Developing a Clinical Prediction Score: Comparing Prediction Accuracy of Integer Scores to Statistical Regression Models

Vigneshwar Subramanian, BA,* Edward J. Mascha, PhD,† and Michael W. Kattan, PhD‡

Researchers often convert prediction tools built on statistical regression models into integer scores and risk classification systems in the name of simplicity. However, this workflow discards useful information and reduces prediction accuracy. We, therefore, investigated the impact on prediction accuracy when researchers simplify a regression model into an integer score using a simulation study and an example clinical data set. Simulated independent training and test sets (n = 1000) were randomly generated such that a logistic regression model would perform at a specified target area under the receiver operating characteristic curve (AUC) of 0.7, 0.8, or 0.9. After fitting a logistic regression with continuous covariates to each data set, continuous variables were dichotomized using data-dependent cut points. A logistic regression was refit, and the coefficients were scaled and rounded to create an integer score. A risk classification system was built by stratifying integer scores into low-, intermediate-, and high-risk tertiles. Discrimination and calibration were assessed by calculating the AUC and index of prediction accuracy (IPA) for each model. The optimism in performance between the training set and test set was calculated for both AUC and IPA. The logistic regression model using the continuous form of covariates outperformed all other models. In the simulation study, converting the logistic regression model to an integer score and subsequent risk classification system incurred an average decrease of 0.057–0.094 in AUC, and an absolute 6.2%–17.5% in IPA. The largest decrease in both AUC and IPA occurred in the dichotomization step. The dichotomization and risk stratification steps also increased the optimism of the resulting models, such that they appeared to be able to predict better than they actually would on new data. In the clinical data set, converting the logistic regression with continuous covariates to an integer score incurred a decrease in externally validated AUC of 0.06 and a decrease in externally validated IPA of 13%. Converting a regression model to an integer score decreases model performance considerably. Therefore, we recommend developing a regression model that incorporates all available information to make the most accurate predictions possible, and using the unaltered regression model when making predictions for individual patients. In all cases, researchers should be mindful that they correctly validate the specific model that is intended for clinical use. (Anesth Analg 2021;132:1603–13)

GLOSSARY

AIC = Akaike information criterion; **ARISCAT** = Assess Respiratory Risk in Surgical Patients in Catalonia; **AUC** = area under the receiver operating characteristic curve; **ICU** = intensive care unit; **IPA** = index of prediction accuracy; **LASSO** = least absolute shrinkage and selection operator; **MCSE** = Monte Carlo standard error; **ROC** = receiver operating characteristic; **SD** = standard deviation; **SE** = standard error; **TRIPOD** = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; **WBC** = white blood cell count

When designing statistical prediction models for clinician use, a key consideration is ease of use and interpretability. Clinicians are hesitant to use a model that requires too many inputs or is difficult to understand. Presumably for these reasons, medical researchers often develop a regression model using a chosen set of predictors and transform it into an integer score.[1–9] To achieve an integer score, a common approach is to categorize continuous covariates on the basis of some "optimal" cut points, which are frequently selected from univariate analyses. A regression model is fitted to the categorized inputs, and the regression coefficients are scaled by an arbitrary factor and rounded to the nearest integer. The integer score is the sum of present coefficients. Patients may subsequently be stratified into risk groups (eg, low, intermediate, and high risk) by separating them into quantiles based on the distribution of mapped risks.

From the *Cleveland Clinic Lerner College of Medicine at Case Western Reserve University, Cleveland, Ohio; and †Departments of Quantitative Health Sciences and Outcomes Research and ‡Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication November 25, 2020.

Funding: None.

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Reprints will not be available from the authors.

Address correspondence to Michael W. Kattan, PhD, Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Ave/JJN3-01, Cleveland, OH 44195. Address e-mail to kattanm@ccf.org.

DOI: 10.1213/ANE.0000000000005362

This approach has several perceived advantages. Critically, it is practical because it ensures that the scoring function produces a nonnegative integer, which can be calculated by summing the weights of each predictor that is present for a given patient. Integer scores can also be mapped to predicted probabilities using a lookup table or computer. If there are only a few predictors, clinicians can often calculate the score by hand.

However, each step in this workflow theoretically reduces prediction accuracy. Categorizing variables, particularly if continuous, discards useful information and decreases power.[10,11] Rounding scaled coefficients to integers introduces a small error.[12] Stratifying patients into broad risk categories decreases the granularity of predictions.[13] Models developed in this manner often generalize more poorly than expected to new patients (ie, they are overly optimistic).[14] Moreover, researchers often mistakenly assume that the predictive accuracy of the integer score is equivalent to that of the underlying regression, and therefore report the wrong metrics when evaluating model performance.

In this article, we empirically illustrate the consequences of converting a regression model to an integer score. Specifically, we use simulated data to build logistic regression models and show how performance changes after dichotomizing variables, rounding regression coefficients, and stratifying into risk categories. We also use a clinical data set from a published risk model[15] to illustrate the effect of simplifying to an integer score on model performance. Finally, we discuss these issues in the context of the Assess Respiratory Risk in Surgical Patients in Catalonia (ARISCAT) tool, a risk score for postoperative pulmonary complications after surgery,[1] and its external validation.[16]

METHODS

Data Simulation

We first simulated 3 series of sample data. Initially, 5 binary predictors and 2 continuous predictors were specified. Binary variables were assumed to be binomially distributed, and continuous variables were assumed to be normally distributed. Distributional assumptions and relevant parameters are listed in Supplemental Digital Content 1, Table 1, http://links.lww.com/AA/D323 (N= 1000 for each of 1000 simulation runs for each target AUC [0.70, 0.80, 0.90]. A random deviate was drawn for all predictors [X_1 through X_7] within each sample. Outcome data [Y] was estimated from a logistic regression with log-odds displayed above, with a multiplier applied to approximate the target AUC. Distributional assumptions were identical for simulations 1 and 2 [base + varying sample sizes]. In simulation 3, the correlation between predictors was varied. In simulation 4, the number of continuous and binary predictors was varied. SD: standard deviation. P[X=1]: probability that X = 1.). Binary outcome data were simulated as follows: probability of outcome was generated from a logistic regression with predictor odds ratios varying from 0.37 to 12.2; a multiplier was applied such that a logistic regression fitted to the data would have an area under the receiver operating characteristic curve (AUC) approximately equal to a specified target: 0.7, 0.8, or 0.9; the outcome was drawn from a binomial distribution with the modified probability. For each value of target AUC, we generated 1000 independent training data sets, each with 1000 observations, using a random seed. We also generated 1000 independent test data sets, each with 1000 observations, to validate the models.

Model Development

Prediction models were built sequentially on simulated data as follows. First, a logistic regression model was fitted to each training set, with continuous covariates retained as continuous. A risk classification system was built by stratifying predicted probabilities for the training set obtained from this logistic regression into low-, medium-, and high-risk tertiles. Continuous covariates were then dichotomized through univariate receiver operating characteristic (ROC) curve analyses. Cut points were selected as the value that maximized the sum of sensitivity and specificity (the Youden[17] J-statistic) over all possible cut point values from the observations specific to each data set. This method is not optimal in practice for multiple reasons: it may result in sensitivities and specificities far different from each other[18]; cut points are data-dependent and may not generalize well to other samples; and the cut points that maximize AUC in univariate analyses likely differ from those that maximize AUC in a multivariable model. However, it is commonly done in the literature,[7–9,19] and it serves our purposes here for comparing methods. For each data set, a second logistic regression was subsequently fitted to the dichotomized inputs to serve as the underlying model for the integer score. As is done in practice, reference levels for each binary covariate were specified such that all regression coefficients were positive, to achieve a nonnegative score. An integer score was developed by scaling and summing the regression coefficients from the logistic regression model with dichotomized covariates. Each regression coefficient was divided by the magnitude of the smallest coefficient, then rounded to the nearest integer. Thus, the covariate with the smallest regression coefficient contributed 1 point to the integer score, and the weights of the other covariates were scaled accordingly. This resulted in scores ranging from 0 points, if no covariate was present, to some maximum positive integer,

if all covariates were present. Integer scores were mapped to predicted probabilities by multiplying the score by the magnitude of the smallest coefficient (to return to the original scale), adding the intercept term, and taking the inverse logit. A second risk classification model was built by stratifying predicted probabilities mapped from integer scores for the training set into tertiles, representing low, intermediate, and high risks.

