American Journal of Epidemiology

Vol. 177, No. 5
DOI: 10.1093/aje/kws241
Advance Access publication:
January 29, 2013

Practice of Epidemiology

Mortality Risk Score Prediction in an Elderly Population Using Machine Learning

Sherri Rose*

* Correspondence to Dr. Sherri Rose, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205 (e-mail: srose@jhsph.edu).

Initially submitted October 11, 2011; accepted for publication April 27, 2012.

Standard practice for prediction often relies on parametric regression methods. Interesting new methods from the machine learning literature have been introduced in epidemiologic studies, such as random forest and neural networks. However, a priori, an investigator will not know which algorithm to select and may wish to try several. Here I apply the super learner, an ensembling machine learning approach that combines multiple algorithms into a single algorithm and returns a prediction function with the best cross-validated mean squared error. Super learning is a generalization of stacking methods. I used super learning in the Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) to predict death among 2,066 residents of Sonoma, California, aged 54 years or more during the period 1993–1999. The super learner for predicting death (risk score) improved upon all single algorithms in the collection of algorithms, although its performance was similar to that of several algorithms. Super learner outperformed the worst algorithm (neural networks) by 44% with respect to estimated cross-validated mean squared error and had an R^2 value of 0.201. The improvement of super learner over random forest with respect to R^2 was approximately 2-fold. Alternatives for risk score prediction include the super learner, which can provide improved performance.

aging; estimation techniques; machine learning; mortality; regression analysis

Abbreviations: CV MSE, cross-validated mean squared error; LASSO, least absolute shrinkage and selection operator; METs, metabolic equivalents; SPPARCS, Study of Physical Performance and Age-Related Changes in Sonomans.

Generating risk scores and risk prediction functions has been an important area of research in medicine and epidemiology. Prediction methods have been used to produce risk scores and risk tables for heart disease (1–6), breast cancer (7–12), stroke (13, 14), age-related macular degeneration (15, 16), and other outcomes. In population-based studies of the comparative effectiveness of treatments, patients may be matched or stratified on the basis of their predicted risk of a disease or death. Standard practice for prediction has relied on parametric regression methods. However, newer methods from the machine learning literature have been introduced in medical and epidemiologic studies for prediction (17–21), such as random forest (22) and neural networks (23). Researchers are left with questions such as, "When should I use random forest instead of standard regression techniques? When should I use neural networks?"

Given the increasing number of covariates available to investigators, including biological, clinical, and genomic data, flexible machine learning procedures are appealing, since they have the ability to discover interaction, nonlinear, and higher-order effects as well as to approximate intricate functions that are not well represented by individual covariate terms or interaction terms. With a parametric regression, the complexity of the model may also become such that there are more parameters than observations. However, the investigator will not know which method (i.e., algorithm) to select a priori and may wish to run many, especially given the varied performance of individual algorithms in different data sets.

Ensembling methods allow researchers to implement multiple algorithms with an a priori benchmark regarding how to arrive at the final algorithm. Investigators therefore do not need to decide beforehand whether to forgo one

technique in favor of another; they can use several. Super learning (24, 25), the method implemented in this paper, is an a priori-specified ensembling machine learning approach that combines multiple algorithms into a single algorithm and returns a prediction function with the best cross-validated mean squared error (CV MSE). Thus, the benefits of super learning include the fact that no single method needs to be selected a priori; by using cross-validation, we can run many algorithms, including flexible machine learning algorithms that may better capture the features of the data than standard parametric regression. This super learner is a generalization of stacking algorithms (26, 27) and has optimality properties (24, 28) that led to the name "super" learner. These properties are discussed in the Materials and Methods section and in the Web Appendix (available at http://aje.oxfordjournals.org/). In an earlier paper, LeBlanc and Tibshirani (29) discussed the relationship of stacking algorithms to model-mix algorithms (30) and predictive sample-reuse methods (31). Additional methods for ensemble learning have also been developed (32–37). Further background information can be found in a 2000 review of ensemble methods (38).

I implemented super learning in a National Institute of Aging-funded study, the Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS), to predict mortality among 2,066 residents of Sonoma, California, aged 54 years or more. The study was initially described in an earlier paper (39). Previous studies of elderly populations in the United States have indicated that gender, smoking status, heart health, physical activity, educational level, income, and weight are among the important predictors of mortality in elderly populations (40–42). Prediction functions for mortality have been generated in an elderly Northern California population (43) and for nursing home residents with advanced dementia (44), the former using super learning. The contributions of this paper include: 1) generating a prediction function for mortality risk score in an elderly population that outperforms a previous function and 2) translating the ensembling method of super learning from the biostatistics literature to the applied epidemiology literature.

MATERIALS AND METHODS

Data

The SPPARCS was a population-based study of health and aging. The study enrolled 2,092 residents of Sonoma, California, and surrounding areas aged 54 years or more. Recruitment occurred between May 1993 and December 1994, with follow-up continuing for approximately 10 years; I considered 5-year mortality (through 1999) in the current study. Approval was obtained from the institutional review board for the protection of human subjects at the main study site (39).

