this strategy to compare the predictive performance of several different statistical modeling techniques, including logistic regression, artificial neural networks, support vector machines, and random forest, to evaluate capability of a set of traditional biomarkers collected at presentation to predict one-year response to therapy in patients with lupus nephritis as a prototypic example of a complex disease with heterogeneous presentations and outcomes.

**Methods**

*Prediction model evaluation strategy:*

The motivating data on 140 subjects includes information on treatment response in patients with lupus nephritis (responder vs. non-responder) and \*\*\* induction therapy type (mycophenolate mofetil, azathioprine, cyclophosphamide, abatacept, or rituximab), 7 biomarkers commonly used in a clinical setting, and 18 novel low abundance urine biomarker. The goal of the study was to evaluate if multivariable modeling would improve the predictive performance of these biomarkers over use of these biomarkers in isolation without modeling. The full data set included 37 responders and 103 non-responders. Given that the number of responders was small, it was important to ensure that a sufficient number of responders were included in both training and test sets to provide sufficient information about prediction performance. Thus the following strategy was employed.

1. Split the data into *m* different training and test sets by randomly selecting 2/3 of the observations for the training set and 1/3 for the test set while keeping the ratio of responders and non-responders constant in the test and training sets.

2. Construct a prediction model using each of the *m* training data sets

3. Evaluate the performance of the prediction models constructed in 2 based on the original training data

4. Evaluate the performance of the prediction models constructed in 2

The primary outcome of interest was treatment response at one year as described by Wofsy et al. from the outcome used the Genentech lupus nephritis and Rituxan (LUNAR) trial ([11](#_ENREF_11)). Univariate associations between demographic factors, therapeutic agent usage, and all biomarkers was initially evaluated using simple logistic regression models. For classification analysis based on multivariable models, logistic regression, artificial neural networks (ANNs), support vector machines (SVMs), and random forest models were considered. Variables in the logistic regression models were selected using stepwise variable selection conducted by choosing the model with the smallest Akaike Information Criterion (AIC). The variable selection procedure and all models were run using the *stepAIC* function in the MASS package in R. Parameters in the random forest models, including the number of variables considered at each step of the algorithm and maximum tree size, were selected using the train function in the caret package in R. All models were developed using the training sets selected as described above and then evaluated for predictive performance on both the training and test sets. The value of *m* was set at 1000 to provide a distribution of different performance characteristics for all models that were considered. Predictive performance of all univariate and multivariable models was assessed using model the area under the curve (AUC) for the receiver operating characteristics curves, accuracy, sensitivity, and specificity. All analyses were conducted in R v 3.1.1.

*Patient populations*

All research was in compliance with the *Helsinki Declaration* and was approved by the Medical University of South Carolina (MUSC) Institutional Review Board. Patients were recruited from four different lupus nephritis populations. The Medical University of South Carolina (MUSC) Division of Rheumatology and Immunology Lupus Erythematosus clinical research group (MUSCLE) assisted in recruitment of participants in Charleston, SC. Participants were also recruited from the Hopkins Lupus Cohort, a well-established and rigorously described cohort begun in 1987 by Dr. Michelle Petri ([7](#_ENREF_7)). A third patient population came from those in the Genentech lupus nephritis study of active class III or IV lupus nephritis entitled “A Study to Evaluate the Efficacy and Safety of Rituximab in Subjects With ISN/RPS Class III or IV Lupus Nephritis (LUNAR)” (NCT00282347) ([8](#_ENREF_8), [9](#_ENREF_9)). A fourth population came from patients participating in the Bristol-Meyers Squibb (BMS) trial of abatacept in lupus nephritis entitled “Efficacy and Safety Study of Abatacept to Treat Lupus Nephritis (Abatacept in LN)” (NCT00430677) ([10](#_ENREF_10), [11](#_ENREF_11)). All subjects met 1997 criteria for SLE ([12](#_ENREF_12)). All subjects had ISN/RPS active class II, III, IV, or V nephritis by biopsy performed by the treating clinician or had a new increase in 24-hour urine protein of 500 mg or of spot urine protein/creatinine ratio of 0.5 within 2 months of the urine collection. Subjects could not have an active infection or ongoing pregnancy or serum Cr > 2.5. LUNAR and BMS participants had more rigorous exclusion criteria described previously ([9](#_ENREF_9), [10](#_ENREF_10" \o "Furie, 2011 #10057)).

