Revisiting Performance Metrics for Prediction with Rare Outcomes

Statistical Methods in Medical Research
XX(X):2–27

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Samrachana Adhikari[1], Sharon-Lise Normand[2,3], Jordan Bloom[4], David Shahian[4], and Sherri Rose[5]

Abstract

Machine learning algorithms are increasingly used in the clinical literature, claiming advantages over logistic regression. However, they are generally designed to maximize area under the receiver operating characteristic curve (AUC). While AUC and other measures of accuracy are commonly reported for evaluating binary prediction problems, these metrics can be misleading. We aim to give clinical and machine learning researchers a realistic medical example of the dangers in relying on a single measure of discriminatory performance to evaluate binary prediction questions. Prediction of medical complications after surgery is a frequent but challenging task because many post-surgery outcomes are rare. We predicted post-surgery mortality among patients in a clinical registry who received at least one aortic valve replacement. Estimation incorporated multiple evaluation metrics and algorithms typically regarded as performing well with rare outcomes, as well as an ensemble and a new extension of the lasso for multiple unordered treatments. Results demonstrated high accuracy for all algorithms with moderate measures of cross-validated AUC. False positive rates were less than 1%, however, true positive rates were less than 7%, even when when paired with a 100% positive predictive value, and graphical representations of calibration were poor. Similar results were seen in simulations, with the addition of high AUC (>90%) accompanying low true positive rates. Clinical studies should not primarily report AUC or accuracy.

Keywords

Prediction, Classification, Machine Learning, Ensembles, Mortality

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

1 Introduction

Prediction with various types of electronic health data has become increasingly common in the clinical literature.[1] Data sources include medical claims, electronic health records, registries, and surveys. While each class of data has benefits and limitations, the growth in data collection and its availability to researchers has provided opportunities to study rare outcomes, such as mortality, in different populations that were previously difficult to examine in smaller epidemiologic studies. Creating mortality risk score functions has evolved in recent years to include *both* electronic health data and modern machine learning techniques. These machine learning methods claim advantages over logistic regression in terms of out-of-sample performance.[2] Previous studies have examined mortality in older adults,[3;4] intensive care units,[5;6], individuals with cardiovascular disease,[7;8] and other settings[9;10] using machine learning.

Machine learning algorithms for binary outcomes are generally designed to minimize prediction error or maximize the area under the receiver operating characteristic curve (AUC). This AUC value, also referred to as the *c*-index or AUROC, is a summary metric of the predictive discrimination of an algorithm, specifically measuring the ranking performance for random discordant pairs. However, when the outcome of interest is rare, or more generally, when there is class imbalance in the outcome, benchmarking the performance of such algorithms is not straightforward, although AUC and accuracy (i.e., the number of correct classifications over sample size) are the standard measures reported.[11–13]

Assessing prediction performance primarily using AUC or accuracy can be misleading and is "ill-advised,"[12] especially for rare outcomes.[13;14] High accuracy can be achieved with a simple rule predicting the majority class for all observations, but this will not preform well for metrics centered on true positives. Previous work has also highlighted that when the outcome is rare, other measures, such as the percentage of true positives among all predicted

[1]Department of Population Health, New York University School of Medicine
[2]Department of Health Care Policy, Harvard Medical School
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[4]Department of Surgery, Massachusetts General Hospital
[5]Center for Health Policy, Stanford University

Corresponding author:
Sherri Rose
Email: sherrirose@stanford.edu

Prepared using sagej.cls

positives (i.e., positive predictive value)[15] and related precision-recall curves (i.e., plots of positive predictive value vs. true positive rate),[13;14] can be more informative. These earlier articles featured hypothetical settings with no real data,[13] simulations only,[15] and a lack of cross-validated metrics.[14] There are also many arguments that measures of calibration (i.e., alignment of predicted probabilities with observed risk) for general classification problems are both more interpretable and better assess future performance.[16;17]

Despite these published warnings, many machine learning competitions assign their leaderboard and winners solely on a single metric, typically AUC for binary outcomes. Work published in machine learning conference proceedings, even those events specifically focused on health care, often only consider AUC to compare methods.[18–20] This also occurs in the medical literature.[10;21] The lack of penetrance of the cautions against using a single metric, especially AUC, may be driven by a relative paucity of attention to this issue in translational papers, particularly work demonstrating these problems in real health data analyses. This is a major concern for the biomedical literature given the growing volume of papers applying machine learning to binary prediction problems in health outcomes.

This article provides a more comprehensive, clinically focused study evaluating prediction performance for rare outcomes incorporating (*i*) analyses in registry data, (*ii*) simulations, (*iii*) multiple metrics, (*iv*) multiple algorithms, and (*v*) cross-validated measures. It was motivated by the prediction of medical complications after major surgery. Examples of such complications in cardiac surgeries include in-hospital mortality after a percutaneous coronary intervention[22] and reoperation after valve surgery.[23] Accurate and timely identification of patients who could have complications post-surgery, using characteristics collected prior to the surgery, has the potential to save lives and health care resources. However, many post-surgery outcomes are rare, making prediction a challenging task.

Aortic valve replacement (AVR) is necessary for many patients with symptomatic aortic valve disease,[24;25] and more than 64,000 AVR procedures were performed in the United States in 2010.[26] Mortality is a major risk factor following AVR surgery. Mechanical prosthetic valves are composed of synthetic material requiring anticoagulants following AVR, whereas bioprosthetic valves use natural (animal) cells as a primary material.[23] There are multiple manufacturers for each valve type and manufacturers introduce new generations of earlier valves over time. In addition to demographic features, comorbidities,

medication history, and surgical urgency, the specific valve used is an important predictor of mortality following AVR.[26] However, earlier work predicting mortality outcomes following AVR has been limited to comparing mechanical vs. bioprosthetic values as valve-specific information is typically unavailable.[26] Recent work also demonstrated that bioprosthetic valves had increased mortality for some age groups.[27]

We predicted 30-day and 1-year mortality among patients from a state-mandated clinical registry in Massachusetts who received at least one AVR between 2002 and 2014. Estimation incorporated multiple algorithms typically regarded as performing well with rare outcomes, as well as an ensemble and a new straightforward extension of the lasso for multiple unordered treatments developed here. Our application also expands on earlier applied work by additionally using manufacturer and generation specific subtypes of mechanical and bioprosthetic valves as predictors of mortality. We include wide-ranging simulation studies designed based on our AVR cohort. Our results demonstrate more extreme findings with respect to discordance along performance metrics than seen previously,[14] and algorithms generally performed poorly on measures focused on true positives. Our goal with this work is to provide machine learning practitioners in clinical research with a clear demonstration of the pitfalls of relying on a single metric, and contribute to the body of literature that articulates the need to declare multiple measures.

2 The Statistical Estimation Problem

Baseline predictors are given by vector X of length p and Y is a post-surgery death outcome such that

$$Y = \begin{cases} 1 & \text{if event occurred by time } t \\ 0 & \text{otherwise,} \end{cases}$$

with $t \in \{30 \text{ days}, 1 \text{ year}\}$. X contains a vector V of binary treatment variables representing the distinct aortic valves. The observational unit is $U = (Y, X)$ in nonparametric model $\mathcal{M}$. The goal is to estimate $\psi_0 = E(Y|X) = \Pr(Y = 1|X)$, where the subscript 0 indicates the unknown true parameter, as a minimizer of an objective function:

$$\psi_0 = \arg\min_{\psi} E(L(U, \psi)),$$

with candidate algorithm ψ. We consider the rank loss function, $L(U, \psi) = 1-\text{AUC}$, as the primary global loss function.[28] For a fixed ψ, let $\hat{Y}_1^z, \ldots, \hat{Y}_s^z$ be the predicted probabilities for s outcomes where $Y = 1$ and $\hat{Y}_1^w, \ldots, \hat{Y}_q^w$ be the predicted probabilities for q outcomes where $Y = 0$. Then the AUC associated with ψ is written as: $AUC = \sum_{o=1}^{s} \sum_{r=1}^{q} \mathbb{I}(\hat{Y}_o^z > \hat{Y}_r^w)/(sq)$, where $\mathbb{I}$ is an indicator function.[29] As a secondary global loss function, we use the negative log-likelihood function: $L(U, \psi) = -\log\left[(\psi)^Y (1-\psi)^{1-Y}\right]$.

2.1 Estimation Methods

We consider multiple candidate algorithms ψ typically regarded as performing well with rare outcomes, as well as an extension of the lasso for multiple unordered treatments and an ensemble[30] of these algorithms that optimizes with respect to a global loss function. The global loss function optimized for the ensemble can be different than the loss function optimized within each candidate algorithm. Existing methods for rare outcomes used in our data analysis include lasso,[31] logistic regression with Firth's bias reduction,[32;33] group lasso,[34] sparse group lasso,[35] random forest,[36;37], and logistic regression. These algorithms plus gradient boosted trees,[38;39] Bayesian additive regression trees (BART),[40] neural networks,[41;42] and support vector machines (SVMs)[43] were implemented in our simulations studies to expand the set of candidate algorithms to reflect additional widely used tools. Our extension of the lasso, used in both the data analysis and simulations, involves excluding the covariates for treatment from the penalty term, and is described in further detail below. Thus, the ensemble averaged over seven algorithms in our data analysis and eleven in our simulations.

The super learner creates an optimal weighted average of learners that will perform as well as or better than all individual algorithms with respect to a global loss function, and has various optimality properties discussed elsewhere.[30] We note that while we consider the rank and negative log-likelihood loss as global loss functions, we could alternatively have chosen to optimize our super super with respect to classification criteria.[44] As is common in the applied literature, our goal is instead to obtain the best estimator of $\Pr(Y = 1|X)$ while still being interested in evaluating performance by mapping the predicted probability into a classifier.

Five of the individual techniques, including our extension, are forms of penalized regression, which have gained traction in the clinical literature for

their potentially beneficial performance for rare outcomes.[45] In penalized regression methods, a penalty function $P(\beta)$ with preceding multiplicative tuning parameter λ is introduced in the objective function for coefficients β to reduce bias at the cost of increased variance: $L(U,\psi) + \lambda P(\beta)$. Lasso regression has a penalty function $P(\beta) = ||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ with l_1 norm $||.||_1$, such that some βs may be shrunk to exactly zero as the value of the tuning parameter λ increases.[31] Previous work targeting causal parameters in comparative effectiveness research for multiple unordered treatments demonstrated that penalized regression methods may shrink all treatment variables V to zero or near zero.[46] Thus, although we target a different parameter, in our extension of the lasso the coefficients for the binary indicators for the multiple treatments V are excluded from the penalty function. Suppose these treatment coefficients within $j = \{1, \ldots, p\}$ are from k to p, we can then write this penalty as: $P(\beta) = \sum_{j=1}^{k-1} |\beta_j|$. We refer to the procedure as the treatment-specific lasso regression.

Of course, logistic regression is a well-known parametric tool for predicting binary outcomes. Logistic regression with Firth's bias correction is a method that can reduce the bias of regression estimates due to separation in small samples with unbalanced classes (e.g., a rare outcome). It produces finite parameter estimates by means of penalization with the Fisher information matrix $I(\beta)$.[32] We can translate this into the above framework with penalty $P(\beta) = \log|I(\beta)|$ and fixed $\lambda = 1/2$. To account for the categorical variables (i.e., sets of binary indicators for the levels in each category of a predictor) found in our application, we also consider group lasso regression[34] and sparse group lasso regression.[35] Group lasso regression enforces regularization on sets of variables (i.e., groups) rather than individual predictors. Let G be the total number of groups in X and p_g be the number of covariates in each group g. The penalty function for the group lasso regression can then be written as $P(\beta) = \sum_{g=1}^{G} \sqrt{p_g}||\beta_g||_2$, where $||.||_2$ denotes an l_2 norm of a vector. For sparse group lasso regression, regularization is enforced both on the entire group of predictors as well as within each group. The penalty term for the sparse group lasso regression is given as $P(\beta) = (1-\alpha)\sum_{g=1}^{G} \sqrt{p_g}||\beta_g||_2 + \alpha||\beta||_1$, where $\alpha \in [0,1]$.

Three of the other techniques are ensembles of classification trees. Classification trees generally rely on recursive binary partitioning of the predictor space to create bins that are highly homogenous for the outcome. A key benefit of tree-based methods comes from their ability to identify nonlinear

relationships between the outcome variable and predictors.[2] Random forests aggregates multiple classification trees, each with a random selection of input predictors. Whereas in gradient boosted trees, multiple trees are trained in an additive and sequential manner based on residuals to improve outcome classification.[38;39] BART is a Bayesian approach leveraging regularization priors and posterior inference to estimate ensembles of trees.[40] The set of priors for the tree structure and leaves prevent any single tree from dominating the overall fit and a probit likelihood is used in the terminal nodes.

The last two algorithms considered are feed-forward neural networks and SVMs, originally called support vector networks. Neural networks follow an iterative procedure to estimate the relationships between variables with layers of nodes, where number of units in the hidden layer is a hyperparameter that needs to be specified.[41] SVMs aim to find optimal partitions of the data across a decision surface using a hinge loss and a pre-specified kernel function.[43] These two algorithms, along with gradient boosted trees and BART described directly above, are applied in our simulation studies only.

Our implementation of these candidate algorithms ψ relies on the `R` packages **glmnet**[47] (lasso and treatment-specific lasso), **brglm**[48] (logistic regression with Firth's bias correction), **gglasso**[49] (group lasso), **msgl**[50] (sparse group lasso), **randomForest**[37] (random forests), **xgboost**[51] (gradient boosted trees), **bartMachine**[52] (BART), **nnet**[53] (neural networks), **kernlab**[54](SVM), and **SuperLearner**[55] (ensemble). Internal tuning parameters for each algorithm were selected using nested cross-validation.[56] Additional details regarding the hyperparameter tuning are included in the supplemental material. As a benchmark for thresholding these prediction algorithms, a naive prediction rule that assigns every patient to the majority class is also considered. In our application, the majority class is surviving up to time $t \in \{30 \text{ days}, 1 \text{ year}\}$. We provide code to implement these estimators in a public GitHub repository `github.com/SamAdhikari/PredictionWithRareOutcomes`.

2.2 Evaluation Measures

We consider a suite of metrics for evaluating algorithms with out-of-sample prediction probabilities in stratified 5-fold nested cross-validation. The stratification refers to distributing our rare outcomes across the cross-validation folds to ensure that a roughly equal number of events occur in each fold. These measures have been previously identified in the literature as capturing differing

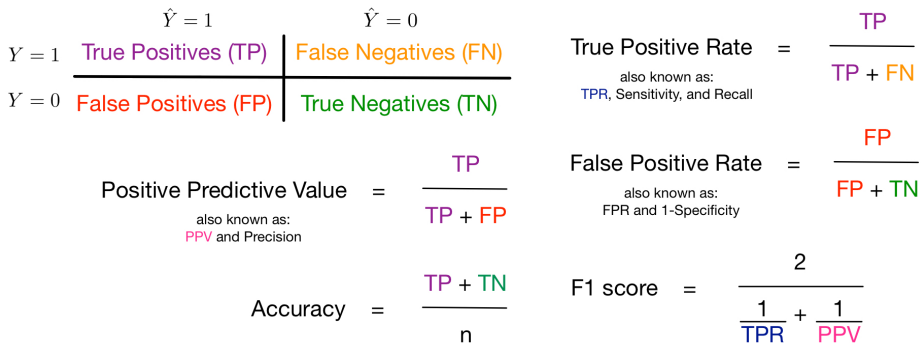

Figure 1. Summary of Several Evaluation Measures under Simple Random Sampling. *Not displayed: area under the receiver operator characteristic curve (also known as c-index), precision-recall plots, and plots of the percentage of true positive outcomes across risk percentiles*

aspects of algorithm performance. We extend the common approach where a fixed probability threshold is used to assign patients into different outcome classes by building a flexible thresholding rule. For a set of candidate thresholds between 0 and 1, accuracy is computed at each threshold. The threshold at which overall accuracy is maximized is then selected as the threshold for assigning a patient to an outcome class for that particular algorithm. If there are multiple thresholds that maximize the accuracy, the minimum of those thresholds is used.

The threshold selected is used to compute multiple evaluation metrics. In Figure 1, we use the notation in the 2×2 contingency table to formally define five of the evaluation metrics, where $\hat{Y}$ is the predicted outcome. Positive predictive value (PPV) is the proportion of true positive outcomes over the number of predicted positive outcomes and accuracy is the overall proportion of true positives and true negatives for n total observations. The true positive rate (TPR) is the proportion of true positive outcomes over the number of observed positive outcomes and the false positive rate (FPR) is the proportion of false positive outcomes over the number of observed negative outcomes. Finally, F_1 score is computed as the harmonic mean of TPR and PPV giving equal weight to precision and recall. We note that our naive prediction rule discussed in Section 2.1 will have high accuracy, but a TPR of zero. (We also consider a threshold where TPR is maximized in sensitivity analyses.) At the specified threshold, we present the mean evaluation metrics averaged over the cross-validation folds as well as the 95% confidence interval around the mean estimates computed using the standard error.

As discussed earlier, many medical machine learning applications claim a good prediction tool will have high AUC and accuracy. However, these results can be misleading, especially for rare outcomes, and it has been argued that use of PPV and precision-recall curves might be more informative.[14] We do consider AUC given its pervasiveness in the literature, as presented in Section 1, but also precision-recall curves, which plot PPV versus TPR. Precision-recall curves are favored in some previous literature due to their focus on true positive classifications among the overall positive classifications, which may lead to better insights into future predictive performance.[14] Assessing calibration is also important in projecting future predictive performance. We consider one variation of a graphical representation of calibration[16;57] with bar plots of the percentage of true positive outcomes across estimated risk percentiles, however alternative model-based approaches also exist.[58] Our set of evaluation measures is not exhaustive and debates regarding the best metrics continue in the scientific discourse. However, this collection represents an entry point for illustrating the perils of relying on AUC and accuracy.

3 Predicting Mortality After AVR

Our study data are from a state-mandated clinical registry coordinated by the Massachusetts Data Analysis Center.[59] The data included all AVRs performed between 2002 and 2014 in all nonfederal acute care Massachusetts hospitals for patients at least 18 years of age, regardless of health insurance status. We considered patients with AVR procedures only as well as other cohorts that included combination procedures. This resulted in four different cohorts: isolated AVR, patients who had either isolated AVR or a combination of AVR and mitral valve replacement procedures, patients who had either isolated AVR or a combination of AVR and coronary bypass surgery procedures, and patients in all of the previous cohorts combined. The endpoints of interest in the analysis are short-term (within 30 days) and long-term (within 1 year) mortality outcomes following AVR, recorded between 2002 and 2015, including patients who had at least one year of follow-up. Loss to follow-up was minimal across the cohorts (i.e., 2-6%) and noninformative (e.g., included patients who had moved out of state).

We focus on the isolated AVR cohort in the main text with the three other cohorts presented in the supplemental material. Table 1 displays the mortality rates overall in the isolated AVR cohort as well as subgroups for mechanical and

	Mortality Rate (%)	
	30 Day	1 Year
Overall Cohort (n)		
Isolated AVR (6472)	1.8	5.3
Mechanical (n)		
Group 1 (34)	5.9	11.7
Group 2 (67)	1.5	2.9
Group 3 (27)	3.7	3.7
Group 4 (685)	1.3	3.8
Group 5 (107)	0.9	2.3
Group 6 (248)	1.6	4.4
Bioprosthetic (n)		
Group 7 (361)	1.7	6.9
Group 8 (*)	*	*
Group 9 (299)	1.7	5.0
Group 10 (505)	1.4	3.9
Group 11 (149)	2.0	4.7
Group 12 (2308)	2.2	6.6
Group 13 (381)	0.5	4.2
Group 14 (1304)	1.8	5.1

Table 1. Observed Mortality Rates in Isolated AVR Cohort. *Cells with* < 10 *events were suppressed and replaced with *.*

bioprosthetic valves. These groupings were constructed based on the features of the devices, such as manufacturer and generation-specific information, and clinical expertise. They also serve as the binary indicator variables for our multiple treatments. We note that our overall 30-day and 1-year mortality rates of 1.8% and 5.3%, respectively, may or may not be deemed 'rare' depending on the definition considered, which can range from well under 1% to over 10%. As discussed in Section 1, the broader setting of imbalanced outcome classes clearly applies here.