The covariate variables $W = \{W_1, \ldots, W_{13}\}$ included in this analysis are listed in Table 1. Age was categorized into multiple indicator variables with age >70–≤80 years treated as the reference group. Self-rated health was also categorized into indicator variables ("excellent," "fair," and "poor"), with "good" treated as the reference group. A leisure-time physical activity score was calculated from baseline questionnaire data on vigorous physical activities (e.g., jogging) completed in the previous 7 days. These activities were assigned standardized intensity values in metabolic equivalents (METs) (45). Current recommendations from the Centers for Disease Control and Prevention correspond to at least 22.5 METs for 1 week, which is equivalent to 30 minutes of vigorous activity on 5 days per week or more (46). Thus, the leisure-time physical activity score variable was an indicator of vigorous physical activity at or above 22.5 METs. "Current smoker," "former smoker," and "chronic health condition" were indicator variables reflecting status at the baseline interview. "Cardiac event" was an indicator variable representing a cardiac event that occurred prior to baseline.

The outcome variable Y was death occurring within 5 years of baseline and was recorded for each subject. Vital status was ascertained using emergency contacts, healthcare providers, and the Social Security Death Index when subjects could not be found. Deaths were confirmed using death certificates (47). The cohort was reduced to a size of 2,066 because 26 subjects were missing leisure-time physical activity score or self-rated health score values (1.2% missing data). Additional discussion of the data appears in the Results section.

Table 1. Characteristics of Participants ($n = 2,066$) in the Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS), 1993–1999

Variable	No.	%
Death occurring within 5 years of baseline (Y)	269	13
Female gender (W_1)	1,225	59
Age at baseline, years		
$54 < x \leq 60$ (W_2)	323	16
$60 < x \leq 70$ (W_3)	749	36
$70 < x \leq 80$ (referent)	1,339	65
$80 < x \leq 90$ (W_4)	245	12
$x > 90$ (W_5)	22	1
Self-rated health at baseline		
Excellent (W_6)	657	32
Good (referent)	1,037	50
Fair (W_7)	309	15
Poor (W_8)	63	3
Leisure-time physical activity score ≥22.5 METs[a] at baseline (W_9)	1,460	71
Current smoker (W_{10})	172	8
Former smoker (W_{11})	1,020	49
Cardiac event prior to baseline (W_{12})	356	17
Chronic health condition at baseline (W_{13})	918	44

Abbreviations: METs, metabolic equivalents.
[a] Leisure-time physical activity score was computed from questionnaire data and based on vigorous physical activities assigned standardized intensity values in METs.

Statistical analysis

We introduce O, a random variable, and write $O \sim P_0$ to indicate that the true probability distribution of O is P_0. The empirical probability distribution is denoted P_n. Our model is nonparametric, and we make the assumption that the actual data as observed in practice can be represented as n independent and identically distributed observations of the random variable O, and $O = (W, Y)$, where W is a vector of covariates and Y is an outcome. For each subject, we measure Y_i and W_i. The regression function $E_0(Y|W)$ (or $P_0(Y = 1|W)$, since our outcome Y is a binary indicator of death within 5 years of baseline) is our function of interest for prediction. We define our parameter of interest $\bar{Q}_0 = E_0(Y|W)$ as $\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L(O, \bar{Q})$, where $L(O, \bar{Q}) = (Y - \bar{Q})^2$ and $\bar{Q}$ is a possible function in the parameter space of functions that map an input W into a predicted value for Y. $E_0 L(O, \bar{Q})$, the expected loss, evaluates the candidate $\bar{Q}$, and it is minimized at the optimal choice $\bar{Q}_0$. Our goal in prediction is to find the best estimator of this regression function, since $\bar{Q}_0$ is unknown.

We implement super learning in the R programming language (48) with the SuperLearner package (49) to estimate $\bar{Q}_0$. The super learner ensembling method allows researchers to use multiple algorithms to improve upon the single algorithms included in the library of algorithms in nonparametric and semiparametric statistical models by using cross-validation and taking a weighted average of the algorithms.

The collection of 12 algorithms used in this analysis included: generalized boosted regression with 10,000 trees and interaction depth = 2 (50, 51), main-terms Bayesian logistic regression using a Cauchy prior with scale = 2.5 (52), penalized regression (least absolute shrinkage and selection operator (LASSO)) (53), generalized additive regression (54), main-terms logistic regression (48), two different implementations of multivariate adaptive regression splines (the R packages earth and polymars) (55–57), bootstrap aggregation of regression trees (bagging) with 100 replications (58, 59), recursive partitioning and regression trees (60), the arithmetic mean assigning the marginal probability of mortality to each subject (48), classification and regression trees using random forest with 1,000 trees (22), and neural networks with 2 units in the hidden layer (23) (see Table 2). "Main-terms" is used to indicate that all covariates were included in the algorithm with no other (e.g., interaction) terms. Therefore, the collection of algorithms contained a broad range of popular learners, such as nonlinear models, cubic splines, and tree-based algorithms. These algorithms were chosen because super learner will perform better when the collection of algorithms includes different approaches for the estimation of $\bar{Q}_0$. The collection was limited to 12 for computational speed.

The super learner algorithm is described below and in Figure 1. I started with the SPPARCS data matrix and the collection of 12 algorithms (step 1). The data were then split into 10 mutually exclusive and exhaustive blocks of approximately equal size in preparation for performing 10-fold cross-validation (step 2). The observations were randomly shuffled before sorting into blocks, and each block was coerced to be similar in distribution to the full sample for the mortality outcome but not each covariate in W. (Each block contained approximately 27 subjects with $Y = 1$.) The V-fold cross-validation procedure involved fitting the data on the "training set" and evaluating the fit in the "validation set" in all of the V folds. For each of the 12 algorithms, the algorithm was fitted on the training set (9 of the 10 blocks) of each fold, and then predicted values were obtained using the validation set (1 of the 10 blocks) of each fold (steps 3 and 4). It is important to stress that the observations in the validation sets were not included in the fitting process for that fold and were used only to evaluate the predictor obtained on its respective training sample. At the end of this fitting and prediction process, I had 12 columns of predicted values obtained from the validation sets, one column for each algorithm $(D_j, j = 1, \ldots, 12)$. At this stage, I calculated the estimated CV MSE,

$$\text{CV MSE} = \frac{\sum_{i=1}^{n} (Y_i - D_{j,i})^2}{n},$$

for each algorithm using the predicted probabilities D_j (step 5).