Model Performance

Each model was externally validated on an independent test set. For models built on dichotomized inputs, continuous covariates in the test set were dichotomized using the cut points derived from the training set. Model discrimination and calibration were assessed by calculating the AUC and the index of prediction accuracy (IPA). The AUC is a commonly reported measure of model discrimination.[20] The IPA is a rescaling of the Brier score, calculated as 1 − (model Brier score/null model Brier score), that reflects both discrimination and calibration. An IPA of 100% indicates a perfect model; 0% indicates a useless model, that is, the null model (model with no predictors) that predicts the event prevalence for all cases; and a negative value indicates a harmful model, that is, a model that performs worse than the null model.[21] Optimism in model performance between the training and test set was also assessed for both the AUC and the IPA by taking the difference between the metric as estimated in the training set and the metric as estimated in the test set. Monte Carlo standard errors (MCSEs) were estimated for AUC, IPA, and the optimism in each metric.

Simulated Scenarios

The above workflow was repeated 4 times to investigate the broader generalizability of findings under a variety of scenarios. In the base simulation, as described above, each data set had 1000 observations (n = 1000), all predictors were assumed to be independent ($r = 0$), and 2 continuous and 5 binary predictors were incorporated. Each subsequent simulation varied 1 condition. In the second simulation, sample size was varied from the base simulation: 1000 training and 1000 test data sets with n = 200 and n = 500 observations, respectively, were generated. In the third simulation, the correlation between predictors was varied from the base simulation: 1000 training and 1000 test data sets with $r = 0.10$ and $r = 0.30$, respectively, were generated. In the fourth simulation, the ratio of continuous to binary predictors was varied from the base simulation: 1000 training and 1000 test data sets with 5 continuous and 2 binary predictors, or 7 continuous predictors, respectively, were generated. In all 4 simulations, models were built, and performance was assessed on the basis of AUC, IPA, and optimism, as described previously.

Application of Methods to Study of Early Prediction of Prognosis in Elderly Acute Stroke Patients

Bautista et al[15] developed a risk score to predict in-hospital mortality in elderly acute stroke patients. We refer the reader to their publication for a complete explanation of methods. In brief, Bautista et al[15] collected data for 2 cohorts. The first sample consisted of baseline clinical data (defined as data available within 12 hours of hospital admission) for 584 elderly stroke patients (age ≥65 years) admitted to the intensive care unit (ICU) from the years 2005 to 2009. Bautista et al[15] split this sample into a training and a test subsample, fit a logistic regression model with potential predictors to the training subsample, and used backward variable selection based on Akaike information criterion (AIC) to select the final predictors. Stroke type and admission year were forced into the final model by default. The model was validated internally on the test subsample and externally on a second cohort of elderly stroke patients admitted to the ICU from 2016 to 2017. The second cohort only contained data for the final predictors. Of note, Bautista et al[15] did not scale or round regression coefficients to create an integer score.

We use the 2 samples (2005–2009 and 2016–2017) from Bautista et al[15] to illustrate the change in performance when simplifying a regression model to an integer score. First, we fit 2 logistic regression models using all of the data from the 2005 to 2009 sample compared to the training subsample used in Bautista et al.[15] With the exception of admission year, all of the covariates initially considered by Bautista et al[15] were included, with restricted cubic splines fit to all continuous covariates to relax the assumption of nonlinearity. We performed Wald tests to check whether linearity was an appropriate assumption for each continuous covariate. We compared the results of backward variable selection based on AIC to the results obtained on the training subsample in Bautista et al.[15] We then fitted a logistic regression model with continuous covariates retained to the entire 2005–2009 sample, using only the covariates that were also available in the 2016–2017 sample. This was done to allow us to externally validate the models. We dichotomized continuous covariates using cut points that maximized the Youden[17] J-statistic in univariate ROC analyses as described previously, with the exception of white blood cell count (WBC), which was dichotomized at 11,000, the cut point used in Bautista et al.[15] We fit another logistic regression model to the dichotomized inputs and scaled and rounded the regression coefficients to obtain an integer score, as discussed in section "Model Development." Finally, we internally validated the models by estimating bootstrap optimism-corrected AUC and IPA using 1000 bootstraps of the 2005–2009 sample, and externally validated

the models by estimating AUC and IPA using the 2016–2017 sample. In brief, bootstrap validation was performed by refitting the model to 1000 bootstrap samples, estimating AUC and IPA of each model using the bootstrap sample and the original sample, estimating optimism as the mean decrease in AUC and IPA, and subtracting the optimism estimate from AUC and IPA estimates obtained from the model fit to the original sample.[14] Confidence intervals for IPA were estimated by rescaling estimates of the mean values and standard errors (SEs) of the Brier score of the model of interest and null model as follows: (1 – [mean null model Brier score estimate + upper ΔBrier]/mean null model Brier score estimate and 1 – [mean null model Brier score estimate + lower ΔBrier]/mean null model Brier score estimate). Upper ΔBrier and lower ΔBrier refer to the upper and lower estimates of the difference in Brier score between the model of interest and the null model at the specified confidence level, that is, mean full model Brier score – mean null model Brier score ± $t_{1-\alpha/2}$ × pooled SE.

Literature Study: Development and Validation of ARISCAT Risk Score

The ARISCAT risk score was developed to predict the risk of postoperative pulmonary complications, a composite outcome of several fatal or nonfatal postoperative events, after surgery. We refer to the publication by Canet et al[1] for a complete explanation of methods. In brief, a sample consisting of 2464 patients undergoing a range of procedures at 59 participating hospitals in Spain was randomly divided into a development subsample, with 66.6% of cases, and a validation subsample, with 33.3% of cases. Potential predictors were identified and cut points for continuous variables were selected on the basis of investigators' clinical consensus. A logistic regression model was fitted using a backward stepwise selection procedure, and a simplified risk score was calculated by multiplying each regression coefficient by 10 and rounding to the nearest integer. Discrimination and calibration were assessed by the AUC and the Hosmer-Lemeshow goodness-of-fit statistic, respectively. In a subsequent study, Mazo et al[16] externally validated the risk score using an external cohort of 5099 patients undergoing surgery at a number of hospitals throughout Spain, Western Europe, and Eastern Europe. Discrimination and calibration were assessed for the originally derived beta coefficients using the AUC and calibration slope, respectively.