Baseline predictors included demographic information (e.g., age, sex, race/ethinicity, and type of health insurance); comorbidities; family history of cardiac problems; cardiac presentation prior to the AVR (e.g., ejection fraction, cardiac shock, and acute coronary syndrome status); procedure-specific information (e.g., type of procedure performed); hospital; and medication history. Key baseline covariates measured prior to the valve replacement surgery are summarized in Table 2 with the complete list of covariates for all four cohorts presented in the supplemental material. We note that use of race and ethnicity in risk prediction algorithms should include thoughtful consideration regarding what these variables represent (e.g., structural racism) and how they may perpetuate health inequities if an algorithm is deployed.[60] Our

Predictors	30-Day Mortality		1-Year Mortality	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
Demographic				
Age (mean, years)	73	68	73	68
Height (mean, cm)	166	168	168	169
Weight (mean, kg)	80	83	80	84
Male (%)	52	58	56	58
Race/Ethnicity (%)				
White	89	92	92	92
Black	4	2	3	2
Hispanic	4	3	2	3
Comorbidities (%)				
Diabetes	39	26	37	26
Hypertension	84	73	78	73
Left main disease	7	2	4	2
Previous cardiovascular Intervention	34	23	32	22
Medication (%)				
Betablocker				
Yes	61	42	51	47
Contraindicated	2	6	5	6
Anticoagulation				
Yes	17	12	23	12
Contraindicated	0	2	1	2

Table 2. Key Baseline Predictors Measured Prior to Surgery in Isolated AVR Cohort.

algorithms had poor performance and we do not recommend them; this issue was important to raise as these variables were included. Continuous covariates were standardized to have mean zero and a standard deviation of one. Covariates with $> 10\%$ missingness were deemed unreliable and excluded from the analyses. For covariates with $< 10\%$ missingness, we introduced a missingness indicator variable.[61]

As described in Section 2.1, the out-of-sample cross-validated predicted probabilities were first deployed to select a prediction threshold that maximizes the accuracy for each algorithm in each fold, with sensitivity analyses in the supplemental material maximizing TPR. Analyses using the rank loss function and negative log-likelihood loss function were performed separately, with negative log-likelihood loss function performance reported in the supplementary material (results were similar). The top row and first panel of the second row in Figure 2 presents the accuracy, FPR, PPV, and TPR evaluation metrics for 20 prediction thresholds between 0 and 1 for the 30-day mortality outcome. We observe that three algorithms underperformed the naive prediction rule

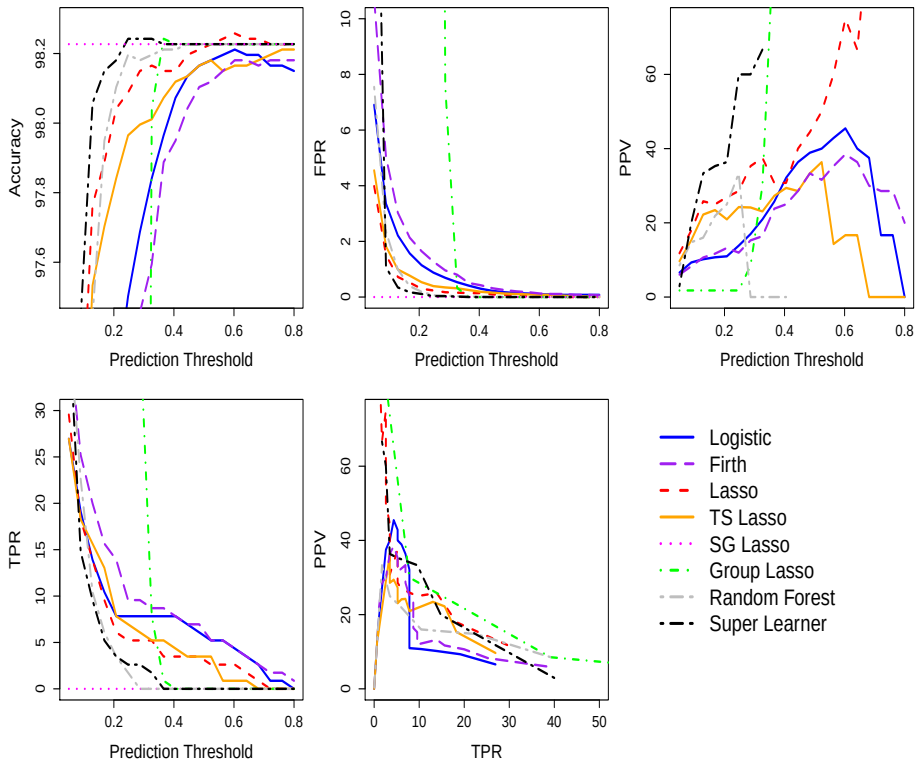

Figure 2. Data Analysis: Cross-Validated Algorithm Performance by Prediction Threshold and Precision-Recall Plot for 30-Day Mortality in Isolated AVR Cohort. *Prediction threshold chosen to maximize accuracy. Plots display the mean over 5-folds at each threshold value. For algorithms with TPR equal to zero, PPV is undefined and not plotted. In this isolated AVR cohort, the naive prediction rule accuracy is the same as the sparse group (SG) lasso. TS is an abbreviation for treatment-specific.*

with respect to accuracy: logistic regression, logistic regression with Firth's bias correction, and the treatment-specific lasso. FPR values quickly converged toward zero as the prediction threshold increased for all algorithms except group lasso. PPV performance varied by algorithm, although multiple algorithms did not exceed a mean of 50% for any threshold. (See supplementary material for 1-year mortality outcome figure.)

Figure 3 displays results at the selected threshold in the isolated AVR cohort for accuracy, AUC, FPR, TPR, PPV and F_1 score. Accuracies were high for both outcomes and flat across all algorithms: 98% for 30-day mortality and 95% for 1-year mortality. AUC ranged from 57 to 74% for 30-day mortality and 73 to

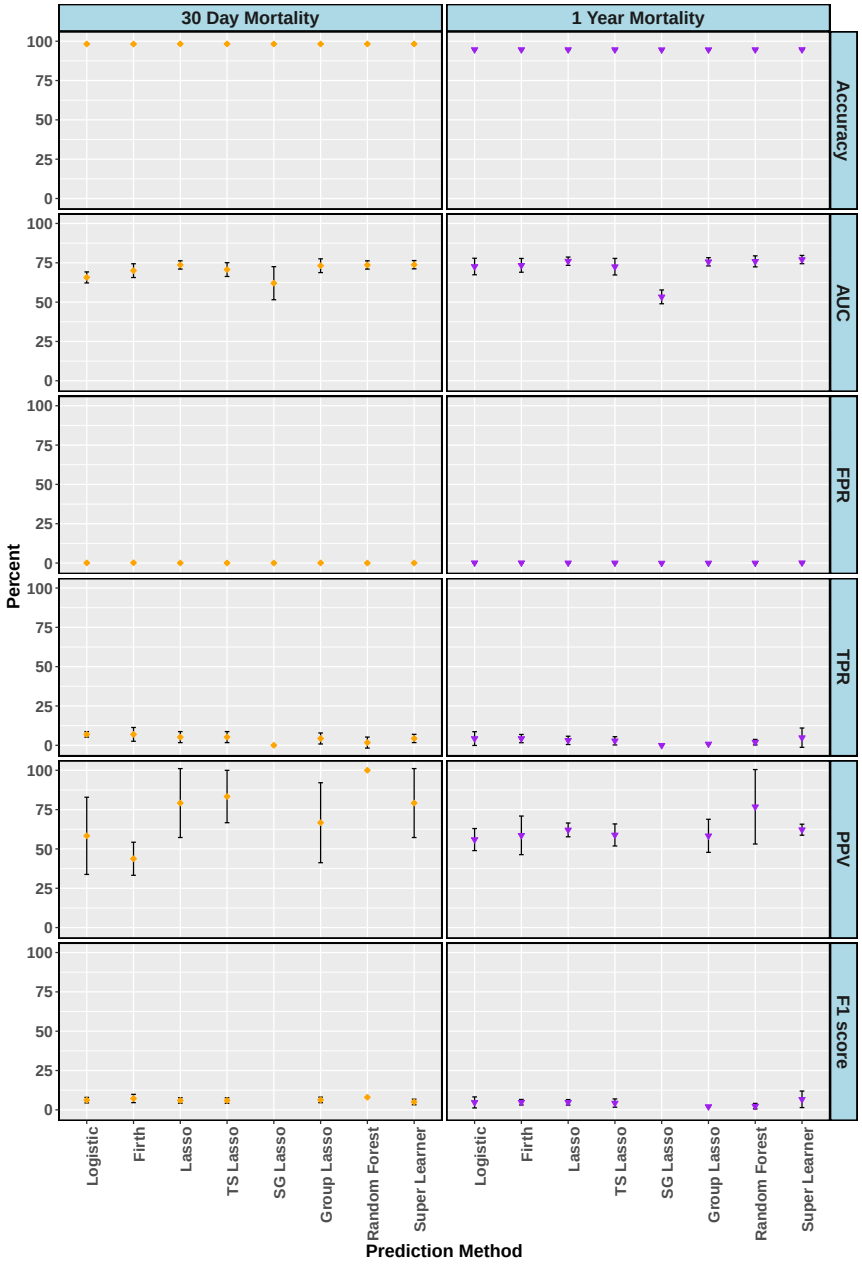

Figure 3. Data Analysis: Cross-Validated Algorithm Performance in Isolated AVR Cohort. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. TS is an abbreviation for treatment-specific and SG is for sparse group.*

76% (except sparse group lasso at 53%) for 1-year mortality. AUC values in the 70s are often reported in published clinical analyses.[10;18–20;62] The FPRs were desirably low; less than or equal to 0.1% for all algorithms and both outcomes. However, Figure 3 also shows extremely poor TPR, with all algorithms less than 7% and sparse group lasso at exactly zero. We found moderate PPVs (only defined for algorithms where the number of predicted positive values was nonzero), with random forests an outlier at 100% PPV for 30-day mortality. This 100% PPV could easily be misinterpreted, however, if not additionally noted that it was paired with a 2% TPR. The precision-recall plot in Figure 2 for 30-day mortality also highlights poor TPR and weak PPV performance. Many algorithms did not have a mean of at least 50% PPV (precision) for any level of TPR (recall). F_1 scores (when not undefined due to undefined PPV) were poor with all values less than 7%. Plots of the percentage of true positive outcomes across risk percentiles show less than 12% of true positive outcomes in the top ventile for 30-day mortality and less than 25% for 1-year mortality (see supplementary material for figures). Thus, overall we found that all algorithms demonstrated poor performance for predicting two mortality outcomes after AVR surgery, despite high accuracy values and moderate measures of AUC.

4 Simulations

A large set of simulation studies for the two mortality outcomes and all four cohorts was designed to evaluate our findings in the context of a known data generating distribution, while also exploring additional settings and algorithms. These simulations were based on the real data from Section 3. For each cohort, a matrix of predictors X_{sim} was simulated to resemble the observed data. Nine positive continuous predictors, representing age, height, and weight, among others, were simulated from a truncated Normal distribution, such that for each continuous covariate h, $X^h_{\text{sim}} \sim N(\mu_h, \sigma^2_h)$ with $X \in [a_h, b_h]$. The means μ_h and variances σ^2_h were estimated as the empirical means and variances within the observed cohorts, while the lower limits a_h and the upper limits b_h were estimated as the minimums and maximums for each variable in the observed data.

Thirty-six binary predictors were simulated using Bernoulli distributions: $X^d_{\text{sim}} \sim \text{Bern}(e_d)$, where e_d is the observed proportion of events for each binary covariate d. Seven categorical predictors with multiple levels, including the valve types, were simulated from multinomial distributions using the observed

proportions as the probabilities for each category. These categorical variables were converted into binary predictors, resulting in 72 total predictors. Finally, simulated outcomes were generated for each of the two endpoints such that $Y_{\text{sim}} \sim \text{Bern}(e_y)$ and $\text{logit}(e_y) = X_{\text{sim}}\beta$, where e_y is the proportion of events and βs were assigned using estimated coefficients from the observed AVR cohorts.

We explored three distinct settings with increasing complexities in data generation and covariate selection choices. In simulation setting 1, we used the same set of predictors for both data generation and prediction, assuming the analyst had access to all the predictors that generated the data. Specifically, all main effects of the available predictors X_{sim} were used to generate the outcome and to estimate ψ_0. In simulation setting 2, complex nonlinear functions of predictors, including interaction terms and quadratic forms, were used for data generation. However, this nonlinearity in predictors was essentially ignored while estimating ψ_0; these specifications were not explicitly provided to algorithms with a strict functional form although the random forests were not restricted from discovering interactions.

In simulation setting 3, we omitted a portion of predictors while generating the data, such that only a subset of X_{sim} was used to create Y_{sim}. The predictors that were omitted (i.e., did not contribute information to the generation of Y_{sim}) were intentionally introduced into the estimation of ψ_0, whereas a separate subset of predictors that were used for data generation were omitted from the estimation step. Estimation in this setting also did involve variables from X_{sim} that were part of both data generation and available for inclusion in the algorithms. This last setting represents the realistic scenario where important true predictors are not available for building a prediction function and uninformative 'noise' variables are included instead.

Further details on the construction of the simulations are available on our companion GitHub page with `R` code `github.com/SamAdhikari/PredictionWithRareOutcomes` as well as in our supplementary material. Mirroring the data analyses in Section 3, simulation results based on the isolated AVR cohort are discussed here with the additional results included in the supplemental material. We also present true conditional risk estimates based on the AUC loss function using the true data-generating probabilities and the simulated outcomes to compare with the cross-validated AUC estimates.

Figures 4 and 5 display evaluation metrics for 30-day and 1-year mortality outcomes, respectively, across the three simulation settings. These metrics were

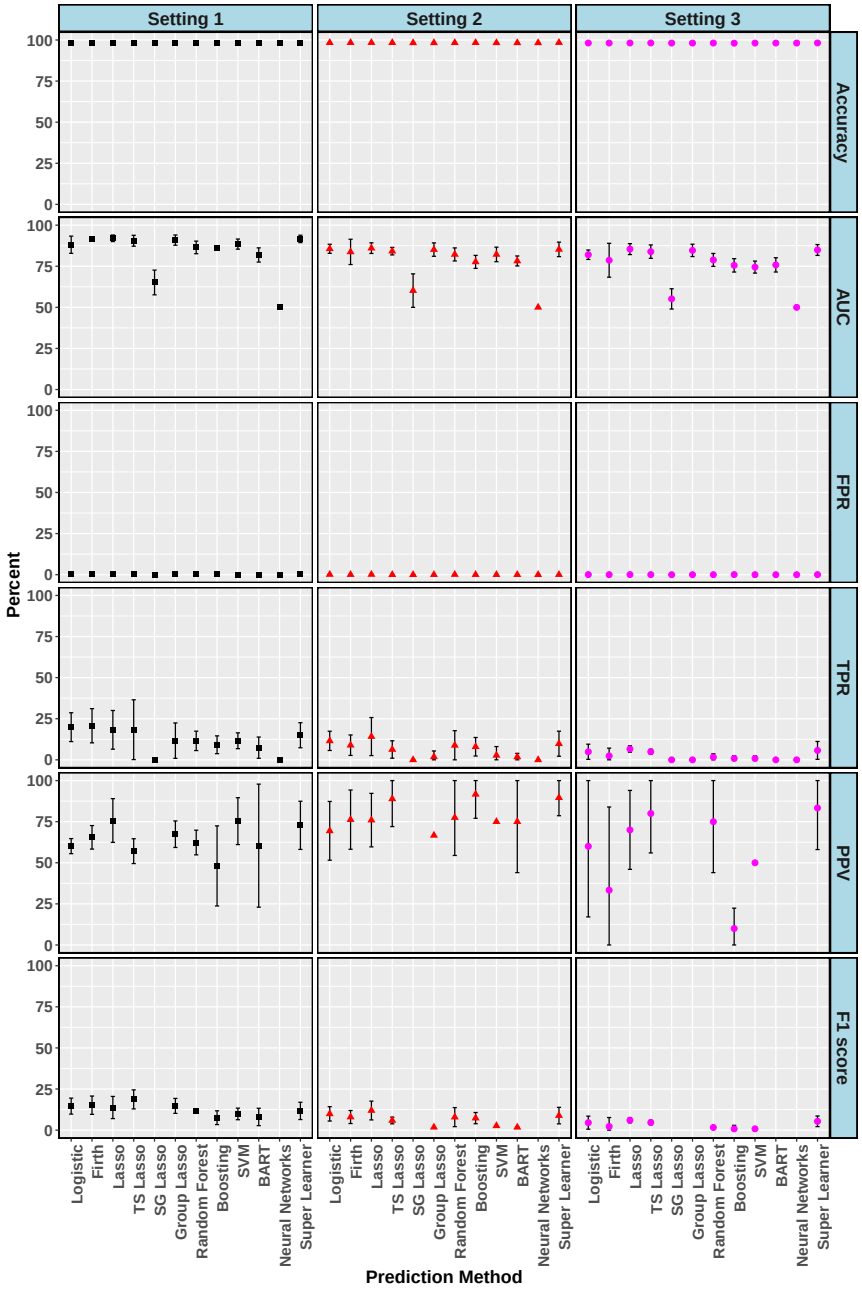

Figure 4. Simulation: Cross-Validated Algorithm Performance for 30-Day Mortality in Isolated AVR Cohort. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 94% for setting 1, 89% for setting 2 and 95% for setting 3. TS is an abbreviation for treatment-specific and SG is for sparse group.*

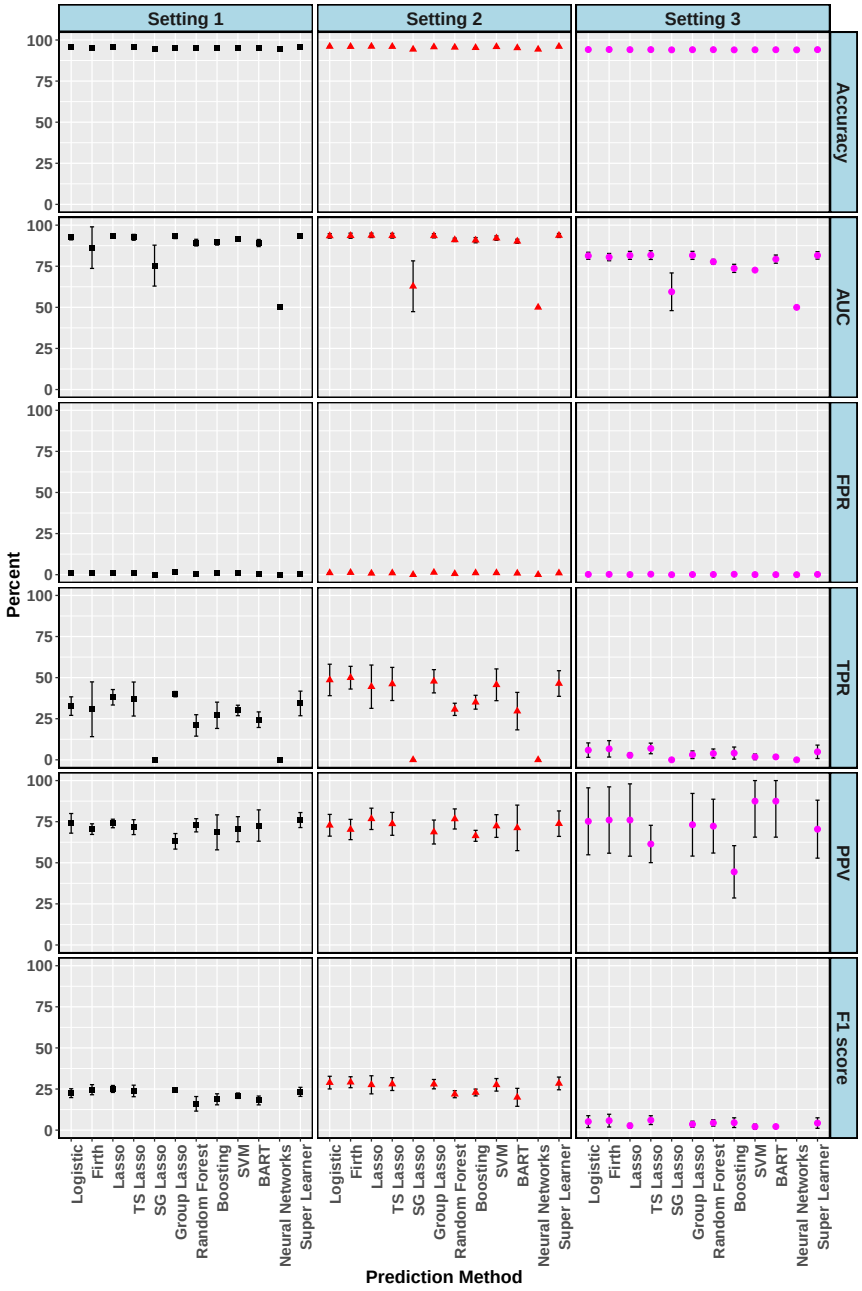

Figure 5. Simulation: Cross-Validated Algorithm Performance for 1-Year Mortality in Isolated AVR Cohort. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 94% for settings 1 and 2 and 95% for setting 3. TS is an abbreviation for treatment-specific and SG for sparse group.*

computed at the best threshold for each algorithm, as in the data analyses. High accuracies and near zero or zero FPRs were seen in all settings and both outcomes. Setting 3, reflecting the realistic scenario with missing predictors and 'noise' variables, had the worst performance for both outcomes with near zero or zero TPR and near zero F_1 scores. Broadly, setting 3 was the most similar to the results found in our data analyses. In settings 1 and 2 for 1-year mortality, AUC hovered around 90% for most algorithms, PPV was about 75%, and F_1 score approximately 25%. TPRs even reached the improved level of 40% in setting 1 and 50% in setting 2. The four additional algorithms added to our simulations did not appreciably improve performance, with gradient boosted trees, BART and SVMs having similar metrics to other algorithms. Neural networks was particularly poor, achieving the worst AUC values paired with 0% TPRs across settings and for both outcomes. In sensitivity analyses (see supplemental material) considering thresholds where TPR is maximized rather than accuracy, we saw improvements in TPR, although all values were less than 75% when considering those with low FPR. Values of 100% TPR were achieved, but only paired with 100% FPR. However, in the realistic setting 3, most TPR values were less than 50%, with two exactly zero. There was also a massive drop in PPV for every algorithm across each setting with all below 25%.