All subjects had cortical renal biopsies performed as part of routine care. Subjects were evaluated with a history and physical exam as well as blood and urine collection at baseline for traditional markers of SLE disease activity and nephritis activity (SLE International Collaborating Clinics (SLICC) renal activity index (([6](#_ENREF_6)))) data elements). Urine was collected by clean catch, centrifuged to remove sediment, and frozen in aliquots at -80°C for batch analysis. If urine was not immediately processed, it was held at 0-4ºC for less than four hours prior to centrifugation and aliquot for freezing. Traditional biomarkers of lupus nephritis, including serum C3, C4, creatinine, anti-dsDNA antibody (positive or negative per the individual laboratory), serum creatinine, and urine for 24 hour protein content or urine protein/creatinine ratio were evaluated at baseline and outcome measures were performed at one year by CLIA certified central laboratories at MUSC, Johns Hopkins, Covance, and Quintiles.

**Results**

*Univariate Predictor Performance*

Performance metrics for the four most predictive factors are shown in Table 2. The primary performance metric of interest was AUC. Area under the curve for the individual clinical and urinary biomarkers on the training data ranged from 0.57 to 0.68 (C4 and urine Pr/Cr respectively), and on the test data ranged from 0.56 to 0.66 (C4 and urine urine Pr/Cr respectively). The most consistent clinical marker across both the test and training data sets was urine protein to creatinine ratio with AUC in training and test sets of 0.65 and 0.66, accuracies of 0.73 and 0.74, sensitivity of 0 and 0, and specificity of 1 and 1 respectively. Sensitivity and specificity were defined based on predictions from the GLMM regression models. For those individuals for whom the estimated probability of response from the GLMM model was >0.5 were predicted as responders. As noted in Table 2, a majority of the factors had sensitivities of 0 in both test and training data and perfect specificities of 1. The occurrence of perfect specificity and zero value for sensitivity indicates that choosing a threshold probability of 0.5 resulted in all observations being predicted to be a zero. The threshold probability can be adjusted to optimize sensitivity and specificity of the model. In the case of these data, adjusting the threshold probability to a lower value resulted in greater sensitivity with some decrease in specificity.

**Table 2:** Mean AUC, accuracy, sensitivity, and specificity for the 4 best clinical markers across 1000 training and test datasets. Values represent the mean (95% bootstrap confidence interval) for each performance metric.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Data Type** | **AUC** | **Accuracy** | **Sensitivity** | **Specificity** |
| age | train | 0.58 (0.51,0.66) | 0.73 (0.68,0.78) | 0 (0,0) | 1 (0.99,1) |
|  | test | 0.57 (0.28,0.71) | 0.74 (0.64,0.83) | 0 (0,0) | 1 (0.97,1) |
| urprcr | train | 0.65 (0.58,0.73) | 0.73 (0.69,0.79) | 0 (0,0) | 1 (1,1) |
|  | test | 0.66 (0.50,0.80) | 0.74 (0.64,0.83) | 0 (0,0) | 1 (1,1) |
| c4c | train | 0.57 (0.50,0.66) | 0.73 (0.69,0.78) | 0 (0,0) | 1 (1,1) |
|  | test | 0.56 (0.30,0.69) | 0.74 (0.64,0.83) | 0 (0,0) | 1 (1,1) |
| egfr | train | 0.62 (0.55,0.68) | 0.73 (0.67,0.78) | 0 (0,0.1) | 0.99 (0.95,1) |
|  | test | 0.62 (0.48,0.76) | 0.72 (0.62,0.83) | 0 (0,0.1) | 0.97 (0.89,1) |