Software

All analyses were conducted using R version 4.0.1 (R Foundation for Statistical Computing, Vienna, Austria) with packages lava,[22] riskRegression,[23] rms,[24] boot,[25,26] and cutpointr.[27]

RESULTS

Simulated Model Performance

Supplemental Digital Content 2, Figure 1, http://links.lww.com/AA/D324, demonstrates the variability in model performance and optimism across replicates for each data series (Cumulative average of model performance and optimism across samples for simulations with target AUC of 0.70 [left], 0.80

Table 1. Simulated AUC, IPA, and Optimism by Model, Base Simulation

Model	AUC (mean ± SE)	ΔAUC	Optimism AUC (mean ± SE)	IPA (mean ± SE, %)	ΔIPA (%)	Optimism IPA ± SE
Target AUC 0.70						
CR	0.71 ± 0.00052	REF	0.0084 ± 0.0007	13 ± 0.067	REF	1.39 ± 0.092
RT CR	0.68 ± 0.00052	−0.027	0.0354 ± 0.0007	10 ± 0.084	−3.2	4.58 ± 0.109
DR	0.67 ± 0.00054	−0.038	0.0261 ± 0.0007	8 ± 0.064	−4.5	3.39 ± 0.087
IS	0.67 ± 0.00054	−0.038	0.0261 ± 0.00069	8 ± 0.064	−4.5	3.39 ± 0.087
RT IS	0.65 ± 0.00052	−0.057	0.0451 ± 0.00068	7 ± 0.07	−6.1	5.03 ± 0.094
Target AUC 0.80						
CR	0.8 ± 0.00043	REF	0.0054 ± 0.0006	26 ± 0.078	REF	1.26 ± 0.108
RT CR	0.76 ± 0.00043	−0.035	0.0402 ± 0.00059	22 ± 0.083	−3.5	4.71 ± 0.112
DR	0.75 ± 0.00048	−0.049	0.0195 ± 0.00063	18 ± 0.078	−7.7	3.41 ± 0.105
IS	0.75 ± 0.00048	−0.049	0.0194 ± 0.00063	18 ± 0.078	−7.8	3.4 ± 0.105
RT IS	0.72 ± 0.00048	−0.076	0.0465 ± 0.00064	16 ± 0.083	−10	6.04 ± 0.111
Target AUC 0.90						
CR	0.9 ± 0.00030	REF	0.0034 ± 0.00041	48 ± 0.083	REF	1.18 ± 0.117
RT CR	0.86 ± 0.00034	−0.039	0.0424 ± 0.00045	43 ± 0.08	−5	6.17 ± 0.113
DR	0.84 ± 0.00042	−0.058	0.0141 ± 0.00055	35 ± 0.095	−13	3.48 ± 0.127
IS	0.84 ± 0.00042	−0.058	0.0141 ± 0.00055	35 ± 0.096	−13	3.47 ± 0.127
RT IS	0.81 ± 0.00043	−0.094	0.0502 ± 0.00056	31 ± 0.094	−18	7.75 ± 0.125

Base case (n = 1000, independent predictors, 2 continuous and 5 binary predictors). AUC, IPA, and their respective optimism were estimated for the 1000 training and test sets in each of the 3 simulated data series (target AUC 0.70, 0.80, and 0.90). Mean metric, change in mean metric across each subsequent step, and mean optimism are shown. Optimism is defined as the difference between the metric estimated by applying the model to the original training set and the metric estimated by applying the model to the independent test set.

Abbreviations: ΔAUC, difference in AUC from regression retaining continuous variables; ΔIPA, difference in IPA from regression retaining continuous variables; AUC, area under the receiver operating characteristic curve; CR, continuous regression; DR, dichotomized regression; IPA, index of prediction accuracy; IS, integer score; REF, reference; RT CR, risk tertiles from continuous regression; RT IS, risk tertiles from integer score; SE, standard error.

Table 2. Simulated AUC, IPA, and Optimism by Model, Varying Sample Size

Model	AUC (mean ± SE)	ΔAUC	Optimism AUC (mean ± SE)	IPA (mean ± SE, %)	ΔIPA (%)	Optimism IPA ± SE
			n = 200			
Target AUC 0.70						
CR	0.70 ± 0.00123	REF	0.0363 ± 0.0016	10 ± 0.176	REF	6.68 ± 0.241
RT CR	0.67 ± 0.00119	−0.025	0.0614 ± 0.00157	7 ± 0.195	−2.6	9.25 ± 0.257
DR	0.65 ± 0.00134	−0.043	0.077 ± 0.00161	5 ± 0.189	−5.4	12.08 ± 0.254
IS	0.65 ± 0.00134	−0.043	0.0769 ± 0.00161	5 ± 0.189	−5.5	12.07 ± 0.254
RT IS	0.64 ± 0.00128	−0.059	0.0931 ± 0.00155	3 ± 0.192	−6.8	13.35 ± 0.252
Target AUC 0.80						
CR	0.78 ± 0.00106	REF	0.028 ± 0.00141	23 ± 0.199	REF	6.65 ± 0.276
RT CR	0.75 ± 0.00108	−0.035	0.0634 ± 0.00143	20 ± 0.209	−3.2	9.86 ± 0.28
DR	0.73 ± 0.0012	−0.054	0.0638 ± 0.0015	14 ± 0.214	−8.7	12.62 ± 0.292
IS	0.73 ± 0.0012	−0.054	0.0637 ± 0.0015	14 ± 0.215	−8.8	12.58 ± 0.292
RT IS	0.71 ± 0.00117	−0.079	0.0885 ± 0.00147	12 ± 0.21	−11	14.65 ± 0.281
Target AUC 0.90						
CR	0.89 ± 0.00075	REF	0.0178 ± 0.00095	46 ± 0.213	REF	6.26 ± 0.281
RT CR	0.85 ± 0.00084	−0.04	0.0577 ± 0.00104	41 ± 0.199	−4.4	10.64 ± 0.265
DR	0.83 ± 0.00103	−0.062	0.044 ± 0.00122	32 ± 0.24	−14	11.51 ± 0.303
IS	0.83 ± 0.00103	−0.062	0.044 ± 0.00122	32 ± 0.242	−14	11.5 ± 0.304
RT IS	0.8 ± 0.00103	−0.098	0.0794 ± 0.00123	28 ± 0.227	−18	15.3 ± 0.285
Model	**AUC (mean ± SE)**	**ΔAUC**	**Optimism AUC (mean ± SE)**	**IPA (mean ± SE, %)**	**ΔIPA (%)**	**Optimism IPA ± SE**
			n = 500			
Target AUC 0.70						
CR	0.71 ± 0.00075	REF	0.0159 ± 0.00101	12 ± 0.1	REF	2.77 ± 0.138
RT CR	0.68 ± 0.00071	−0.027	0.0428 ± 0.00098	9 ± 0.115	−3.1	5.85 ± 0.154
DR	0.67 ± 0.00078	−0.039	0.0408 ± 0.00099	8 ± 0.098	−4.6	5.77 ± 0.136
IS	0.67 ± 0.00078	−0.039	0.0408 ± 0.00099	8 ± 0.099	−4.7	5.76 ± 0.136
RT IS	0.65 ± 0.00074	−0.056	0.0585 ± 0.00096	6 ± 0.104	−6.2	7.27 ± 0.143
Target AUC 0.80						
CR	0.79 ± 0.00062	REF	0.0113 ± 0.00085	25 ± 0.114	REF	2.63 ± 0.158
RT CR	0.76 ± 0.00064	−0.035	0.0462 ± 0.00087	22 ± 0.124	−3.4	6.04 ± 0.166
DR	0.74 ± 0.0007	−0.05	0.032 ± 0.00092	17 ± 0.118	−7.9	5.94 ± 0.163
IS	0.74 ± 0.0007	−0.05	0.032 ± 0.00092	17 ± 0.119	−8	5.94 ± 0.163
RT IS	0.72 ± 0.00068	−0.077	0.0587 ± 0.00091	15 ± 0.12	−11	8.51 ± 0.164
Target AUC 0.90						
CR	0.90 ± 0.00044	REF	0.0071 ± 0.00062	48 ± 0.127	REF	2.45 ± 0.177
RT CR	0.86 ± 0.0005	−0.039	0.0462 ± 0.00067	43 ± 0.119	−4.8	7.28 ± 0.168
DR	0.84 ± 0.00058	−0.06	0.0226 ± 0.00077	34 ± 0.135	−13	5.67 ± 0.184
IS	0.84 ± 0.00058	−0.06	0.0225 ± 0.00077	34 ± 0.135	−13	5.66 ± 0.184
RT IS	0.80 ± 0.0006	−0.095	0.0581 ± 0.00079	30 ± 0.132	−18	9.78 ± 0.182

