Comparison of standard and penalized logistic regression in risk model development

Yan Yan, MD, PhD,[a] Zhizhou Yang, BA,[b] Tara R. Semenkovich, MD, MPHS,[b] Benjamin D. Kozower, MD, MPH,[b] Bryan F. Meyers, MD, MPH,[b] Ruben G. Nava, MD,[b] Daniel Kreisel, MD, PhD,[b] and Varun Puri, MD, MSCI[b]

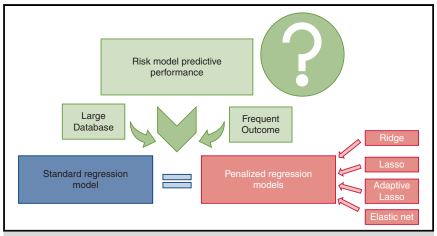
ABSTRACT

Objective: Regression models are ubiquitous in thoracic surgical research. We aimed to compare the value of standard logistic regression with the more complex but increasingly used penalized regression models using a recently published risk model as an example.

Methods: Using a standardized data set of clinical T1-3N0 esophageal cancer patients, we created models to predict the likelihood of unexpected pathologic nodal disease after surgical resection. Models were fitted using standard logistic regression or penalized regression (ridge, lasso, elastic net, and adaptive lasso). We compared the model performance (Brier score, calibration slope, C statistic, and overfitting) of standard regression with penalized regression models.

Results: Among 3206 patients with clinical T1-3N0 esophageal cancer, 668 (22%) had unexpected pathologic nodal disease. Of the 15 candidate variables considered in the models, the key predictors of nodal disease included clinical tumor stage, tumor size, grade, and presence of lymphovascular invasion. The standard regression model and all 4 penalized logistic regression models had virtually identical performance with Brier score ranging from 0.138 to 0.141, concordance index ranging from 0.775 to 0.788, and calibration slope from 0.965 to 1.05.

Conclusions: For predictive modeling in surgical outcomes research, when the data set is large and the outcome of interest is relatively frequent, standard regression models and the more complicated penalized models are very likely to have similar predictive performance. The choice of statistical methods for risk model development should be on the basis of the nature of the data at hand and good statistical practice, rather than the novelty or complexity of statistical models. (JTCVS Open 2022;9:303-16)


Risk model predictive performance
?
Large Database
Frequent Outcome
Standard regression model
Penalized regression models
Ridge
Lasso
Adaptive Lasso
Elastic net

Standard and penalized models had similar performance with frequent outcome of interest in a large data set.

CENTRAL MESSAGE

The choice of statistical methodology for risk model development should focus on the nature of the data and good statistical practice rather than method complexity.

PERSPECTIVE

Penalized regression models might be advantageous when the number of events is small relative to the number of potential predictors. However, they are being increasingly used in surgical outcomes research even in data sets in which the event is not rare. In this study, we evaluated the predictive performance of penalized models versus standard regression when the outcome of interest is frequent.

Descriptive modeling, the most commonly applied approach in medicine, involves summarizing data and describing associations between dependent and independent variables. Predictive modeling, in contrast, aims to predict new or future observations. An exceedingly small fraction of the numerous multivariable models published annually in surgical outcomes research become used clinically. Predictive modeling seeks to overcome this shortcoming by delivering tools applicable in routine clinical practice.

Predictive modeling is on the basis of developing rules to estimate the probability of the presence of a specific disease in diagnostic research or the probability of the occurrence of a future event in prognostic research. In thoracic surgery or oncology, for example, models are frequently developed to predict the likelihood of cancer recurrence or the probability of a postoperative complication.[1-3] When the clinical outcome of interest is binary (yes/no), the prediction rule is commonly developed using a multivariable logistic regression model.[4]

From the Divisions of [a]Public Health Sciences and [b]Cardiothoracic Surgery, Washington University School of Medicine, St Louis, Mo.

This work was funded by Veterans Affairs I01 HX002475-03 (Puri, Yan), National Heart, Lung, and Blood Institute R01 HL146856-02 (Puri, Yan), and NCI Cancer Center Support Grant #P30 CA091842, Eberlein, PI (Yan).

Received for publication May 5, 2021; accepted for publication Jan 13, 2022; available ahead of print Feb 24, 2022.

Address for reprints: Varun Puri, MD, MSCI, 660 S Euclid Ave, Campus Box 8234, St Louis, MO 63110 (E-mail: varunpuri@wustl.edu).

2666-2736

https://doi.org/10.1016/j.xjon.2022.01.016

Abbreviations and Acronyms

cvMSE	= cross-validated mean square error
EPV	= events per variable
MLE	= maximum likelihood estimation
MSE	= mean square error
NCDB	= National Cancer Database
PRM	= penalized regression model

A logistic regression model can be fitted using 1 of 2 approaches: the more commonly utilized maximum likelihood estimation (MLE) (standard approach),[5,6] or the newer, penalized regression model (PRM) approach.[7,8] With epidemiological and biostatistics journals often reporting improved predictive accuracy using PRMs,[9-12] this approach is being increasingly adopted in thoracic surgery publications.[7,8,13,14] However, PRM analyses might have some disadvantages, including arbitrary variable selection and discordance of output from different software.[15-17] Furthermore, PRMs are conceptually more complex to build and require specific software programs.

Despite the appropriate and timely emphasis that has been given to statistical approach and rigor in clinical journals over the past decade, many clinicians find study methodology difficult to understand. Multivariable models are the most commonly used statistical tools in health services research. In the setting of newer available methods to build multivariable models, we aim to: (1) provide a brief description of the PRM approach in contrast to standard MLE regression models, and (2) compare the performance of PRM and standard logistic regression models using a standardized data set and a clinical question relevant to outcomes researchers.

Because risk model development, validation, and application are highly pertinent topics in surgical research, we anticipate that this work will be useful for those reading the literature to judge if a valid statistical approach is used in a report as well as for those reviewing manuscripts to know the range of statistical options available for analysis. Even more importantly, for clinical researchers, our analyses will provide guidance on the advantages and disadvantages of certain statistical techniques and will help identify when it is ideal to collaborate with an expert biostatistician.

METHODS

Development of the Question

In a recent publication, we used a multivariable logistic regression model to predict the likelihood of occult lymph node metastases in patients with surgically resectable esophageal cancer who were clinically staged as N0.[18] Our logistic regression model was fitted using the standard MLE approach. Patients in the study were selected from the National Cancer Database (NCDB). Of 3186 patients in our study, 688 (22%) had pathologic lymph node involvement. The presence of the outcome of interest (ie, pathologic lymph node involvement) in our study, is deemed an "event." Hence, we had 688 events in our analysis. Potential predictors included histology, stage, tumor size, grade, presence of lymphovascular invasion, and other demographic or socioeconomic variables. In multivariable models, each level of such predictors (eg, stage might have 4 levels) leads to the generation of its individual regression coefficient. Therefore, the total number of regression coefficients in our previous study was approximately 30. The ratio of the number of events to the number of regression coefficients in the model is termed the "events per variable" or EPV. A well accepted rule of thumb is to have at least 10 to 15 EPV in any model.[19,20] Our previous study with 688 events and approximately 30 regression coefficients had an EPV of >20. For a data set with such a high EPV, it is not clear if more recently developed modeling methods such as PRM provide better predictive performance of models than the standard MLE approach. We designed the following analysis to investigate this question.