I next calculated the optimal weighted combination of the 12 algorithms from a proposed family of weighted combinations indexed by the weight vector α. The proposed family of weighted combinations included only those α vectors that summed to 1 and for which the weight was greater than or equal to 0. This restriction to a convex combination increases the stability of the super learner (25). My goal was to determine which weighted combination minimized the CV MSE over the family of weighted combinations. The predicted probabilities D_j for each algorithm were

Table 2. Algorithms Included in the Super Learner, Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS), Sonoma, California, 1993–1999

Algorithm	Description
bayesglm	Bayesian main-terms logistic regression
glmnet	LASSO
gam	Generalized additive regression
glm	Main-terms logistic regression
gbm	Generalized boosted regression
earth	Multivariate adaptive regression splines
polymars	Multivariate adaptive polynomial spline regression
ipredbagg	Bagging for classification, regression and survival trees
randomForest	Classification and regression with random forest
rpart	Recursive partitioning and regression trees
mean	Arithmetic mean
nnet	Neural network

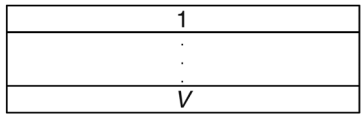
Abbreviation: LASSO, least absolute shrinkage and selection operator.

1. Start with the SPPARCS data and a collection of M algorithms. In this analysis M = 12 and n = 2,066.

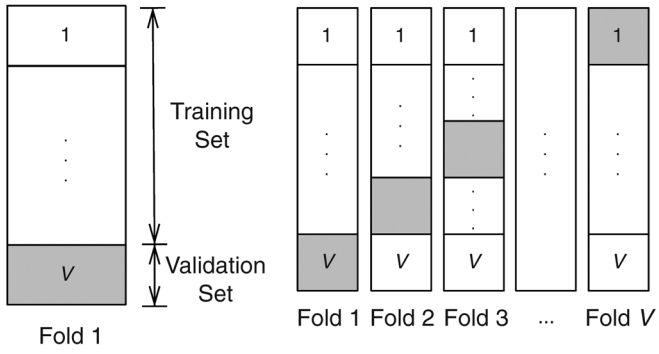
ID	$W1$	...	$W12$	$W13$	Y
1	1	...	0	1	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	0	...	1	1	1

bayesglm
glmnet
.
.
.
nnet

2. Split the SPPARCS data into V mutually exclusive and exhaustive blocks of equal or approximately equal size. Here V = 10.



3. Fit each algorithm on the training set for each V fold. For example, in fold 1, our training set could be blocks 1–9, where block 10 will be the validation set. Each algorithm is fit on blocks 1–9. In fold 2, our training set might be blocks 1–8 and block 10 with block 9 serving as the validation set, and so on. At the end of this stage you have V fits for each algorithm.



4. For each algorithm, predict the outcome Y using the validation set in each fold, based on the corresponding training set fit for that fold. You now have a vector of predicted values D_j, j = 1, ..., M for each algorithm.

ID	D_{bayesglm}	...	D_{nnet}
1	0.54	...	0.42
.	.	.	.
.	.	.	.
n	0.09	...	0.12

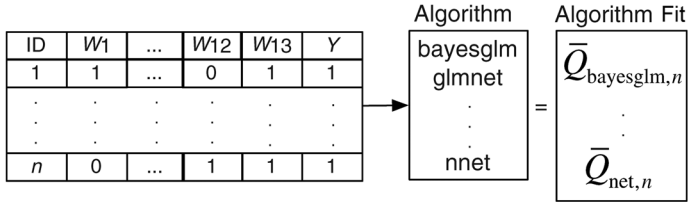
5. Compute the estimated CV MSE for each algorithm using the predicted values D_j calculated from the validation sets.

$$\text{CV MSE}_j = \frac{\sum_{i=1}^{n}(Y_i - D_{j,i})^2}{n}$$

6. Calculate the optimal weighted combination of M algorithms from a family of weighted combinations indexed by the weight vector α. This is done by performing a regression of Y on the predicted values D to estimate the vector α. This calculation determines the combination that minimizes the CV risk over the family of weighted combinations.

$$P_n(Y = 1 \mid D) = \text{expit}(\alpha_{\text{bayesglm},n} D_{\text{bayesglm}} + \ldots + \alpha_{\text{nnet},n} D_{\text{nnet}})$$

7. Fit each of the M algorithms on the complete data set. These fits combined with the estimated weights form the super learner function that can be used for prediction.