Varying sample size (n = 200, 500, independent predictors, 2 continuous and 5 binary predictors). AUC, IPA, and their respective optimism were estimated for the 1000 training and test sets in each of the 3 simulated data series (target AUC 0.70, 0.80, and 0.90). Mean metric, change in mean metric across each subsequent step, and mean optimism are shown. Optimism is defined as the difference between the metric estimated by applying the model to the original training set and the metric estimated by applying the model to the independent test set.

Abbreviations: ΔAUC, difference in AUC from regression retaining continuous variables; ΔIPA, difference in IPA from regression retaining continuous variables; AUC, area under the receiver operating characteristic curve; CR, continuous regression; DR, dichotomized regression; IPA, index of prediction accuracy; IS, integer score; REF, reference; RT CR, risk tertiles from continuous regression; RT IS, risk tertiles from integer score; SE, standard error.

[middle], and 0.90 [right]. a-c, cumulative average of AUC; d-f, cumulative average of IPA; g-i, cumulative average of optimism in AUC; j-l, cumulative average of optimism in IPA. Results for the regression model with dichotomized variables were omitted for visual clarity, as they overlapped entirely with their counterparts for the integer score. AUC: area under the receiver operating characteristic curve; IPA: index of prediction accuracy.). Monte Carlo estimates were reasonably stable after 250 iterations for every model in all of the simulations. Tables 1–4 present mean ± MCSE estimates of AUC, IPA, optimism in AUC and IPA, and the change in each metric between each step of model development for each respective simulation. Overall, the logistic regression model retaining continuous covariates outperformed all other models in both AUC and IPA. Optimism was also smallest for the logistic regression model retaining continuous covariates. The largest drop in model performance was observed in the dichotomization step; conversely, rounding regression coefficients to generate an integer score after the dichotomization step had little to no impact on model performance. Boxplots of mean AUC and mean IPA for each series across each step of model development in all 4 simulations are depicted in the Figure.

Table 3. Simulated AUC, IPA, and Optimism by Model, Varying Correlation

Model	AUC (mean ± SE)	ΔAUC	Optimism AUC (Mean ± SE)	IPA (mean ± SE, %)	ΔIPA (%)	Optimism IPA ± SE
			***r* = 0.10**			
Target AUC 0.70						
CR	0.69 ± 0.00052	REF	0.0108 ± 0.00075	11 ± 0.062	REF	1.65 ± 0.092
RT CR	0.67 ± 0.00052	−0.025	0.0358 ± 0.00074	8 ± 0.081	−3.1	4.76 ± 0.111
DR	0.66 ± 0.00055	−0.035	0.0279 ± 0.00074	7 ± 0.062	−3.8	3.46 ± 0.088
IS	0.66 ± 0.00055	−0.035	0.0279 ± 0.00074	7 ± 0.061	−3.9	3.46 ± 0.088
RT IS	0.64 ± 0.00052	−0.052	0.045 ± 0.00072	6 ± 0.065	−5.3	4.89 ± 0.094
Target AUC 0.80						
CR	0.81 ± 0.00043	REF	0.0059 ± 0.0006	28 ± 0.082	REF	1.39 ± 0.115
RT CR	0.77 ± 0.00044	−0.036	0.0415 ± 0.0006	25 ± 0.086	−3.6	4.94 ± 0.118
DR	0.76 ± 0.00049	−0.05	0.0193 ± 0.00066	20 ± 0.084	−8.3	3.57 ± 0.116
IS	0.76 ± 0.00049	−0.05	0.0193 ± 0.00066	20 ± 0.084	−8.4	3.57 ± 0.116
RT IS	0.73 ± 0.0005	−0.079	0.0479 ± 0.00067	17 ± 0.088	−11	6.45 ± 0.121
Target AUC 0.90						
CR	0.89 ± 0.00031	REF	0.0032 ± 0.00045	47 ± 0.082	REF	1.06 ± 0.12
RT CR	0.85 ± 0.00035	−0.039	0.0420 ± 0.00048	42 ± 0.08	−4.8	5.85 ± 0.117
DR	0.84 ± 0.00041	−0.057	0.0140 ± 0.00055	34 ± 0.092	−13	3.32 ± 0.126
IS	0.84 ± 0.00041	−0.057	0.0140 ± 0.00055	34 ± 0.093	−13	3.32 ± 0.126
RT IS	0.80 ± 0.00043	−0.093	0.0499 ± 0.00057	30 ± 0.094	−17	7.53 ± 0.127
Model	**AUC (mean ± SE)**	**ΔAUC**	**Optimism AUC (mean ± SE)**	**IPA (mean ± SE, %)**	**ΔIPA (%)**	**Optimism IPA ± SE**
			***r* = 0.30**			
Target AUC 0.70						
CR	0.70 ± 0.00053	REF	0.0091 ± 0.00076	12 ± 0.065	REF	1.45 ± 0.094
RT CR	0.67 ± 0.00052	−0.026	0.0349 ± 0.00075	9 ± 0.083	−3.1	4.58 ± 0.113
DR	0.66 ± 0.00055	−0.036	0.0265 ± 0.00075	8 ± 0.063	−4.1	3.35 ± 0.092
IS	0.66 ± 0.00055	−0.036	0.0264 ± 0.00075	8 ± 0.064	−4.1	3.34 ± 0.092
RT IS	0.65 ± 0.00052	−0.054	0.0444 ± 0.00072	6 ± 0.068	−5.6	4.85 ± 0.096
Target AUC 0.80						
CR	0.80 ± 0.00044	REF	0.0060 ± 0.00061	27 ± 0.082	REF	1.39 ± 0.114
RT CR	0.77 ± 0.00045	−0.035	0.0413 ± 0.00062	24 ± 0.089	−3.5	4.91 ± 0.12
DR	0.75 ± 0.00049	−0.05	0.0199 ± 0.00066	19 ± 0.082	−8.1	3.61 ± 0.115
IS	0.75 ± 0.00049	−0.05	0.0199 ± 0.00066	19 ± 0.083	−8.2	3.61 ± 0.115
RT IS	0.73 ± 0.00049	−0.078	0.0477 ± 0.00066	16 ± 0.087	−11	6.41 ± 0.119
Target AUC 0.90						
CR	0.91 ± 0.0003	REF	0.0032 ± 0.00042	50 ± 0.085	REF	1.14 ± 0.12
RT CR	0.87 ± 0.00034	−0.039	0.0422 ± 0.00046	44 ± 0.081	−5.1	6.26 ± 0.116
DR	0.85 ± 0.0004	−0.058	0.0140 ± 0.00053	36 ± 0.095	−13	3.5 ± 0.126
IS	0.85 ± 0.0004	−0.058	0.0140 ± 0.00053	36 ± 0.095	−14	3.49 ± 0.126
RT IS	0.81 ± 0.00043	−0.095	0.0507 ± 0.00056	32 ± 0.095	−18	7.95 ± 0.127

Varying correlation (n = 1000, *r* = 0.10, *r* = 0.30; 2 continuous and 5 binary predictors). AUC, IPA, and their respective optimism were estimated for the 1000 training and test sets in each of the 3 simulated data series (target AUC 0.70, 0.80, and 0.90). Mean metric, change in mean metric across each subsequent step, and mean optimism are shown. Optimism is defined as the difference between the metric estimated by applying the model to the original training set and the metric estimated by applying the model to the independent test set.

