Statistical Strategies for Developing Classification Algorithms with Application to Insulin Sensitivity Status

Wenting Xie[1], William D Johnson[2], Charmaine S Tam[2], Bin Li[3]
[1]Wenting Xie's Affiliation, Pennington Biomedical Research Center, 6400 Perkins Road, Baton Rouge, LA 70808-4124 USA
[2]William D Johnson's Affiliation, Pennington Biomedical Research Center, 6400 Perkins Road, Baton Rouge, LA 70808-4124 USA
[2]Charmaine S Tam's Affiliation, The Charles Perkins Centre and School of Biological Sciences, University of Sydney, NSW 2006 Australia
[3]Bin Li's Affiliation, Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803-5606 USA

Abstract:
Insulin resistance is a strong precursor to the development of the metabolic syndrome and type 2 diabetes. The hyperinsulinemic-euglycemic clamp, the gold standard for assessing insulin resistance in humans, is labor-intensive and expensive and thus examining surrogate markers for insulin resistance is necessary. In this paper, we incorporated the newer statistical algorithms to boost accuracy of insulin prediction. Data including subject characteristics (age, ethnicity, sex), body composition (BMI) and blood biochemistry (glucose, insulin) were obtained from 270 individuals participating in research studies at the Pennington Biomedical Research Center in Louisiana between 2001 and 2011. Using these data, we applied and compared four statistical methods to predict insulin resistance including classical logistic regression, and the newer methods of single classification tree, boosted regression tree (BRT) and random forest (RF) as well as a novel approach of combining logistic regression and featured selection from BRT or RF. Random forest (AUC=0.858) and boosted regression tree (AUC=0.845) gave the best prediction performance for predicting insulin resistance. This was followed by logistic regression method combined with feature selection technique from BRT or RF (AUC=0.763) and finally single classification tree (AUC=0.741). However, when using variables without a large portion of missing values we found that logistic regression (AUC=0.84) gave the best prediction performance. The result shows that boosted regression tree and random forest approaches may provide better algorithms where missing data may be an issue. We also found an appropriate combination of traditional logistic regression and variable selection from BRT or RF may improve model performance. Logistic regression is still appropriate when missing data may not be a factor. In conclusion, we have illustrated the exploration of different statistical models when determining prediction performance in biomedical studies.

Keywords: Boosted regression tree; Random Forest; Tree based methods; Logistical regression; Insulin Sensitivity Status; metabolic markers;

1. Introduction

Type 2 diabetes, a metabolic condition characterized by high blood glucose affects 11.3% of adults over the age of 20 with prediabetes or insulin resistance affecting 35% of Americans [1]. Alarmingly, 27% of patients with diabetes are not aware that they have the disease. Insulin resistance is a strong precursor to the development of the metabolic syndrome and type 2 diabetes. Developed by De Fronzo and colleagues [2], the hyperinsulinemic-euglycemic clamp technique is the gold standard for assessing insulin resistance in humans and is typically used to test the effect of interventions (weight loss, weight gain or pharmacological treatment) on insulin sensitivity. However, the clamp technique is time-consuming (varying from 2-8 hours), burdensome for research participants, labor-intensive requiring several highly-trained personnel and expensive. As such, there is a strong need to develop accurate statistical models to predict insulin resistance from other clinical biomarkers that are both easier to obtain and less expensive to measure. Such measures include subject demographics, blood chemistry and body composition. Most importantly, prediction of a subject's insulin sensitivity status may assist clinicians in earlier detection of patients who are at a special risk of developing type 2 diabetes.

The aim of this paper was to apply and compare conventional and modern statistical modeling techniques to predict insulin resistance in humans. We applied three newer statistical methods including single classification tree, random forest (RF) and boosted regression tree (BRT), to predict subjects' insulin resistance status, and compared these models with the traditional logistic regression approach. In addition, we tested a novel approach of combining logistical regression and variable ranking feature from BRT or RF and found it was superior to traditional logistic regression based on stepwise selection. We also performed further model comparisons after excluding one variable with significant missing data and showed that when missing data were not an issue, the best performance was still realized using the traditional logistic regression.