8. To obtain predicted values for the SPPARCS data, run it through the super learner function.

$$\bar{Q}_{\text{SL},n} = 0.4614\bar{Q}_{\text{bayesglm},n} = 0.496\bar{Q}_{\text{gbm},n} + 0.044\bar{Q}_{\text{mean},n}$$

Figure 1. Super learner (SL) algorithm applied to data from the Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS), Sonoma, California, 1993–1999. CV, cross-validated; CV MSE, cross-validated mean squared error; ID, identification number.

used as inputs in a regression to predict the outcome Y. Therefore, I had a regression of Y on the predicted values D with 12 coefficients $\alpha = \{\alpha_{\text{bayesglm}}, \alpha_{\text{glmnet}}, \ldots, \alpha_{\text{nnet}}\}$, one for each of the 12 algorithms. Selecting the weights that minimized the CV MSE was therefore this simple minimization problem, formulated as this regression of the outcomes Y on the predicted values of the algorithms according to the user-supplied parametric family of weighted combinations:

$$P_n(Y = 1|D) = \text{expit}(\alpha_{\text{bayesglm},n} D_{\text{bayesglm}} + \alpha_{\text{glmnet},n} D_{\text{glmnet}} + \cdots + \alpha_{\text{nnet},n} D_{\text{nnet}}), \quad (1)$$

where the subscript n indicates estimated values. The weighted combination with the smallest CV MSE was the best estimator according to our prespecified criterion: minimizing the estimated expected squared error loss function (step 6). The only algorithms with nonzero estimated weights were Bayesian logistic regression, generalized boosted regression, and the arithmetic mean.

Lastly, I fitted each of the 12 algorithms on the full SPPARCS data set. These algorithm fits combined with the weight vector estimated in equation 1 generated the super learner prediction function:

$$\bar{Q}_{\text{SL},n} = 0.461\bar{Q}_{\text{bayesglm},n} + 0.496\bar{Q}_{\text{gbm},n} + 0.044\bar{Q}_{\text{mean},n}, \quad (2)$$

where algorithms with a zero weight were excluded. This prediction function was the weighted combination of the candidate algorithms applied to the whole data set (step 7). To obtain predicted values for the SPPARCS data set, we input the SPPARCS data into equation 2, the super learner prediction function (step 8). In order to calculate the estimated CV MSE for the super learner, the super learner algorithm itself was cross-validated with the CV.SuperLearner function in the SuperLearner package. Using a server with dual quad-core processors running at 2.6 GHz and with 32 GB of memory, my analysis using SuperLearner and CV.SuperLearner took 3 hours to complete. Running the SuperLearner function alone, the function needed to generate equation 2, took 18 minutes.

Demonstrations of the super learner's finite sample performance in simulations and data have been presented elsewhere (24, 25, 43, 61, 62). Asymptotic results have also been discussed (24, 63). In brief, these results prove that in realistic scenarios (where none of the algorithms represent a correctly specified parametric regression), the cross-validated selector performs asymptotically as well as the oracle, which we define as the best estimator given the algorithms in the collection of algorithms. Consequently, the super learner performs asymptotically as well as the best choice among the family of weighted combinations of estimators. Thus, adding additional algorithms improves the performance of the super learner. The asymptotic equivalence remains true if the number of algorithms in the library grows very fast with sample size. In the unlikely situation that the collection of algorithms contains a correctly specified parametric regression, the super learner will approximate the truth as fast as the correctly specified regression. However, the super learner, in this case, will be more variable than the correctly specified regression (24, 63). See the Web Appendix for the precise finite sample oracle inequality.

RESULTS

A summary of SPPARCS variables can be found in Table 1. Of the 2,066 study subjects, there were 269 deaths within 5 years of baseline (13% of subjects). The majority of subjects (59%) were female. The largest age group was >70–≤80 years (65% of the study subjects). Half of the subjects (50%) had self-rated health classified as "good," with 32% rating their own health as "excellent," 15% as "fair," and 3% as "poor." The leisure-time physical activity score was greater than or equal to 22.5 METs for 71% of the subjects. Only 8% of participants were current smokers, with 49% reporting formerly smoking. Seventeen percent of the subjects had had a cardiac event prior to baseline, and 44% had a chronic health condition at baseline.

The super learner algorithm for predicting mortality risk score in the SPPARCS data improved upon all 12 single algorithms, with an R^2 value of 0.201 and an estimated CV MSE of 9.04×10^{-2}. Thus, the super learner had a 20.1% gain relative to assigning 13% to all subjects. The performance of the super learner was similar to several algorithms in the collection of algorithms while also being a vast improvement over others. Results are presented in Table 3. Relative efficiency for each of the k algorithms was defined as RE = CV MSE(k)/CV MSE(super learner). For the 12 single algorithms, the estimated CV MSE ranged from 9.11×10^{-2} to 1.30×10^{-1}, and R^2 ranged from 0.195 to −0.150. (R^2 values can occur outside the range of [0, 1] in cross-validated data. A negative R^2 indicates that the marginal probability of mortality provides a better predictor than the algorithm, that serious overfitting of the data has occurred, and that the cross-validated residual sum of squares is larger than the total sum of squares.)