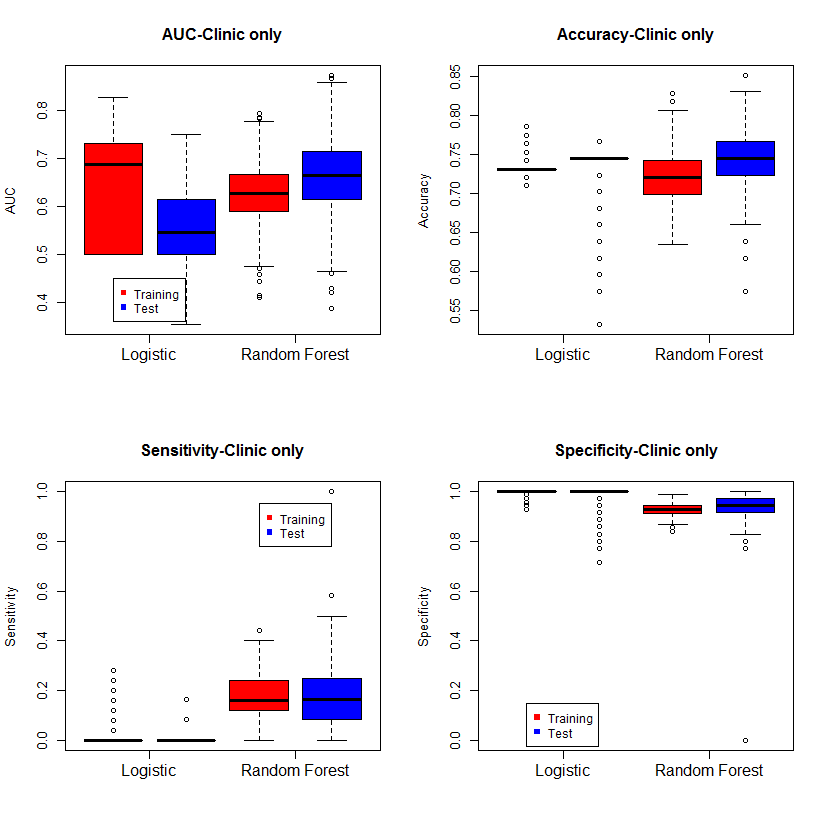
***Multivariable Results****:*

The predictive performance for multivariable logistic regression and random forest models based on (1) traditional clinical and demographic markers is shown in Table 3.

**Table 3:** Mean AUC, accuracy, sensitivity, and specificity for multivariable logistic regression and random forest models that include (1) traditional clinical markers and (2) traditional markers plus novel biomarkers across 1000 training and test datasets. Values represent the mean (95% bootstrap confidence interval) for each performance metric.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Markers** | **Model** | **Metric** | **Training Results** | **Test Results** | **Training-Test** |
| List markers here |  | AUC | 0.72 (0.66, 0.79) | 0.61 (0.45, 0.75) | 0.11 (-0.08, 0.32) |
| **Logistic** | Accuracy | 0.74 (0.70, 0.77) | 0.71 (0.62, 0.77) | 0.03 (-0.06, 0.16) |
| **Regression** | Sensitivity | 0.14 (0.00, 0.36) | 0.08 (0.00, 0.25) | 0.05 (-0.13, 0.27) |
|  | Specificity | 0.96 (0.90, 1.00) | 0.92 (0.77, 1.00) | 0.04 (-0.04, 0.7) |
|  | AUC | 0.63 (0.51, 0.73) | 0.66 (0.51, 0.80) | -0.03 (-0.27, 0.21) |
| **Random** | Accuracy | 0.72 (0.67, 0.78) | 0.74 (0.66, 0.81) | -0.02 (-0.13, 0.09) |
| **Forest** | Sensitivity | 0.16 (0.00, 0.36) | 0.20 (0.00, 0.50) | -0.02 (-0.33, 0.24) |
|  | Specificity | 0.93 (0.88, 0.97) | 0.92 (0.80, 1.00) | -0.00 (-0.10, 0.11) |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**Figure 1**: Performance of logistic regression and random forest models in test and training data in models including only clinical markers.