Data and Patients

Data for this study come from the NCDB Participant Use File for esophageal cancer. The NCDB is a retrospective data set from the American Cancer Society and the American College of Surgeons that captures >70% of all new nationwide cancer diagnoses from >1500 hospitals accredited by the Commission on Cancer. Patients with clinically localized esophageal cancer (cT1-3N0M0) who underwent esophagectomy from 2004 to 2014 were included in the study. Patients were excluded if: (1) they underwent any type of neoadjuvant therapy (radiation, chemotherapy, hormone, therapy, or other systemic therapy), (2) were documented as having clinical T0 (no evidence of a primary tumor) or Tis (high-grade dysplasia) disease, or (3) had missing data on the timing of their surgery, tumor stage, tumor size, pathologic lymph node staging, or histology.[18] This study was exempt from Washington University's Institutional Review Board approval because the data set is deidentified. All computations in our analysis were performed using R "base," "glmnet," and "rms" packages (R Foundation for Statistical Computing).

Variables

The outcome variable was whether a patient was pathologically node negative (N0) or positive (N+). Potential predictors included the following demographic and tumor variables: age, sex, race (White vs non-White), insurance status (private vs public), median income according to zip code (lowest quartile of <$38,000 vs >$38,000), education status according to zip code (lowest quartile of >21% without a high school degree vs the remainder), population according to zip code (>250,000 vs <250,000), Charlson–Deyo Score (0, 1, ≥2), treatment center type (academic vs nonacademic), year of diagnosis, histology (squamous cell vs adenocarcinoma), tumor stage, tumor size, grade, and presence of lymphovascular invasion. Data on these variables are largely available preoperatively, with the possible exception of lymphovascular invasion. We used a missing category in that context. In our previous study, to evaluate the model built around these variables, we performed additional analyses with 3 clinically relevant variations and compared their predictive accuracy. First, we excluded the variable, lymphovascular invasion, because it is not always available preoperatively. Second, we excluded the cT3N0 patients, because many such individuals are prescribed induction chemoradiation on the basis of the depth of invasion. Finally, we divided the cT1N0 patients into T1a, T1b, and T1 not otherwise specified populations. Because none of these variations improved the predictive performance of the initial model and because the "missing" category in lymphovascular invasion allowed for its use in all patients, we retained the original for creation of the nomogram presented in our previous publication.[18]

Statistical Models

Logistic regression with MLE (standard model). Details about model development are presented in our recent work.[18] Briefly, we started by identifying all potential predictors associated with pathologically positive lymph nodes on the basis of clinical knowledge. Next, we used multivariable logistic regression to develop the risk model. The model was internally validated using bootstrapping techniques. We selected predictors with $P < .05$ and clinical importance criteria in the final model. The regression coefficients were estimated using MLE, a probabilistic framework for estimating the parameters of a model.[21]

Logistic regression with penalization (PRMs). Different from standard regression, penalized logistic regression has a shrinkage term in the model. The purpose of the shrinkage term is to prevent extreme values of regression coefficients in model development, so that the possibility of overfitting is reduced. An overfitted model predicts very well in the patients from the model development data set, but has poor predictive performance for patients in other data sets.

Depending on the form of shrinkage term, there are different types of PRMs. The most popular types include ridge, lasso, adaptive lasso, and elastic net. In brief: ridge includes the summation of squares of regression coefficients as the shrinkage term. Lasso contains the summation of the absolute value of regression coefficients as the shrinkage term. Elastic net has a mixture of ridge and lasso as the shrinkage term. Adaptive lasso is a variant of lasso, which allows different weights for shrinking different regression coefficients. Details about these shrinkage methods are provided in Appendix 1. Details about the computation of these models are provided in Appendices 2 and 3.

Comparison of predictive performance between MLE (standard) and PRMs. The predictive performance of the models was assessed using 3 measures:

1. Brier score for overall performance: the Brier score is the average squared difference between the observed outcome and the predicted probabilities. The lower the Brier score, the greater the predictive accuracy of the model.
2. Calibration slope for calibration: model calibration is the process of adjustment of the model parameters and forcing within the margins of the uncertainties to obtain a model representation of the processes of interest that satisfies preagreed criteria (goodness of fit). Calibration is a measure of fit of the model and is estimated by plotting predicted probabilities against the actual probabilities (Figure 1). A slope of 1, as shown in Figure 1, suggests perfect calibration. A slope of <1 suggests overfitting; slope >1 suggests underfitting.
3. C statistic for discrimination: the C statistic is the probability that the patient who experienced the event has a higher predicted value than the patient without the event. A value of 0.5 suggests that the model has no discrimination ability, whereas a value of 1 suggests that the model can discriminate perfectly between higher-risk and lower-risk patients. The C statistic is derived from the area under the receiver operating characteristic curve (Figure 2). A higher C statistic is desirable.

We elected not to use the binary classification method to compare predictive model performance in this study because a clinically meaningful cutoff value cannot be identified. Additionally, because the output for a logistic regression model is a probability, classification might lead to a loss of valuable clinical information and is a less sensitive way of comparison.

Original sample. We evaluated the performance of standard MLE and PRMs in the original NCDB sample. The Brier score, concordance index, and calibration slope of the standard MLE model and various PRMs were compared.

Bootstrap sample. Next, we created a distribution of 1000 bootstrap samples. In bootstrapping, random samples are drawn with replacement from the original data set.[22] From bootstrap samples, we obtained an "apparent" predictive accuracy distribution by using the bootstrap samples to develop the model and evaluate the performance. Finally, we also obtained an "internally validated" predictive accuracy distribution by using the bootstrap samples to develop the model and the original sample to evaluate the performance.

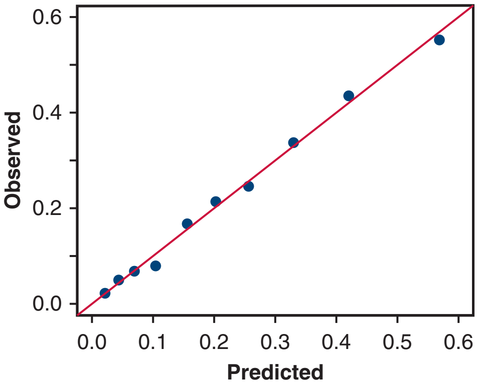


FIGURE 1. An example of a calibration plot. A perfectly accurate prediction model would result in a plot in which the plotted observed and predicted probabilities follow the 45° line (slope = 1). The x coordinate of the dot represents predicted probability, and the y coordinate represents observed probability. *Each dot* represents the performance of a subset of data, ordered by the predicted probability. The *solid line* represents the regression line of the dots and corresponds to a slope of 1 in this case.

