Supplemental Appendix E1: Random Forest Methodology Details

Predictors of pN+

Predictors of pN+ were identified in 2 steps using Random Forest technology: 1) building a forest by regressing pN status (pN0, pN+) on cancer and other characteristics, and 2) using the forest for discovering the importance of variables and their relationships. Random forest classification and regression was implemented using the randomForestSRC R package.[10]

The first step was building a forest of 1,000 random bootstrap classification trees, grown under Gini index splitting with random input selection. Each random classification tree was constructed by sampling the data with replacement to build a new data set of size equal to the original. This bootstrap sampling procedure on average included 63% of the patients (some patient data are duplicates); the remaining 37%, referred to as out-of-bag (OOB) data, were used to construct OOB (cross-validated) estimates of a patient's probability of being pN+. Trees were grown as deeply as possible under the restriction that terminal nodes contained no fewer than 2 patients (nodesize). Nodesize was determined by fitting Random Forest systematically under different nodesize specifications and choosing the nodesize value minimizing OOB misclassification error rate, where OOB error was calculated using OOB predicted pN+. All other Random Forest parameters were set to default settings of the software. Missing data were imputed using the Random Forest method described by Ishwaran and colleagues.[8]

The second step was to quantify the predictive importance of each variable, estimated using Breiman permutation variable importance (VIMP).[3] On a scale from

-100% to 100%, VIMP estimates the expected change in misclassification error of predicted pN+ classification if the variable were removed from the multivariable forest analysis. To determine whether a variable's predictiveness was statistically significant, minimal depth variable selection was used.[9] Minimal depth equals the shortest distance from the root of a tree to the first node where a given variable produces a split (branch). Shorter distances indicate more predictiveness. Forest-averaged minimal depth for each variable was compared with a threshold value determined from a null minimal depth distribution to test whether the variable was predictive.[9]

Predictors of Number of Positive Nodes

Random Forest nonparametric regression, implemented by the randomForestSRC R package,[10] was used to regress number of regional lymph nodes, as for pN+. A forest of 1,000 Random Forest regression trees was grown using weighted mean-squared error splitting with random input selection to obtain OOB estimates for the predicted number of nodes. The OOB optimized nodesize value of 6 was used; all other Random Forest parameters were set to default values of the software.

Predictors of pN Classification

A Random Forest strategy similar to that for pN+ was used to regress pN classification (N0, N1, N2, and N3) on the 31 independent variables to obtain OOB estimates for pN classification and their probabilities. Identical Random Forest tuning parameters were used.