SUPPLEMENTARY METHODS FOR "MULTIVARIABLE MODELING OF PHENOTYPIC RISK FACTORS FOR FIRST-ONSET TMD: THE OPPERA PROSPECTIVE COHORT STUDY"

1. Overview

This appendix provides a more detailed description of the statistical methods used in the manuscript entitled "Multivariable modeling of phenotypic risk factors for first-onset TMD: the OPPERA prospective cohort study." For a more detailed description of these methods, see Hastie et al. (2009).[16]

Throughout this appendix, we will assume that the objective is to predict an outcome variable y based on a series of predictor variables denoted by $x_1, x_2, \ldots, x_p$. Let n denote the number of observations, let y_i denote the value of the response variable for observation i, and let x_{ij} denote the value of predictor variable j for observation i. For simplicity we will assume that y is continuous throughout most of the discussion below, although these methods can be generalized to the case where the outcome is a (possibly censored) survival time (as is the case for OPPERA).

2. Lasso Regression

Lasso regression[41] is a multivariable regression method that is useful for predicting an outcome measure in the presence of a large number of correlated predictor variables. Conventional least squares regression models minimize the sum of the squared differences between the predicted and actual values of the outcome variable. In mathematical terms, the least squares solution is the set of β's that minimizes

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{1}$$

Conventional least squares regression is a powerful tool that is useful in many situations, but it has several important shortcomings. In general all of the least squares regression coefficients will be nonzero, meaning that one cannot easily identify a list of the "most important" variables. Also, conventional least squares models can produce unreliable results when there are a large number of correlated predictors (as is the case in OPPERA). Finally, conventional least squares model are prone to overfitting, meaning that the model produces accurate results on the data set used to build the model but produces poor results on future data sets.

Lasso regression was proposed to overcome these shortcomings of ordinary least squares. The motivation for lasso is that a regression model that contains numerous large coefficients is more likely to be overfit. Thus, rather than finding the model that minimizes the sum of squares (1, lasso favors models with fewer large coefficients. Specifically, the lasso regression coefficients minimize the following criteria:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

Note that (2) is equal to the sum of squared errors (1) plus the (absolute) sum of the coefficients times a constant λ. The lasso criteria will prefer a less accurate model with small coefficients over a slightly more accurate model with numerous large coefficients.

Lasso regression has several advantages over conventional least squares regression. As the tuning parameter λ increases, an increasing number of the lasso regression coefficients will be equal to 0. Thus, lasso performs variables selection, meaning that one can reduce a large set of predictor variables down to a smaller set of "important" variables. Also, lasso gives more reliable results than conventional least squares when there are a large number of correlated predictors and is less prone to overfitting. See Tibshirani (1996)[41] or Hastie et al. (2009)[16] for details.

The lasso regression coefficients depend on the value of the tuning parameter λ in (2). To estimate the optimal value of λ, one may use the following procedure (known as cross-validation):

(1) Randomly partition the data into 10 partitions of approximately equal size.
(2) For each partition, fit a lasso model using the 90% of the data not contained in the partition for a large number of possible values of λ.
(3) Predict the value of y on the remaining 10% of the data (that was not used to fit the model) for each possible value of λ.
(4) Repeat steps 2 and 3 for all 10 partitions of the data.
(5) Calculate the average prediction error for each possible value of λ. Choose the λ that gives the lowest prediction error.

Note that the prediction error is estimated based on each model's predictive accuracy on data points that were not used to fit the model. Thus, choosing λ using cross-validation minimizes the risk of overfitting.

Throughout this discussion, we have assumed that the outcome variable y is continuous. However, lasso can be generalized to cases where y is a censored survival time. See Tibshirani (1997)[40] for details.

3. Random Forests

A random forest is a predictive model that is created by averaging the results of a series of decision trees. This section begins with a description of decision trees and then describes how to build a random forest using a large set of decision trees. The final two subsections describe how random forests can be used to identify important variables in a predictive model and evaluate the effect of a given variable after adjusting for other variables.