Overfitting. An overfitted model classifies very accurately in the training data set (which is used to develop the model) but has poor performance in the validation data set used to test model performance. Overfitting in a model is an undesirable characteristic and is assessed via "optimism" in all 3 performance measures (ie, Brier score, calibration slope, and C statistic). "Optimism" is defined as the difference between "apparent" performance and "internally validated" performance. The larger the "optimism" value, the more overfitting in the model. A 0 value indicates no optimism and no overfitting. An optimism value close to 0 is desirable.

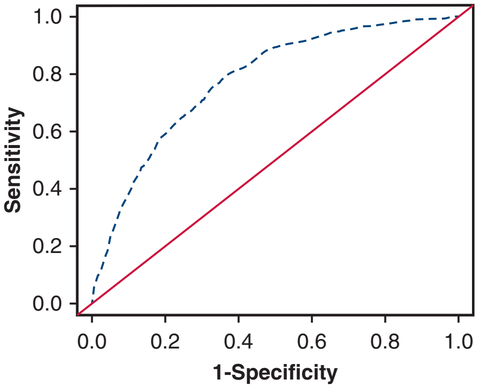


FIGURE 2. An example of a receiver operating characteristic (ROC) curve. The *dashed line* represents the ROC curve of the predictive model and the *solid line* represents the performance of a random classifier. The larger the area under the ROC curve (C statistic), the higher the discrimination of the model.

TABLE 1. Demographic and tumor characteristics in the cohort

Variable	Path N0 (n = 2498)	Path N+ (n = 688)	*P* value
Mean age (± SD), y	65.0 ± 9.9	65.2 ± 10.7	.66
Male sex	1980 (79.3)	565 (82.1)	.10
White race	2286 (91.5)	639 (92.9)	.25
Private insurance	996 (39.9)	276 (40.1)	.91
Population of area >250,000	1789 (71.6)	477 (69.3)	.24
Median income <$38,000	365 (14.6)	107 (15.6)	.54
>21% without high school diploma	324 (13.0)	75 (10.9)	.15
Charlson–Deyo Score			.48
0	1699 (68.0)	459 (66.7)	
1	632 (25.3)	174 (25.3)	
≥2	167 (6.7)	55 (8.0)	
Academic center	1562 (62.5)	413 (60.0)	.23
Diagnosis year			.15
2004	163 (6.5)	38 (5.5)	
2005	157 (6.3)	54 (7.9)	
2006	184 (7.4)	54 (7.9)	
2007	208 (8.3)	67 (9.7)	
2008	306 (12.3)	98 (14.2)	
2009	334 (13.4)	77 (11.2)	
2010	273 (10.9)	71 (10.3)	
2011	253 (10.1)	69 (10.0)	
2012	209 (8.4)	69 (10.0)	
2013	221 (8.9)	43 (6.3)	
2014	190 (7.6)	48 (7.0)	
Histology			.05
Squamous	548 (21.9)	127 (18.5)	
Adenocarcinoma	1950 (78.1)	561 (81.5)	
T stage			<.001
T1	1602 (64.1)	238 (34.6)	
T2	639 (25.6)	287 (41.7)	
T3	257 (10.3)	163 (23.7)	
Tumor Size			<.001
<1 cm	539 (21.6)	24 (3.5)	
1 to <2 cm	605 (24.2)	83 (12.1)	
2 to <3 cm	531 (21.3)	151 (22.0)	
3 to <4 cm	343 (13.7)	166 (24.1)	
4 to <5 cm	216 (8.7)	115 (16.7)	
≥5 cm	264 (10.6)	149 (21.7)	
Grade			<.001
1	371 (14.9)	20 (2.9)	
2	1103 (44.2)	253 (36.8)	
3	780 (31.2)	373 (54.2)	
4	35 (1.4)	18 (2.62)	
Unknown	209 (8.4)	24 (3.5)	
Lymphovascular invasion			<.001
Yes	171 (6.9)	155 (22.5)	
No	846 (33.9)	99 (14.4)	
Missing	1481 (59.3)	434 (63.1)	

Data are presented as n (%) except where otherwise noted.

TABLE 2. Comparison of standard MLE and PRM in the original data

Type of model	Brier score	C index	Calibration slope
MLE (standard)	0.140	0.779	1.000
Ridge	0.140	0.780	1.119
Lasso	0.140	0.782	1.021
Adaptive lasso	0.140	0.780	1.034
Elastic net	0.140	0.782	1.021

Ridge, lasso, adaptive lasso, and elastic net are the PRMs. *MLE*, Maximum likelihood estimation; *PRM*, penalized regression model.

RESULTS

The distribution of potential predictors of pathologic lymph node positive/negative status is shown in Table 1. Each level of a variable (eg, T stage has 3 levels: T1, T2, and T3) generates its own regression coefficient/parameter in the models. With year of diagnosis as a continuous variable in the model, we had 25 parameters in the model. Because 688 patients had unexpected nodal disease (the outcome of interest), we had 688 events. This results in a ratio of 668/25 or 27.5 EPV.

The predictive performance of the standard MLE and PRMs in the original sample is shown in Table 2. In the original sample, the Brier scores for all models are indistinguishable to within 3 digits (0.140), indicating that overall performance is identical for these models. The C statistics for all models are also very similar, from 0.779 for standard MLE to 0.782 for lasso and elastic net models. All models with PRM are underfitted (calibration slope >1), whereas the MLE model has perfect calibration (calibration slope = 1).

The predictive performance of the standard MLE and PRMs in the bootstrap samples is shown in Table 3. This represents the "apparent" predictive accuracy of the model. Again, similar to the model performance in the original data, the standard MLE and the various PRMs have virtually identical performance.

The "internally validated" performance is the result of developing the model in the bootstrap sample and testing model performance in the original cohort. The "internally validated performance" of the standard MLE and PRMs is shown in Table 4. The standard MLE and PRMs have very similar performance.

The optimism values for the Brier score, C index, and calibration slopes for the various models are shown in Table 5. Optimism reflects the degree of overfitting in the models, and values close to 0 are more desirable. The optimism values for all 3 measures of model performance are closer to 0 for the standard MLE models than for any of the PRMs.

Next, in an exploratory analysis, we used the current data set to perform simulations to evaluate the sample size and outcome frequency needed for standard MLE to perform similarly to PRMs on the basis of the C index, Brier score, and calibration slope. At sample size >20% to 30% of the total sample size (approximately 600-900 patients), standard MLE was noninferior to PRMs across all 3 metrics (Figure E1). Outcome frequency simulation was performed by setting the outcome of interest (lymph node involvement by cancer) at 5%, 10%, 20%, 40%, and 80% using a sample size of 800 patients. Performance of standard MLE approached that of the PRMs at an outcome frequency >20% (Figure E2).

A classification table was additionally generated using an arbitrary but clinical reasonable cutoff at 30% probability of occult nodal metastasis. Sensitivity and specificity were similar for the standard MLE and any PRM model (Table E1).