Lastly, for the isolated AVR cohort, we made modifications to the generation of the outcome in each of the three simulation settings to investigate whether performance improved as class imbalance for the outcome decreased. Simulated data were created with five different mortality rates ranging from 10 to 50%, with 50% representing no class imbalance. For simulation setting 3, shown in Figure 6, TPR was always less than 75% when the mortality rate $\leq 40\%$ and less than 50% when the mortality rate $\leq 30\%$. With only one exception (10% rate with super learner algorithm), PPV was also less than 75% for all mortality rates, even with no class imbalance. Further graphical results are displayed in the supplementary material and we summarize several additional key findings from those results here. In simulation settings 1 and 2, TPRs improved with increasing mortality rate and were even around 75% for mortality rates $\geq 40\%$ for the majority of the algorithms, although still paired with low F_1 scores. Plots of the percentage of true positive outcomes across risk percentiles for setting 1 showed improved calibration as mortality rate increased. At 10% mortality, most algorithms approached 80% of true positive outcomes in the top ventile and, at 50% mortality, these values approached 100%. We also explored the impact of

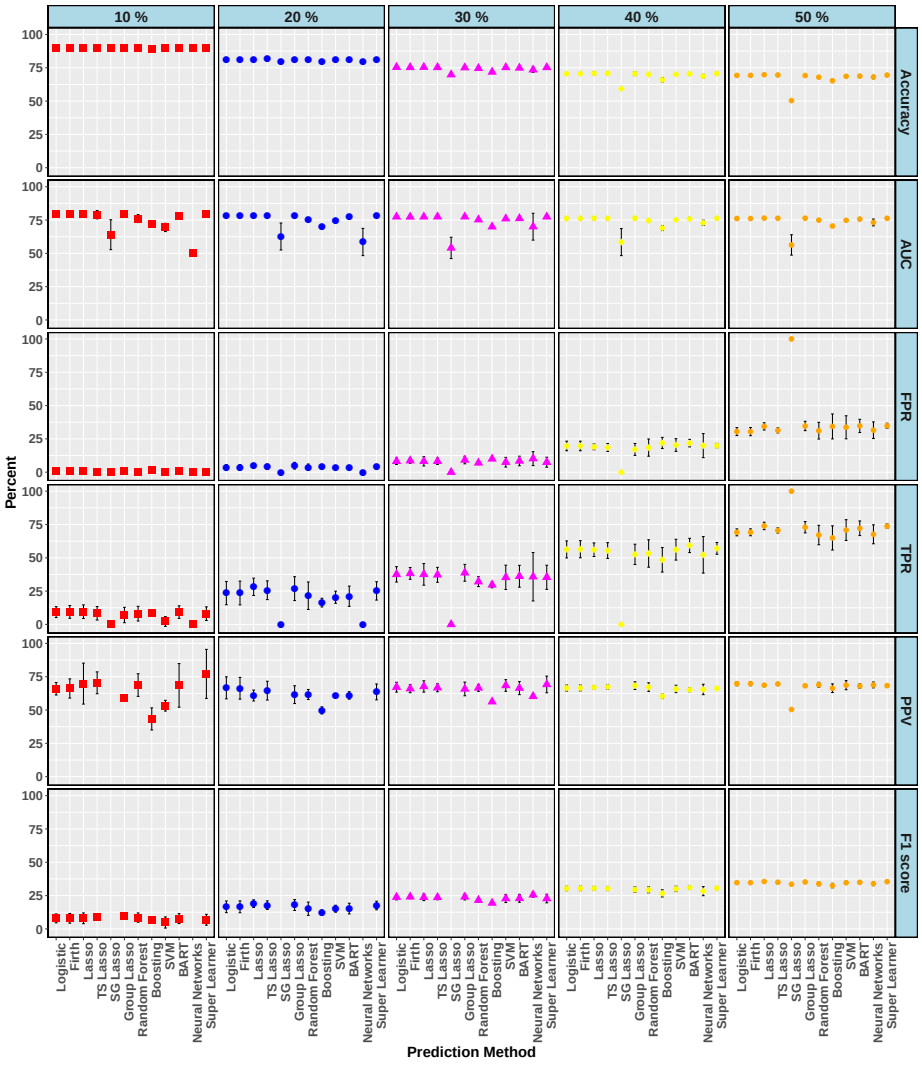

Figure 6. Simulation: Cross-Validated Algorithm Performance for Varied Class Balance in Simulation Setting 3. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 94% (for 10% event rate), 92% (for 20%, 30%, and 50% event rates), and 91% (for 40% event rate). TS is an abbreviation for treatment-specific and SG is for sparse group.*

the number of cross-validation folds (5, 10, or 15) and found that the results were similar across fold choice.

5 Discussion

AUC and accuracy measures are commonly reported in medical and machine learning applications to assess prediction functions for binary outcomes. However, when the outcome of interest is rare, prediction performance with these metrics can be misleading; evaluations using one or two metrics will not be sufficient. Even when both AUC and accuracy are high, we found that TPR, PPV, F_1 score or graphical presentations of the percentage of true positive outcomes across risk percentiles can be poor. The TPR, F_1 score, precision-recall curves, and percentile plots were the most consistent metrics for correctly identifying the poor performance in our data analyses and simulations. Although it should be noted that F_1 score may still be low with no class imbalance, low FPR, and high TPR, PPV, AUC, and accuracy. We found that PPV (i.e., not paired with TPR in a precision-recall curve) was sometimes misleading (e.g., 100% when TPR was near zero) and not necessarily more informative than AUC and accuracy, which further distinguishes our work from some previous studies.[15] Overall, our results are also more extreme with respect to discordance between measures (e.g., near perfect accuracy paired with near-zero TPR) compared to earlier works.

As one might imagine, this endeavor began as a study to design new prediction functions for 30-day and 1-year mortality in four AVR cohorts created from registry data. We aimed to build a tool leveraging the best existing algorithm options for rare outcomes while also proposing a new methodological extension of our own that was specific to multiple unordered treatments. What we found was unexpected – none of the methods yielded a usable prediction function, and far from it. Futhermore, this was only discovered because we considered a large suite of evaluation metrics. Had we been functioning in the common scenario where only AUC or accuracy were measured, we would have declared strong performance, and perhaps suggested that our tools had practical relevance for applied settings. As our data analyses and simulations demonstrated, high AUC and accuracy can be accompanied by extremely low TPR for predicting both short-term and long-term mortality following AVR.

One of the "strengths" of registry data, often touted as an advantage over claims databases, is the availability of detailed clinical information. We had access to dozens of relevant variables for prediction of mortality following AVR and were unable to develop a prediction function we could recommend. The prediction of medical complications after major surgery is a critical need, as

accurate identification of patients at risk for serious complications post-surgery could save lives and preserve health care resources. Thus, it is regrettable we do not offer tools to contribute to this crucial area. However, it is always important to recognize that all data sources will not be appropriate for solving all research questions; this registry and other registries have proven valuable for many other settings.

Where our contribution does lie is in providing a more comprehensive, clinically focused evaluation of prediction performance with rare outcomes featuring (*i*) a relevant AVR data set, (*ii*) an array of simulations, (*iii*) multiple varied evaluation measures, (*iv*) parametric and machine learning algorithms, and (*v*) cross-validated metrics. We aimed to give clinical and machine learning researchers a realistic medical example of the dangers in relying on a single measure of discriminatory performance to evaluate performance. Additionally, we provide reproducible `R` code for our simulation study algorithms and evaluation measures on a companion GitHub page.

Another way the machine learning literature has dealt with class imbalance is by resampling the original dataset, either oversampling the minority (i.e., rare) outcome class or undersampling the majority class.[63–66] The synthetic minority oversampling technique (SMOTE) is one popular approach among these methods, relying on k-nearest neighbors to oversample observations with the rare outcome.[63] However, the feasibility of SMOTE and similar techniques in high-dimensional data is not clear, which is why we did not consider them here.[65] Other considerations include the lack of practical guidelines or procedures to select the rates of oversampling, especially when the outcome is extremely rare.

We also only explored global fit measures. Particularly when making claims that a tool is ready to be deployed in practice, developers must evaluate whether the algorithm has the potential to cause harm, especially to marginalized groups. Additionally calculating group fit measures is a critical (but not sufficient) step in assessing algorithms for fairness.[67] Constrained and penalty regression methods have been developed in this literature that aim to balance more than one metric, such as an overall fit metric and a group fit metric.[68–70] These techniques have largely been applied in other fields, such as criminal justice, with limited use in health care.[71] Related work in constrained binary classification is highly relevant, where algorithms can be designed to optimize TPR subject to a maximum level of positive predictions, for example.[44] Partial AUC methods can also restrict to a range of acceptable values of TPR or FPR.[72]

We recommend that medical studies focused on prediction, particularly with rare outcomes, should not report AUC or accuracy as a primary metric and minimally report a suite of metrics to have a more complete understanding of algorithm performance. Our simulation studies varying the level of class imbalance in the outcome indicate that any class imbalance can lead to problematic performance. Finally, methodological development of additional algorithms for rare outcomes targeting constrained loss functions optimizing multiple metrics as well as resampling-based approaches are promising future directions.

6 Acknowledgement

We are indebted to the Massachusetts Department of Public Health (MDPH) and the Massachusetts Center for Health Information and Analysis Case Mix Databases (CHIA) for the use of their data.This work was supported by NIH grant number R01-GM111339 from the National Institute of General Medical Sciences in the United States.

References

1. S Rose. Machine learning for prediction in electronic health data. *JAMA Netw Open*, 1(4):e181404, 2018.
2. J Friedman, T Hastie, and T Tibshirani. *The Elements of Statistical Learning*. New York: Springer, 2001.
3. S Rose. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*, 177(5):443–452, 2013.
4. M Makar et al. Short-term mortality prediction for elderly patients using medicare claims data. *Int J Mach Learn Comput*, 5(3):192, 2015.
5. T Verplancke et al. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Med Inf Decis Making*, 8(1):56, 2008.
6. R Pirracchio et al. Mortality prediction in intensive care units with the super icu learner algorithm (SICULA): a population-based study. *Lancet Respir Med*, 3(1): 42–52, 2015.
7. PC Austin et al. Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biom J*, 54(5):657–673, 2012.

8. M Motwani et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J*, 38(7):500–507, 2016.
9. R Shouval et al. Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: a European group for blood and marrow transplantation acute leukemia working party retrospective data mining study. *J Clin Oncol*, 33(28):3144–3152, 2015.
10. RA Taylor et al. Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach. *Acad Emerg Med*, 23(3):269–278, 2016.
11. JA Hanley and BJ McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.
12. NR Cook. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935, 2007.
13. J Lever, M Krzywinski, and N Altman. Points of significance: Classification evaluation. *Nature Methods*, 13:603–604, 2016.
14. T Saito and M Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432, 2015.
15. V Labatut and H Cherifi. Evaluation of performance measures for classifiers comparison. *arXiv preprint arXiv:1112.4133*, 2011.
16. GA Diamond. What price perfection? calibration and discrimination of clinical prediction models. *J Clin Epidemiol*, 45(1):85–89, 1992.
17. B Van Calster et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*, 74:167–176, 2016.
18. AEW Johnson, TJ Pollard, and RG Mark. Reproducibility in critical care: a mortality prediction case study. *Proc of Mach Learn Res*, 68:361–376, 2017.
19. JC Forte et al. Predicting long-term mortality with first week post-operative data after coronary artery bypass grafting using machine learning models. *Proc of Mach Learn Res*, 68:39–58, 2017.
20. H Suresh et al. Clinical intervention prediction and understanding with deep neural networks. *Proc of Mach Learn Res*, 68:322–337, 2017.
21. J Wu, J Roy, and WF Stewart. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care*, 48(6): S106–S113, 2010.
22. RS Resnic et al. Simplified risk score models accurately predict the risk of major in-hospital complications following percutaneous coronary intervention. *Am J*

Cardiol, 88(1):5–9, 2001.

23. SA Huygens et al. Contemporary outcomes after surgical aortic valve replacement with bioprostheses and allografts: a systematic review and meta-analysis. *Eur J Cardiothorac Surg*, 50(4):605–616, 2016.
24. ER Bates. Treatment options in severe aortic stenosis. *Circulation*, 124(3):355–359, 2011.
25. J Joseph et al. Aortic stenosis: pathophysiology, diagnosis, and therapy. *Am J Med*, 130(3):253–263, 2017.
26. DT Du et al. Early mortality after aortic valve replacement with mechanical prosthetic vs bioprosthetic valves among medicare beneficiaries: a population-based cohort study. *JAMA Intern Med*, 174(11):1788–1795, 2014.
27. AB Goldstone et al. Mechanical or biologic prostheses for aortic-valve and mitral-valve replacement. *N Engl J Med*, 377(19):1847–1857, 2017.
28. E LeDell, MJ van der Laan, and M Petersen. AUC-maximizing ensembles through metalearning. *Int J Biostat*, 12(1):203–218, 2016.
29. C Cortes and M Mohri. AUC optimization vs. error rate minimization. *Advances in neural information processing systems*. 313–320. 2004.
30. MJ van der Laan, EC Polley, and AE Hubbard. Super learner. *Stat Appl Genet Mol Biol*, 6(1), 2007.
31. R Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B*, 58(1):267–288, 1996.
32. G Heinze and M Schemper. A solution to the problem of separation in logistic regression. *Stat Med*, 21(16):2409–2419, 2002.
33. G Heinze et al. *logistf: Firth's bias reduced logistic regression*, 2013. URL https://CRAN.R-project.org/package=logistf. R package version 1.0.
34. M Yuan and Y Lin. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B*, 68(1):49–67, 2006.
35. N Simon, J Friedman, T Hastie, and R Tibshirani. A sparse-group lasso. *J Comput Graph Stat*, 22(2):231–245, 2013.
36. L Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.
37. A Liaw and M Wiener. Classification and regression by randomforest. *R News*, 2 (3):18–22, 2002.
38. J Friedman. Stochastic gradient boosting. *Comput Stat Data An*, 38(4):367-378, 2002.
39. J Friedman. Greedy function approximation: a gradient boosting machine. *Ann Stat*, 1189-1232, 2001.

40. H Chipman et al. BART: Bayesian additive regression trees. *Ann Appl Stat*, 4(1): 266-298, 2010.
41. C Bishop and others. *Neural networks for pattern recognition*. Oxford university press, 1995.
42. B Ripley. *Pattern recognition and neural networks*. Cambridge, 1996.
43. H Drucker, C Burges, L Kaufman, A Smola, V Vapnik. Support vector regression machines. *Adv Neural Inf Process Syst*, 155-161, 1997.
44. W Zheng et al. Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies. *Stat Med*, 37(2):261-279, 2018.
45. M Pavlou et al. How to develop a more accurate risk prediction model when there are few events. *BMJ*, 351:h3868, 2015.
46. S Rose and SL Normand. Double robust estimation for multiple unordered treatments and clustered observations: Evaluating drug-eluting coronary artery stents. *Biometrics*, 75:(1):289-296, 2019.
47. J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.
48. I Kosmidis. *brglm: Bias Reduction in Binary-Response Generalized Linear Models*, 2017. URL https://CRAN.R-project.org/package=brglm. R package version 0.6.1.
49. Y Yang and H Zou. *gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm*, 2017. URL https://CRAN.R-project.org/package=gglasso. R package version 1.4.
50. M Vincent. *msgl: Multinomial sparse group lasso*, 2017. URL https://CRAN.R-project.org/package=msgl. R package version 2.3.6.
51. T Chen, T He, M Benesty. *xgboost: eXtreme Gradient Boosting*, 2019. URL https://CRAN.R-project.org/package=xgboost. R package version 0.90.0.2.
52. A Kapelner and J Bleich. *bartMachine: Bayesian Additive Regression Trees*, 2018. URL https://CRAN.R-project.org/package=bartMachine. R package version 1.2.4.2.
53. B Ripley and W Venables. *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*, 2016. URL https://CRAN.R-project.org/package=nnet. R package version 7.3-12.
54. A Karatzoglou and others. *kernlab: Kernel-Based Machine Learning Lab*, 2019. URL https://CRAN.R-project.org/package=kernlab. R package version 0.9-29.
55. E Polley, E LeDell, C Kennedy, and M van der Laan. *SuperLearner: Super Learner Prediction*, 2018. URL https://CRAN.R-project.org/package=SuperLearner. R

package version 2.0-23.

56. GC Cawley and NLC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*, 11(Jul):2079–2107, 2010.
57. EW Steyerberg et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21(1):128, 2010.
58. CS Crowson et al. Assessing calibration of prognostic risk scores. *Statistical methods in medical research*, 25(4):1692-1706, 2016.
59. L Mauri et al. Drug-eluting or bare-metal stents for acute myocardial infarction. *N Engl J Med*, 359(13):1330–1342, 2008.
60. D Vyas et al. Hidden in Plain Sight – Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med*, doi:10.1056/NEJMms2004740, 2019.
61. N Horton and K Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models *Am Stat*, 61: (1):79-90, 2007.
62. SL Bergquist et al. Classifying lung cancer severity with ensemble machine learning in health care claims data. *Proc Mach Learn Res*, 68:25–38, 2017.
63. N Chawla et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*, 16321-357, 2002.
64. X-Y Liu et al. Exploratory undersampling for class-imbalance learning. *IEEE T Syst Man Cy B*, 39:(2):539-550, 2008.
65. R Blagus and L Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14:(106):1471-2105, 2013.
66. M Koziarski. Radial-Based undersampling for imbalanced data classification. *arXiv preprint arXiv:1906.00452*, 2019.
67. A Chouldechova and A Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.
68. M Zafar et al. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *arXiv pre-print arXiv:1610.08452*, 2017.
69. M Zafar et al. Fairness constraints: Mechanisms for fair classification. *arXiv pre-print arXiv:1507.05259*, 2017.
70. A Zink and S Rose. Fair regression for health care spending. *Biometrics*, 76:(3): 973-982, 2020.
71. I Chen et al. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci*, 10.1146/annurev-biodatasci-092820-114757, 2021.