3.1. **Decision Trees.** Decision trees are predictive models that partition all possible values of the predictor variables into rectangular regions and then approximate the response variable as a constant within each such region. One of the main reasons for the popularity of decision trees is that the resulting predictive model can be visualized as a tree diagram. See Figure 1 for an example of a decision tree that was fit to an artificial data set. In this data set, the objective is to predict an outcome y based on 4 predictor variables (denoted by x_1, x_2, x_3, and x_4). At each node of the tree, one moves to either the left or right child node depending on the value of the predictor variable. For example, suppose that $x_1 = 7$ and $x_3 = 6$. At the top node of the tree, one would move to the right child node, since $x_1 = 7 > 5.513$. At the second node, one would move to the left child node, since $x_1 = 7 < 7.817$. Finally, since $x_3 = 6 < 6.504$, one would approximate y as $\hat{y} = 3.863$.

For details on fitting decision tree models, see Breiman et al. (1984)[7] or Hastie et al. (2009)[16].

3.2. **Random Forests.** Decision trees have several advantages compared to conventional linear models.[16] They can be applied to problems with large numbers of correlated predictor variables with minimal loss in accuracy. Also, they can easily model nonlinear relationships between the predictors and the response and higher-order interactions. There are also methods for handling missing values in decision trees models. Unfortunately, decision trees have high variance compared to other types of predictive models, so predictions based on decision trees are often inaccurate.

Random forests are a method for reducing the variance of decision trees. Rather than predicting y based on a single decision tree, a random forest model is calculated by fitting a series of decision trees to the same data set and predicting y by averaging the predictions of each individual tree. The algorithm for fitting a random forest model is described below:

(1) Generate a sample of size n from the data with replacement. (Since sampling is performed with replacement, some observations will be included in the data set more than once and some observations will not be included at all.)
(2) Fit a random forest decision tree to the sample generated in step 1. A random forest decision tree is identical to a regular decision tree with one exception: When fitting a regular decision tree, the split in each node may be based on any of the p predictor variables. In a random forest decision tree, m of the p predictor variables are selected at random at each node. (Typically $m = \sqrt{p}$, although other values of m are possible.) The split associated with the node must be based on one of these m variables.