2. Methods and Theory

2.1 Data Source and Study Population

We compiled data from individuals involved in research studies at the Pennington Biomedical Research Center, Baton Rouge, Louisiana between 2001 and 2011 (n=270). Input predictors included subject characteristics [sex, ethnicity, age], body composition [BMI] and serum metabolic markers [fasting insulin and glucose, HbA1C (only available for 166 out of 270 participants)]. BMI was calculated as weight (kg)/height $(m)^2$. Subjects were defined as 'Insulin Resistant' or 'Insulin Sensitive' based on self-reported diabetes status or a fasting insulin level≥15uU/ml. Based on these criteria, 147 people out of 270 were classified as being insulin resistant and 123 subjects were classified as being insulin sensitive.

2.2 Logistic Regression

Let G denote the dependent random variable where G=1 if a given subject is insulin resistant and G=0 if the subject is insulin sensitive. Further, let x_i represent the predictor vector for the i^{th} subject where i =1, 2, 3…, n. A logistic regression model can be expressed as shown in formula (1) where the log odds of G being "1" equals to the linear combination of weighted x_i where the coefficient vector β contains the weights and α is the intercept. Based on the maximum likelihood method, we can get the parameter

estimates of α and β by solving the conditional likelihood equation (2). Finally, the estimated probability of outcome "1" for any subject can be calculated by equations (3) and (4) with estimated α and β.

$$\text{Log}\left(\frac{\Pr(G=1 \mid X=x_i)}{\Pr(G=0 \mid X=x_i)}\right) = \alpha + \beta x_i \quad (1)$$

(α is the intercept, β is coefficient vectors, x_i is *i*'s x input vector);

$$\boldsymbol{L}(\alpha,\beta) = \sum_{i=1}^{n} \log \Pr(G = g_i \mid X = x_i) = \sum_{i=1}^{n} \log Pg_i(x_i;\alpha,\beta) \quad (2)$$

$$P(x;\ \alpha,\beta) = \Pr(G=1|X=x) = \frac{\exp(\alpha+\beta^T x)}{1+\exp(\alpha+\beta^T x)} \quad (3)$$

$$1- P(x;\ \alpha,\beta) = \Pr(G=0|X=x) = \frac{1}{1+\exp(\alpha+\beta^T x)} \quad (4)$$

2.3 Tree –based Classification

Tree-based classification methods partition the feature space into separate hyper rectangular regions (i.e., the regions to which observations belong) [4]. Next, a constant (in regression problems) or a class label (in classification problems) is fitted to each region guided by a set of decision rules developed during the fitting process. Figure 1 illustrates the structure of a recursive binary tree model. With three splitting variables X1~X3 and three splitting points t1~t3, four terminal nodes Y1~Y4 are produced. For instance, an observation with $X1 \leq t1$ and $X2 \leq t2$ will be assigned to region "Y1" based on this decision tree. Splitting variables and splitting points which develop the decision rule are selected to minimize prediction errors. A single tree can be further pruned based on cost-complexity criteria that balance the trade-off between tree complexity and goodness of fit.

2.4 Boosted Regression Trees

The application of boosting in machine learning comes from an idea that searching and combining many moderately inaccurate rules is better than producing a single, highly accurate prediction rule [4]. Discrete AdaBoost is one of the most popular boosting procedures for classification problems. Suppose the training sample consists of n observations (X1, Y1), (X2, Y2), … , (Xn, Yn) with Xi as an input vector and Yi as a binary response taking value either -1 or +1. Then the final prediction is defined as the sign of $F(x) = \sum_1^M c_m f_m(x)$ where every fm(x) is a classifier and c_m is the constant constraint for that classifier [5]. At $(m-1)^{th}$ fitting, AdaBoost attempts to put more weight on samples that are poorly predicted by current classifiers and produce a reweighted version of the sample for m^{th} iteration. The boosting technique applied in our study is called "gradient boosting". During each iteration, gradient boosting sequentially fits a parameterized function (base learner) to the current residuals which are the gradient of the loss function being minimized and evaluated over all training data in current step [6]. After each iteration, the newly fitted function is added to the model and the residual is recalculated by the updated model for preparing next iteration. Finally, an addition-formed model including all fitted functions (based learners) is constructed. In case of boosted regression trees, each base learner can be viewed as a regression tree where splitting variables and splitting points are parameters. Different from AdaBoost in which reweighted versions of sample are computed to fit the base learner, gradient boosting injects extra randomness into this procedure by randomly selecting a subsample (without replacement) to fit the base learner and compute the model update for the current