72. LE Dodd and MS Pepe. Partial AUC estimation and regression. *Biometrics*, 59 (3):614-623, 2003.

Statistical Methods in Medical Research
XX(X):1–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Supplemental Material: Revisiting Performance Metrics for Prediction with Rare Outcomes

Samrachana Adhikari[1], Sharon-Lise Normand[2,3], Jordan Bloom[4], David Shahian[4], and Sherri Rose[5]

[1]Department of Population Health, New York University School of Medicine
[2]Department of Health Care Policy, Harvard Medical School
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[4]Department of Surgery, Massachusetts General Hospital
[5]Center for Health Policy, Stanford University

Corresponding author:
Sherri Rose
Email: sherrirose@stanford.edu

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

Supplemental Material: Additional Simulation Study Details

We include additional details of the simulation study modeled after the isolated AVR cohort for 30-day mortality outcome. Our data-generation process for the other cohort simulations is consistent with this general description. Variables with observed rate of less than 1% in our data analyses were not included in the simulations. Further details are found in Section 4 of the manuscript and all simulation code is available on our companion GitHub page `github.com/SamAdhikari/PredictionWithRareOutcomes`.

The nine continuous variables described in Section 4 of our manuscript represented: age, height, weight, cross clamp time, perfuse time, hemodialysis ejection fraction, creatine level, body surface area, and body size. The 36 binary covariates represented: sex, government insurance, HMO insurance, commercial insurance, Medicaid, state-specific government insurance, Medicare, self-insurance/no insurance, first surgery, hypertension, family history of coronary arterial disease, chronic lung disease, immunosuppressant, pulmonary valve disease, coronary valve disease, diabetes, previous cardiovascular intervention, endocarditis, treated endocarditis, previous myocardial infarction, hemodialysis ejection fraction done, congestive heart failure, aortic valve insufficiency, tricuspid valve insufficiency, mitral valve insufficiency, pulmonary valve insufficiency, other cardiac procedure, intraoperative blood products used, left main disease, adenosine disphosphate inhibitors, aspirins, beta blockers, inotropes, steroids, lipid lowering drugs, and anticoagulants.

Seven categorical variables (including the valves) with more than two categories were generated from multinomial distributions, as described in Section 4 of the manuscript. The covariates were race/ethnicity (caucasian, Black, hispanic, other), number of diseased veins (0, 1, 2, 3), New York Heart Association class (Class 1, Class 2, Class 3), ejection fraction (< 30, 30, 40+), hemodialysis ejection fraction method (left ventricular, echo, other), and surgical urgency (elective, urgent, emergent/salvage), along with the 11 valve groups. Outcomes for simulation settings 1, 2, and 3 were generated from Bern(e_{y^1}), Bern(e_{y^2}), and Bern(e_{y^3}), where $e_{y^{(\cdot)}}$ is the proportion of events, respectively with equations shown below.

$$
\begin{aligned}
\text{logit}(e_{y^1}) = & - 3.75 + 0.45\ \text{Age} + 0.13\ \text{Sex} + 0.22\ \text{Caucasian} + 0.24\ \text{Black} \\
& - 0.29\ \text{Hispanic} + 0.06\ \text{Government Insurance} - 0.13\ \text{Government, medicaid} \\
& + 0.05\ \text{Government, medicare} + 0.10\ \text{Government, state specific} \\
& - 0.21\ \text{Commercial Insurance} - 0.19\ \text{HMO Insurance} \\
& + 0.22\ \text{None/Self Insurance} + 0.00\ \text{Family history of CAD} \\
& + 0.05\ \text{Hypertension} + 0.57\ \text{Chronic lung disease} \\
& + 0.15\ \text{Immunosuppressant} + 0.40\ \text{PV disease} \\
& + 0.38\ \text{Coronary valve disease} + 0.42\ \text{Diabetes} \\
& + 0.43\ \text{Endocarditis} - 0.11\ \text{Treated Endocarditis} \\
& + 0.03\ \text{Previous CV Intervention} - 0.01\ \text{Previous MI} \\
& + 0.60\ \text{Congestive heart failure} - 0.11\ \text{NYHA Class 1} \\
& - 0.38\ \text{NYHA Class 2} - 0.37\ \text{NYHA Class 3} - 1.12\ \text{HDEF done} \\
& + 0.01\ \text{HDEF} + 0.32\ \text{HDEF Method LV} + 0.30\ \text{HDEF Method Echo} \\
& - 0.21\ \text{HDEF Method Other} + 0.46\ \text{EF Category} > 40 \\
& + 0.42\ \text{EF Category } 30 - 0.08\ \#\ \text{diseased vein 1} \\
& - 0.05\ \#\ \text{diseased vein 2} - 0.42\ \#\ \text{diseased vein 3} \\
& - 0.11\ \text{AV insufficiency} - 0.12\ \text{MV insufficiency} + 0.16\ \text{TV insufficiency} \\
& - 0.26\ \text{PV insufficiency} + 0.10\ \text{First surgery} - 1.11\ \text{Elective Surgery} \\
& - 0.63\ \text{Urgent Surgery} + 0.26\ \text{ADP Inhibitors} + 0.18\ \text{Anticoagulants} \\
& - 0.23\ \text{Aspirin} + 0.04\ \text{Betablocker} + 0.01\ \text{Inotropes} - 0.62\ \text{Lipid lowering} \\
& + 0.70\ \text{Steroids} - 0.25\ \text{Other Cardiac Procedure} + 0.41\ \text{IBPR} \\
& + 0.22\ \text{Left main disease} - 2.13\ \text{Body surface area} - 0.01\ \text{Body size.} \\
& + 0.79\ \text{Height} + 1.49\ \text{Weight} \\
& + 0.13\ \text{Creatinine} + 0.66\ \text{Perfus time} - 0.42\ \text{Cross clamp time} \\
& - 1.50\ \text{Valve 2} - 1.02\ \text{Valve 4} - 1.51\text{Valve 5} \\
& + 0.00\ \text{Valve 6} - 1.01\ \text{Valve 7} - 0.87\ \text{Valve 9} \\
& + 1.25\ \text{Valve 10} - 1.26\ \text{Valve 11} - 0.93\ \text{Valve 12} \\
& - 1.02\ \text{Valve 13} - 0.84\ \text{Valve 14.}
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}(e_{y^2}) = & - 2.75 + 2.13\ \text{Age} + 0.27\ \text{Sex} + 0.00\ \text{Caucasian} \\
& + 0.19\ \text{Black} + 0.68\ \text{Hispanic} \\
& + 0.48\ \text{Government insurance} - 0.50\ \text{Government, medicaid} \\
& + 0.13\ \text{Government, medicare} - 0.68\ \text{Government, state specific} \\
& + 0.00\ \text{Commercial Insurance} + 0.20\ \text{HMO Insurance} \\
& + 0.53\ \text{None/Self Insurance} + 0.20\ \text{Family history of CAD} \\
& + 0.12\ \text{Hypertension} + 0.00\ \text{Chronic lung disease} \\
& - 0.14\ \text{Immunosuppressant} + 0.35\ \text{PV disease} \\
& + 0.50\ \text{Coronary valve disease} + 0.65\ \text{Diabetes} \\
& + 0.76\ \text{Endocarditis} - 1.28\ \text{Treated Endocarditis} \\
& - 0.37\ \text{Previous CV Intervention} + 0.48\ \text{Previous MI} \\
& + 0.16\ \text{Congestive heart failure} - 0.20\ \text{NYHA Class 1} \\
& - 0.30\text{NYHA Class 2} - 0.34\ \text{NYHA Class 3} \\
& - 0.82\ \text{HDEF done} + 0.00\ \text{HDEF} - 0.13\ \text{HDEF Method LV} \\
& - 0.32\ \text{HDEF Method Echo} - 0.47\ \text{HDEF Method Other} \\
& + 0.68\ \text{EF Category} > 40 + 1.28\ \text{EF Category 30} \\
& + 0.19\ \#\ \text{diseased vein 1} - 0.46\ \#\ \text{diseased vein 2} \\
& - 1.55\ \#\ \text{diseased vein 3} + 0.00\ \text{AV insufficiency} \\
& + 0.32\ \text{MV insufficiency} + 0.00\ \text{TV insufficiency} \\
& - 0.24\ \text{PV insufficiency} + 0.00\ \text{First surgery} - 1.74\ \text{Elective Surgery} \\
& - 1.28\ \text{Urgent Surgery} + 0.53\ \text{ADP Inhibitors} - 0.28\ \text{Anticoagulants} \\
& - 0.26\ \text{Aspirin} + 0.32\ \text{Betablocker} - 0.96\ \text{Inotropes} - 0.58\ \text{Lipid lowering} \\
& + 0.85\ \text{Steroids} - 0.46\ \text{Other Cardiac Procedure} + 0.28\ \text{IBPR} \\
& + 1.12\ \text{Left main disease} - 0.11\ \text{Body surface area} + 0.13\ \text{Body size.} \\
& - 0.24\ \text{Height} + 0.86\ \text{Weight} + 0.13\ \text{Creatinine} \\
& + 1.02\ \text{Perfus time} - 0.57\ \text{Cross clamp time} - 1.48\ \text{Valve 2} \\
& - 1.67\ \text{Valve 4} - 1.81\text{Valve 5} + 0.00\ \text{Valve 6} - 2.34\ \text{Valve 7} \\
& - 1.73\ \text{Valve 9} + 0.75\ \text{Valve 10} - 1.74\ \text{Valve 11} - 1.64\ \text{Valve 12} \\
& - 2.53\ \text{Valve 13} - 1.57\ \text{Valve 14} - 0.39\ \text{Sex} \times \text{Diabetes} - 1.57\ \text{Age}^2 \\
& - 0.81\ \text{Weight} \times \text{Hypertension} + 0.45\ \text{Age} \times \text{Congestive heart failure.}
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}(e_{y^3}) = & - 4.25 - 0.25 \text{ Previous CV Intervention} \\
& + 0.50 \text{ Previous MI} + 0.44 \text{ Congestive heart failure} \\
& - 0.11 \text{ NYHA Class 1} - 0.28 \text{NYHA Class 2} \\
& - 0.28 \text{ NYHA Class 3} - 1.15 \text{ HDEF done} \\
& + 0.11 \text{ HDEF} + 0.05 \text{ HDEF Method LV} \\
& - 0.14 \text{ HDEF Method Echo} - 0.05 \text{ HDEF Method Other} \\
& + 0.73 \text{ EF Category} > 40 + 1.38 \text{ EF Category 30} \\
& + 0.37 \text{ \# diseased vein 1} - 0.11 \text{ \# diseased vein 2} \\
& - 1.18 \text{ \# diseased vein 3} - 0.01 \text{ AV insufficiency} \\
& + 0.45 \text{ MV insufficiency} - 0.11 \text{ TV insufficiency} \\
& - 0.24 \text{ PV insufficiency} - 0.04 \text{ First surgery} \\
& - 1.63 \text{ Elective Surgery} - 1.18 \text{ Urgent Surgery} \\
& + 0.62 \text{ ADP Inhibitors} - 0.08 \text{ Anticoagulants} \\
& - 0.16 \text{ Aspirin} + 0.45 \text{ Betablocker} - 0.81 \text{ Inotropes} \\
& - 0.46 \text{ Lipid lowering} + 0.83 \text{ Steroids} \\
& - 0.45 \text{ Other Cardiac Procedure} + 0.40 \text{ IBPR} \\
& + 1.15 \text{ Left main disease} + 1.43 \text{ Body surface area} \\
& + 0.31 \text{ Body size} - 0.87 \text{ Height} - 0.87 \text{ Weight} \\
& + 0.13 \text{ Creatinine} + 1.01 \text{ Perfus time} - 0.60 \text{ Cross clamp time} \\
& - 0.00 \text{ Valve1} - 1.50 \text{ Valve 2} - 1.57 \text{ Valve 4} \\
& - 1.82 \text{ Valve 5} - 1.23 \text{ Valve 6} - 1.58 \text{ Valve 7} \\
& - 0.99 \text{ Valve 9} + 1.25 \text{ Valve 10} - 1.00 \text{ Valve 11} \\
& - 0.88 \text{ Valve 12} - 1.86 \text{Valve 13} + 0.25 \text{ Valve 14.}
\end{aligned}
$$

Supplemental Material: Additional Tables & Figures

Predictors	Isolated AVR	AVR or AVR & CABG	AVR or AVR & MVR	Any AVR
Age (mean, years)	68	70	68	71
Male (%)	58	63	58	62
Race (%)				
Caucasian	92	93	91	93
Black	2	2	2	2
Hispanic	3	2	3	2
Other	3	4	3	4
Latino (%)	3	3	3	3
Body surface area (mean, m^2)	2	2	2	2
Body Size (mean, cm/kg)	2	2	2	2
Height (mean, cm)	169	169	169	169
Weight (mean, kg)	84	83	84	83
Creatinine (mean, mg/dL)	1	1	1	1
Perfus Time (mean, min)	110	130	115	132
Cross Clamp Time (mean, min)	80	98	84	100
Government Insurance (%)	63	66	63	66
Commercial Insurance (%)	42	40	41	40
HMO Insurance (%)	18	16	17	16
None/Self Insurance (%)	2	2	2	2
Government payor (%)				
Military	1	1	1	1
State specific plan	5	4	5	4
Medicare	50	55	50	55
Medicaid	7	6	7	6
None	37	34	37	34
Medicare Fee-for-Service (%)	13	13	13	12
Hospital ID (%)				
A	15	17	15	14
B	10	10	10	10
C	7	7	7	7
D	16	14	16	14
E	2	2	2	2
F	2	3	2	3
G	6	6	6	6
H	13	13	13	13
I	3	3	3	3
J	7	7	7	7
K	3	3	3	3
L	5	5	5	5
M	4	4	4	4
N	7	6	7	6

Table 1. Baseline Covariates. *Features observed at baseline for each cohort. HMO: health maintenance organization*

Predictors (%)	Isolated AVR	AVR or AVR & CABG	AVR or AVR & MVR	Any AVR
ADP Inhibitors				
Yes	1	2	1	2
Contraindicated	2	1	2	1
Anticoagulants				
Yes	12	16	13	17
Contraindicated	2	1	2	1
Aspirins				
Yes	49	57	49	57
Contraindicated	1	1	1	1
Beta blockers				
Yes	48	56	48	55
Contraindicated	6	5	6	5
Inotropes				
Yes	1	1	1	1
Contraindicated	2	1	2	1
Steroids				
Yes	3	4	3	4
Contraindicated	2	1	2	1
Coumadin	1	1	1	2
Lipid Lowering	41	46	41	45
Intravenous Nitrates	1	2	1	2

Table 2. Medication-Related Baseline Covariates. *Medication used at baseline for each cohort.*

Predictors	Isolated AVR	AVR or AVR & CABG	AVR or AVR & MVR	Any AVR
Family History CAD (%)	17	20	17	20
Hypertension (%)	73	78	73	78
Chronic Lung Disease (%)	17	18	17	18
Immunosuppressant (%)	4	4	4	4
Pulmonary Valve Disease (%)	7	12	7	12
Coronary Valve Disease (%)	13	15	13	15
Diabetes (%)	26	31	26	30
Endocarditis (%)	6	4	7	4
Treated Endocarditis (%)	3	2	3	2
Previous CV Intervention (%)	23	25	23	25
Previous MI (%)	11	19	10	19
Previous MI (within 7 days, %)	11	19	10	19
CHF (%)	36	38	38	38
NYHA Class (%)				
1	5	4	5	4
2	20	19	20	19
3	26	29	27	29
4	49	48	48	48
Cardiogenic Shock (%)	1	1	1	1
Other Cardiac Procedure (%)	6	4	6	4
IBPR (%)	32	37	33	38
Left Main Disease (%)	2	9	2	9
HDEF Done (%)	97	97	97	97
HDEF Method (%)				
Left ventiricular	19	23	19	23
Echo	72	68	72	68
Other	3	4	3	4
EF Category (%)				
<30	7	7	7	8
30	5	7	5	6
40+	88	86	88	86
HDEF (mean)	55	54	55	54
# of Diseased Veins				
0	80	45	80	48
1	10	18	10	17
2	4	16	4	15
3	6	21	6	20
Aortic Valve Insufficiency (%)	69	67	70	67
Mitral Valve Insufficiency (%)	75	74	76	74
Tricuspid Valve Insufficiency (%)	67	65	67	65
Pulmonary Valve Insufficiency (%)	46	43	46	43
First surgery (%)	72	71	71	71
Surgical urgency (%)				
Elective	77	69	76	69
Urgent	22	30	23	30
Emergent or salvage	1	1	1	1

Table 3. Comorbidity-Related Baseline Covariates. *Comorbidities observed at baseline for each cohort. CAD: coronary arterial disease; CV: cardiovascular; MI: myocardial infarction; CHF: congestive heart failure; NYHA: New York Heart Association; IBPR: intraoperative blood products refused; HDEF: hemo data-ejection fraction; EF: Ejection fraction; 'Other Cardiac Procedure' refers to cardiac procedures other than coronary artery bypass grafting (CABG) or valve procedures.*

Prepared using sagej.cls

	Valve Type, % (n)			
	Isolated AVR	AVR or AVR & CABG	AVR or AVR & MVR	Any AVR
Mechanical				
Group 1	0.5 (34)	0.4 (49)	0.6 (42)	0.5 (57)
Group 2	1.0 (67)	0.8 (87)	1.2 (85)	0.9 (105)
Group 3	0.4 (27)	0.3 (32)	0.4 (27)	0.3 (32)
Group 4	11 (685)	8.0 (926)	11 (779)	8.6 (1020)
Group 5	1.7 (107)	1.3 (151)	1.8 (126)	1.4 (170)
Group 6	3.8 (248)	2.9 (332)	3.9 (269)	2.9 (353)
Bioprosthetic				
Group 7	5.6 (361)	5.7 (660)	5.5 (376)	5.7 (675)
Group 8	*	*	*	*
Group 9	4.6 (299)	4.4 (509)	4.7 (312)	4.4 (522)
Group 10	7.8 (505)	8.2 (940)	7.8 (531)	8.1 (966)
Group 11	2.3 (149)	3.0 (350)	2.3 (156)	3.0 (357)
Group 12	36 (2308)	40 (4660)	35 (2385)	40 (4737)
Group 13	5.9 (381)	5.9 (682)	5.9 (402)	5.8 (703)
Group 14	20 (1304)	18 (2127)	20 (1339)	18 (2162)

Table 4. Percentage of Types of Valves in Each Cohort. *These valves are grouped by manufacturer and generation specific subtypes . Cells with < 10 events were suppressed and replaced with *.*

Cohort	n	30 Day (%)	1 Year (%)
AVR or AVR & CABG	11502	2.4	6.9
AVR or AVR & MVR	6824	1.8	5.7
Any AVR	11854	2.4	7.1

Table 5. 30-Day and 1-Year Mortality Rates for Three Cohorts.