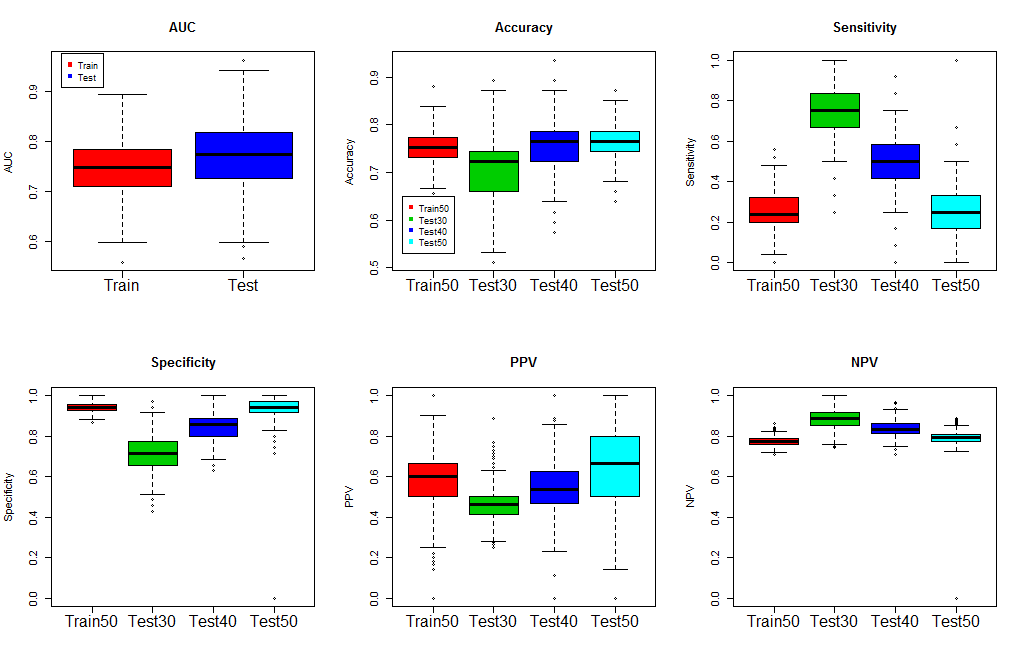


*Random forest vs. Logistic regression*:

Logistic regression showed superior predictive performance when evaluating the training data relative to random forest (Figure 1). However, when evaluated on new test data, the performance of logistic regression showed a marked declined with the mean AUC decreasing from 0.72 in training data to 0.63 in test data. However, the performance of random forest was very similar for both test and training data. The decrease in performance of the logistic regression models on new data is likely a result of over fitting of the model, which is known to be a weakness of logistic regression. However, as has been reported in the literature, random forest models yield an unbiased estimate of predictive performance thus we expect the performance on training data and independent test data to be very similar.

Area under the curve is determined by considering all possible thresholds for a single classifier or values from a model with multiple predictors. In the case of logistic regression, the AUC is developed using the predicted probability of being classified as a responder and considering all predicted probabilities observed in the data as possible cutoffs. In the case of the random forest model the AUC is determined by thresholding the number of trees that predict the subject to be a responder. While AUC provides and overall estimate of the predictive performance of a classification model, it is not useful in clinical practice where a decision must be made about subject treatment based on the prediction model. Therefore the “best” threshold must be selected to allow a clinician to determine course of treatment. The choice of the “best” threshold should provide a desired level of sensitivity and specificity (which of course vary with choice of threshold). The default choice for threshold for predicting disease in logistic regression is 0.50, which indicated the model predicts a 50% or greater chance that the individual has disease. Similarly, in random forest, the default prediction of response is determined by majority vote, that is, at least 50% of the trees in the forest classify the subject as a responder. However, the default choice of majority vote may not provide the optimum sensitivity or specificity. This is clearly evident in Table 4 where each model has >90% specificity but no greater than 51% sensitivity. This suggests that while we do a good job of correctly classifying non-responders, the ability to correctly classify responders is not better than flipping a coin. Thus it is often worthwhile to examine the impact of varying the number of trees required in the model to determine response status. We examined the impact of varying the number of trees necessary to classify a subject as a responder in the random forest model with all clinical and novel urine biomarkers on accuracy, sensitivity, specificity , PPV, and NPV (Figure X).