iteration, instead of using the whole training data in each iteration. Results from some examples showed that bringing in randomness can improve the model performance [7]. In sum, the method of boosted regression trees is a numeric optimization technique that aims to minimize the loss function by increasing tree classifiers.

The "learning rate" is introduced as a tuning parameter in BRT. A small learning rate is often preferred within the budget of computation time. The interaction effects, accounted by the BRT model through tree depth, also influences the fitting process and should reflect the true interaction levels among predictors in theory. In summary, the learning rate and interaction factor determine the number and complexity of trees in BRT.

2.5 Random Forest

Besides boosting, bagging is another technique incorporated in "ensemble trees" methods by growing trees on independently bootstrap samples. Random Forests proposed by Breiman [8] can be viewed as an improved version of bagging in which additional randomness is brought in by selecting random subsets of predictors. From step 1 to B (total number of trees), a tree is grown from a bootstrap sample of the training data. The tree is fitted so that at each node, the best split variable and splitting point are picked from m randomly selected variables out of total p ($p \geq m$) predictors. The final prediction is a combination of trees: $\{Tb\}_1^B$ (average for regression or majority votes for classification). Moreover, the prediction performance of random forest is very insensitive to the two tuning parameters including the size of the subset predictors and the number of trees. Similar to bagging and other "ensemble methods", random forest improves prediction by reducing variance.

Before RF, missing values in predictors can be replaced with imputed values using proximity. A proximity matrix indicates the similarity among pairs of observations in terms of fractions of trees that two observations fall in the same terminal node [9]. In other words, the more terminal nodes two observations fall in at the same time during the fitting process, the more similar they are and the larger proximity value they will get in proximity matrix. For continuous variables, missing values are replaced in the imputation by the weighted average of non-missing value where the weights were proximities corresponding to the similarity between missing observation and each non-missing observation. For categorical variables, proximities among the missing and the non-missing observations were calculated and averaged. After that, the category with group of observations having the largest average proximity to the missing observation is assigned [9].

2.6 Statistical Methods

To compare prediction performance by assessing internal validation across the four statistical models, the sample was separated into a randomly selected training set of 202 cases and a testing set of 68 cases. Logistic regression, single classification tree, boosted regression tree and random forest were applied and compared. The area under the curve (AUC), sensitivity, specificity, and misclassification percentages were calculated for the testing set and used to evaluate model performance. The cutoff probability which represented the predictive probability of being insulin resistant was arbitrarily chosen to be 0.5 for calculating corresponding sensitivity, specificity and misclassification percentage. In other words, we classified a subject to be insulin resistant if its predictive probability of being insulin resistant based on any model was 0.5 or larger. Different cutoff points may lead to different sensitivity, specificity and misclassification

percentage, so we mainly ranked model performance by AUC scores that were independent of the cutoff probability.

Models were fitted in R version 2.11.1. GLM and CART packages were used to fit logistic regression models and to invoke single classification tree methods respectively. GBM [10] and random forest packages were used to fit BRT and RF respectively.

3. Results and Discussion

3.1 Data Description

Demographics for insulin resistant and insulin sensitive subjects are summarized in Table 1. As expected, subjects classified as insulin resistant were significantly older (P<0.0001), and had a higher BMI (P<0.0001), higher fasting glucose (P<0.0001) and higher HbA1C level (P<0.0001). Sex and ethnic differences between the two groups were not statistically significant.