	Valve Type (%)			
	30 Day		1 Year	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
Mechanical				
Group 1	1.7	0.5	1.1	0.5
Group 2	0.9	1.0	0.6	1.1
Group 3	0.9	0.4	0.3	0.4
Group 4	7.8	1.1	7.4	11
Group 5	0.9	1.6	0.9	1.7
Group 6	3.4	3.8	3.1	3.9
Bioprosthetic				
Group 7	5.2	5.5	7.1	5.5
Group 8	*	*	*	*
Group 9	4.3	4.6	4.3	4.6
Group 10	6.1	7.8	5.7	7.9
Group 11	2.6	2.3	2.0	2.3
Group 12	43	35	44	35
Group 13	1.7	5.9	4.6	5.9
Group 14	21	20	19	20

Table 6. Percentage of Valves by Mortality Outcome in Isolated AVR Cohort. *Cells with < 10 events were suppressed and replaced with *.*

	Algorithm	Hyperparameters	Tuning method
1.	Logistic regression		
	i. without penalty		
	ii. with Firth's correction		
	iii. with lasso penalty	variable-wise sparsity	CV
	iv. with TS lasso penalty	variable-wise sparsity	CV
	v. with group lasso penalty	group-wise sparsity	CV
	vi. with SG lasso penalty,		
	a. group sparsity = 0.15	group-wise sparsity	CV
	b. group sparsity = 0.50	group-wise sparsity	CV
	c. group sparsity = 0.85	group-wise sparsity	CV
2.	Random forest		
	i. node size = 1	# predictors in tree	
		size of the tree	OOB
	ii. node size = 50	# predictors in tree	
		size of the tree	OOB
	iii. node size = 100	# predictors in tree	
		size of the tree	OOB
3.	Gradient boosted trees		
	i. step size shrinkage = 0.3		
	maximum tree depth = 6		
	ii. step size shrinkage = 0.7		
	maximum tree depth = 6		
	iii. step size shrinkage= 0.3		
	maximum tree depth = 15		
	iv. step size shrinkage = 0.7		
	maximum tree depth = 15		
4.	BART		
	i. number of trees = 50		
	base = 0.95		
	power = 2		
	k = 2		
	quantile of the prior = 0.9		
5.	Neural networks		
	i. # units in hidden layer = 1		
	ii. # unit in hidden layer = 3		
	iii. # units hidden layer = 4		
6.	SVM (radial kernel)	cost parameter	CV

Table 7. Hyperparameters and Related Tuning Methods for Algorithms in the Extended Ensemble. *SG is sparse group, TS is treatment-specific, SVM is support vector machine, CV is cross-validation, OOB is out of bag, and for BART, k determines the prior probability that $E(Y|X)$ is between (-3, 3).*

Simulation setting	Data generation	Predictors for fitting
1	$\text{logit}(e_{y^1}) = X_{\text{sim}}\beta$	X_{sim}
2	$\text{logit}(e_{y^2}) = X_{\text{sim}}\beta + X_{\text{sim},I}\beta_I$	X_{sim}
3	$\text{logit}(e_{y^3}) = X_{\text{sim},1}\beta_1 + X_{\text{sim},2}\beta_2$	$X_{\text{sim},1}$ and $X_{\text{sim},3}$

Table 8. Data Generation and Predictors Under Different Simulation Settings. X *denotes full set of covariates;* X_{sim} *is the union of* $X_{sim,1}$, $X_{sim,2}$, *and* $X_{sim,3}$. $X_{sim,I}$ *includes interactions between selected variables from* X_{sim}.

	Mortality Rate (%)					
	Setting 1		Setting 2		Setting 3	
Cohort	30 Day	1 Year	30 Day	1 Year	30 Day	1 Year
Isolated AVR	1.9	5.7	1.7	5.7	1.9	5.9
AVR or AVR & MVR	2.0	5.9	2.0	5.8	2.0	6.2
AVR or AVR & CABG	2.6	6.6	2.9	7.2	2.9	7.3
Any AVR	2.7	7.3	2.4	7.0	2.9	7.7

Table 9. Mortality Rates in Simulated Data.

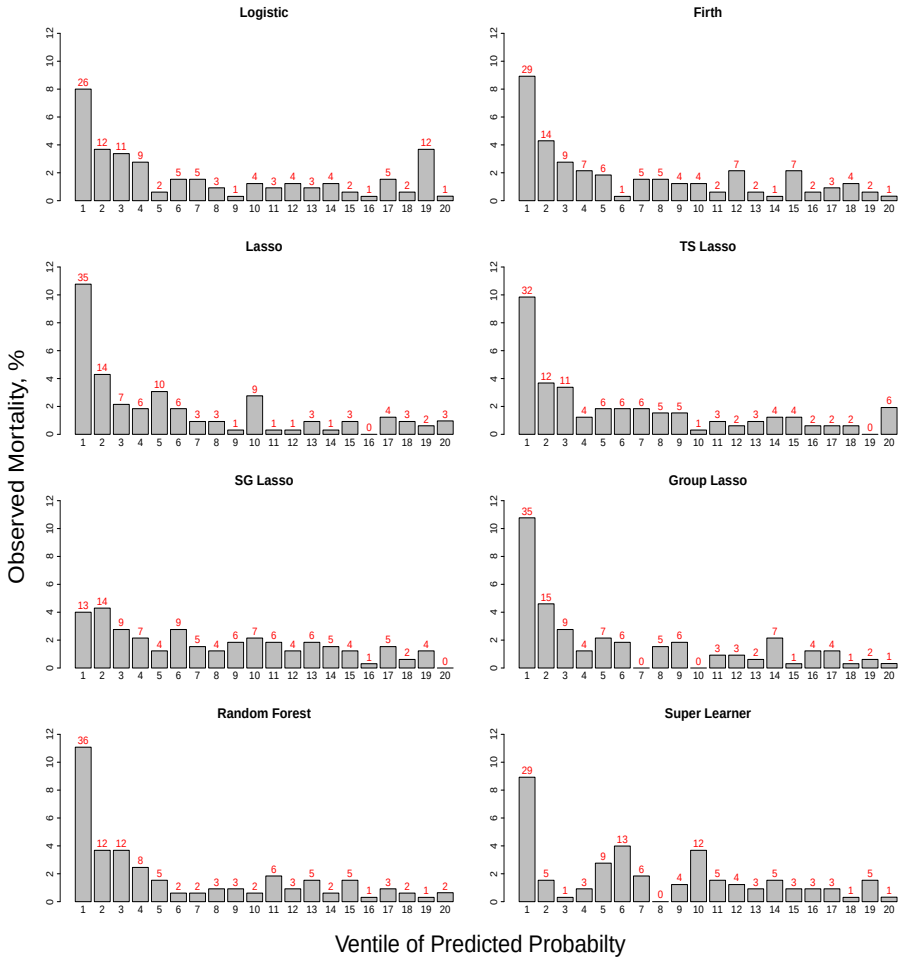

Figure 1. Data Analysis: Rate of Observed 30-Day Mortality within each Ventile of Predicted Mortality Risk for Different Algorithms in Isolated AVR Cohort. *The predicted mortality risks are in decreasing order and red values are the number of events in each ventile. TS is an abbreviation for treatment-specific and SG is sparse group.*

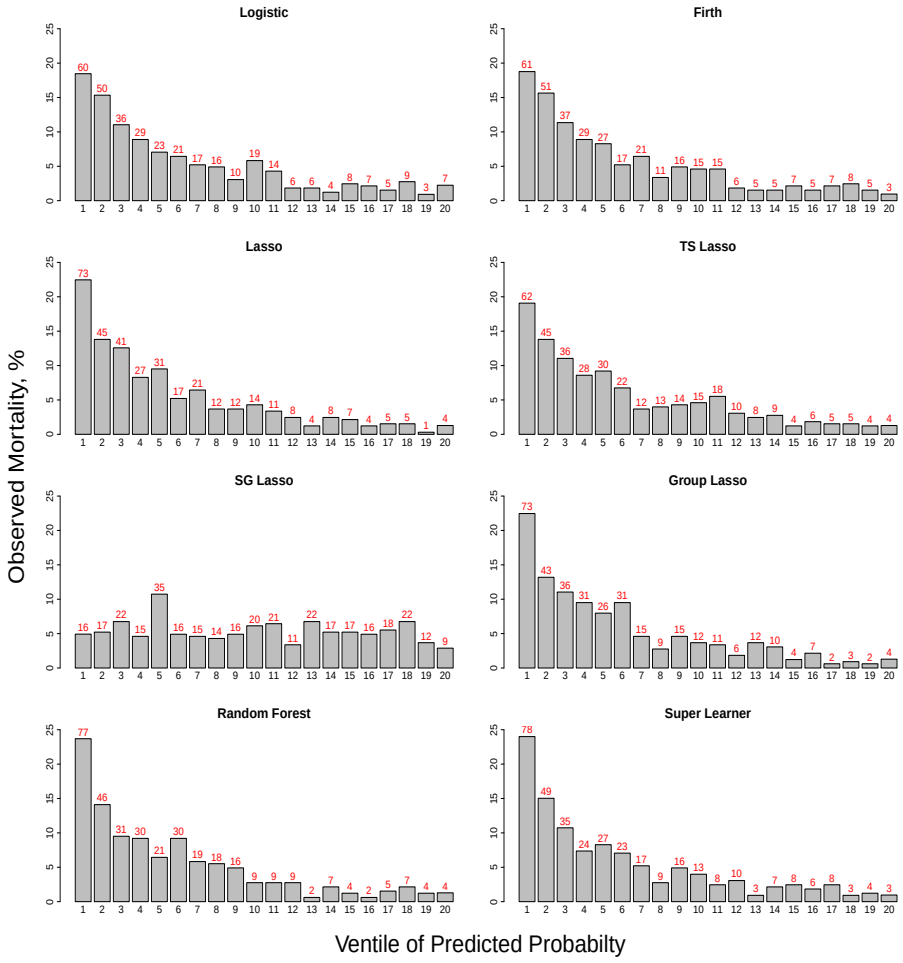

Figure 2. Data Analysis: Rate of Observed 1-Year Mortality within each Ventile of Predicted Mortality Risk for Different Algorithms in Isolated AVR Cohort. *The predicted mortality risks are in decreasing order and red values are the number of events in each ventile. TS is an abbreviation for treatment-specific and SG is sparse group.*

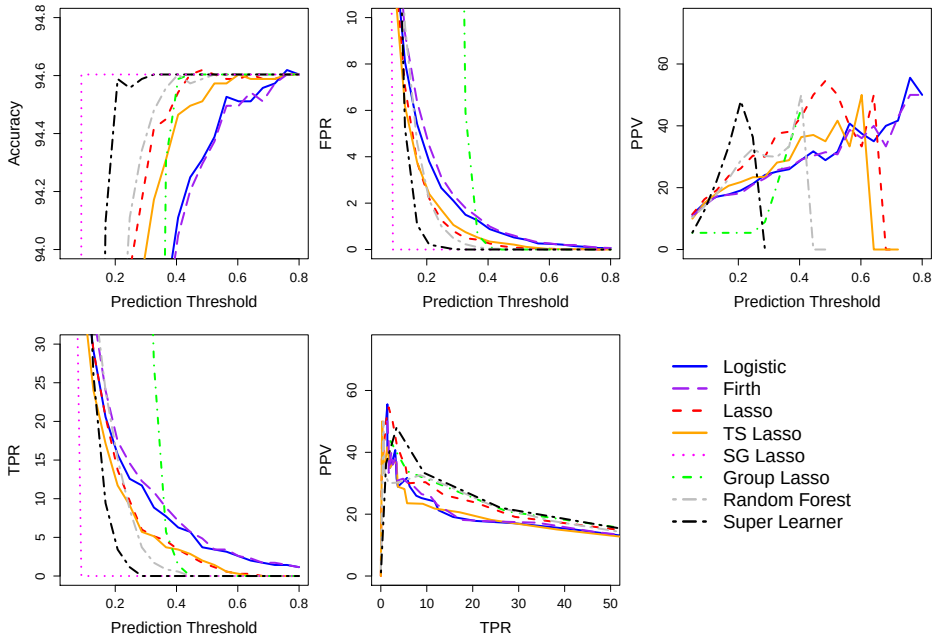

Figure 3. Data Analysis: Cross-Validated Algorithm Performance by Prediction Threshold and Precision-Recall Plot for 1-Year Mortality in Isolated AVR Cohort. *For algorithms with TPR equal to zero, PPV is undefined and not plotted. TS is an abbreviation for treatment-specific and SG is sparse group.*

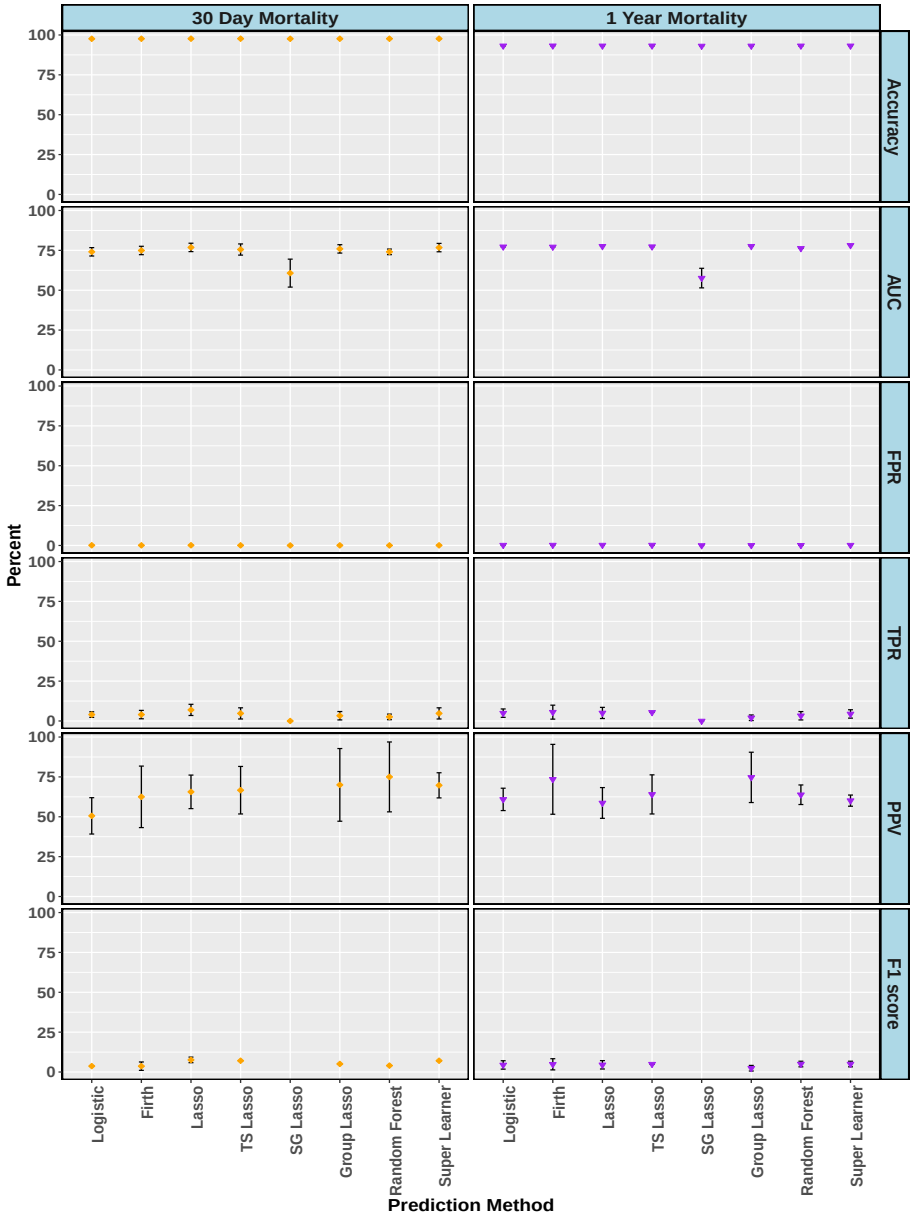

Figure 4. Data Analysis: Cross-Validated Algorithm Performance with 95% Confidence Intervals in AVR or AVR & CABG Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore* F_1 *score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. TS is an abbreviation for treatment-specific and SG is sparse group.*

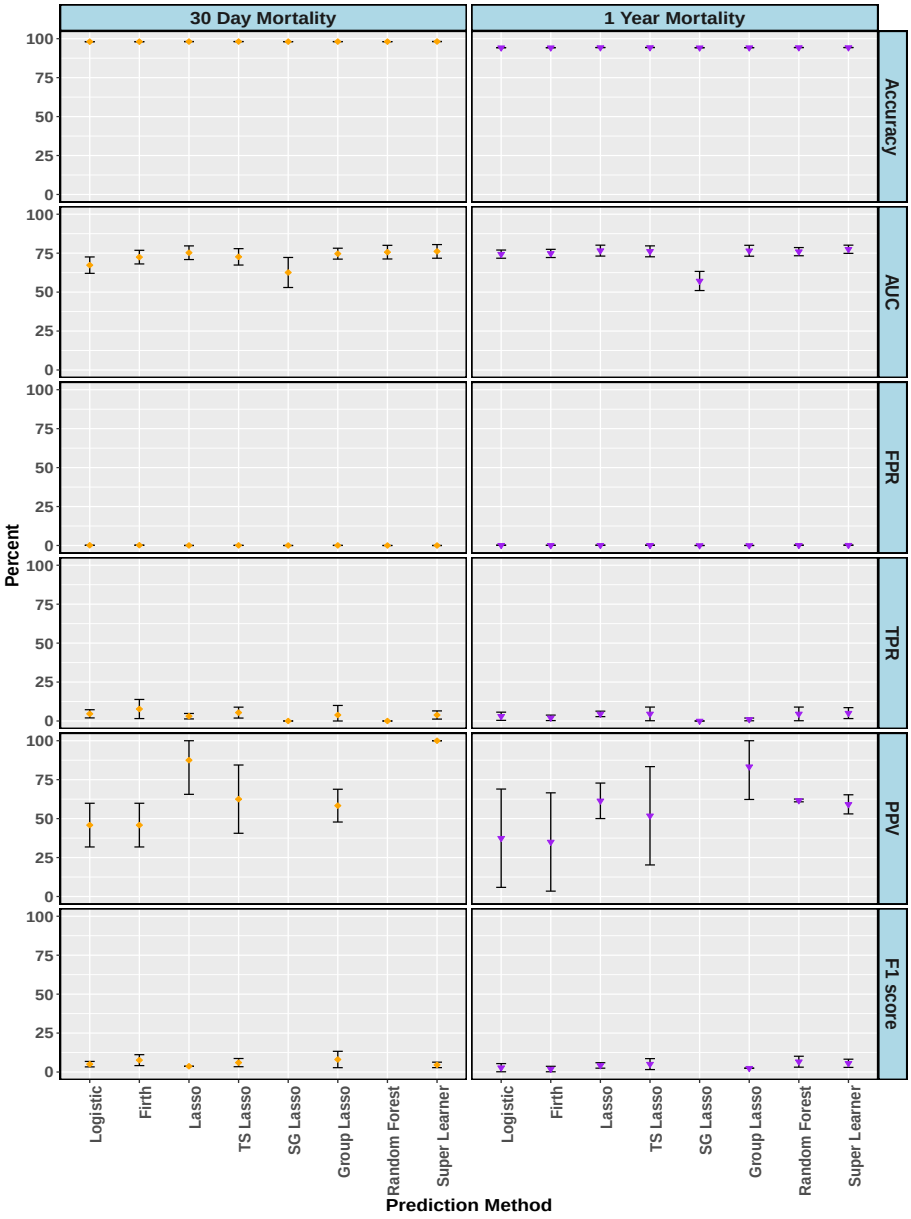

Figure 5. Data Analysis: Cross-Validated Algorithm Performance with 95% Confidence Intervals in AVR or AVR & MVR Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore* F_1 *score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. TS is an abbreviation for treatment-specific and SG is sparse group.*