DISCUSSION

In this analysis, we used a risk model for predicting the probability of occult lymph node metastases in patients with surgically resectable esophageal cancer as a case study to compare standard logistic regression (MLE) and the more complicated penalized logistic regression methods. Predictive performance (Brier score, concordance index, and calibration slope) was quantified in the original sample and in bootstrap samples. In this case study, the standard MLE model had virtually identical predictive accuracy as the more sophisticated PRMs. Moreover, overfitting from the MLE model was no greater than that from the PRMs (Figures 3 and 4).

Generalizability of a risk model from a derivation sample to new patients depends on several factors[23-27]: similarity between derivation and new patient cohort in the definition of predictors and outcome, patient selection (case mix), the prevalence/incidence of the outcome, and

TABLE 3. Comparison of "apparent" predictive accuracy between MLE and PRM in the bootstrap sample

Type of model	Brier score	C index	Calibration slope
MLE (standard)	0.140 (0.132, 0.147)	0.782 (0.763, 0.801)	1.000 (1.000, 1.000)
Ridge	0.139 (0.131, 0.146)	0.786 (0.768, 0.803)	1.122 (1.100, 1.145)
Lasso	0.138 (0.130, 0.146)	0.788 (0.769, 0.804)	1.036 (1.012, 1.090)
Adaptive lasso	0.138 (0.131, 0.146)	0.787 (0.767, 0.804)	1.017 (1.004, 1.057)
Elastic net	0.138 (0.130, 0.146)	0.788 (0.769, 0.804)	1.040 (1.012, 1.091)

Ridge, lasso, adaptive lasso, and elastic net are the PRMs. Values shown are median (2.5 percentile, 97.5 percentile). *MLE*, Maximum likelihood estimation; *PRM*, penalized regression model.

TABLE 4. Comparison of predictive accuracy between MLE and PRM with the model developed in bootstrap samples and tested in the original cohort ("internally validated" performance)

Type of model	Brier score	C index	Calibration slope
MLE (standard)	0.141 (0.141, 0.142)	0.777 (0.774, 0.779)	0.968 (0.875, 1.078)
Ridge	0.141 (0.141, 0.142)	0.775 (0.771, 0.778)	1.059 (0.974, 1.158)
Lasso	0.141 (0.140, 0.142)	0.777 (0.773, 0.779)	0.976 (0.883, 1.096)
Adaptive lasso	0.141 (0.140, 0.142)	0.777 (0.772, 0.780)	0.965 (0.872, 1.079)
Elastic net	0.141 (0.140, 0.142)	0.777 (0.772, 0.779)	0.978 (0.883, 1.094)

Ridge, lasso, adaptive lasso, and elastic net are the PRMs. Values shown are median (2.5 percentile, 97.5 percentile). *MLE*, Maximum likelihood estimation; *PRM*, penalized regression model.

overfitting of the risk model. An overfitted model tends to contain the true predictor–outcome relationship and the noise–bizarre relationship in the derivation data set. When overfitting is present, predictive performance is exaggerated in the derivation data set and the observed future patient outcomes do not agree with predicted values.

Shrinkage is a very popular way to alleviate the problem of model overfitting by shrinking the regression coefficients toward 0. It has the effect of moving extreme predicted values toward the average risk, therefore making more accurate predictions when the model is applied in new patients. Although we may apply shrinkage after estimation of regression coefficients using a shrinkage factor,[28] applying shrinkage during estimation using a penalized method is a novel approach to reduce model overfitting. However, in our case study, none of the 4 PRMs outperformed a standard logistic regression model.

There are several explanations for our findings. First, we have a relatively large data set with a high event rate (688/3186; 22% of patients with occult lymph node metastases) and a modest number of candidate predictors (15 variables with 25 regression parameters). Hence our EPV of 27.5 (688 events divided by 25 parameters) is significantly larger than the rule of thumb of needing approximately 10 to 15 EPV in the model. Second, the relationship of most predictors (stage, grade, tumor size, etc) with the outcome is well understood from clinical experience and previous cohort studies, thus making variable selection more intuitive. Third, we adhered to recommended good practices in building our standard logistic regression predictive model. We carefully examined the distribution of each potential predictor, meaningfully recoded some of the variables as necessary, used clinical knowledge supplemented by statistical methods for variable selection, and tested clinically important interactions. A combination of a sizable data set, clinical knowledge about possible predictors, and appropriate statistical methods allowed us to fit a risk model using standard logistic regression which has a predictive accuracy similar to that of more complicated PRMs.

Despite our findings, there are certain situations in surgical outcomes research in which PRMs are better suited for developing risk prediction rules. In the event of a relatively small data set, a rare outcome, and a large number of potential predictors (thus a low EPV) with possible correlations necessitating an elaborate statistical approach for variable selection, PRMs are advantageous over standard regression models because of their ability to shrink extreme coefficients that cause overfitting. This scenario is likely to arise when analyzing institutional data sets with granular variables or studying a relatively uncommon event. Conversely, for large data sets with a relatively frequent outcome of interest (leading to a high EPV) and a moderate number of potential predictors, standard regression models are very likely to have similar predictive performance as the more complicated PRMs. In addition, in the clinical setting, it is important to minimize the number of inputs in the predictive model to create a practical user interface. Although one might expect the PRM methods to create models with fewer variables, in our study, ridge, lasso, and elastic net did not shrink any parameter estimates to 0. Therefore, all 15

TABLE 5. Comparison of optimism in the measures of model performance between the standard MLE and the PRM

Type of model	Optimism of Brier score	Optimism of C index	Optimism of calibration slope
MLE (standard)	−0.002 (−0.010, 0.005)	0.006 (−0.013, 0.024)	0.032 (−0.078, 0.125)
Ridge	−0.003 (−0.011, 0.005)	0.011 (−0.008, 0.028)	0.063 (−0.045, 0.152)
Lasso	−0.003 (−0.011, 0.004)	0.011 (−0.008, 0.028)	0.062 (−0.042, 0.148)
Adaptive lasso	−0.003 (−0.011, 0.005)	0.010 (−0.009, 0.028)	0.056 (−0.049, 0.143)
Elastic net	−0.003 (−0.011, 0.004)	0.011 (−0.008, 0.028)	0.062 (−0.042, 0.148)

Ridge, lasso, adaptive lasso, and elastic net are the PRMs. Values shown are median (2.5 percentile, 97.5 percentile). Values closer to 0 are more desirable. *MLE*, Maximum likelihood estimation; *PRM*, penalized regression model.