3.2 Model Performance

3.2.1 Logistic Regression

First, the result from the logistic regression model showed that HbA1C (p-value=0.05) with estimated coefficient 1.99 was the only significant predictor. Thus, a dimension reduced logistic regression with one predictor HbA1C was fitted and resulted in an AUC score of 0.684, indicating poor prediction performance.

3.2.2 Single Classification Tree

A full tree was performed in the training set and variables including fasting-glucose, age and BMI were selected by the final tree model. As a result, we reached an AUC score of 0.741, a sensitivity of 0.784, a specificity of 0.516 and a misclassification error percentage of 0.338, which did not show very balanced sensitivity and specificity.

3.2.3 Boosted Regression Trees

A boosted regression tree model with learning rate 0.001 and a bag fraction 0.75 was fitted in our training data. *Bernoulli* was chosen as the fitting distribution. A BRT with all variables produced an AUC score of 0.845, a misclassification error rate of 0.25, a sensitivity of 0.784 and a specificity of 0.710. We also investigated the importance of each predictor by measuring its relative influence; a concept developed Friedman [4], to select important variables in prediction. Relative influence of a single variable x_j calculates the total improvement on reducing prediction error by splitting on variable x_j over all trees$\{\mathrm{Tm}\}_1^M$, averaged by total number of trees [4]. Table 2 showed relative influence of all predictors in BRT fitting. The relative influence of each predictor was scaled to have a sum of 100%. BMI with relative influence 33.88%, fasting glucose with 30.73 %, HbA1C with 28.97% and Age with 5.82% were the top four most important variables. In contrast ethnicity and sex had very little influence (< 0.5%). A reduced BRT model with those four most important predictors resulted in an AUC score of 0.845, a sensitivity of 0.757, a specificity of 0.710 and a misclassification error portion of 0.265, and thus a reduced BRT model performed as well as full BRT model.

In addition, we can interpret the dependence relationship between response and predictors by partial dependence in BRT, which measures the marginal effect of a variable after adjusting average effects from all other variables in model [4]. Partial dependent plots in Figure 2 indicates that people had increased probability of being more insulin resistant in

the following conditions: 1) being overweight (BMI≥25); 2) high fasting glucose level(fasting glucose ≥100); 3)HbA1C≥5.3; 4) and being older than after 53 years. The plots were not very smoothing due to the use of tree-based methods.

3.2.4 Random Forest

A random forest model with all predictors was conducted. The sensitivity, specificity, AUC and misclassification error were 0.757, 0.742, 0.826 and 0.25. In addition, we ranked the variable importance by measuring the mean decrease in node impurities over all trees when splitting on that variable. Gini index was used as a measure of node impurity with a higher Gini index suggesting larger node impurities and a lower Gini index indicating lower node impurities. After ranking variable importance, four most essential variables were selected as HbA1C, fasting glucose, BMI and age (see Figure 3). In addition, a RF model with these four variables were performed and resulted in a sensitivity of 0.730, a specificity of 0.806, a AUC score of 0.858 and a misclassification error of 0.235.There was slight difference in the variable importance rank between RF and BRT. Nevertheless, BMI, fasting glucose, HbA1C and age were the four most significant variables in both models.

3.2.5 Logistic Regression Combined with Feature Selection

In addition, a novel approach that applying feature selection from RF or BRT to logistic regression was inspired. Thus, a logistic regression model with those four variables selected from BRT and RF was fit and obtained a higher AUC at 0.763 compared to logistical regression based on stepwise selection (AUC = 0.684).

3.2.6 Model Comparison

In summary, the overall prediction performance achieved by RF and BRT were better than the single classification tree and logistic regression when HbA1C was considered as candidate predictor. Table 3 summarized the prediction performance among four final models ranked by AUC. Fasting glucose, HbA1C, BMI and age were predictors in these final models except for the single tree model that deselected HbA1C. RF showed the highest AUC, relatively balanced sensitivity and specificity could be chosen as the final model in this case. In addition, the logistic regression model in table 3 was based on the novel approach that combined variable ranking feature from RF or BRT.