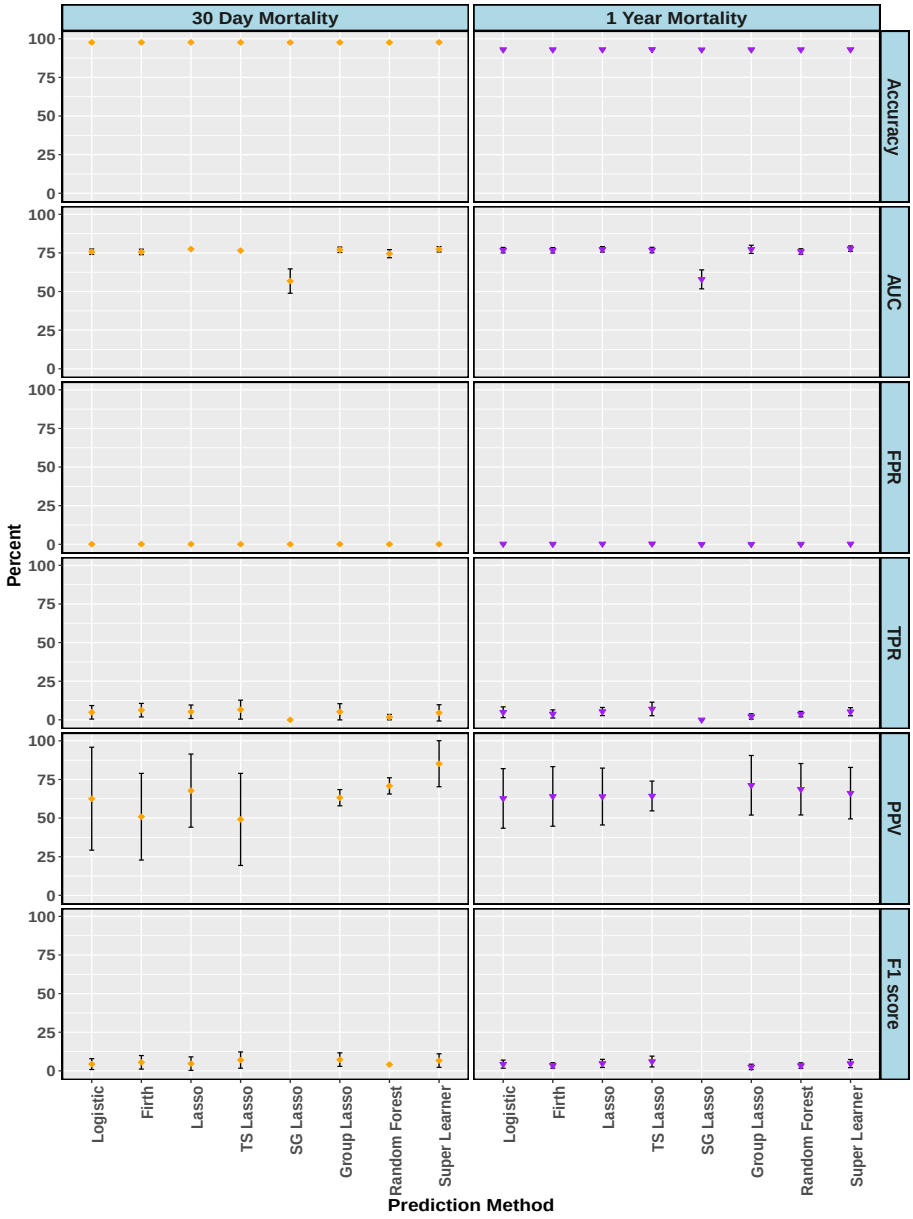

Figure 6. Data Analysis: Cross-Validated Algorithm Performance with 95% Confidence Intervals in any AVR Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. TS is an abbreviation for treatment-specific and SG is sparse group.*

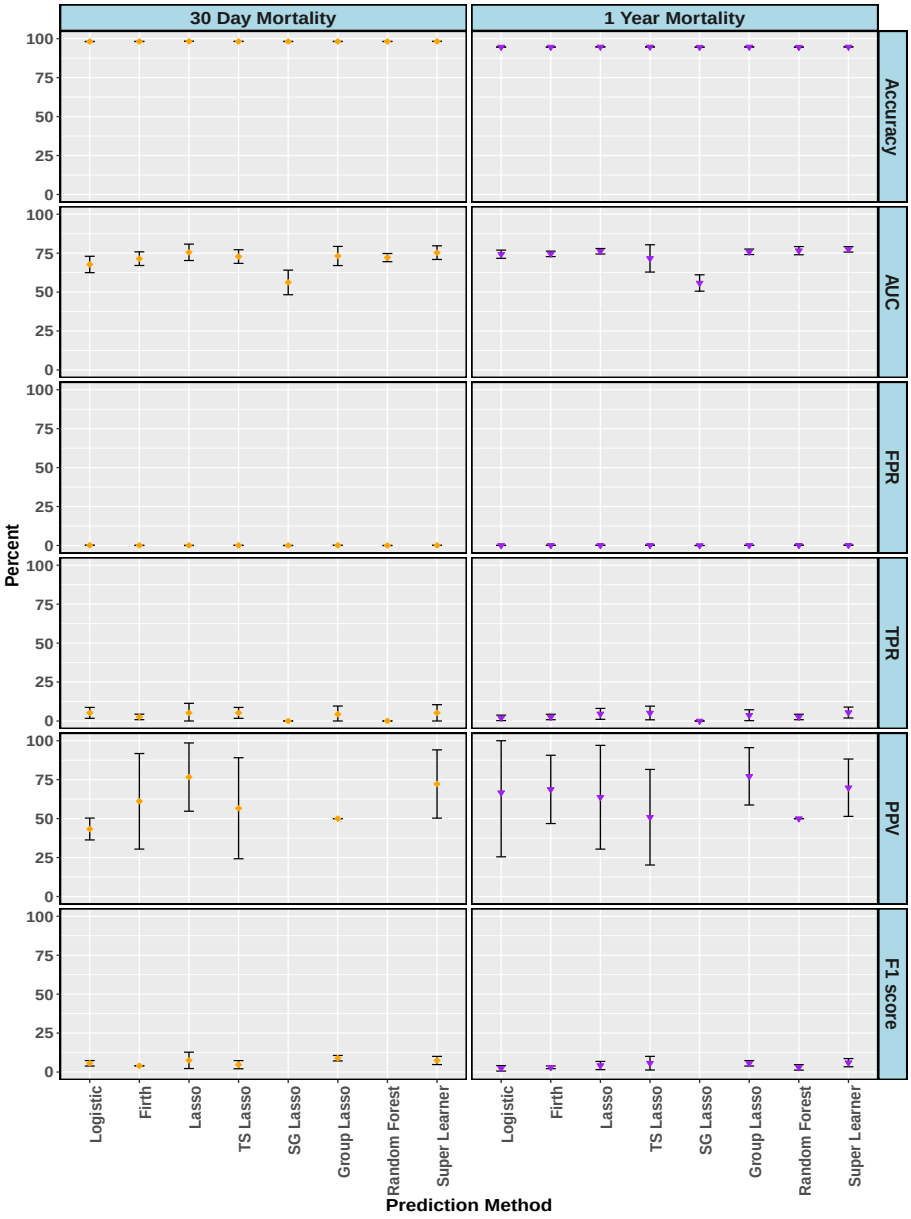

Figure 7. Data Analysis: Cross-Validated Algorithm Performance with 95% Confidence Intervals in Isolated AVR Cohort using Negative Log-Likelihood Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. TS is an abbreviation for treatment-specific and SG is group.*

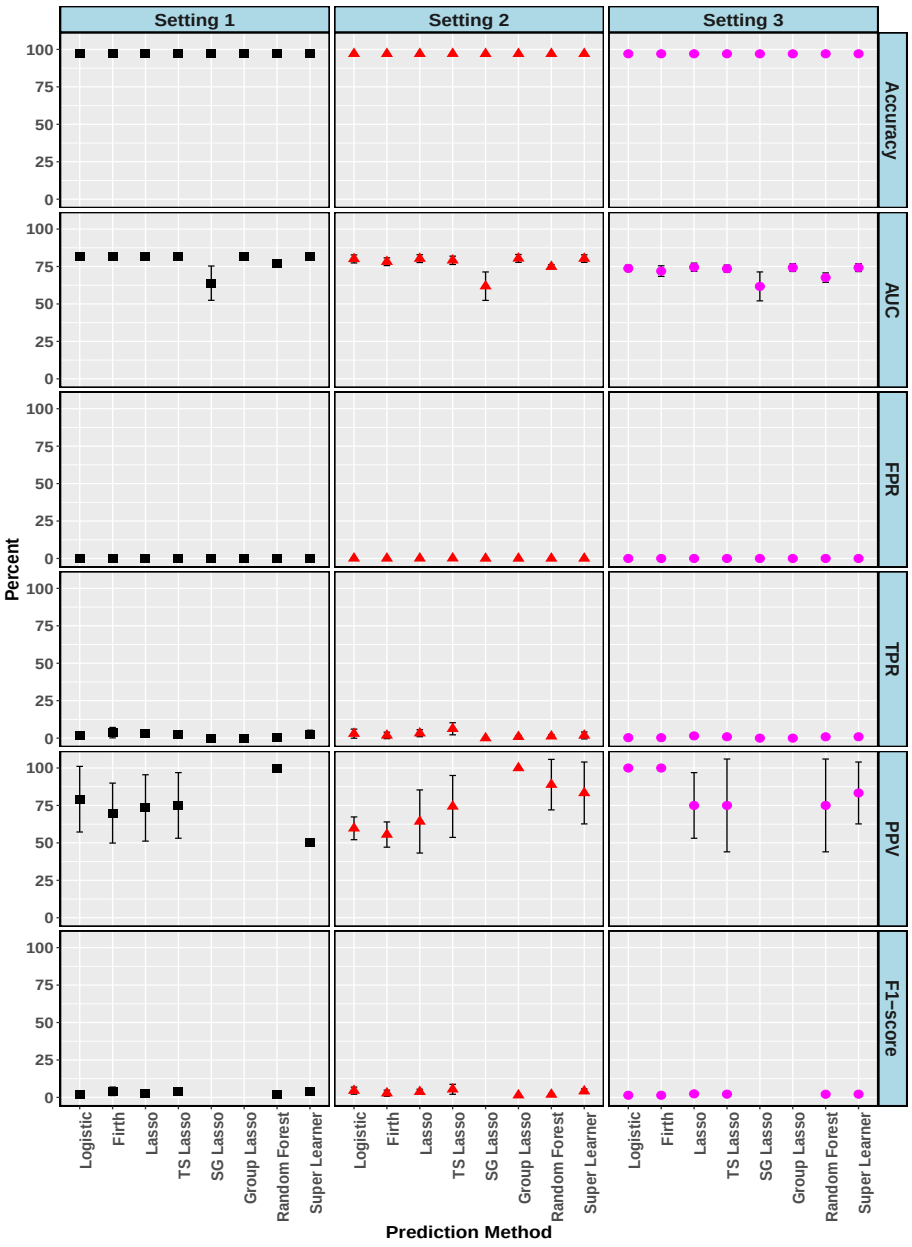

Figure 8. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 30-Day Mortality in AVR or AVR & CABG Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 84% for setting 1, 85% for setting 2, and 77% for setting 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

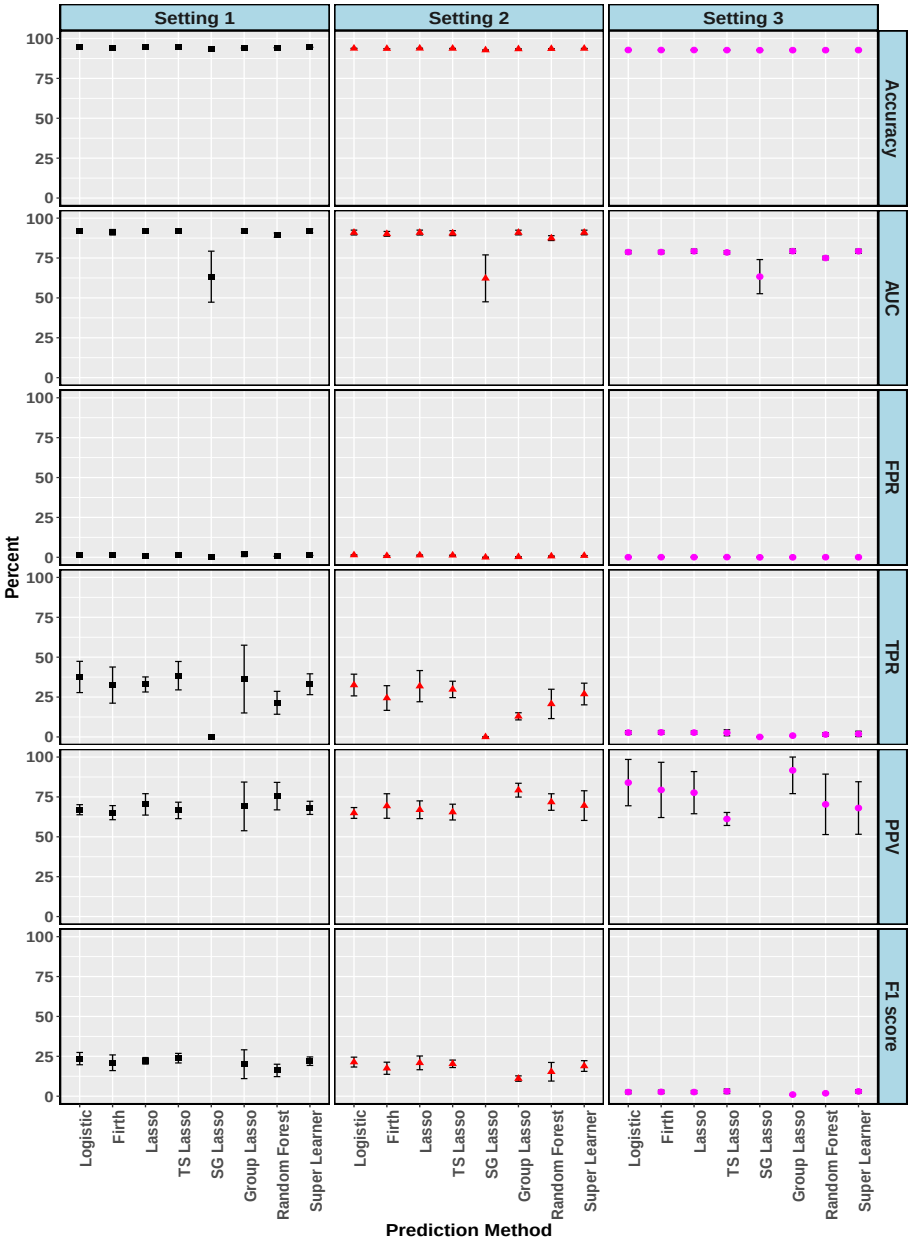

Figure 9. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 1-Year Mortality in AVR or AVR & CABG Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown.True conditional risk estimate based on AUC loss is 93% for setting 1, 91% for setting 2, and 92% for setting 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

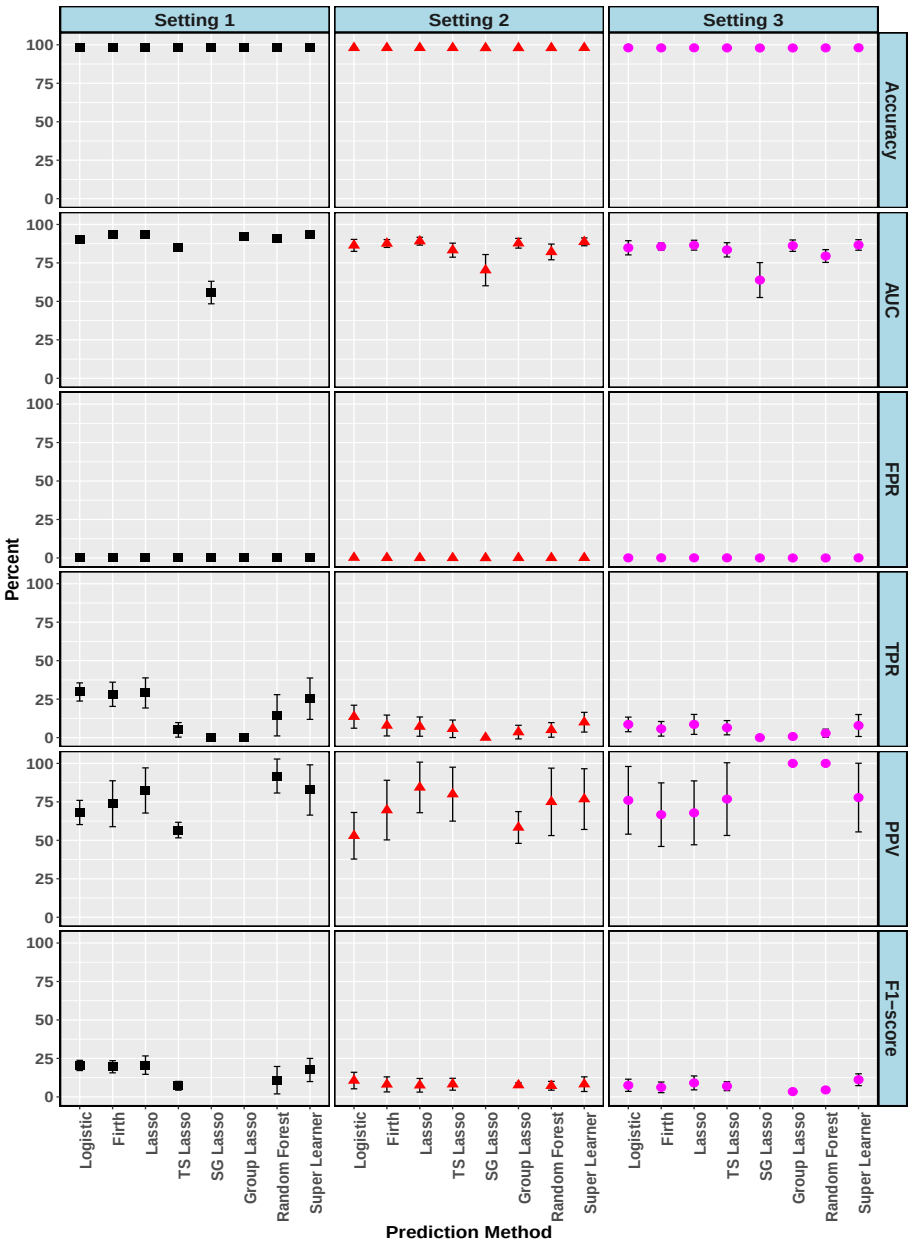

Figure 10. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 30-Day Mortality using in AVR or AVR & MVR Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 95% for setting 1, 91% for setting 2, and 94% for setting 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