Abbreviations: ΔAUC, difference in AUC from regression retaining continuous variables; ΔIPA, difference in IPA from regression retaining continuous variables; AUC, area under the receiver operating characteristic curve; CR, continuous regression; DR, dichotomized regression; IPA, index of prediction accuracy; IS, integer score; REF, reference; RT CR, risk tertiles from continuous regression; RT IS, risk tertiles from integer score; SE, standard error.

Base Simulation (n = 1000, *r* = 0; 2 Continuous + 5 Binary Predictors)

In the first simulation (Table 1; Figure A), with target AUC of 0.70, 0.80, and 0.90, respectively, the continuous regression had a mean ± MCSE AUC of 0.71 ± 0.00052, 0.80 ± 0.00043, and 0.90 ± 0.00030; a mean IPA of 13% ± 0.067%, 26% ± 0.078%, and 48% ± 0.083%; an observed optimism in AUC of 0.008 ± 0.00070, 0.019 ± 0.00060, and 0.003 ± 0.00041; and an observed optimism in IPA of 1.4% ± 0.09%, 1.3% ± 0.11%, and 1.2% ± 0.12%. The total loss in model performance when going from the logistic regression model retaining continuous variables to the risk stratification model built on the integer score was a decrease in mean AUC of 0.057, 0.076, and 0.094, and a decrease in mean IPA of 6.2%, 10.4%, and 17.5% in each simulation, respectively.

Simulation Varying Sample Size (n = 200, n = 500)

In the second simulation (Table 2; Figure B, C), in the n = 200 condition, with target AUC of 0.70, 0.80, and 0.90, respectively, the continuous regression had a mean AUC of 0.70, 0.78, and 0.89; a mean IPA of 10%, 23%, and 46%; and an observed optimism in AUC of 0.0363, 0.028, and 0.0178; and an observed optimism in IPA of 6.68%, 6.65%, and 6.26%. In the n = 500 condition, the continuous regression had a mean AUC of

Table 4. Simulated AUC, IPA, and Optimism by Model, Varying Number of Continuous Predictors

Model	AUC (mean ± SE)	ΔAUC	Optimism AUC (mean ± SE)	IPA (mean ± SE, %)	ΔIPA (%)	Optimism IPA ± SE
5 continuous predictors +2 binary predictors						
Target AUC 0.70						
CR	0.70 ± 0.00055	REF	0.0092 ± 0.00077	10 ± 0.063	REF	1.38 ± 0.09
RT CR	0.67 ± 0.00053	−0.026	0.0353 ± 0.00077	6 ± 0.09	−4.2	5.57 ± 0.12
DR	0.66 ± 0.0006	−0.04	0.0418 ± 0.00077	6 ± 0.064	−4.5	4.85 ± 0.091
IS	0.66 ± 0.0006	−0.041	0.0418 ± 0.00077	6 ± 0.065	−4.5	4.85 ± 0.092
RT IS	0.64 ± 0.00058	−0.059	0.0607 ± 0.00076	4 ± 0.075	−6.3	6.61 ± 0.103
Target AUC 0.80						
CR	0.80 ± 0.00049	REF	0.0067 ± 0.00072	23 ± 0.088	REF	1.54 ± 0.13
RT CR	0.76 ± 0.00049	−0.04	0.0468 ± 0.00072	15 ± 0.113	−7.8	9.31 ± 0.151
DR	0.74 ± 0.00057	−0.058	0.036 ± 0.00076	14 ± 0.086	−9.2	5.73 ± 0.122
IS	0.74 ± 0.00056	−0.058	0.0359 ± 0.00076	14 ± 0.086	−9.3	5.71 ± 0.122
RT IS	0.72 ± 0.00055	−0.088	0.0650 ± 0.00075	10 ± 0.096	−13	9.19 ± 0.134
Target AUC 0.90						
CR	0.90 ± 0.00037	REF	0.0052 ± 0.00051	41 ± 0.107	REF	1.81 ± 0.15
RT CR	0.84 ± 0.00042	−0.057	0.0626 ± 0.00055	27 ± 0.134	−14	15.8 ± 0.167
DR	0.83 ± 0.0005	−0.072	0.0310 ± 0.00065	24 ± 0.107	−17	6.86 ± 0.145
IS	0.83 ± 0.00051	−0.072	0.0310 ± 0.00065	24 ± 0.109	−17	6.85 ± 0.146
RT IS	0.79 ± 0.00051	−0.11	0.0726 ± 0.00068	16 ± 0.127	−24	14.19 ± 0.171
Model	**AUC (mean ± SE)**	**ΔAUC**	**Optimism AUC (mean ± SE)**	**IPA (mean ± SE, %)**	**ΔIPA (%)**	**Optimism IPA ± SE**
7 continuous predictors						
Target AUC 0.70						
CR	0.69 ± 0.00052	REF	0.0099 ± 0.00074	11 ± 0.063	REF	1.53 ± 0.091
RT CR	0.67 ± 0.0005	−0.025	0.0352 ± 0.00073	8 ± 0.08	−3.1	4.6 ± 0.109
DR	0.65 ± 0.00056	−0.043	0.0483 ± 0.00071	6 ± 0.065	−5.1	5.96 ± 0.09
IS	0.65 ± 0.00056	−0.045	0.0479 ± 0.00071	5 ± 0.071	−5.5	5.93 ± 0.092
RT IS	0.63 ± 0.00055	−0.065	0.0675 ± 0.00069	3 ± 0.086	−8	8.37 ± 0.104
Target AUC 0.80						
CR	0.80 ± 0.00043	REF	0.0058 ± 0.0006	26 ± 0.077	REF	1.31 ± 0.109
RT CR	0.76 ± 0.00044	−0.034	0.0399 ± 0.00061	23 ± 0.085	−3.2	4.52 ± 0.114
DR	0.73 ± 0.00049	−0.063	0.0377 ± 0.00065	16 ± 0.079	−10	6.26 ± 0.109
IS	0.73 ± 0.0005	−0.065	0.0370 ± 0.00065	15 ± 0.087	−11	6.16 ± 0.111
RT IS	0.70 ± 0.0005	−0.095	0.0665 ± 0.00064	12 ± 0.101	−14	9.45 ± 0.119
Target AUC 0.90						
CR	0.90 ± 0.0003	REF	0.0034 ± 0.00041	48 ± 0.081	REF	1.17 ± 0.114
RT CR	0.86 ± 0.00034	−0.038	0.0416 ± 0.00045	44 ± 0.078	−4.7	5.86 ± 0.111
DR	0.81 ± 0.00042	−0.085	0.0299 ± 0.00056	29 ± 0.088	−19	6.46 ± 0.122
IS	0.81 ± 0.00043	−0.088	0.0293 ± 0.00057	28 ± 0.102	−21	6.32 ± 0.125
RT IS	0.78 ± 0.00045	−0.12	0.0661 ± 0.00058	23 ± 0.124	−25	10.81 ± 0.133

Varying number of continuous predictors (n = 1000, independent predictors, 5 continuous and 2 binary predictors, and 7 continuous predictors). AUC, IPA, and their respective optimism were estimated for the 1000 training and test sets in each of the 3 simulated data series (target AUC 0.70, 0.80, and 0.90). Mean metric, change in mean metric across each subsequent step, and mean optimism are shown. Optimism is defined as the difference between the metric estimated by applying the model to the original training set and the metric estimated by applying the model to the independent test set.