The super learner improved upon the worst algorithm, neural networks (nnet), by 44% with respect to CV MSE. The super learner also improved on the 2 implementations of multivariate adaptive regression splines (polymars and earth) by 5% and 6%, respectively, bagging (ipredbagg) by 9%, classification and regression trees using random forest (randomForest) by 12%, recursive partitioning and regression trees (rpart) by 15%, and the arithmetic mean (mean) by 25%, all with respect to CV MSE. The 5 algorithms that performed very similarly to the super learner were LASSO (glmnet), generalized boosted regression (gbm), main-terms Bayesian logistic regression (bayesglm), generalized additive regression (gam), and main-terms logistic regression (glm). The super learner improved upon these algorithms with respect to CV MSE by only 1%. When R^2 values were evaluated, these same 5 algorithms had decreases in R^2 between 0.6% and 0.7% in comparison with the super learner. The R^2 results were more drastic for

Table 3. Relative Efficiency[a] With Respect to Cross-Validated Mean Squared Error and R^2 for Each Algorithm, Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS), Sonoma, California, 1993–1999

Algorithm	CV MSE	RE	R^2
SuperLearner	9.04×10^{-2}	—[b]	0.201
glmnet	9.11×10^{-2}	1.01	0.195
gbm	9.11×10^{-2}	1.01	0.195
bayesglm	9.12×10^{-2}	1.01	0.195
gam	9.12×10^{-2}	1.01	0.194
glm	9.12×10^{-2}	1.01	0.194
polymars	9.52×10^{-2}	1.05	0.159
earth	9.55×10^{-2}	1.06	0.157
ipredbagg	9.84×10^{-2}	1.09	0.131
randomForest	1.01×10^{-1}	1.12	0.105
rpart	1.04×10^{-1}	1.15	0.078
mean	1.11×10^{-1}	1.25	0.000
nnet	1.30×10^{-1}	1.44	−0.150

Abbreviations: CV MSE, cross-validated mean squared error; RE, relative efficiency.

[a] RE = CV MSE(algorithm)/CV MSE(SuperLearner).

[b] Referent.

other algorithms—for example, the super learner surpassed randomForest by 2-fold with respect to R^2 and by almost 3-fold for rpart.

Figure 2 illustrates the differences in predicted probabilities obtained from the super learner in comparison with 2 algorithms in the library, glm and randomForest. The predicted values for glm were similar to those for the super learner; hence the grouping of values around zero. However, the histogram comparing randomForest and super learner shows a wider spread of numbers ranging from −0.39 to 0.47, with fewer near zero. This display agrees with the CV MSE results discussed above, where glm performed almost as well as super learner and randomForest performed poorly in comparison with super learner.

Had one simply run the glm algorithm, the probability of death would have been estimated by

$$\begin{aligned} P_n(Y=1|W) = \text{expit}(&-1.97 - 0.55W_1 - 1.94W_2 \\ &- 0.93W_3 + 1.33W_4 + 2.13W_5 - 0.37W_6 \\ &+ 0.79W_7 + 1.42W_8 - 0.49W_9 + 1.30W_{10} \\ &+ 0.34W_{11} + 0.45W_{12} + 0.41W_{13}), \end{aligned} \quad (3)$$

with the variables as described in Table 1. A mean squared error based on this full data fit in equation 3 could have been calculated; however, this metric is subject to the potential overfitting of the algorithm. Thus, cross-validation would still be performed to calculate CV MSEs. Equation 3 is the same fit obtained in step 7 of the super learner algorithm described in Figure 1. This latter step of the super learner involves fitting each algorithm in the library with the full data. The key is that the weight vector obtained on the basis of the cross-validated data is applied to these full data fits to obtain predicted values for the super learner.

DISCUSSION

Developing functions for risk score prediction has been an area of significant research in medicine and epidemiology. They have been used to generate risk scores for various diseases, and researchers are now interested in incorporating machine learning methods to augment standard analyses. As data collected on study subjects become increasingly complex, with typical characteristics such as gender and age now being supplemented with possibly biological, genomic, and proteomic measures, the examination of machine learning methods to aid in prediction is a natural step. However, any single algorithm may be outperformed by another algorithm in a given data set, and this cannot be ascertained prior to analysis.

The ensembling super learner algorithm allows an investigator to run multiple algorithms and combine them into an improved estimator, returning a function that can be used for prediction. Each algorithm within the collection of algorithms in the super learner is cross-validated in order to prevent overfitting. The super learner is also indexed by a loss function defined by the desired measure of performance. Thus, other loss functions can be selected. Super learning in this analysis was optimized with respect to a specific loss function; it was tailored to achieve the best estimator based on the CV MSE.

The super learner yielded improved performance over the 12 algorithms studied with respect to estimated CV MSE. While an R^2 value of 0.201 indicates that the SPPARCS data have a somewhat weak signal, it is approximately 2-fold better than the R^2 generated in a previous study of mortality among elderly Kaiser Permanente Northern California members using super learning (43). This previous study had more than 10 times the number of covariates available as the current study, where all were disease indicators except for age and gender, and over 13 times the number of subjects. Thus, it is possible that the inclusion of leisure-time physical activity score and self-rated health increased the predictive ability of the SPPARCS analysis. Previous literature indicates that perception of health in elderly adults may be as important as less subjective measures when assessing later outcomes (64, 65). Likewise, benefits of physical activity in older populations have also been shown (66).