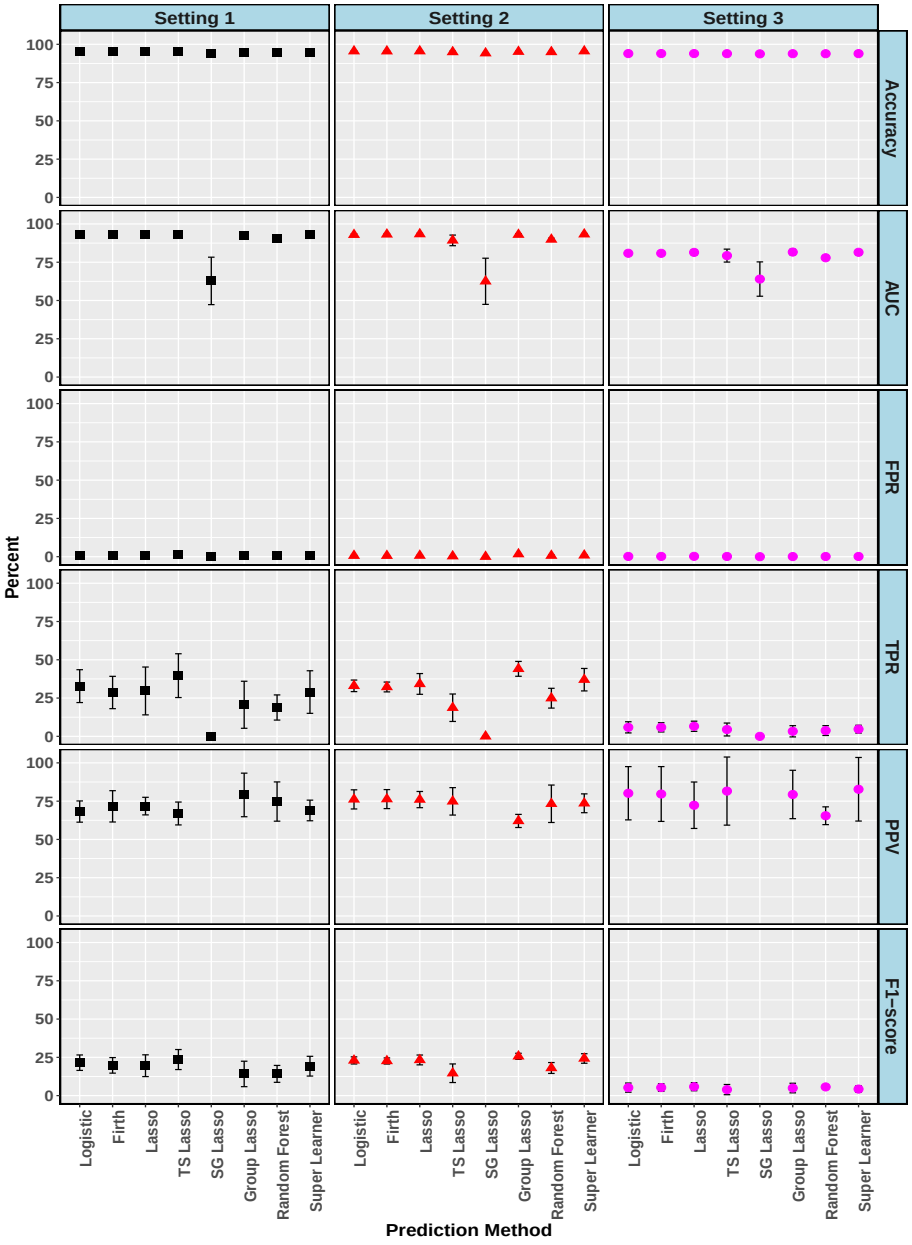

Figure 11. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 1-Year Mortality in AVR or AVR & MVR Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 94% for settings 1 and 2 and 93% for setting 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

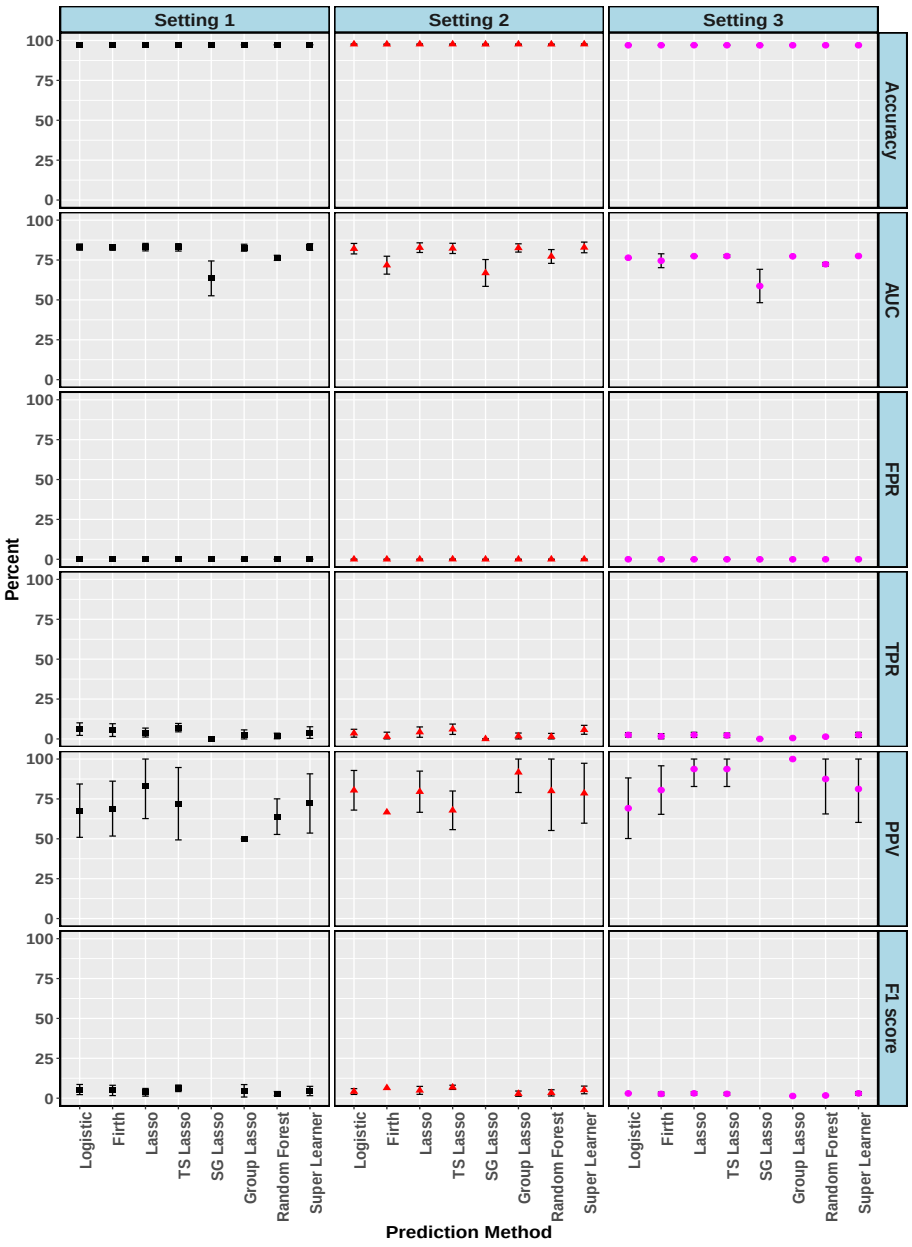

Figure 12. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 30-Day Mortality in any AVR Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 85% for setting 1, 84% for setting 2 and 79% for setting 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

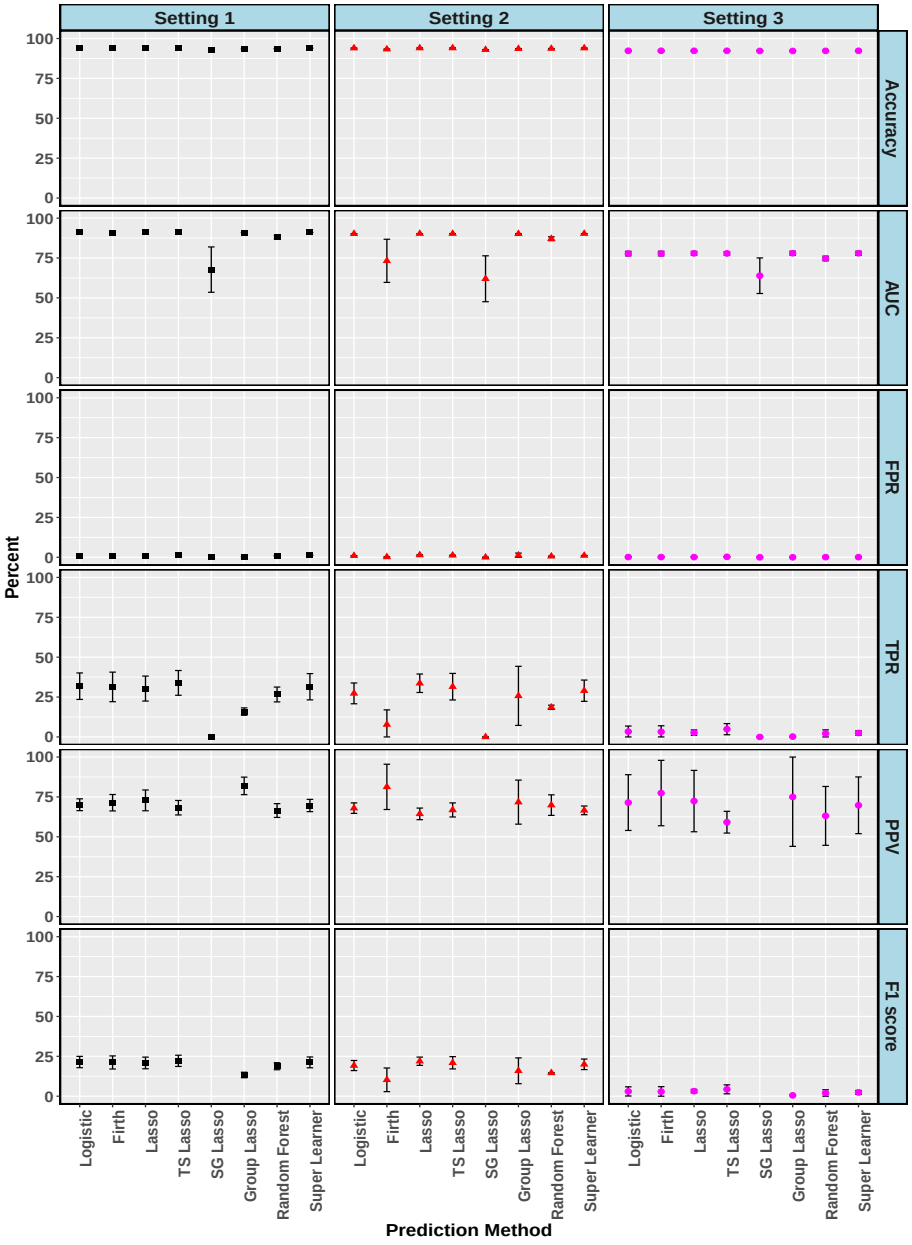

Figure 13. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 1-Year Mortality in any AVR Cohort using AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 92% for settings 1 and 3 and 91% for setting 2. TS is an abbreviation for treatment-specific and SG is sparse group.*

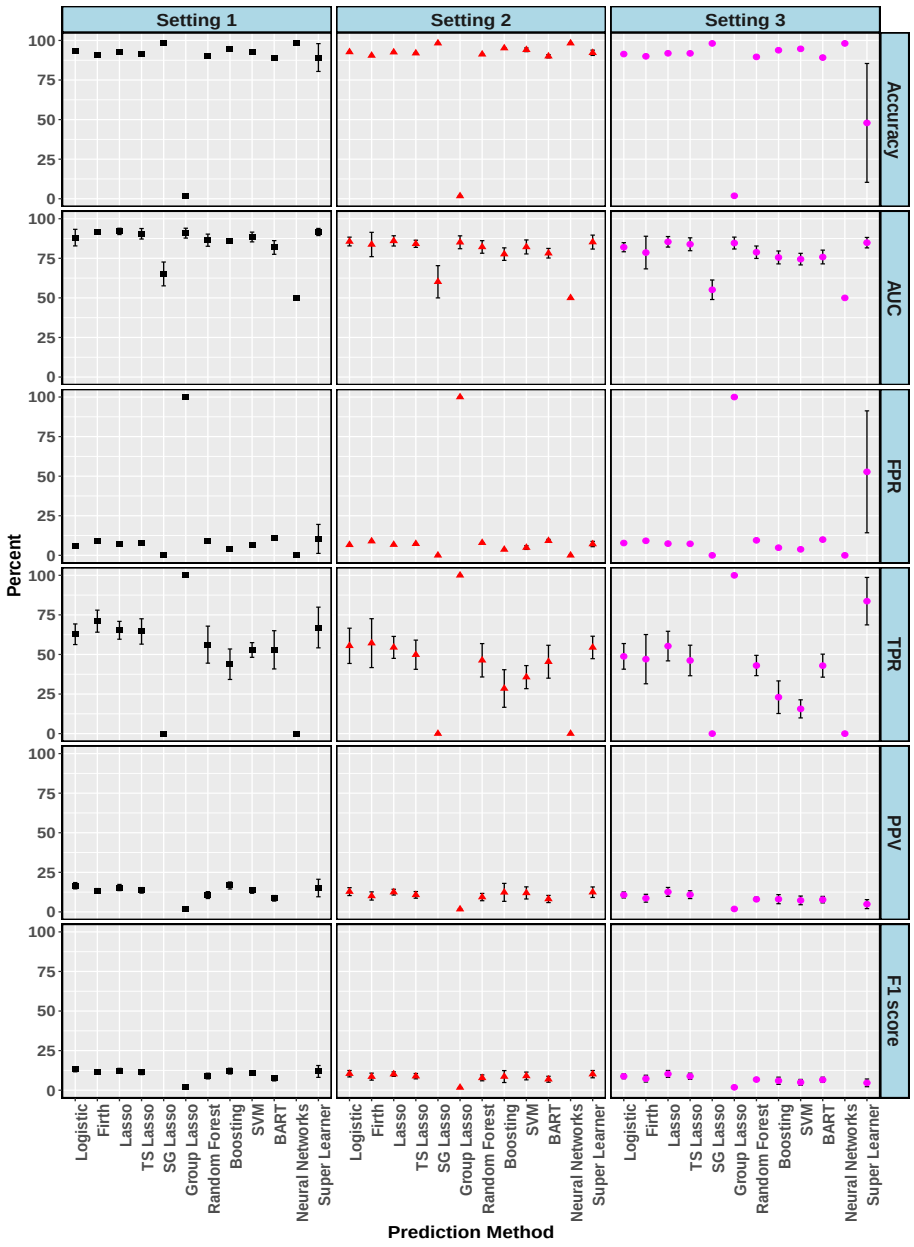

Figure 14. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 30-Day Mortality in Isolated AVR Cohort using AUC Loss Function Maximizing TPR. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 94% for setting 1, 89% for setting 2 and 95% for setting 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

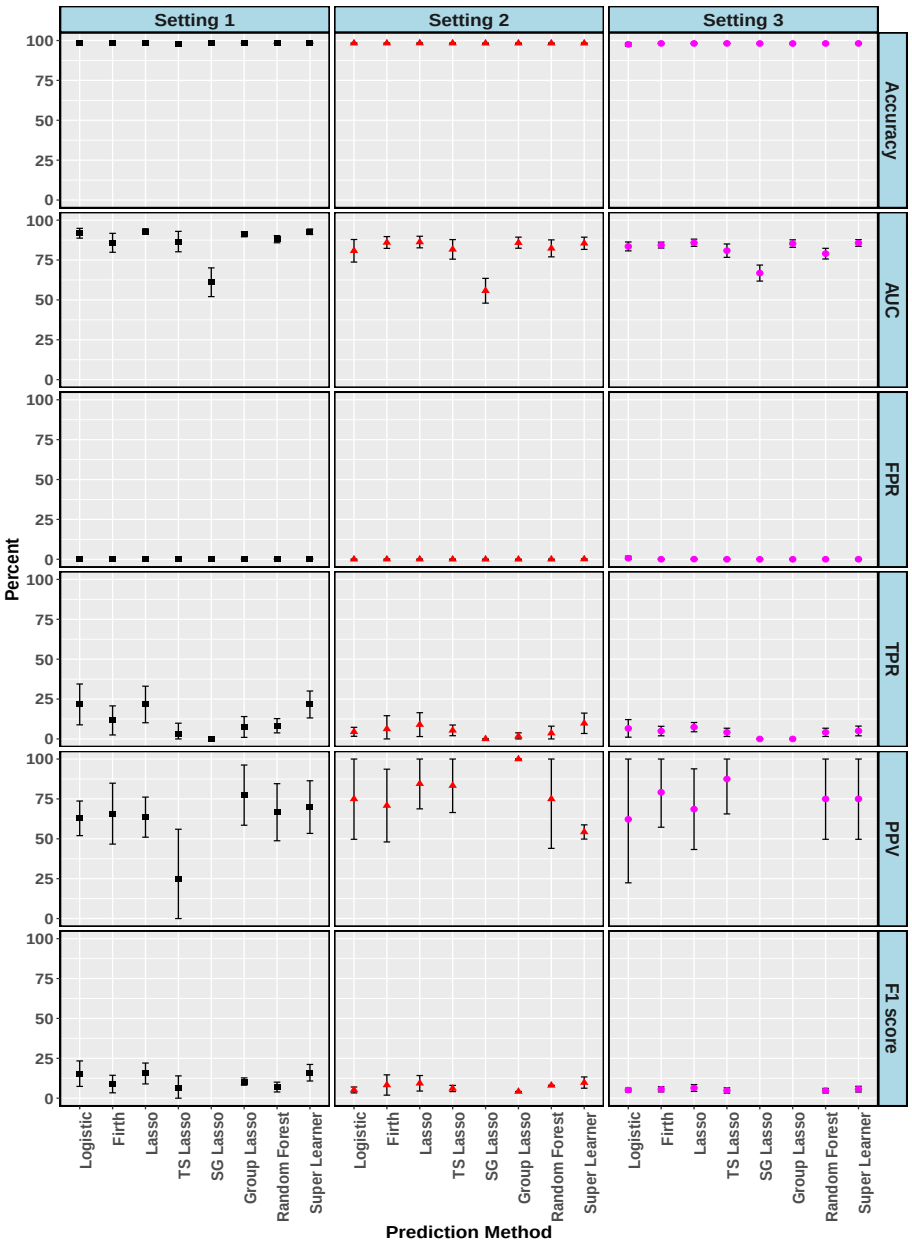

Figure 15. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 30-Day Mortality in Isolated AVR Cohort using Negative Log-Likelihood Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on average log loss (negative log likelihood divided by number of observations) is 0.05 for settings 1 and 3 and 0.07 for setting 2. For comparison, mean cross-validated log loss for super learner was 0.06 in setting 1 and 0.07 in settings 2 and 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

Prepared using sagej.cls

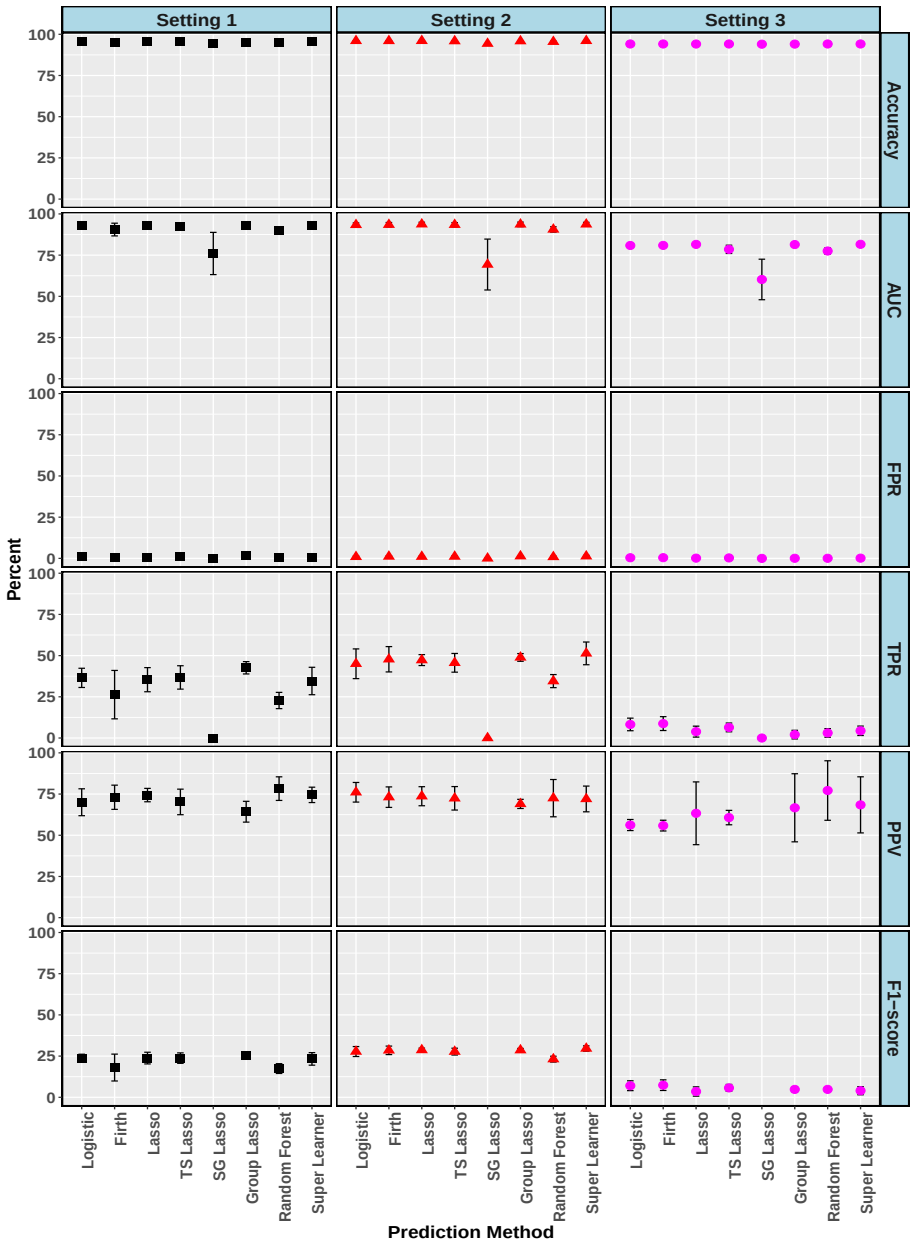

Figure 16. Simulation: Cross-validated Algorithm Performance with 95% Confidence Intervals for 1-Year Mortality in Isolated AVR using Negative Log-Likelihood Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown True conditional risk estimate based on average log loss (negative log likelihood divided by number of observations) is 0.12 for settings 1 and 3 and 0.11 for setting 2. For comparison, mean cross-validated log loss for super learner was 0.13 in setting 1, 0.12 in setting 2 and 0.19 in setting 3. TS is an abbreviation for treatment-specific and SG is sparse group.*