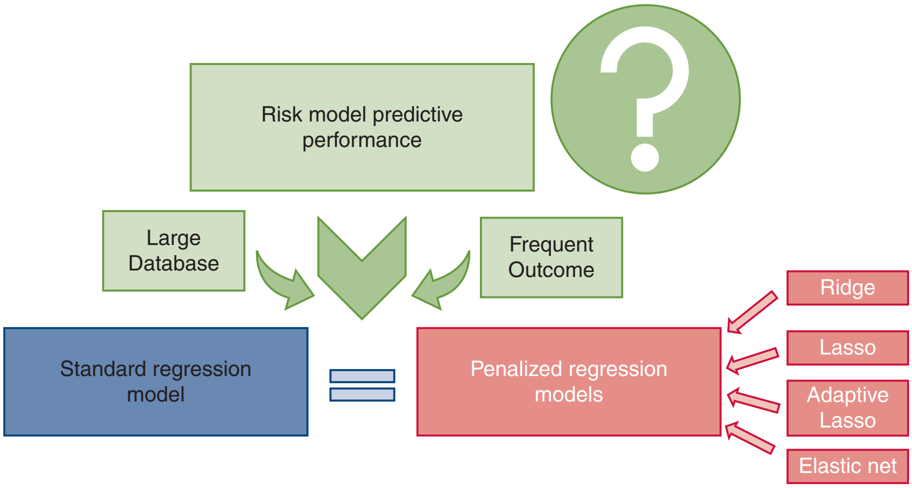
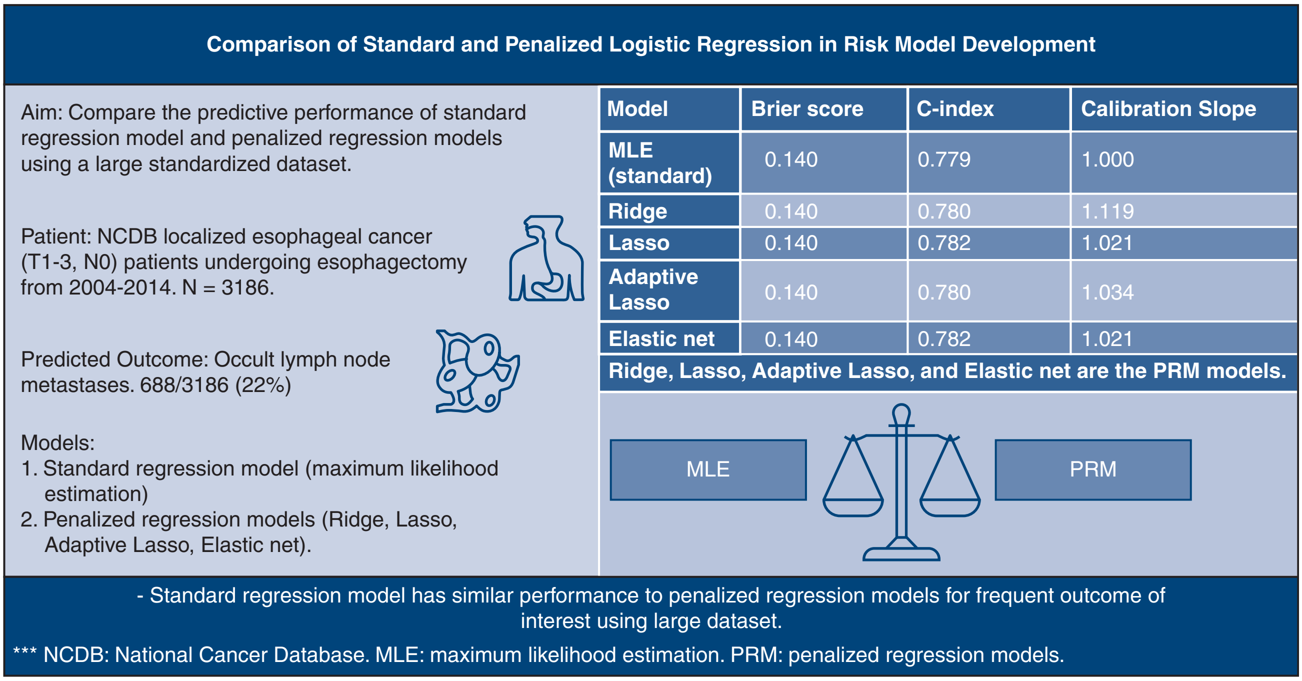

Risk model predictive performance
Large Database
Frequent Outcome
Standard regression model
Penalized regression models
Ridge
Lasso
Adaptive Lasso
Elastic net

FIGURE 3. The predictive performance of standard regression model is similar to penalized regression models when predicting a frequent outcome using a large database.

variables were kept in the models. Adaptive lasso selected 6 variables. In contrast, standard MLE only selected 5 variables in the final model and is in fact more easily applicable in the clinical setting.

In simulations using our data set we noted that when the sample size was >600 to 900 and the frequency of the outcome was greater than 20% to 30%, the standard MLE model performed as well as the more complex PRMs. Because extensive knowledge regarding the clinical relevance of candidate predictors is necessary to appropriately select variables included in regression models, we did not perform simulations for the number of candidate variables. Furthermore, these simulations remain specific to the analyzed data set.

A caveat in using a large clinical or administrative database to compare different statistical methods should be noted. That is, if the variables in a database do not contain detailed clinical information necessary for high-quality risk prediction



Comparison of Standard and Penalized Logistic Regression in Risk Model Development

Aim: Compare the predictive performance of standard regression model and penalized regression models using a large standardized dataset.

Patient: NCDB localized esophageal cancer (T1-3, N0) patients undergoing esophagectomy from 2004-2014. N = 3186.

Predicted Outcome: Occult lymph node metastases. 688/3186 (22%)

Models:
1. Standard regression model (maximum likelihood estimation)
2. Penalized regression models (Ridge, Lasso, Adaptive Lasso, Elastic net).

Model	Brier score	C-index	Calibration Slope
MLE (standard)	0.140	0.779	1.000
Ridge	0.140	0.780	1.119
Lasso	0.140	0.782	1.021
Adaptive Lasso	0.140	0.780	1.034
Elastic net	0.140	0.782	1.021

Ridge, Lasso, Adaptive Lasso, and Elastic net are the PRM models.

- Standard regression model has similar performance to penalized regression models for frequent outcome of interest using large dataset.

*** NCDB: National Cancer Database. MLE: maximum likelihood estimation. PRM: penalized regression models.

FIGURE 4. The predictive performance of a standard logistic regression model and a penalized regression model (*PRM*) was compared using National Cancer Database (*NCDB*) esophagectomy data regarding occult nodal metastasis. In this scenario, the standard logistic regression model is noninferior to the PRM. *MLE*, maximum likelihood estimation.

models, the difference in model predictive performance from different statistical methods will be obscured. The NCDB used in this study is a robust national database, but more granular information including patient comorbidities, laboratory values, and modality of clinical staging were lacking. It is possible that the model performance was limited by the quality of data itself rather than the methods.

CONCLUSIONS

The choice of statistical methods for risk model development should be on the basis of the nature of the data at hand and good practice rather than the complexity or novelty of statistical models. Close collaboration between clinicians who will be end users of the models, and the statistical team is critical to inform this process.