Figure X: Boxplots of performance metrics with changing cutoff- Random forest models



Decreasing the number of trees required to predict subject response increase sensitivity and decreased specificity. The impact on accuracy was only evident when the threshold was set to 30% of trees to make a prediction of treatment response. Based on the results shown in Figure X, a threshold of 40% of trees to make a prediction of treatment response is a good compromise between increasing sensitivity and decreased specificity. At a threshold of 40%, the accuracy is similar to the accuracy observed for a threshold of 50%, the sensitivity increases from 25% to 45% while maintaining a specificity > 80%.

*Random Forest Variable Importance Measures*

In addition to providing an unbiased estimate of the prediction error, random forest models also provide a quantitative measure of variable importance for all variables considered in the model. The importance of each variable is determined by comparing a measure of predictive performance, the Gini Index, before and after permuting the variable of interest. The variable importance for each variable in the data set was ranked for the 100 random forest models according to the magnitude of their variable importance scores. Boxplots of the importance score ranks across all 1000 training data sets are shown in Figure 4. In this figure, smaller ranks indicate greater importance. Figure 4 suggests \*\*\*

*Random Forest Model on All Data*

Because random forest models yield and unbiased estimate of predictive performance, we repeated the analysis including all the data (i.e. we did not split the data into test and training data). Random forest also provides a measure of variable importance for all variables evaluated in the model. Thus we also wanted to use the model constructed on all of the data to examine the relative importance of all clinical and novel biomarkers. The variable importance plot for the random forest model fit to all the data is shown in Figure 3.

**This plot will be traded out for the distribution of importance scores plots, I may also add plots showing the proportion of times logistic regression included the different markers (as this is the LR version of importance)**

Table 4: Comparison of performance metrics for random forest models with different predictor sets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **AUC** | **Accuracy** | **Sensitivity** | **Specificity** |
| Full model  (30 Variables) | 0.78 | 0.24 | 0.23 | 0.96 |
| 5 Most Important Predictors | 0.77 | 0.21 | 0.43 | 0.92 |
| 13 Most Important Predictors | 0.80 | 0.20 | 0.35 | 0.95 |
| Clinical Variables + 4 Most Important Biomarkers | 0.79 | 0.19 | 0.51 | 0.91 |

**Discussion**

The main purpose of this paper was to present a strategy for evaluating the predictive capability of a set of clinically available variables in data with a small sample and a limited number of responders. Using a clinical dataset, we demonstrate how this strategy can be used to compare predictive performance of different statistical methods, evaluate prediction performance of the set of markers on new data, and determine potential candidate markers for further evaluation.

The comparison of logistic regression and random forest models also demonstrated differences in prediction performance that must be considered when selecting the appropriate statistical procedure for evaluating a set of novel biomakers to improve the predictive power of the model. In training data, the AUCs for the traditional multivariable logistic regression models are larger than random forest models. However, the AUCs for the test data show a larger decrease relative to training data for the logistic regression models while AUCs for random forest models are similar for test and training data. In test data, the AUCs for random forest models were larger relative to logistic regression models suggesting that random forest is less prone to over fitting and demonstrates less bias relative to logistic regression, which is consistent with previous studies of the prediction performance in random forest (ref). For treatment response in patient with lupus nephritis, the results of this “pilot” study suggest that use of random forest modeling does have the potential to improve predictive performance of a model for predicting response to therapy in patients with lupus nephritis.

Random forest has many desirable features for evaluating novel biomarkers for prediction of disease outcomes. Random forest is more robust to noise in the data and irrelevant predictors than logistic regression (ref?). Additionally, random forest also provides a quantitative measure of variable importance which can be useful in determining which particular biomarkers from a larger set of biomarkers is most predictive of the outcome and an unbiased internal measure of prediction accuracy, referred to as out of bag error, and thus as demonstrated in this study, can be used to model the full dataset.