Prepared using sagej.cls

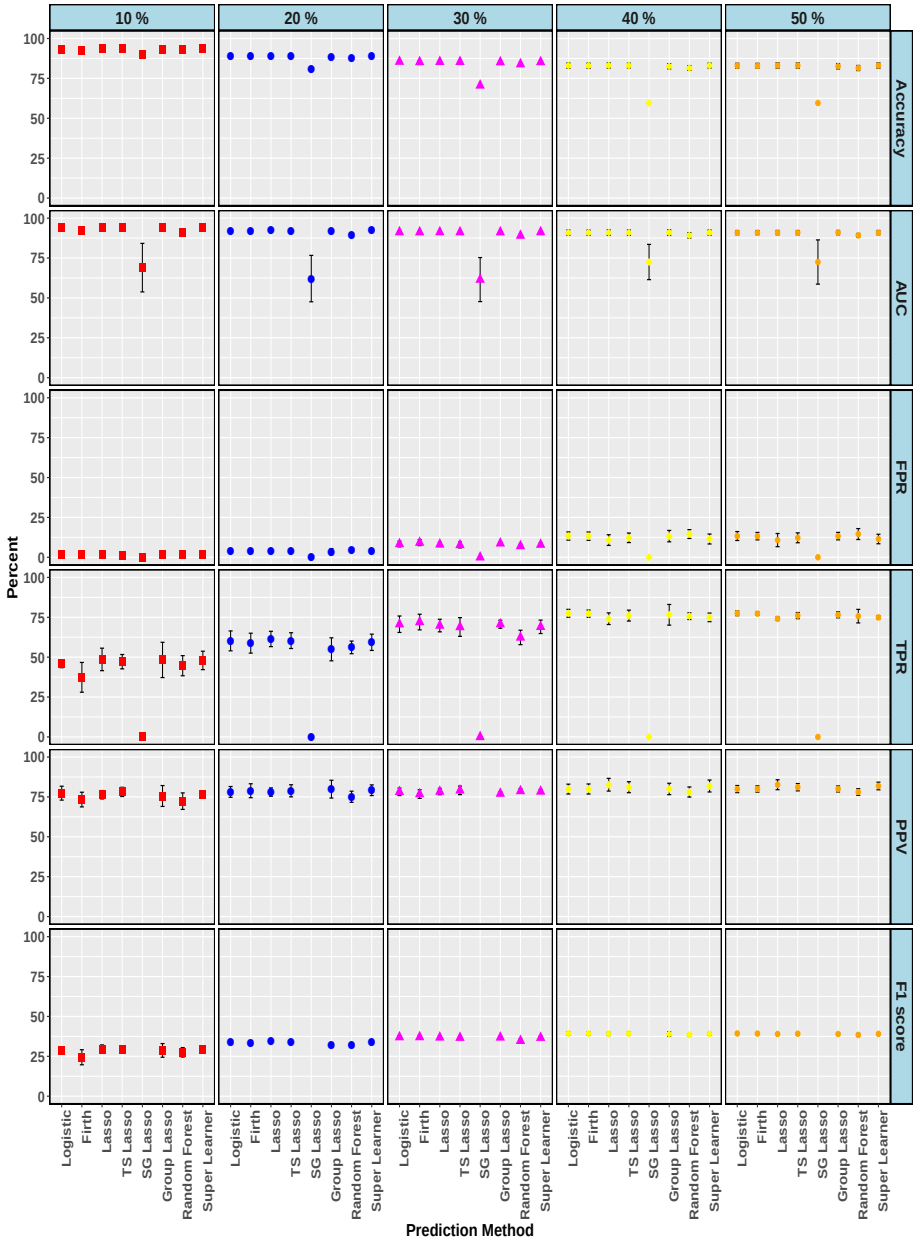

Figure 17. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for Isolated AVR and Different Mortality Rates using Simulation Setting 1 and AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 94% (for 10% event rate), 93% (for 20% and 50% event rates), 92% for (30% event rate), and 91% (for 40% event rate). TS is an abbreviation for treatment-specific and SG is sparse group.*

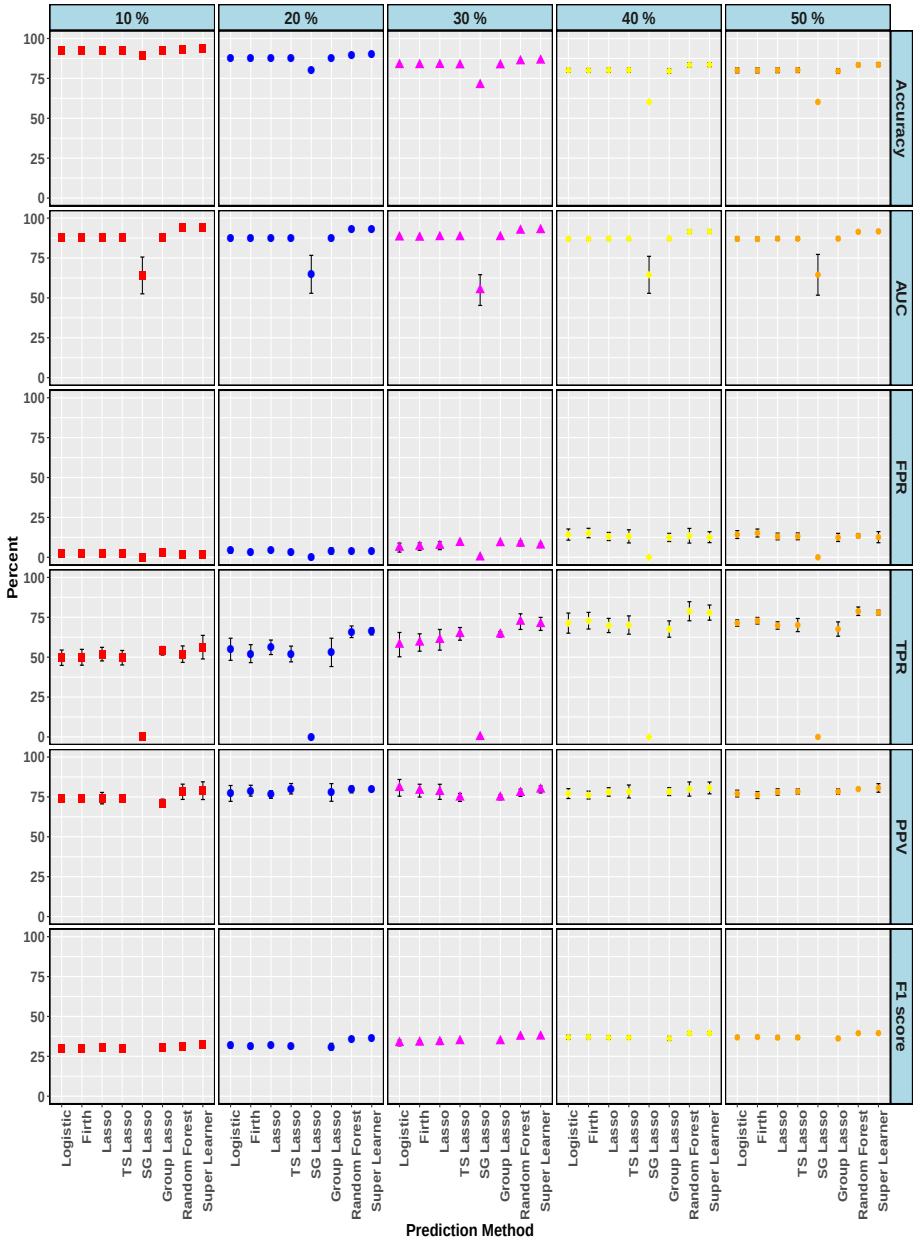

Figure 18. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for Isolated AVR and Different Mortality Rates using Simulation Setting 2 and AUC Loss Function. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. True conditional risk estimate based on AUC loss is 96% (for 10% event rate), 95% (for 20% event rate), 94% for (30% event rate), 94% (for 40% event rate), and 93% (for 50% event rate). TS is an abbreviation for treatment-specific and SG is sparse group.*

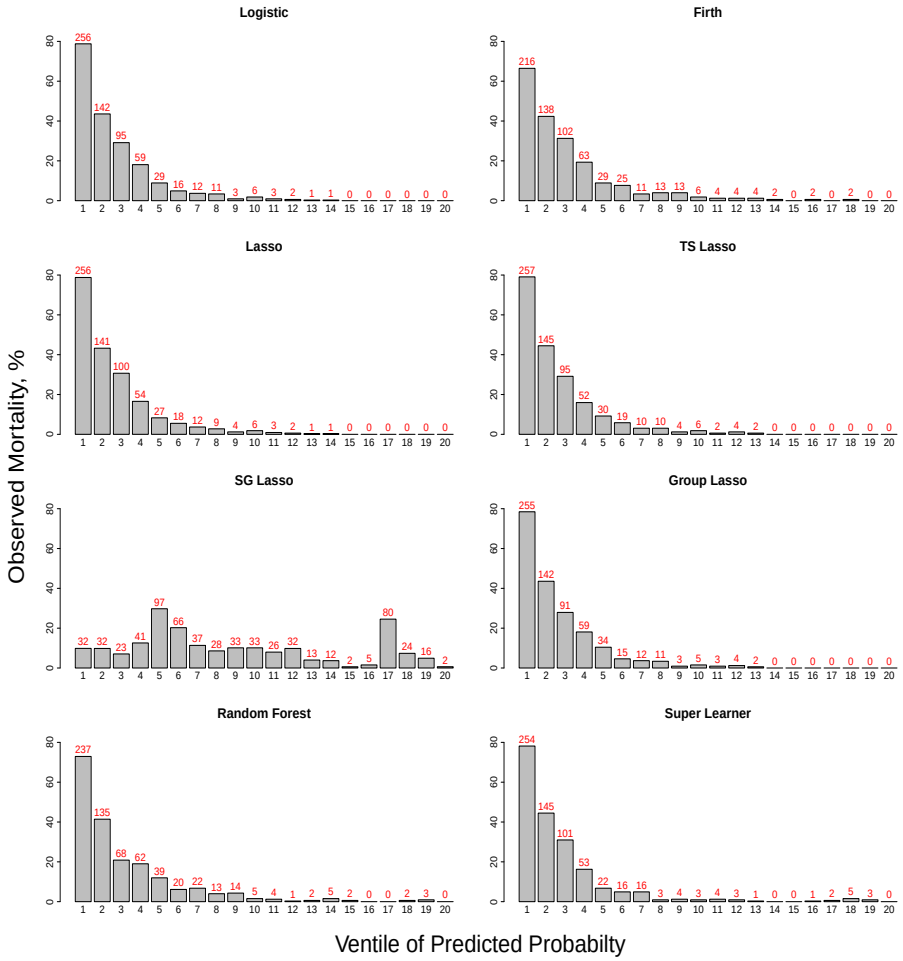

Figure 19. Mortality Rates within each Ventile of Predicted Mortality Risk for Different Algorithms in Simulated Data with 10% Mortality Rate under Simulation Setting 1. *The predicted mortality risks are in decreasing order and red values are the number of events in each ventile. TS is an abbreviation for treatment-specific and SG is sparse group.*

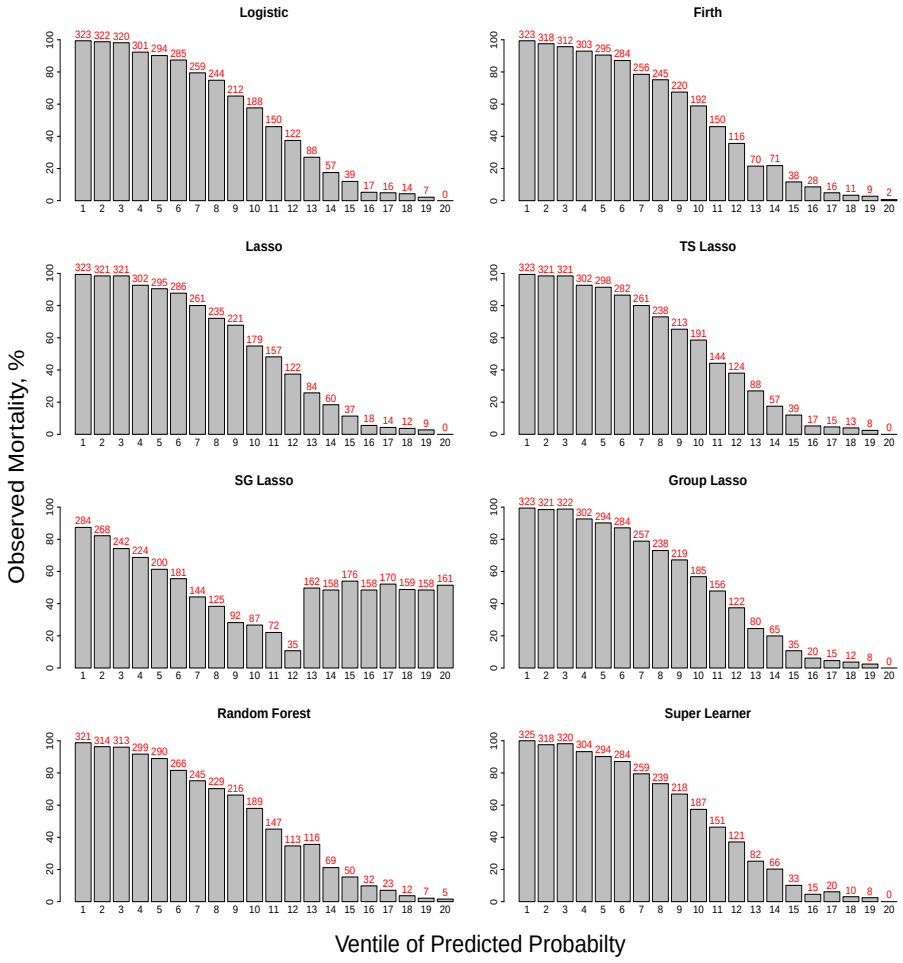

Figure 20. Mortality Rates within each Ventile of Predicted Mortality Risk for Different Algorithms in Simulated Data with 50% Mortality Rate under Simulation Setting 1. *The predicted mortality risks are in decreasing order and red values are the number of events in each ventile. TS is an abbreviation for treatment-specific and SG is sparse group.*

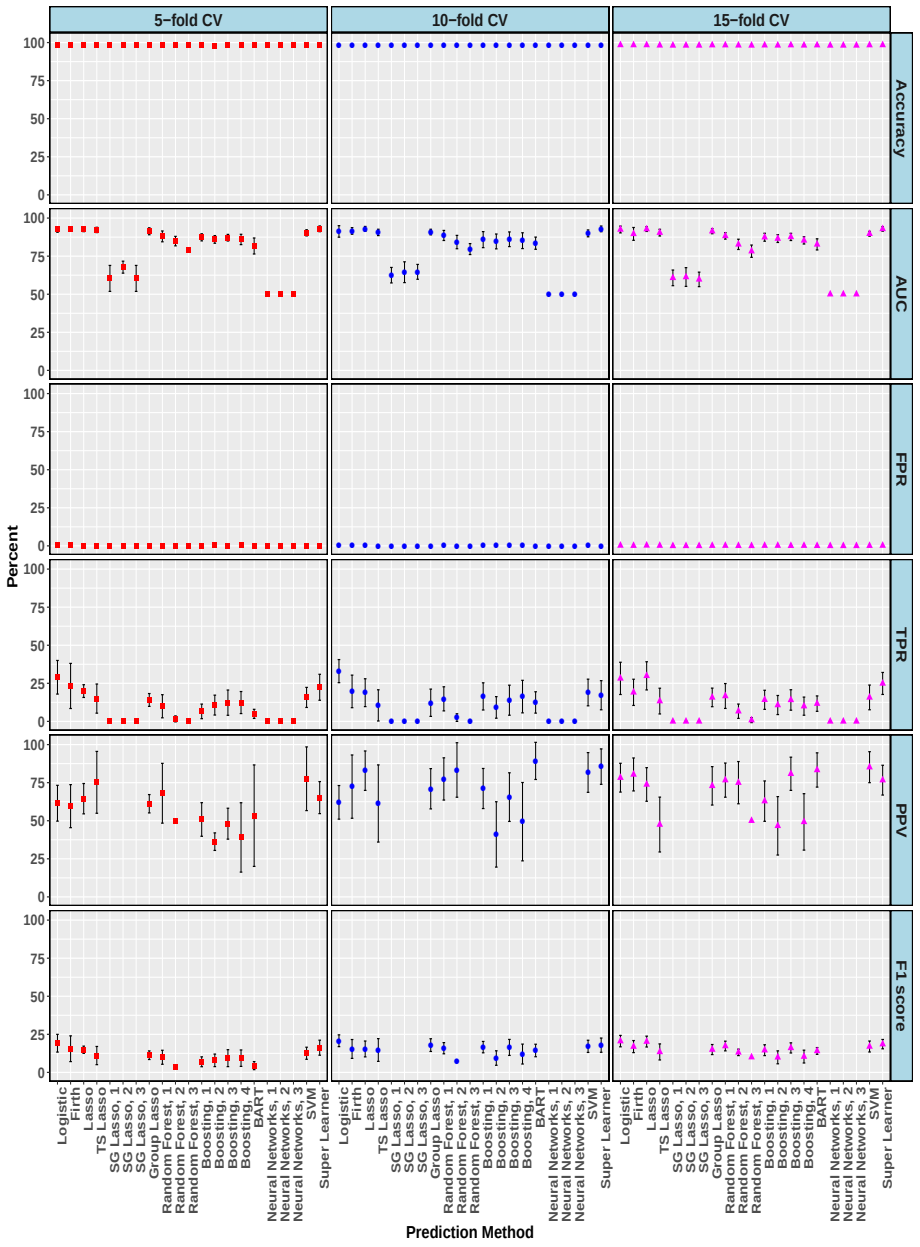

Figure 21. Simulation: Cross-Validated Algorithm Performance with 95% Confidence Intervals for 30-Day Mortality in Isolated AVR using Simulation Setting 1 and AUC Loss Function with Varied Cross-Validation Folds and Extended Algorithms with Different Hyperparameters in the Ensemble. *For algorithms with zero predicted positive values, PPV is undefined and not plotted, and therefore F_1 score is also undefined and not plotted. 95% confidence intervals for estimates with standard errors less than 1% are not shown. TS is an abbreviation for treatment-specific and SG is sparse group.*