Abbreviations: ΔAUC, difference in AUC from regression retaining continuous variables; ΔIPA, difference in IPA from regression retaining continuous variables; AUC, area under the receiver operating characteristic curve; CR, continuous regression; DR, dichotomized regression; IPA, index of prediction accuracy; IS, integer score; REF, reference; RT CR, risk tertiles from continuous regression; RT IS, risk tertiles from integer score; SE, standard error.

0.71, 0.79, and 0.90; a mean IPA of 12%, 25%, and 48%; an observed optimism in AUC of 0.0159, 0.0113, and 0.0071; and an observed optimism in IPA of 2.77%, 2.63%, and 2.45%.

Simulation Varying Correlations (r = 0.10, r = 0.30)

In the third simulation (Table 3; Figure D, E), in the r = 0.10 condition, with target AUC of 0.70, 0.80, and 0.90, respectively, the continuous regression had a mean AUC of 0.69, 0.81, and 0.89; a mean IPA of 11%, 28%, and 47%; an observed optimism in AUC of 0.0108, 0.0059, and 0.0032; and an observed optimism in IPA of 1.65%, 1.39%, and 1.06%. In the r = 0.30 condition, the continuous regression had a mean AUC of 0.70, 0.80, and 0.91; a mean IPA of 12%, 27%, and 50%; an observed optimism in AUC of 0.0091, 0.006, and 0.0032; and an observed optimism in IPA of 1.45%, 1.39%, and 1.14%.

Simulation Varying Ratio of Continuous to Binary Predictors

In the fourth simulation (Table 4; Figure F, G), in the condition with 5 continuous and 2 binary predictors, with target AUC of 0.70, 0.80, and 0.90, respectively, the continuous regression had a mean AUC of 0.70, 0.80,

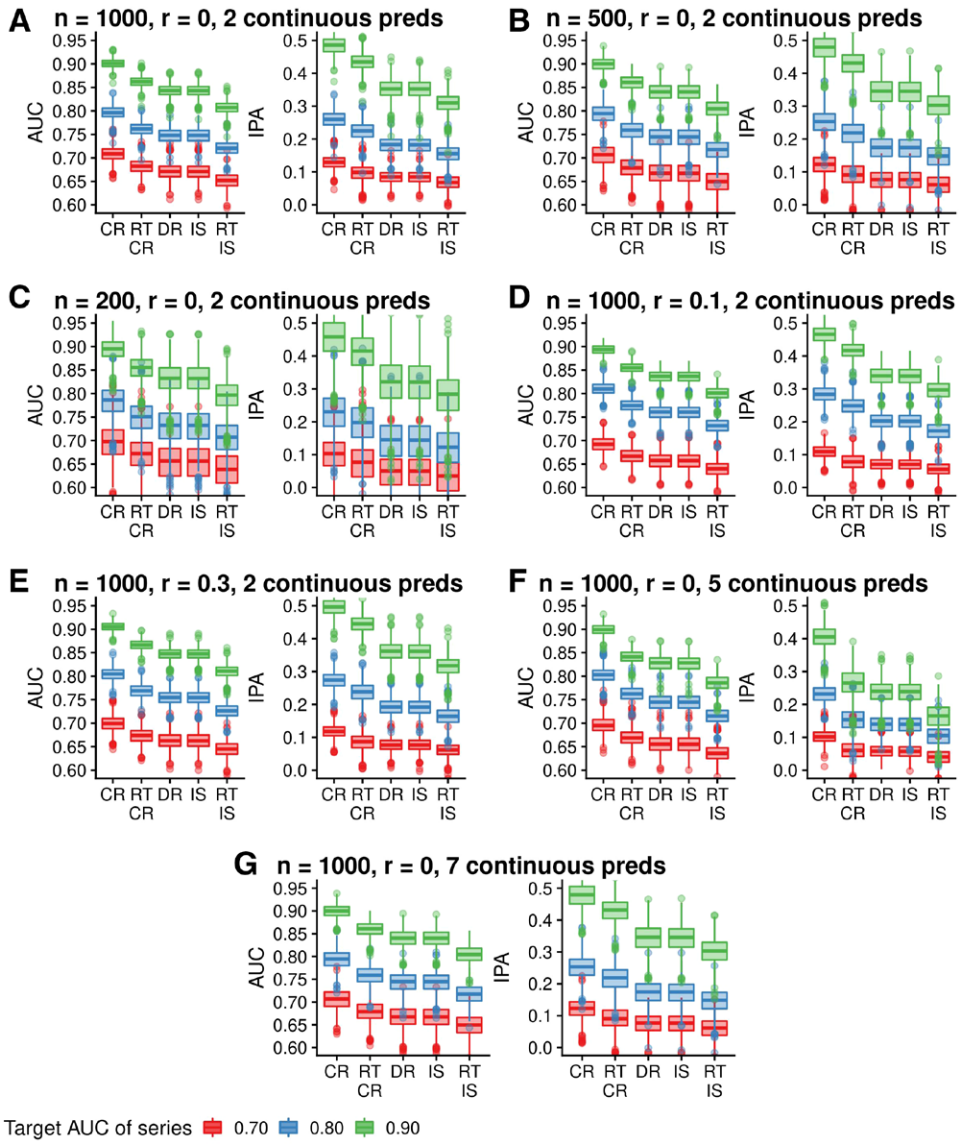

Figure. Boxplots of model performance metrics for simulations with target AUC of 0.70 (red), 0.80 (blue), and 0.90 (green). Mean AUC and mean IPA are shown for: (A) n = 1000, 2 continuous +5 binary predictors, r = 0; (B) n = 500, 2 continuous +5 binary predictors, r = 0; (C) n = 200, 2 continuous +5 binary predictors, r = 0; (D) n = 1000, 2 continuous +5 binary predictors, r = 0.1; (E) n = 1000, 2 continuous +5 binary predictors, r = 0.3; (F) n = 1000, 5 continuous +2 binary predictors, r = 0; (G) n = 1000, 7 continuous predictors, r = 0. AUC indicates area under the receiver operating characteristic curve; CR, continuous regression; DR, dichotomized regression; IPA, index of prediction accuracy; IS, integer score; RT CR, risk tertiles from continuous regression; RT IS, risk tertiles from integer score.

and 0.90; a mean IPA of 10%, 23%, and 41%; an observed optimism in AUC of 0.0092, 0.0067, and 0.0052; and an observed optimism in IPA of 1.38%, 1.54%, and 1.81%. In the condition with 7 continuous predictors, the continuous regression had a mean AUC of 0.69, 0.80, and 0.90; a mean IPA of 11%, 26%, and 48%; an observed optimism in AUC of 0.0099, 0.0058, and 0.0034; and an observed optimism in IPA of 1.53%, 1.31%, and 1.17%.