The super learner showed nontrivial improvements over neural nets ($R^2 = -0.150$) and random forest ($R^2 = 0.105$), two algorithms that have been used in the applied epidemiology and clinical literature for prediction. Interestingly, the main-terms Bayesian logistic regression, which performed almost as well as the super learner in this analysis, with $R^2 = 0.195$, performed poorly in the previously discussed Kaiser Permanente study, with $R^2 = -0.005$ (43). This highlights the issue researchers are faced with when selecting an algorithm for generating prediction functions. Bayesian logistic regression performed well here, but it performed poorly in the Kaiser Permanente study. The benefit of

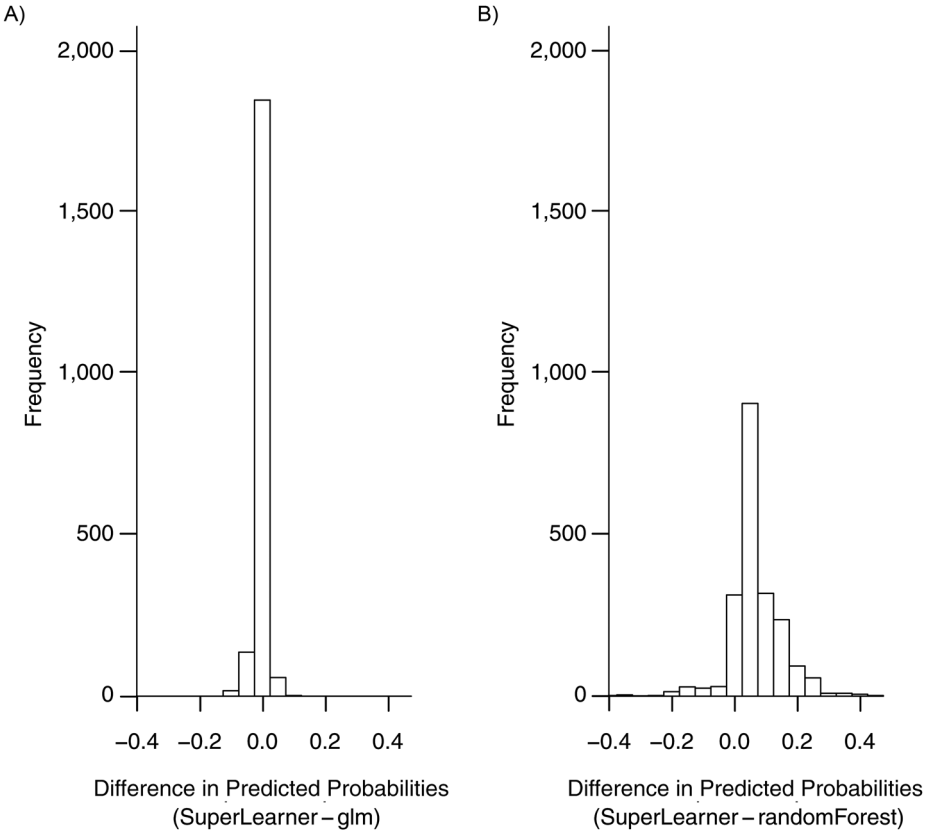


Figure 2. Differences between SuperLearner and glm (A) and randomForest (B) in cross-validated predicted probabilities of mortality, Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS), Sonoma, California, 1993–1999.

ensembling these algorithms in the super learner is then clear: Super learner provides a tool with which researchers can run many algorithms and obtain a prediction function with the best CV MSE, avoiding the need to commit to a single algorithm. Even when the result is a small improvement relative to the best algorithms in the collection of algorithms, as we see in the SPPARCS analysis, we are still able to leverage the flexibility of the super learner to avoid a priori selection of a poor algorithm. In other data applications using super learner (25), larger improvements in CV MSE have also been seen.

As with any approach for prediction, the super learner algorithm should be specified a priori in order to accurately assess its performance with cross-validation. The process of running an algorithm (e.g., logistic regression, super learner, random forest, etc.), adjusting the tuning parameters or included variables, and then rerunning the algorithm can lead to overfitting and misleading results. For super learning, this means the input variables, collection of algorithms, and tuning parameters, such as the number of folds, should be decided before analysis begins.

One might counter that one's own procedure would involve implementing multiple algorithms using cross-validation, calculating the CV MSE (i.e., stopping after step 5 in Figure 1), and then selecting the one algorithm with the best CV MSE. This procedure is itself a super-learning algorithm—the discrete super learner (24, 25). Once a discrete super learner has been implemented, only the relatively trivial calculation of the weight vector needs to be completed to implement the super learner.

Super learning is an effective method for prediction, but it also has applications in effect estimation. Super learning can be used to estimate the relevant portions of the likelihood in an effort to achieve minimal bias and variance for effect parameters in practice (63)—for example, within a maximum likelihood-based substitution estimator of the g-formula (also referred to as G-computation) (67), a targeted maximum likelihood estimator (68, 63), or an inverse probability-weighted estimator (69). As discussed in 2 recent articles (70, 71), researchers are frequently concerned about parametric model misspecification within effect estimation procedures and may wish to implement methods such as super learning.

We must also clarify the distinction between research questions focused on *prediction* and those focused on an *effect*. If we are interested in a risk score and our goal is, for example, the best estimator of $E_0(Y|W)$, this is a prediction research question. However, if we are interested in finding the variables in W that have the greatest effect on Y, that is an effect question. Thus, the lack of supposed

"clinical interpretability" of the super learner prediction function should not be considered a drawback, since interest in discovering variables with the greatest effect on the outcome is a separate research question with a distinct target parameter and bias-variance trade-off.

Limitations of super learning include increased computing time and memory in comparison with running a single algorithm, although with each subsequent generation of computers the upper bound of what is achievable increases substantially. Additionally, its current application as a package in the R programming language and not in the more widely used SAS or STATA is a significant drawback. The performance of super learner also depends on the algorithms included in the collection of algorithms. If there is an algorithm that is vastly superior to other algorithms and it is not included in the collection of algorithms, the super learner is not likely to outperform this algorithm. The selection of algorithms for the super learner should be based on the structure of the data (e.g., binary outcome), computational resources, and time. In other words, if one has unlimited computational resources and time, all suitable algorithms available should be included. In R, there are over 40 packages for prediction. Algorithms with various tuning parameters can also be included within the super learner as separate algorithms. By adjusting tuning parameters (or adding dimension reduction), one can generate an even larger library. For example, random forest with 1,500 trees and random forest with 1,000 trees would be treated as separate algorithms.