Conflict of Interest Statement

The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

References

1. Liang W, Zhang L, Jiang G, Wang Q, Liu L, Liu D, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. *J Clin Oncol*. 2015;33:861-9.
2. Kent MS, Mandrekar SJ, Landreneau R, Nichols F, Foster NR, DiPetrillo TA, et al. A nomogram to predict recurrence and survival of high-risk patients undergoing sublobar resection for lung cancer: an analysis of a multicenter prospective study (ACOSOG Z4032). *Ann Thorac Surg*. 2016;102:239-46.
3. Ohkura Y, Miyata H, Konno H, Udagawa H, Ueno M, Shindoh J, et al. Development of a model predicting the risk of eight major postoperative complications after esophagectomy based on 10 826 cases in the Japan National Clinical Database. *J Surg Oncol*. December 10, 2019 [Epub ahead of print].
4. Foster CA, Charles EJ, Turrentine FE, Sohn MW, Kron IL, Jones RS, et al. Development and validation of procedure-specific risk score for predicting postoperative pulmonary complication: a NSQIP analysis. *J Am Coll Surg*. 2019;229: 355-65.e3.
5. Puri V, Patel AP, Crabtree TD, Bell JM, Broderick SR, Kreisel D, et al. Unexpected readmission after lung cancer surgery: a benign event. *J Thorac Cardiovasc Surg*. 2015;150:1496-504. 1505.e1-5, discussion: 1504-5.
6. Samson P, Puri V, Robinson C, Lockhart C, Carpenter D, Broderick S, et al. Clinical T2N0 esophageal cancer: identifying pretreatment characteristics associated with pathologic upstaging and the potential role for induction therapy. *Ann Thorac Surg*. 2016;101:2102-11.
7. Safi S, Benner A, Walloschek J, Renner M, op den Winkel J, Muley T, et al. Development and validation of a risk score for predicting death after pneumonectomy. *PLoS One*. 2015;10:e0121295.
8. Kaiser HA, Saied NN, Kokoefer AS, Saffour L, Zoller JK, Helwani MA, et al. Incidence and prediction of intraoperative and postoperative cardiac arrest requiring cardiopulmonary resuscitation and 30-day mortality in non-cardiac surgical patients. *PLoS One*. 2020;15:e0225939.
9. Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol*. 2004;57:1262-70.
10. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med*. 2016;35:1159-77.
11. Chen Q, Nian H, Zhu Y, Talbot HK, Griffin MR, Harrell FE Jr. Too many covariates and too few cases? - a comparative study. *Stat Med*. 2016;35:4546-58.
12. Ambler G, Seaman S, Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. *Stat Med*. 2011;31:1150-61.
13. Schweiger T, Liebmann-Reindl S, Glueck O, Starlinger P, Laengle J, Birner P, et al. Mutational profile of colorectal cancer lung metastases and paired primary tumors by targeted next generation sequencing: implications on clinical outcome after surgery. *J Thorac Dis*. 2018;10:6147-57.
14. Buccheri S, Capodanno D, Barbanti M, Popolo Rubbio A, Di Salvo ME, Scandura S, et al. A risk model for prediction of 1-year mortality in patients undergoing MitraClip implantation. *Am J Cardiol*. 2017;119:1443-9.
15. Greenwood CJ, Youssef GJ, Letcher P, Macdonald JA, Hagg LJ, Sanson A, et al. A comparison of penalised regression methods for informing the selection of predictive markers. *PLoS One*. 2020;15:e0242730.
16. Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Stat Methods Med Res*. 2020;29:3166-78.
17. Desboulets LDD. A review on variable selection in regression analysis. *Econometrics*. 2018;6:45.
18. Semenkovich TR, Yan Y, Subramanian M, Meyers BF, Kozower BD, Nava R, et al. A clinical nomogram for predicting node-positive disease in esophageal cancer. *Ann Surg*. 2021;273:e214-21.
19. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48:1503-10.
20. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer; 2009.
21. Myung IJ. Tutorial on maximum likelihood estimation. *J Math Psychol*. 2003;47: 90-100.
22. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol*. 2008;26:1364-70.
23. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol*. 2003;56:826-32.
24. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA*. 1997;277:488-94.
25. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130:515-24.
26. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med*. 1999;33:437-47.
27. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19:453-73.
28. Harrell F. Chapter 4.5 Shrinkage. In: *Regression Modeling Strategies*. 2nd ed. Springer; 2015.

Key Words: regression, predictive modeling, esophagectomy, outcome research

APPENDIX 1. DESCRIPTION OF 4 COMMON PENALIZED METHODS AND OTHER SHRINKAGE METHODS

1. Ridge or $L2$ penalization places a constraint on the sum of squares of regression coefficients $(\lambda\sum_{j=1}^{p}\beta_j^2)$. Ridge was initially designed to deal with issues of many highly correlated variables in a model. It shrinks the regression coefficients toward 0 (but not exactly 0) and has been shown to perform well in scenarios with correlated predictors.
2. Lasso (least absolute shrinkage and selection operator) or $L1$ penalization imposes a constraint on the sum of the absolute value of regression coefficients, and the penalty term has the form $(\lambda\sum_{j=1}^{p}|\beta_j|)$. Lasso was developed for variable selection in high-dimensional data analyses when the number of parameters (p) is much larger than the sample size (p >> n). In lasso, the model can yield 0 estimates when the parameter values are close to 0 and hence can perform variable selection, resulting in parsimonious models. This method is not optimal in the presence of many correlated predictors. It might select 1 at random from a group of highly correlated predictors, which can affect the interpretation of the model and compromise its predictive accuracy.
3. Elastic net is a mixture of ridge and lasso. It has a penalty with ridge and lasso components $(\lambda[\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2])$. The α can be considered as a mixing parameter, describing relative contribution of ridge and lasso to the penalty term. Elastic net combines the strengths of ridge and lasso: it can produce more parsimonious models than ridge by performing variable selection while also tending to select or omit highly correlated predictors as a group.
4. Adaptive lasso is a variant of lasso. It extends lasso by allowing a different weight for each parameter in the penalty term $(\lambda\sum_{j=1}^{p}\omega_j|\beta_j|)$. The weights ω_j allow us to shrink different coefficients differently: coefficients of strong predictors are shrunk less and coefficients of weak predictors are shrunk more. These weights are data-dependent and can be derived from the inverse of the corresponding coefficient from ridge regression.

PRMs provide a better way to obtain shrunk regression parameter estimates because shrinkage is built into the estimation process. Another approach to alleviating overfitting is the use of a shrinkage coefficient after obtaining regression parameter estimates. The shrinkage coefficient can be estimated from the original model fit. For generalized linear models, it can be estimated from (model $\chi^2 -$ p)/model χ^2, where model χ^2 is the likelihood ratio χ^2 statistic for testing the global null hypothesis of all predictors simultaneously, and p is the total degrees of freedom for the predictors including those in the final model and those tested but not in the final model. For an ordinary linear model, it can be estimated from $(n - p - 1)/(n - 1)*R_{adj}^2/R^2$. The shrinkage coefficient can also be estimated as the average calibration slope using bootstrap or cross-validation.

After we obtain the shrinkage coefficient estimate, $\widehat{\gamma}$, we make adjustments to original regression parameter estimates $(\widehat{\beta}_0, \widehat{\beta}_j)$ as

$$\widehat{\beta}_0^s = (1-\widehat{\gamma})\overline{Y}+\widehat{\gamma}\widehat{\beta}_0$$
$$\widehat{\beta}_j^s = \widehat{\gamma}\widehat{\beta}_j, j = 1, \ldots, p.$$

where $\overline{Y}$ is the mean of the response vector[10].