APPLICATIONS

Application of Methods to Study of Early Prediction of Prognosis in Elderly Acute Stroke Patients

When performing backward selection on the basis of AIC using the development subset of the 2005–2009 sample used by Bautista et al,[15] the final multivariable regression model included stroke type, age, smoking, mean arterial pressure <60 mm·Hg, Glasgow Coma Scale score, WBC, serum creatinine, history of congestive heart failure, and warfarin use. When performing backward selection on the basis of AIC using the entire 2005–2009 sample, the final multivariable regression model included the history of chronic obstructive pulmonary disease and serum glucose in addition to all of the predictors selected using the development subset. There was no evidence of violation of the linearity assumption for all continuous predictors: Wald P = .16 for age, P = .31 for serum glucose, P = .86 for serum creatinine, and P = .41 for the overall assumption when using the entire 2005–2009 cohort to train the model; and Wald P = .82 for age, P = .10 for serum creatinine, and P = .24 for the overall assumption when using the development subset used by Bautista et al[15] to train the model.

Logistic regression retaining continuous covariates as continuous outperformed the integer score. Logistic regression had an internally validated AUC and IPA of 0.83 (0.79–0.87) and 29% (20%–37%) versus an internally validated AUC and IPA of 0.80

(0.76–0.84) and 24% (15%–32%) for the integer score. Logistic regression had an externally validated AUC and IPA of 0.88 (0.81–0.95) and 43% (25%–61%) versus an externally validated AUC and IPA of 0.82 (0.73–0.91) and 30% (13%–47%) for the integer score.

Literature Study: ARISCAT Risk Score

AUC values are reported with 95% confidence intervals. The original study[1] reported an AUC of 0.90 (0.85–0.94) in the development subsample and 0.88 (0.84–0.93) in the validation subsample for the logistic regression model; and an AUC of 0.89 (0.83–0.93) for the development subsample and 0.84 (0.77–0.90) for the validation subsample for the simplified risk score. The optimism in AUC was 0.02 for the logistic regression and 0.05 for the risk score. The Hosmer-Lemeshow goodness-of-fit test was $P = .45$ for the logistic regression; calibration was not reported separately for the risk score. In their external validation of the risk score, Mazo et al[16] reported an overall AUC of 0.80, with subgroup AUCs of 0.80, 0.87, and 0.76 for patients in Spain, Western Europe, and Eastern Europe, respectively. Calibration varied widely across subgroups, with an overall calibration slope of 0.63 (intercept 0.66) and slopes of 0.62 (intercept 0.06), 0.81 (intercept 0.65), and 0.58 (intercept 1.44) for patients in Spain, Western Europe, and Eastern Europe, respectively.

DISCUSSION

We have demonstrated how model discrimination and calibration decrease when developing an integer score and risk stratification system from a regression model. Scaling and rounding regression coefficients had a negligible cost, likely because the approach used introduced minimal rounding error. However, nontrivial hits to performance were incurred in dichotomization and stratification across all simulations. When altering sample size or including mild to moderate correlation between predictors, observed decrements were essentially identical to the base simulation. Interestingly, observed decrements increased as the ratio of continuous to binary predictors increased, although this effect was small within the constraints of these simulations. Categorization of continuous predictor variables is necessary to develop an integer score but brought the greatest detriment to performance. Notably, when stratification was done directly on the regression retaining continuous covariates, the resulting classifier performed better than the regression with dichotomized inputs. Similarly, when applying these methods to the acute stroke data set, simplifying to an integer score tangibly diminished performance.

These findings are consistent with previous studies discussing issues introduced by categorizing variables. These include loss of information and generalizability,[10] misleading P values, overly narrow confidence intervals, and inappropriate variable selection.[11] Adding groups reduces but does not entirely eliminate these effects.[10] Categorizing multiple predictors can result in misleadingly significant interactions and relationships between predictors and outcome, in addition to decreased prediction accuracy, depending on the correlations between the predictors.[10] As illustrated here, these issues are also magnified when data-dependent cut points are used, which is common. The increase in optimism observed after dichotomization and stratification likely reflects that the cut points derived from the training set were not adequately generalizable to the test set.

Nonlinear associations between predictor variables and the outcome were not present here, but often exist. If continuous variables are retained, several approaches can capture nonlinear effects, such as fitting splines.[28] Categorization prevents relaxation of the assumption of linearity and further assumes that a predictor's effect can be fully captured within a few groups. Intuitively, this is unrealistic; individuals with similar values who are on opposite sides of a cut point are characterized as having different outcomes under this paradigm.[10]

The ARISCAT risk score demonstrates some of these effects in a published anesthesia prediction tool. Of note, Canet et al[1] adjusted for several methodological concerns discussed previously: cut points were selected by clinician consensus instead of being data-derived, and some variables were categorized into more than 2 groups. There was still a clear decrease in performance incurred during simplification and variable generalizability, although a regression retaining continuous predictors was not reported.[1] Mazo et al[16] discussed that these findings reflect both the optimism inherent to prediction models and real differences in predictor effects in external populations. Additional contributors include categorization of predictors and coercion to an integer score. The Hosmer-Lemeshow test used to assess calibration has fallen out of favor due to low power, poor interpretability, and arbitrary choice of number of groups.[28] Mazo et al[16] instead reported the calibration slope, which on its own does not measure calibration and should be reported with the intercept (which was done here).[29]

We prefer to avoid stepwise selection of variables as used by Canet et al[1] and Bautista et al.[15] Choice of final predictors is not robust, such as when splitting data into training and test subsamples. Stepwise selection also considerably increases the risk of overfit.[30] Many excellent alternatives exist, such as specifying a biologically plausible relationship, least absolute shrinkage and selection operator (LASSO) regression,[31] spike and slab selection,[32] and Bayesian model averaging.[33]

Dichotomization and coercion to integer score are often done because model simplicity is desired. We argue that ease of use should be differentiated from the simplicity of the underlying model. Today prediction models are usually published online; for example, the Cleveland Clinic publishes a freely available library at https://riskcalc.org.[34] Although the underlying model is described,[35–37] it is typically not immediately visible to users.[38–40] Categorization of predictors may simplify the model in the background but does not alter the user experience, which more directly influences a model's adoption.

Interpretability of the output is also important. Although clinicians may memorize a simple integer score, the score has no intuitive meaning without a threshold to categorize it or mapping to a predicted probability, either of which is very challenging to memorize. The latter is preferable if accuracy is prioritized, but as illustrated here, both predicted probabilities and risk categories can be more accurately obtained from a regression model that retains continuous covariates than from an integer score. Additionally, in practice, an integer score can span an extensive range: the ARISCAT risk score ranges in discrete steps from 0 to 123 total points.

We recommend that when designing a prediction model, researchers should carefully consider their desired application and clinical need. An integer score may suffice if approximations are sufficient to make decisions. When prediction accuracy is paramount, we strongly prefer a regression model in which continuous covariates are kept continuous and can have nonlinear associations with the outcome. We recommend describing performance metrics for both the underlying regression and the simplified model to transparently report the change in prediction accuracy.

If a method to compute predictions without a computer is desired, we advocate the use of nomograms. These are graphical representations of regressions in which points are separately calculated for each predictor, and the sum of the points is mapped to a predicted probability.[41] Nomograms can be computed on paper and have the ancillary benefit of demonstrating how a prediction was obtained.