3.3 Models without HbA1C

Considering the large portion of missing value in HbA1C, models without variable HbA1C were also conducted and similar approaches as above were performed. As a result, these four methods had similar sample size at this condition. Results from logistic regression showed that fasting glucose, BMI and age has significant p value (p-v <0.05). Furthermore, variable importance rank from BRT (table 4) and RF (figure 4) both confirmed that those three were relatively significant predictors and thus were used in prediction. Table 5 summarized the prediction performance across the four models with those three predictors. Logistic regression model rather than BRT or RF had highest AUC at 0.84, lowest misclassification error at 0.269, sensitivity at 0.73 and specificity at 0.733, and could be chosen as the best model. However BRT and RF still performed better than the single classification tree. Thus, we can conclude that logistic regression outperformed other three models when HbA1C was excluded which resulted in more available data for building logistical regression model. Our results highlight that cautious model comparisons need be considered when choosing the appropriate prediction model for difference situation.

4. Conclusion

In sum, we conclude that advanced modeling methods like RF and BRT may produce higher prediction accuracy for data with missing values or large dimension, compared to traditional methods like logistic regression. Possible reasons may include: 1) logistic regression cannot incorporate missing values and thus leads to reduced information and sample size during analysis 2) Machine learning methods equipped with techniques like boosting and bagging have unique algorithms that give better predictive accuracy in cases where missing data are an issue. In addition, advanced machine learning methods also provide useful tools to select significant predictors and often lead to a better understanding of the underlying mechanisms that enhance classification. And we demonstrate that variable ranking feature from BRT and RF can be combined with traditional logistic regression to boost model performance. Incorporating powerful techniques into traditional methods has been investigated and lead to development of many newer statistical algorithms in recent decades. For instance, BRT and RF are applications of boosting and bagging technique in tree-based methods. Lasso regression and ridge regression are modified linear regression methods. But newer statistical algorithms have more complex modeling algorithm, require more sophisticated programming and are sometimes more difficult to interpret. Therefore, they are in a slow progress of being accepted by people who get used to working with easy-accessed and easy-interpreted traditional modeling. Moreover, traditional modeling methods have competitive performance in data without missing values or complex structure. As such, the novel approach in our example applying features like variable ranking from newer statistical methods into logistical regression methods enables us to keep the simplicity and interpretability of traditional models as well as making up for its incapability of handling missing values and complex data, and results in better model performance. Overall, our results demonstrate that statistical modeling can help us find alternative methods for disease diagnostic at cheaper expense, improve diagnosis accuracy, and better understand the mechanism and risk factors of disease. Furthermore, our results set up examples of newer statistical algorithm application in other disease diagnosis or preventive medicine research. We know that evaluating model performance using an independent data is more appropriate for assessing external validity of any statistical model. However, limited to the sample we have, we can not conduct an external validation but an internal validation through splitting the same sample. Thus, further study of a diverse population can provide a more representative sample to investigate and assess model performance. Above all, we encourage the application of newer statistical methods and blending features from newer algorithms with traditional methods.

Reference:

[1] http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf (accessed July 1, 2012)
[2] DeFronzo, R. A, Tobin, J. D, and Andres R. 1979. Glucose clamp technique: a method for quantifying insulin secretion and resistance. *Am J Physiol*. 237:E214-E223.
[3] McCullagh, P, Nelder, and A.J. 1989. Generalized Linear Models. Second Edition. Florida: Chapman and Hall/CRC.
[4] Hastie, T., Tibshirani, R., and Friedman J. 2009. The elements of statistical learning. second edition. New York: Springer.
[5] Friedman, J. H., Hastie, T., and Tibshirani, R. 2000. Additive logistic regression: a

statistical view of boosting. *Annals of Statistics*. 28: 337-407.
[6] Friedman, J. H. 2002. Stochastic Gradient Boosting. *Computational Statistics and Data Analysis*. 38: 367-378.
[7] Friedman, J. H. 2001. Greedy Function Approximation: A Gradient Boosting *Machine*. *Annals of Statistics*.29: 1189-1232.
[8]http://www.springerlink.com/content/u0p06167n6173512/fulltext.pdf(accessed December 1, 2011).
[9] http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf (accessed December 1, 2011)
[10]http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf(accessed December1, 2011)