Adjusting the model parameters to fit the purpose of predictive modeling is essential. In a clinical setting, thresholds can be adjusted to maximize the desired clinical utility of the model. If the desire is to identify as many non-responders as possible to increase the aggressiveness of therapy at presentation, then losing sensitivity is less of a problem. However, if one does not want to treat as many responders unnecessarily in the case where increased aggressiveness of therapy also carries with it increased toxicity, then setting a threshold with greater sensitivity of responders is desired. If multiple treatments with varying risk to benefit ratios are available to clinicians, then presentation of different thresholds may be desirable. Clearly, the predictive power of the presented random forest models using traditional biomarkers and demographic features are not optimal for clinical decision-making but are superior to univariate modeling or logistic regression modeling. The strategy for developing models based on limited datasets presented here holds promise as a method for development of novel biomarker models of clinical outcomes. The goal of this development would be to produce more predictive biomarker models that could be used in the clinic or at the bedside as decision support tools to improve patient outcomes.

1a Lawrence RC, Helmick CG, Arnett FC, Deyo RA, Felson DT, Giannini EH, Heyse SP, Hirsch R, Hochberg MC, Hunder GG, Liang MH, Pillemer SR, Steen VD, Wolfe F. Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. Arthritis Rheum. 1998;41(5):778

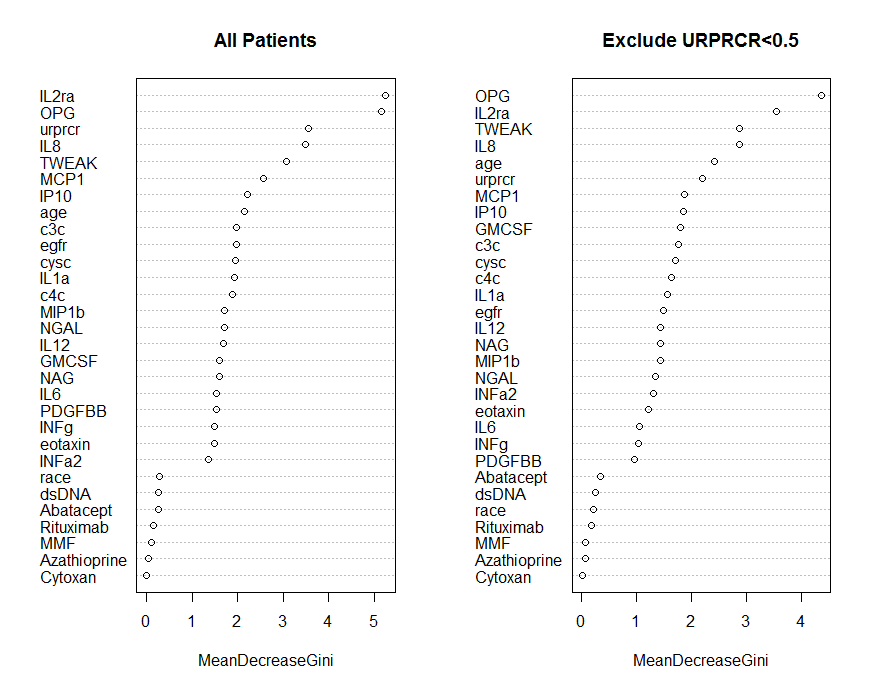
1b Pons-Estel GJ, Alarcón GS, Scofield L, Reinlib L, Cooper GS. Understanding the epidemiology and progression of systemic lupus erythematosus. Semin Arthritis Rheum. 2010;39(4):257.