In all situations, it is important to validate the model intended for end users. For example, reporting an AUC for the regression is misleading if clinicians are actually adopting the integer score. If using *k*-fold cross-validation or bootstrapping to internally validate the model,[42] all steps should be repeated for each case: predictors should be selected, a regression should be fitted, and the model should be validated repeatedly across each fold or bootstrap replicate. If the model is prespecified outside of each replicate, bootstrapping does not properly correct for optimism bias.[43] To ensure proper reporting, we suggest referring to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines.[44]

ACKNOWLEDGMENTS

The authors thank Dongsheng Yang, MS, and Xinge Ji, MS, for their assistance with methods; Ozan Akca, MD, and Alexander F. Bautista, MD, for permission to use their data for our clinical example; and Stephanie S. Kocian, MAT, MA, for her assistance with the article.

DISCLOSURES

Name: Vigneshwar Subramanian, BA.
Contribution: This author helped perform the analysis, interpret data, and draft manuscript.
Name: Edward J. Mascha, PhD.
Contribution: This author helped with study concept and design, interpretation of data, data acquisition, and critical revision.
Name: Michael W. Kattan, PhD.
Contribution: This author helped with study concept and design, interpretation of data, data acquisition, and critical revision.
This manuscript was handled by: Thomas R. Vetter, MD, MPH.

REFERENCES

1. Canet J, Gallart L, Gomar C, et al; ARISCAT Group. Prediction of postoperative pulmonary complications in a population-based surgical cohort. *Anesthesiology.* 2010;113:1338–1350.
2. Pocock SJ, Ariti CA, McMurray JJ, et al; Meta-Analysis Global Group in Chronic Heart Failure. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J.* 2013;34:1404–1413.
3. Mehran R, Pocock SJ, Nikolsky E, et al. A risk score to predict bleeding in patients with acute coronary syndromes. *J Am Coll Cardiol.* 2010;55:2556–2566.
4. Halkin A, Singh M, Nikolsky E, et al. Prediction of mortality after primary percutaneous coronary intervention for acute myocardial infarction: the CADILLAC risk score. *J Am Coll Cardiol.* 2005;45:1397–1405.
5. Singh M, Lennon RJ, Holmes DR Jr, Bell MR, Rihal CS. Correlates of procedural complications and a simple integer risk score for percutaneous coronary intervention. *J Am Coll Cardiol.* 2002;40:387–393.
6. Nasr VG, DiNardo JA, Faraoni D. Development of a pediatric risk assessment score to predict perioperative mortality in children undergoing noncardiac surgery. *Anesth Analg.* 2017;124:1514–1519.
7. Robinson WP, Schanzer A, Li Y, et al. Derivation and validation of a practical risk score for prediction of mortality after open repair of ruptured abdominal aortic aneurysms in a US regional cohort and comparison to existing scoring systems. *J Vasc Surg.* 2013;57:354–361.
8. Guo L, Wei D, Zhang X, et al. Clinical features predicting mortality risk in patients with viral pneumonia: the MuLBSTA score. *Front Microbiol.* 2019;10:2752.
9. Bendapudi PK, Hurwitz S, Fry A, et al. Derivation and external validation of the PLASMIC score for rapid assessment of adults with thrombotic microangiopathies: a cohort study. *Lancet Haematol.* 2017;4:e157–e164.
10. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.* 2006;25:127–141.

11. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst*. 1994;86:829–835.
12. Cole TJ. Scaling and rounding regression coefficients to integers. *Appl Stat*. 1993;42:261.
13. Kattan MW. Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: preoperative application in prostate cancer. *Curr Opin Urol*. 2003;13:111–116.
14. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361–387.
15. Bautista AF, Lenhardt R, Yang D, et al. Early prediction of prognosis in elderly acute stroke patients. *Crit Care Explor*. 2019;1:e0007.
16. Mazo V, Sabaté S, Canet J, et al. Prospective external validation of a predictive score for postoperative pulmonary complications. *Anesthesiology*. 2014;121:219–231.
17. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32–35.
18. Mascha EJ. Identifying the best cut-point for a biomarker, or not. *Anesth Analg*. 2018;127:820–822.
19. Gomez-Builes JC, Acuna SA, Nascimento B, Madotto F, Rizoli SB. Harmful or physiologic: diagnosing fibrinolysis shutdown in a trauma cohort with rotational thromboelastometry. *Anesth Analg*. 2018;127:840–849.
20. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
21. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagn Progn Res*. 2018;2:7.
22. Holst KK, Budtz-Jørgensen E. Linear latent variable models: the lava-package. *Comput Stat*. 2013;28:1385–1452.
23. Ozenne B, Lyngholm SA, Scheike T, Torp-Pedersen C, Alexander GT. riskRegression: predicting the risk of an event using Cox regression models. *R J*. 2017;9:440.
24. Harrell FE. Rms: Regression Modeling Strategies in R. 2019. Available at: https://CRAN.R-project.org/package=rms. Accessed May 28, 2020.
25. Canty A, Ripley B. Boot: Bootstrap R (S-Plus) Functions. 2020. Available at: https://cran.r-project.org/web/packages/boot. Accessed May 28, 2020.
26. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Cambridge University Press; 1997.
27. Thiele C, Hirschfeld G. Cutpointr: improved estimation and validation of optimal cutpoints in R. arXiv. 2020. Accessed April 25, 2020. Available at: https://arxiv.org/abs/2002.09209v1.
28. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Springer; 2015.
29. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? *J Clin Epidemiol*. 2020;118:93–99.
30. Assel M, Sjoberg D, Elders A, et al. Guidelines for reporting of statistics for clinical research in urology. *BJU Int*. 2019;123:401–410.
31. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B Methodol*. 1996;58:267–288.
32. Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann Stat*. 2005;33:730–773.
33. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *J Am Stat Assoc*. 1997;92:179–191.
34. Department of Quantitative Health Sciences. Cleveland Clinic Risk Calculator Library. Available at: http://riskcalc.org:3838/. Accessed April 24, 2018.
35. Bochner BH, Kattan MW, Vora KC; International Bladder Cancer Nomogram Consortium. Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer. *J Clin Oncol Off J Am Soc Clin Oncol*. 2006;24:3967–3972.
36. Slawin KM, Kattan MW, Roehrborn CG, Wilson T. Development of nomogram to predict acute urinary retention or surgical intervention, with or without dutasteride therapy, in men with benign prostatic hyperplasia. *Urology*. 2006;67:84–88.
37. Van Zee KJ, Manasseh DM, Bevilacqua JL, et al. A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy. *Ann Surg Oncol*. 2003;10:1140–1151.
38. Predicting 5-Year Recurrence-Free Survival after Radical Cystectomy for Bladder Cancer. CCF Risk Calculator. Available at: https://riskcalc.org/bladderCancer/. Accessed May 28, 2020.
39. Predicting Acute Urinary Retention or Surgical Intervention within 2 years (with or without Dutasteride). CCF Risk Calculator. Available at: https://riskcalc.org/BenignProstaticHyperplasia/. Accessed May 28, 2020.
40. Predicting Positive Additional Non-Sentinel Lymph Node in Patients with Breast Cancer with Positive SLN with or without Frozen Section Info. CCF Risk Calculator. Available at: https://riskcalc.org/BreastCancerPosNonSentinelLymphNode/. Accessed May 28, 2020.
41. Kattan MW, Marasco J. What is a real nomogram? *Semin Oncol*. 2010;37:23–26.
42. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54:774–781.
43. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol*. 2003;56:441–447.
44. Collins GS, Reitsma JB, Altman DG, Moons KGM; members of the TRIPOD group. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol*. 2015;67:1142–1151.