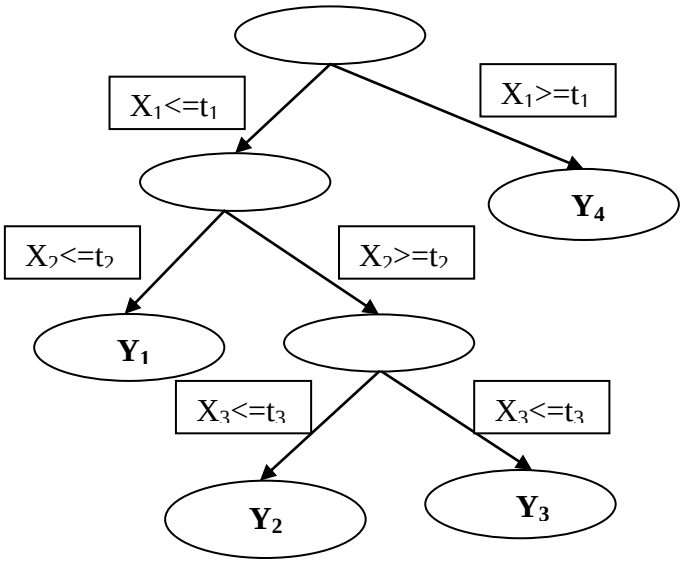

Figure 1: A single decision tree: X_1, X_2 and X_3 are three input variables corresponding to splitting points t_1, t_2 and t_3. Four terminal nodes were Y_1~Y_4