APPENDIX 2. COMPUTATION METHODS

For all 4 penalized methods, λ is an important tuning parameter. $\lambda = 0$ Corresponds to standard maximum likelihood estimation, and as λ increases, the effect of shrinkage penalty grows. When λ is sufficiently large, the regression coefficients will approach (or equal) 0. In addition to tuning parameter λ, we need to find appropriate weights ω_j for adaptive lasso, and optimal mixing parameter α for elastic net.

For ridge and lasso, we use 10-fold cross-validation to identify the optimal value of λ.

1. We partition our data into 10 mutually exclusive blocks of equal size; the same partition of data into 10-fold are maintained across different penalized methods to avoid random sampling variations.
2. In the first fold, we use the first 9 blocks as training data to fit the ridge (lasso) model with each candidate value λ_j (j = 1,2,3,…,m), and use the fit to generate predicted values for each subject in the last block. Thus, after fold 1 is completed, we have predicted values for one-tenth of the data for each λ_j value. Note these predicted values are generated on data not used to fit the model.
3. Repeating the process in step 2, we rotate the validation set—block 10, 9,…, 1, such that each block serves as the

Patient ID	λ_1	λ_2	…	…	…	λ_m	Y
1	0.18	0.23				0.35	0
2	0.54	0.76				0.45	1
…							…
n							
	cvMSE(λ_1)	cvMSE(λ_2)				cvMSE(λ_m)	

validation data once. At the end of complete 10-fold cross-validation, we have predicted values for each subject under each λ_j value. These predicted values use the held-out validation sample from each fold.

4. We calculate the cross-validated mean square error (cvMSE; Brier score) for each λ_j using predicted values and observed outcome as:

$$cvMSE(\lambda_j) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_{ji} - Y_i\right)^2$$

The optimal λ value is chosen as the minimum of cvMSE(λj) – the λ_j at which the cvMSE achieves the minimum.

For adaptive lasso, we first perform ridge regression, using the λ value minimizing the mean square error (MSE) metric in 10-fold cross-validation, to obtain coefficients for each predictor. Then we fit lasso with the inverse of coefficients from ridge regression as the weights ω_j, using the λ value minimizing the MSE metric in 10-fold cross-validation.

For elastic net, the optimal value of α and λ is determined by searching the grid (α = seq(0.1, 0.9, 0.05), $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$). That is, for each value of α from 0.1 to 0.9 at 0.05 interval, we search for optimal value of λ_j minimizing the MSE metric in 10-fold cross-validation.

APPENDIX 3. REGRESSION PARAMETER ESTIMATES FROM MLE AND PRM

	MLE	MLE-full	Ridge	Lasso	Adaptive lasso	Elastic net
Age		−0.007	−0.005 (0.23)	−0.006 (0.06)	0.000 (1.00)	−0.006 (0.06)
Sex		−0.088	−0.097 (0.10)	−0.083 (0.06)	0.000 (1.00)	−0.084 (0.05)
Race		0.041	0.051 (0.25)	0.029 (0.28)	0.000 (1.00)	0.031 (0.25)
Insurance		0.042	0.050 (0.18)	0.039 (0.09)	0.000 (1.00)	0.039 (0.08)
Population of area		0.046	0.054 (0.17)	0.043 (0.06)	0.000 (1.00)	0.044 (0.05)
Median income		−0.278	−0.234 (0.16)	−0.260 (0.07)	−0.192 (0.31)	−0.261 (0.06)
% Without high school diploma		0.395	0.336 (0.15)	0.375 (0.05)	0.295 (0.25)	0.376 (0.05)
Charlson–Deyo score (1 vs 0)		0.068	0.050 (0.26)	0.056 (0.18)	0.000 (1.00)	0.057 (0.17)
Charlson–Deyo score (≥2 vs 0)		0.066	0.060 (0.09)	0.050 (0.25)	0.000 (1.00)	0.051 (0.23)
Facility type		−0.076	−0.049 (0.35)	−0.065 (0.14)	0.000 (1.00)	−0.066 (0.13)
Diagnosis year		0.026	0.011 (0.60)	0.021 (0.18)	0.000 (1.00)	0.022 (0.17)
Histology	0.558	0.523	0.450 (0.14)	0.513 (0.02)	0.519 (0.01)	0.513 (0.02)
Tumor stage (T2 vs T1)	0.639	0.660	0.668 (0.01)	0.659 (0.00)	0.648 (0.02)	0.659 (0.00)
Tumor stage (T3 vs T1)	0.775	0.822	0.828 (0.01)	0.817 (0.01)	0.777 (0.05)	0.818 (0.01)
Tumor size (1-2 vs <1)	0.812	0.828	0.234 (0.72)	0.704 (0.15)	0.501 (0.39)	0.707 (0.15)
Tumor size (2-3 vs <1)	1.342	1.371	0.758 (0.45)	1.250 (0.09)	1.076 (0.22)	1.252 (0.09)
Tumor size (3-4 vs <1)	1.686	1.715	1.103 (0.36)	1.595 (0.07)	1.430 (0.17)	1.598 (0.07)
Tumor size (4-5 vs <1)	1.734	1.753	1.140 (0.35)	1.634 (0.07)	1.473 (0.16)	1.636 (0.07)
Tumor size (≥5 vs <1)	1.796	1.824	1.189 (0.35)	1.702 (0.07)	1.527 (0.16)	1.704 (0.07)
Grade (2 vs 1)	0.964	0.969	0.496 (0.49)	0.870 (0.10)	0.600 (0.38)	0.873 (0.10)
Grade (3 vs 1)	1.480	1.495	1.024 (0.31)	1.398 (0.06)	1.137 (0.24)	1.402 (0.06)
Grade (4 vs 1)	1.423	1.422	0.957 (0.33)	1.314 (0.08)	1.048 (0.26)	1.319 (0.07)
Grade (unknown vs 1)	0.628	0.625	0.099 (0.84)	0.501 (0.20)	0.000 (1.00)	0.506 (0.19)
Lymphovascular invasion (yes vs no)	1.548	1.539	1.446 (0.06)	1.530 (0.01)	1.544 (0.00)	1.529 (0.01)
Lymphovascular invasion (missing vs no)	0.625	0.740	0.598 (0.19)	0.710 (0.04)	0.619 (0.16)	0.711 (0.04)

MLE, Maximum likelihood estimation; *PRM*, penalized regression model.

Parameter estimates for the MLE model are from our recent publication,[18] and MLE-full is from the model with all potential predictors. Percentage shrinkage is the difference in the parameter estimate between MLE-full model and PRM divided by the parameter estimate from the MLE-full model. For example, the parameter estimate of age from the ridge model is -0.005 and from the MLE-full model is -0.007. Then the percentage shrinkage for the parameter estimate of age from the ridge model is $[-0.007 - (-0.005)]/-0.007 = 0.23$, or 23%.