Alternatives to parametric regression for risk score prediction include the ensembling machine learning algorithm super learner, which can provide improved performance and allows researchers to specify a priori an algorithm that uses multiple algorithms for generating a prediction function. The results presented in this paper further demonstrate the promise of the super learner illustrated in previous publications. Additional work assessing the performance of super learning in epidemiologic data is warranted.

ACKNOWLEDGMENTS

Author affiliation: Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland (Sherri Rose).

This work was supported by the National Science Foundation (grant DMS-1103901).

Conflict of interest: none declared.

REFERENCES

1. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol*. 1976; 38(1):46–51.
2. Anderson KM, Wilson PWF, Odell PM, et al. An updated coronary risk profile: a statement for health professionals. *Circulation*. 1991;83(1):356–362.
3. Wilson PWF, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837–1847.
4. Ramsay LE, Haq IU, Jackson PR, et al. Sheffield risk and treatment table for cholesterol lowering for primary prevention of coronary heart disease. *Lancet*. 1995; 346(8988):1467–1471.
5. Ramsay LE, Haq IU, Jackson PR, et al. The Sheffield table for primary prevention of coronary heart disease: corrected. *Lancet*. 1996;348(9036):1251.
6. Jackson R. Updated New Zealand cardiovascular disease risk-benefit prediction guide. *Br Med J*. 2000;320(7236):709–710.
7. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81(24):1879–1886.
8. Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst*. 1999;91(18):1541–1548.
9. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med*. 2004;23(7):1111–1130.
10. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst*. 2006;98(17): 1204–1214.
11. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst*. 2008;100(14):1037–1041.
12. Wacholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med*. 2010;362(11):986–993.
13. Saposnik G, Kapral MK, Liu Y, et al. IScore: a risk score to predict death early after hospitalization for an acute ischemic stroke. *Circulation*. 2011;123(7):739–749.
14. Saposnik G, Raptis S, Kapral M, et al. The iScore predicts poor functional outcomes early after hospitalization for an acute ischemic stroke. *Stroke*. 2011;42(12):3421–3428.
15. Seddon JM, Reynolds R, Maller J, et al. Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Invest Opthalmol Vis Sci*. 2009; 50(5):2044–2053.
16. Seddon JM, Reynolds R, Yu Y, et al. Risk models for progression to advanced age-related macular degeneration using demographic, environmental, genetic, and ocular factors. *Opthalmology*. 2011;118(11):2203–2211.
17. Stassen HH, Szegedi A, Scharfetter C. Modeling activation of inflammation response system: a molecular-genetic neural network analysis. *BMC Proc*. 2007;1(suppl 1):S61.
18. Sun YV, Cai Z, Desai K, et al. Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. *BMC Proc*. 2007;1(suppl 1):S62.
19. Ziegler A, DeStefano AL, Konig IR, et al. Data mining, neural nets, trees—problems 2 and 3 of Genetic Analysis Workshop 15. *Genet Epidemiol*. 2007;31(suppl 1):S51–S60.
20. Austin PC, Tu JV, Lee DS. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *J Clin Epidemiol*. 2010;63(10):1145–1155.
21. Peng SY, Chuang YC, Kang TW, et al. Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. *Eur J Neurol*. 2010;17(7):945–950.

22. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
23. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer-Verlag; 2002.
24. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article 25.
25. Polley EC, Rose S, van der Laan MJ. Super learning. In: van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011:43–66.
26. Wolpert DH. Stacked generalization. *Neural Netw*. 1992; 5(2):241–259.
27. Breiman L. Stacked regressions. *Mach Learn*. 1996;24(1): 49–64.
28. van der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. (UC Berkeley Division of Biostatistics Working Paper 130). Berkeley, CA: University of California, Berkeley; 2003. (http://www.bepress.com/ucbbiostat/paper130/). (Accessed October 6, 2011).
29. LeBlanc M, Tibshirani R. Combining estimates in regression and classification. *J Am Stat Assoc*. 1996;91(436):1641–1650.
30. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Ser B*. 1974;36(2):111–147.
31. Geisser S. The predictive sample reuse method with applications. *J Am Stat Assoc*. 1975;70(35):320–328.
32. Tsybakov AB. Optimal rates of aggregation. In: Schölkopf B, Warmuth MK, eds. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24–27, 2003, Proceedings*. (Lecture Notes in Computer Science, vol 2777). New York, NY: Springer Publishing Company; 2003:303–313.
33. Juditsky A, Nazin AV, Tsybakov AB, et al. *Generalization Error Bounds for Aggregation by Mirror Descent Averaging*. (Advances in Neural Information Processing Systems, vol 18). Cambridge, MA: MIT Press; 2005.
34. Bunea F, Tsybakov AB, Wegkamp MH. Aggregation and sparsity via l_1 penalized least squares. In: Lugosi G, Simon H-U, eds. *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22–25, 2006, Proceedings*. (Lecture Notes in Computer Science, vol 4005). New York, NY: Springer Publishing Company; 2006:379–391.
35. Bunea F, Tsybakov AB, Wegkamp MH. Aggregation for Gaussian regression. *Ann Stat*. 2007;34(5):1674–1697.
36. Dalayan AS, Tsybakov AB. Aggregation by exponential weighting and sharp oracle inequalities. In: Bshouty NH, Gentile C, eds. *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA, June 13–15, 2007, Proceedings*. (Lecture Notes in Computer Science, vol 4539). New York, NY: Springer Publishing Company; 2007:97–111.
37. Dalayan AS, Tsybakov AB. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach Learn*. 2008;72(1-2):39–61.
38. Dietterich TG. Ensemble methods in machine learning. In: Kittler J, Roli F, eds. *Multiple Classifier Systems: First International Workshop, MCS 2000, Cagliari, Italy, June 21–23, 2000, Proceedings*. (Lecture Notes in Computer Science, vol 1857). New York, NY: Springer Publishing Company; 2000:1–15.
39. Tager I, Hollenberg M, Satariano W. Association between self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population. *Am J Epidemiol*. 1998;147(10):921–931.
40. Scotta WK, Macera CA, Cornmanb CB, et al. Functional health status as a predictor of mortality in men and women over 65. *J Clin Epidemiol*. 1997;50(3):291–296.
41. Fried LP, Fronmal RA, Newman AB, et al. Risk factors for 5-year mortality in older adults: the Cardiovascular Health Study. *J Am Med Assoc*. 1998;279(8):585–592.
42. Terracciano A, Lockenhoff CE, Zonderman AB, et al. Personality predictors of longevity: activity, emotional stability, and conscientiousness. *Psychosom Med*. 2008; 70(6):621–627.
43. Rose S, Fireman B, van der Laan MJ. Nested case-control risk score prediction. In: van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011:43–66.
44. Mitchell SL, Miller SC, Teno JM, et al. The advanced dementia prognostic tool: a risk score to estimate survival in nursing home residents. *J Pain Symptom Manage*. 2010; 40(5):639–651.
45. Ainsworth B, Haskell W, Leon A, et al. Compedium of Physical Activities: classification of energy costs of human physical activities. *Med Sci Sports and Exerc*. 1993; 25(1):71–80.
46. Centers for Disease Control and Prevention. *Physical Activity and Health: A Report of the Surgeon General*. Atlanta, GA: Centers for Disease Control and Prevention; 1996.
47. Bembom O, van der Laan MJ, Haight T, et al. Leisure-time physical activity and all-cause mortality in an elderly cohort. *Epidemiology*. 2009;20(3):424–430.
48. R Foundation for Statistical Computing. *R: A Language and Environment for Statistical Computing. Version 2.13.0*. Vienna, Austria: R Foundation for Statistical Computing; 2011.
49. Polley EC, van der Laan MJ. *SuperLearner: Super Learner Prediction, Package Version 2.0–4*. Vienna, Austria: R Foundation for Statistical Computing; 2011. (http://cran.r-project.org/web/packages/SuperLearner/). (Accessed October 1, 2011).
50. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232.
51. Ridgeway G. *gbm: Generalized Boosted Regression Models, Package Version 1.6–3*. Vienna, Austria: R Foundation for Statistical Computing; 2011. (http://cran.r-project.org/web/packages/gbm/). (Accessed October 1, 2011).
52. Gelman A, Jakulin A, Pittau MG, et al. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2009;2(3):1360–1383.
53. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
54. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. New York, NY: Chapman & Hall; 1990.
55. Friedman JH. Multivariate adaptive regression splines. *Ann Stat*. 1991;19(1):1–141.
56. Kooperberg C. *polspline: Polynomial Spline Routines, Package Version 1.15*. Vienna, Austria: R Foundation for Statistical Computing; 2010. (http://cran.r-project.org/web/packages/polspline/). (Accessed October 1, 2011).
57. Milborrow S. *Earth: Multivariate Adaptive Regression Spline Models, Package Version 3.2-1*. Vienna, Austria: R Foundation for Statistical Computing; 2011. (http://cran.r-project.org/web/packages/earth/). (Accessed October 1, 2011).

58. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123–140.
59. Peters A, Hothorn T. *Ipred: Improved Predictors, Package Version 0.8–11*. Vienna, Austria: R Foundation for Statistical Computing; 2011. (http://cran.r-project.org/web/packages/ipred/). (Accessed October 1, 2011).
60. Breiman L, Friedman JH, Olshen R, et al. *Classification and Regression Trees*. Boca Raton, FL: CRS Press; 1984.
61. Polley EC, van der Laan MJ. Predicting optimal treatment assignment based on prognostic factors in cancer patients. In: Peace KE, ed. *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*. Boca Raton, FL: CRC Press; 2009:441–454.
62. Polley EC, van der Laan MJ. Super learning for right-censored data. In: van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011:249–258.
63. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011.
64. Idler E, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. *J Health Soc Behav*. 1997;38(1):21–37.
65. Blazer DG. How do you feel about…? Health outcomes in late life and self-perceptions of health and well-being. *Gerontologist*. 2008;48(4):415–422.
66. Danaei G, Ding EL, Mozaffarian D, et al. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLoS Med*. 2009;6(4):e1000058.
67. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986;7(9–12):1393–1512.
68. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):Article 11.
69. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5):561–570.
70. Sudat SE, Calton EJ, Seto EY, et al. Using variable importance measures from causal inference to rank risk factors of schistosomiasis infection in a rural setting in China. *Epidemiol Perspect Innov*. 2010;7:Article 3.
71. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7):731–738.