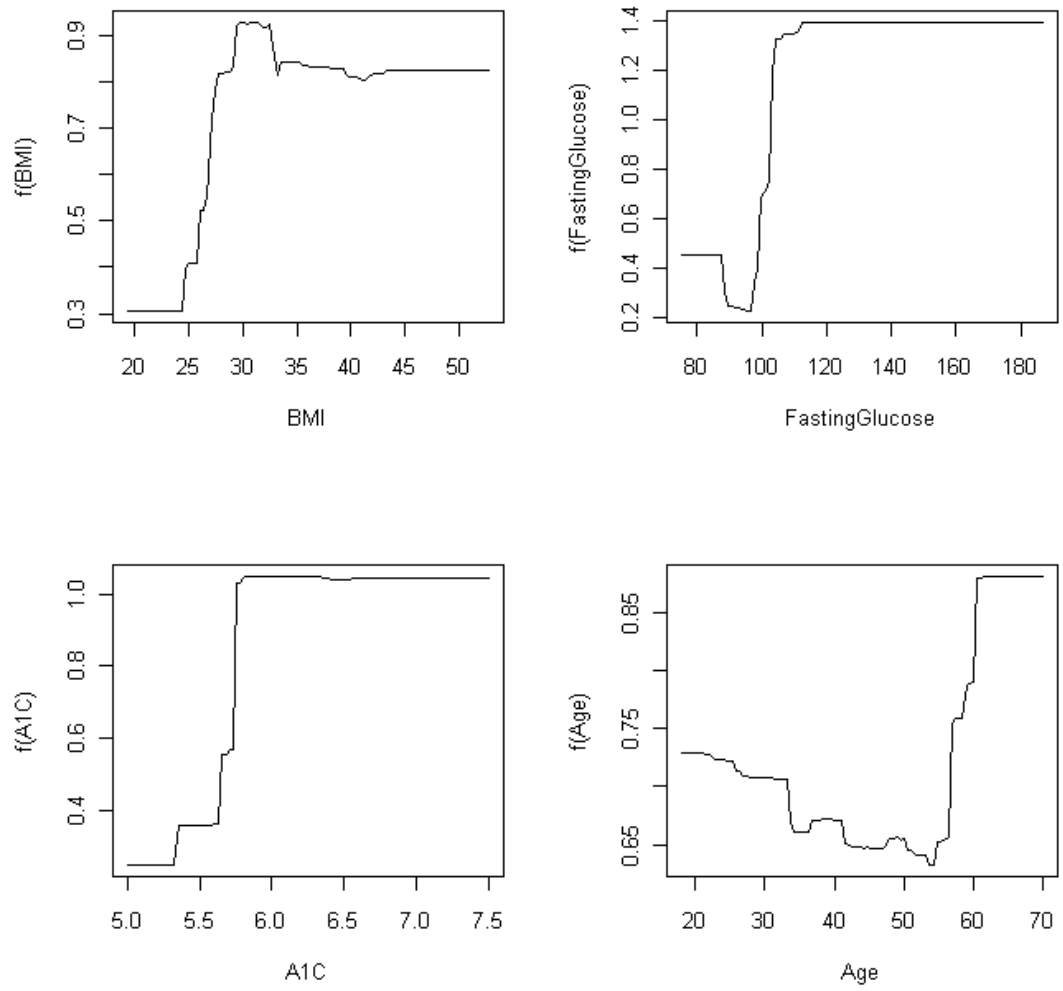

Figure 2: Partial dependence plots of logit(p) scale of being insulin resistant on top four important predictors.

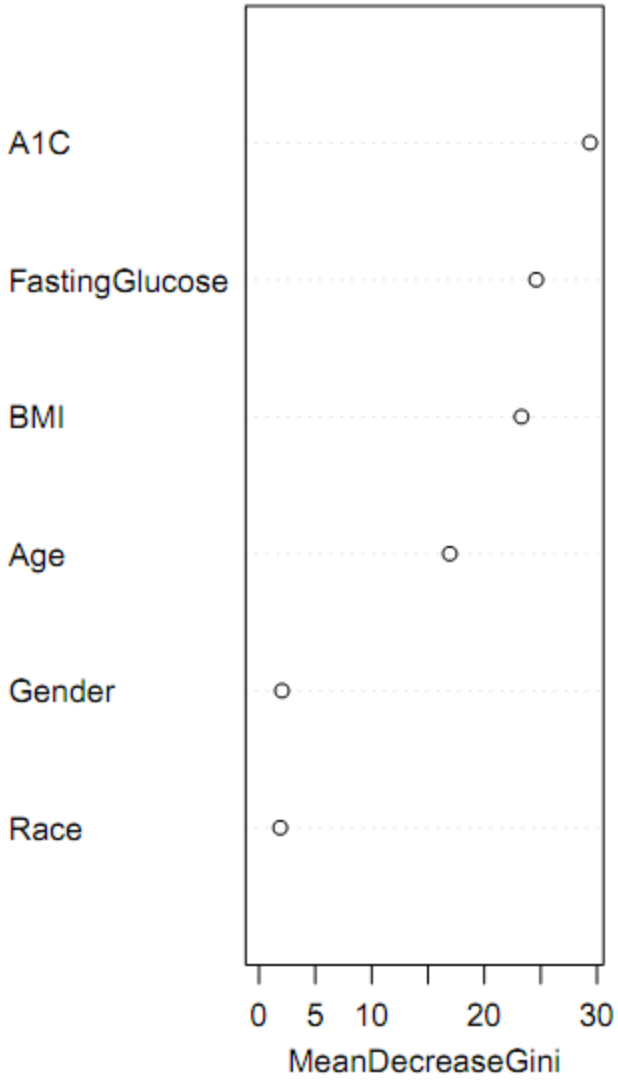

Figure 3: Variable Importance for Predicting Insulin Sensitivity Status Predictors by Random Forest