**This will go in the clinical paper but not the stats paper**

**Additional analyses we discussed (note I only did this for the RF model on the full data set):**

(1) Leaving out those subjects with urprcr < 0.5

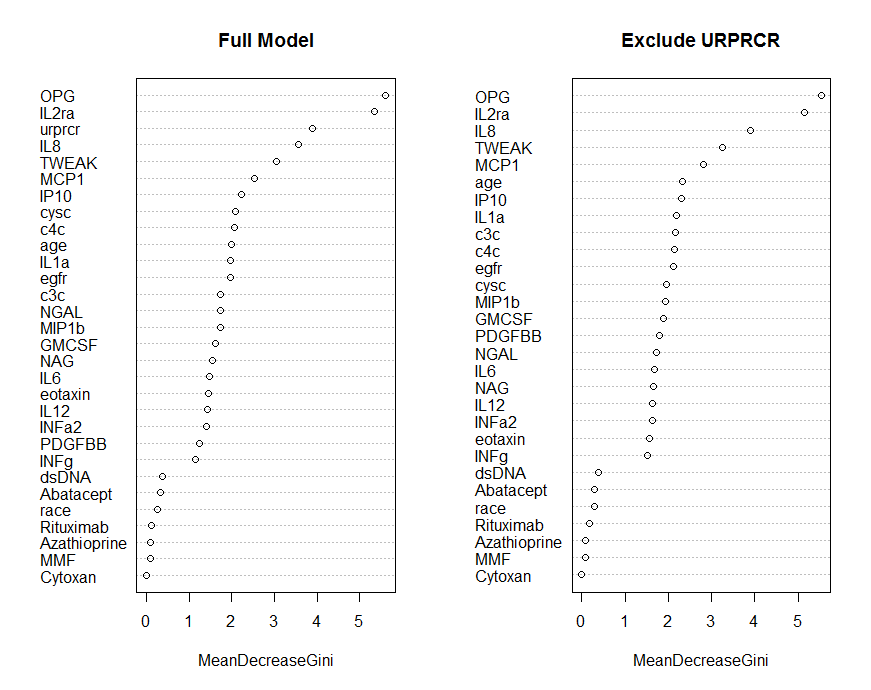
We reran the random forest model excluding those patients for whom urprcr was less than 0.5 at the baseline visit. This resulted in the exclusion of 14 patients, 6 of whom were responders. Exclusion of the 14 patients with urpcr < 0.5, the estimated accuracy is 79% and the AUC is 0.74 compared to an accuracy of 75% and AUC of 0.77 in the random forest model on the full data. The variable importance plots for the “full” model including the 14 patients and the “reduced” model excluding the 14 patients are shown below.



The ranking of the variable importance is very similar for the two models. Both models include IL2ra, OPG, IL8, and TWEAK as some of the most important markers. The rank of GMCSF increases when these 14 patients are excluded from the data (though it is still among a large group of variables with similar importance values- so perhaps this is not so surprising or relevant).

(2) Comparing RF models with and without URPRCR included in the model

We reran the random forest model excluding urprcr from the model. As mentioned previously, the full model (with urprcr included as a variable) had an accuracy of 75% and AUC of 0.77. the model excluding urprcr in the analysis had an accuracy of 78% and an AUC of 0.76. The variable importance plots for the random forest models with and without urprcr are shown below.



The order of the variables by their importance scores changes very little when urprcr is excluded from the model. The biomarkers OPG, IL2ra, IL8, TWEAK, and MCP1 are still have the largest variable importance score.

(3) Table of performance metrics with a threshold of 50% of the trees versus. 40% of the trees

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance Metric | 50% Threshold | | 40% Threshold | |
| Training | Test | Training | Test |
| Accuracy | 0.75 (0.70, 0.81) | 0.77 (0.70, 0.83) | 0.74 (0.67, 0.82) | 0.76 (0.66, 0.85) |
| Sensitivity | 0.19 (0.04, 0.36) | 0.19 (0.00, 0.42) | 0.42 (0.24, 0.60) | 0.44 (0.17, 0.75) |
| Specificity | 0.96 (0.91, 1.00) | 0.97 (0.89, 1.00) | 0.86 (0.81, 0.93) | 0.87 (0.74, 0.97) |

We can see that we gain a great deal of sensitivity by dropping the number of trees that classify the subject as a responder to 40% from 50% without losing a great deal of specificity or accuracy.

**Adding these to the stats paper**

Figure X: Distribution of importance score rankings for all variables in the RF model with all markers. Note variables are ordered by mean rank with lower ranks implying greater importance