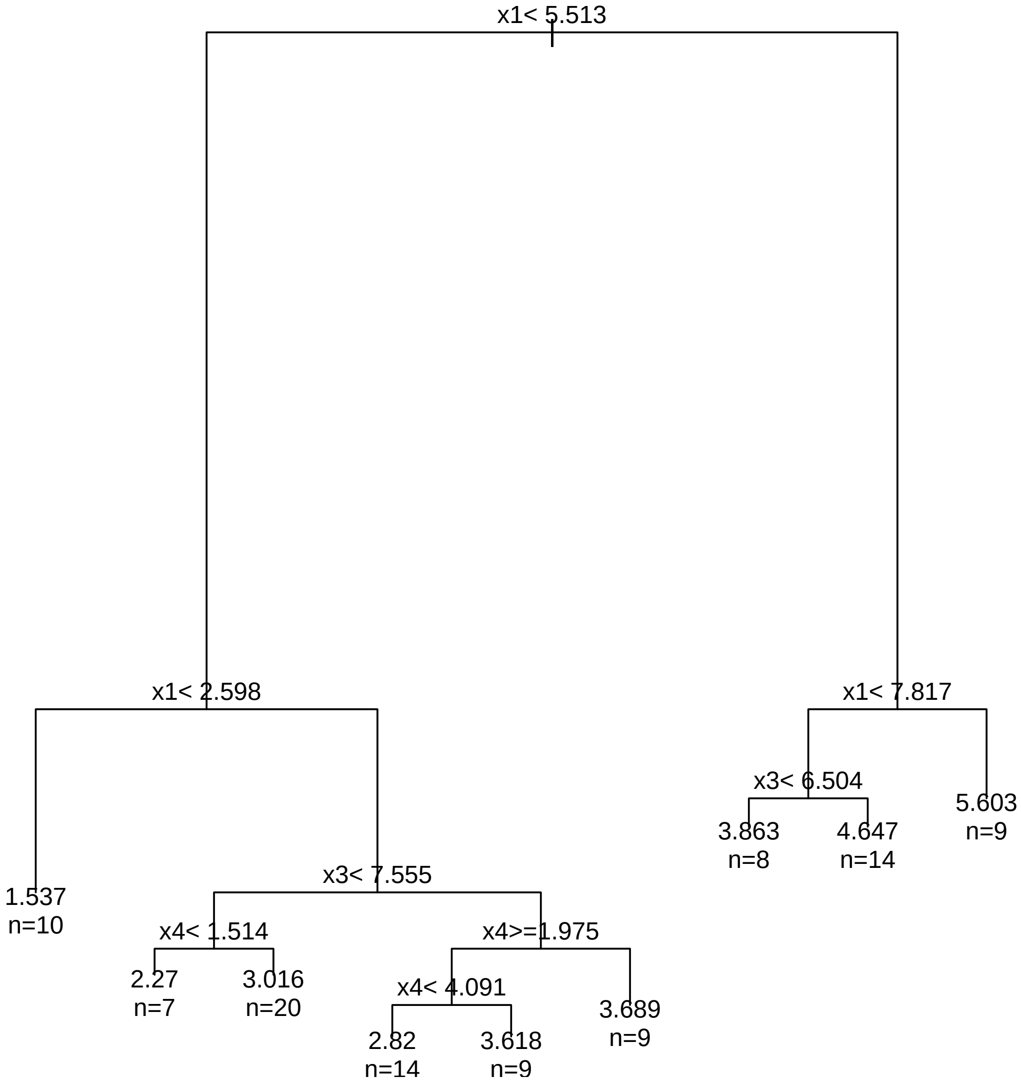

Figure 1. Decision tree fit to an artificial data set. Suppose that $x_1 = 7$ and $x_3 = 6$. At the top node of the tree, one would move to the right child node, since $x_1 = 7 > 5.513$. At the second node, one would move to the left child node, since $x_1 = 7 < 7.817$. Finally, since $x_3 = 6 < 6.504$, one would approximate y as $\hat{y} = 3.863$.

(3) Repeat steps 1 and 2 1000 times. Let T_k be the kth random forest decision tree produced by this procedure.
(4) For a given value of x, the estimated value of y is

$$\hat{y} = \sum_{k=1}^{1000} T_k(x) \tag{3}$$

In other words, at each step of the procedure, one resamples from the data with replacement and fits a random forest decision tree to this resampled version of the data. A random forest decision tree is identical to a regular decision tree except that each node randomly selects m of the p possible predictor variables and must split using one of these m variables. The random forest prediction is obtained by averaging over 1000 random forest decision trees. For more details on fitting random forest models, see Breiman (2001)[6] or Hastie et al. (2009)[16]. See Ishwaran et al. (2008)[21] for a description of how random forests can be applied to censored survival data.

Random forests have the same desirable properties as decision trees in that they can be applied to problems with large numbers of correlated predictor variables and handle nonlinear associations and interactions. The main advantage of random forests compared to decision trees is that random forests are much more accurate. The most significant disadvantage is that random forests are more difficult to interpret since the model can no longer be represented as a tree diagram.

Although random forests can produce accurate predictions, predicting first-onset TMD is not the primary goal of the random forest models in the present study. (Random forests do not perform variable selection. Thus, predicting first-onset TMD using a random forest model would require measuring all 202 variables collected in OPPERA, which is not practical in most situations.) Instead, random forests will be used to answer the two following questions: 1) Which variables are the most important predictors of first-onset TMD? 2) What is the relationship between a given variable and first-onset TMD after adjusting for the effects of the other variables measured in OPPERA? These questions may be answered using variable importance scores and partial dependence plots, as described below.

3.3. **Variable Importance Scores.** The relative importance of variables in a random forest model can be evaluated using variable importance scores (VIS). Recall that random forest models are calculated by averaging a series of decision trees. At each step of the procedure, a sample of n observations is selected (with replacement) from the set of all observations. Note that since the sampling is performed with replacement, some observations will not be included in the sample. These observations are called out of bag (OOB) observations in the random forest literature. Note that each decision tree in the model will have a different set of OOB observations.

To calculate the VIS for variable x, at each step of the model, one uses the decision tree generated at that step to predict the outcome for each OOB observation. However, whenever a split in the tree for variable x is encountered, a daughter node is chosen randomly (independent of the actual value of x). This process is repeated for each tree in the forest model. All the trees in the forest are then averaged to obtain an estimate of the outcome.