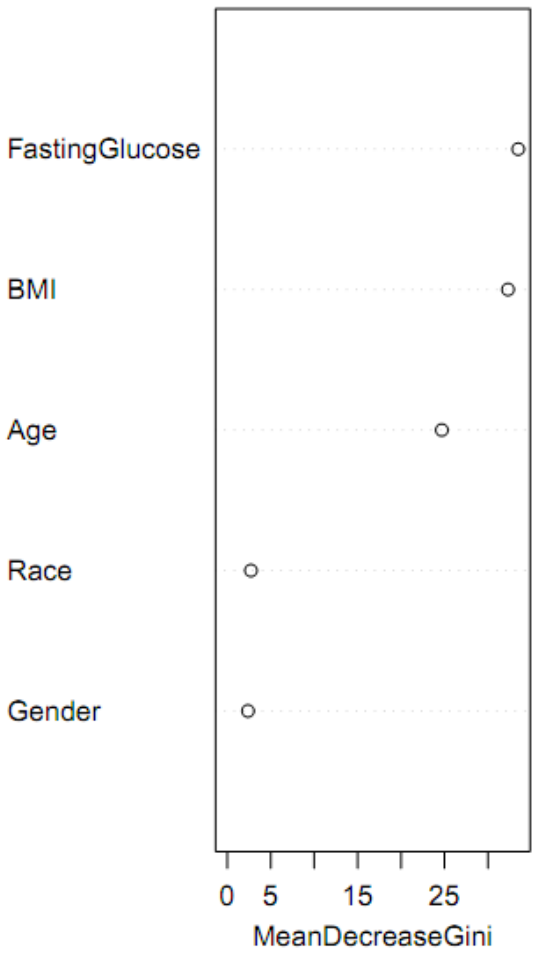

Figure 4: Variable Importance for Predicting Insulin Sensitivity Status Predictors by Random Forest when HbA1C is excluded

Table 1: Characteristics for 270 participants in this cohort study

	Subjects classified as Insulin Sensitive n=123	*Subjects with Insulin Resistance N=147*	*P-value*
Sex	78 Females, 45 Males	88 Females,59Males	0.5505[a]
Race	49 Black, 74 White	52 Black, 95 White	0.4504[a]
Age	41.6 ± 12.5	49.0 ± 13.4	<0.0001[b]
BMI kg/m^2	30.9 ± 7.5	35.3 ± 5.8	<0.0001[b]
Fasting Glucose, mg/dl	93.7 ± 8.7	106.4 ± 17.5	<0.0001[b]
HbA1C, %	5.5 ± 0.4	6.0 ± 0.6	<0.0001[**b**]

Insulin sensitivity status was defined by self-report or fasting insulin values 15 uUnits/ml. Data are presented as mean ± SD. [a]Chi-square test .[b] independent t test

Table 2: Relative Influence of Input Variables in BRT

Variable	*Relative Influence*
1 BMI	33.88
2 Fasting Glucose	30.73
3 HbA1C	28.97
4 Age	5.82
5 Race	0.39
6 Gender	0.21

Table 3: Model Comparison when HbA1C is included

Model Performance (Testing set n=68)	*Sensitivity*	*Specificity*	*AUC*	*Mis.error Percentage*
Random Forest	0.730	0.806	0.858	0.235
Boosted Regression Trees	0.757	0.710	0.845	0.265
Logistic Regression	0.895	0.5	0.763	0.143
Single Classification Tree	0.784	0.516	0.741	0.338

Table 4: Relative Influence of Input Variables in BRT when HbA1C is excluded

Variable	*Relative Influence*
1 Fasting glucose	53.41
2 BMI	32.52
3 Age	13.36
4 Race	0.54
5 Gender	0.17

Table 5: Model Comparison when HbA1C is excluded

Model Performance (Testing set n=68)	*Sensitivity*	*Specificity*	*AUC*	*Mis.error Percentage*
Logistic regression	0.730	0.733	0.840	0.269
Boosted Regression Trees	0.703	0.710	0.799	0.294
Random Forest	0.730	0.613	0.782	0.324
Single Classification Tree	0.784	0.516	0.741	0.339