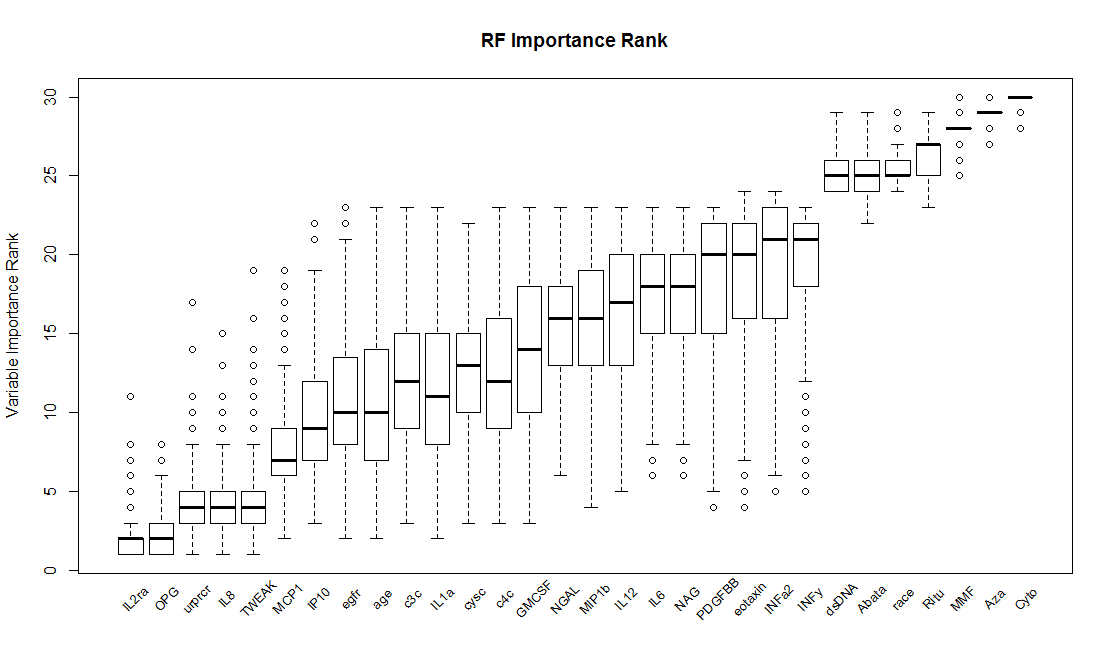
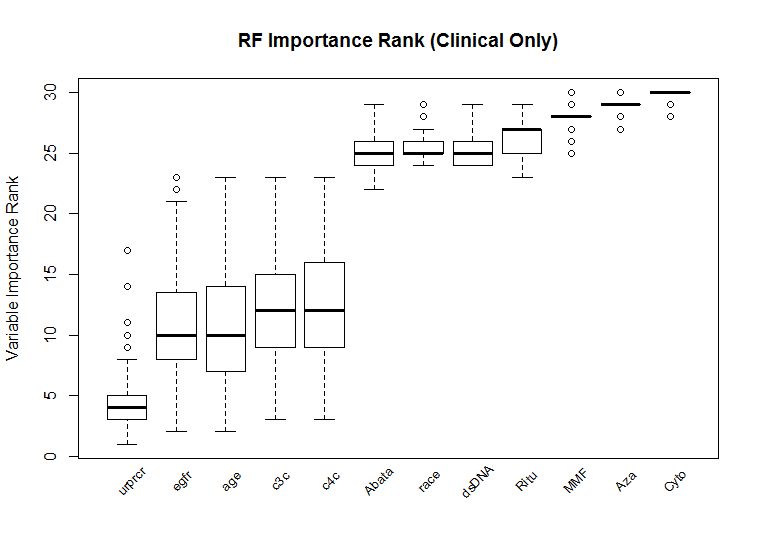


Figure Y: Distribution of importance score rankings for all variables in the RF model with only clinical markers. Note variables are ordered by mean rank with lower ranks implying greater importance



Discovery pipelines suggested in the literature (8, 73, 79) all include discovery, qualification (i.e. prioritization), verification, and validation phases. The Food and Drug Administration’s Critical Path Initiative emphasizes the importance of biomarker qualification as part of the discovery process (94, 101, woodcock and Woolsey?). The goal of qualification phase is to assess the association of hundreds or tens of candidate markers identified during the discovery phase with the biological process or clinical endpoint of interest (97). The sample size in the qualification phase is typically larger than the number of potential biomarkers but still insufficient to properly validate the candidate markers. It is designed to further reduce the number of candidate markers to be evaluated during the validation and optimization phases by selecting only the most promising markers of disease outcome.

Although the qualification stage of biomarker discovery pipeline is extremely important, there are many potential issues.