Lasso does not shrink any regression coefficient to 0, nor does elastic net in this data set. Lasso and elastic net have a very similar shrinkage pattern. Regarding the shrinkage properties of adaptive lasso, it is observed that small coefficients tend to be shrunk to 0, whereas large coefficients are shrunk less than small coefficients. For example, the coefficients of all demographic and socioeconomic variables are shrunk to 0, the coefficients of tumor size and grade are shrunk much less than in ridge (for tumor size shrunk by 15% to 39% for adaptive lasso vs 35% to 72% for ridge).

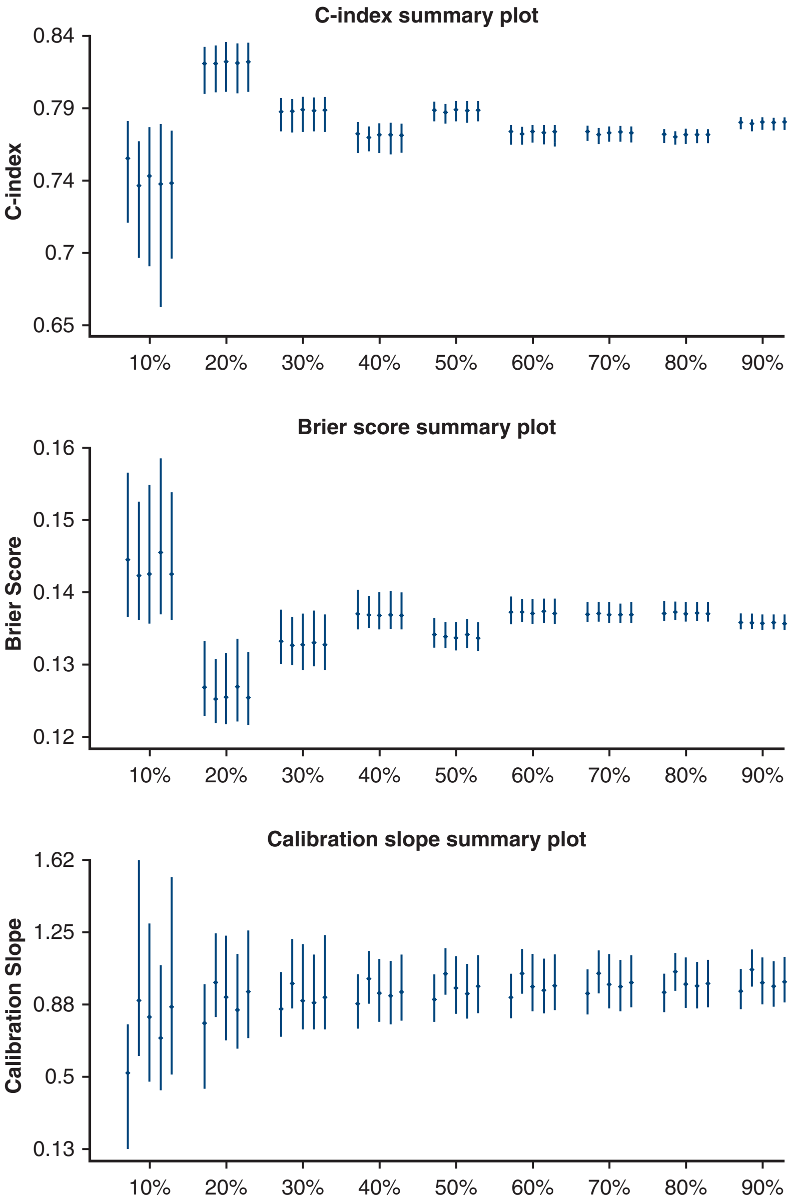


FIGURE E1. C index, Brier score, and calibration slope for simulation of sample size from 10% to 90% of total sample size. The first *vertical line* in each cluster represents the standard maximum likelihood estimation model, followed by ridge, lasso, adaptive lasso, and elastic net models. Length of vertical lines corresponds to the 95% confidence interval on the basis of the distribution of 500 bootstrap samples, and the *diamond* represents the median.

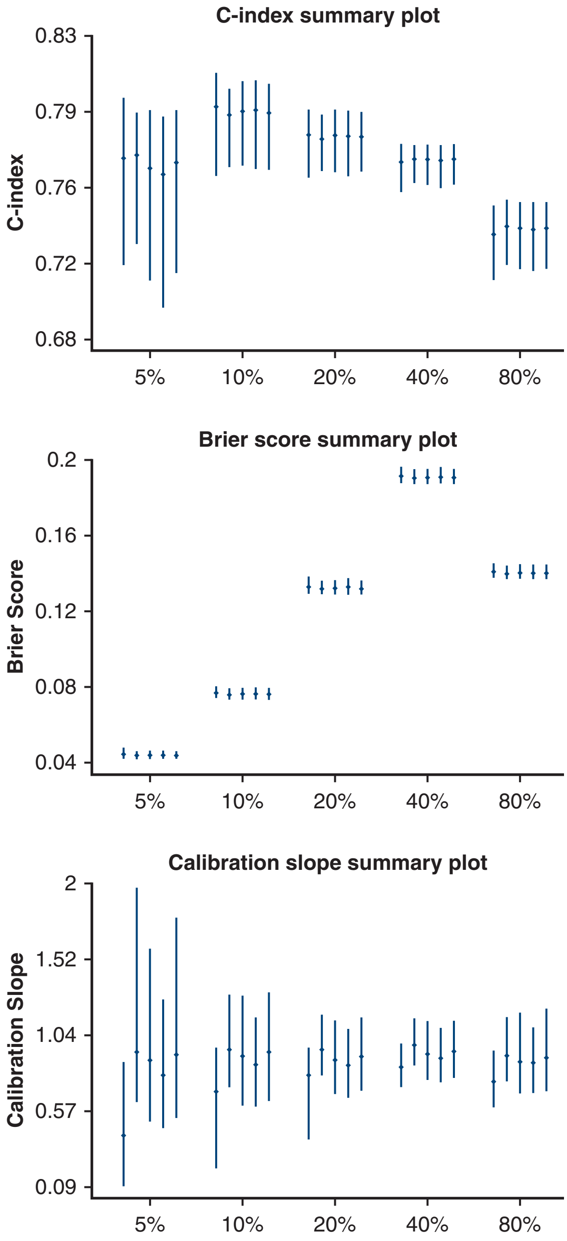


FIGURE E2. C index, Brier score, and calibration slope for simulation of outcome frequency at 5%, 10%, 20%, 40%, and 80% with 800 patients. The *first vertical line* in each cluster represents the standard maximum likelihood estimation model, followed by ridge, lasso, adaptive lasso, and elastic net models. Length of vertical lines corresponds to the 95% confidence interval on the basis of the distribution of 500 bootstrap samples, and the *diamond* represents the median.

TABLE E1. Classification table of the 5 prediction models with cutoff at 30% outcome probability

Type of model	Sensitivity	Specificity
MLE (standard)	59.16%	79.82%
Ridge	57.70%	81.02%
Lasso	59.59%	80.06%
Adaptive lasso	58.28%	80.34%
Elastic net	59.59%	80.06%

MLE, Maximum likelihood estimation.