The purpose of this procedure is to estimate the decrease in the predictive accuracy of the model when x is measured with error. If x is measured with error, then the model may select an incorrect daughter node whenever it encounters a split in a tree for variable x. The VIS is defined to be the difference in the mean prediction error between this new version of the model where x is measured with error and the original version of the model. Thus, the VIS is an estimate of the decrease in the predictive accuracy of the model when variable x is measured with error. Note that it is possible for the VIS to be negative, indicating that predictive accuracy actually increases when variable x is measured with error. See Ishwaran et al. (2008)[21] for more details on calculating variable importance scores.

Note that the VIS does not measure the decrease in predictive accuracy when the variable is not available but rather the decrease in predictive accuracy when the variable is measured with error. This is an important distinction. If variable x_1 is not available but another variable x_2 is highly correlated with x_1, then there may be little or no decrease in predictive accuracy when x_1 is not available, since x_2 can replace x_1 in the model. Thus, the VIS measures the decrease in predictive accuracy when x_1 is measured with error rather than the decrease in predictive accuracy when x_1 is not available.

3.4. **Partial Dependence Plots.** Another important objective of the present study is to evaluate the association between given variables and first-onset TMD after adjusting for the effects of other variables. For example, previous analysis revealed an association between race and lifetime U.S. residence and first-onset TMD.[36] It seems unlikely that lifetime U.S. residence directly causes TMD. Thus, one may wish to determine if this association can be explained by other variables measured in OPPERA.

Suppose we wish to evaluate the association between x_1 and y after adjusting for $x_2, x_3, \ldots x_p$. One may estimate this association by letting $x_1 = x_{10}$ in the random forest model and averaging over all possible values of the remaining variables. In mathematical terms, the adjusted estimate of y when $x_1 = x_{10}$ is

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_{10}, x_{2i}, x_{3i}, \ldots, x_{pi}) \tag{4}$$

Here $\hat{f}(x_{10}, x_{20}, \ldots, x_{p0})$ is the random forest estimate of y when $x_1 = x_{10}, x_2 = x_{20}, \ldots, x_p = x_{p0}$.

To visualize the association between x_1 and y after adjusting for the other variables, the adjusted estimate of y is calculated for 25 different values in the range of x_1. The resulting estimates of y are plotted against the values of x_1, which is known as a partial dependence plot. See Hastie et al. (2009)[16] for a more detailed description of partial dependence plots. In the present study, the association between the values of x_1 and y was modeled using loess regression (see below) to visualize the association and to obtain an approximate confidence interval for the estimated values of y.

4. Loess Regression

Loess regression is useful for predicting a response variable y based on a predictor variable x when the association between x and y is nonlinear. Loess regression models this nonlinear association by using weighted least squares regression. Suppose one wishes to predict the value of y when $x = x_0$. Loess fits a weighted linear regression model that gives greater weight to values of x that are close to x_0 and less weight to values that are further away. A separate regression model is fit for each possible value of x_0, which allows for the possibility of a nonlinear association between x and y. Mathematically, rather than minimizing the conventional least squares criteria (1), loess minimizes the following weighted least squares criteria:

$$\sum_{i=1}^{n} D(x_0, x_i)\left(y_i - \beta_{0,x_0} - \beta_{1,x_0} x_i\right)^2 \tag{5}$$

Here $D(x_1, x_2)$ is a function whose value decreases as the distance between x_1 and x_2 increases and whose value is equal to 0 when the distance between x_1 and x_2 is sufficiently large. The loess models in the present study use the Tukey biweight function for D:

$$D(x_1, x_2) = \begin{cases} \left(1 - \left(\frac{x_1 - x_2}{\lambda}\right)^2\right)^2 & \text{if } |x_1 - x_2| \le \lambda \\ 0 & \text{if } |x_1 - x_2| > \lambda \end{cases} \tag{6}$$

Here λ is a tuning parameter that controls the amount of smoothing. Smaller values of λ give greater weight to points closer to x_0.

Other forms of loess regression are possible, but they are not used in the present study. See Loader (1999)[23] or Hastie et al. (2009)[16] for a more detailed description of loess regression.