Supplementary Text. Data preprocessing, model generation, and model tuning

Five different machine learning (ML) models were created and evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. The models used were logistic regression (LR), penalized LR (chosen through elastic net variants of least absolute shrinkage and selection operator [LASSO]), random forest (RF), stochastic gradient boosting machine (GBM), and extreme gradient boosting (XGBoost). Patients who had missing data for any predictor were excluded from training and testing of all models, which resulted in 3,215 omitted patients. Categorical predictors were transformed into binary format using one-hot encoding.[1]

The data used in the study were divided into 2 sets using a random partitioning technique. The training set consisted of 50% of the data, while the remaining 50% made up the test set. The partitioning was done in a way that ensured both sets had almost equal numbers of patients who were readmitted and those who were not. The training set was used to estimate model parameters and fine-tune the models, while the test set was exclusively used to validate the performance of the models.

To ensure optimal performance, hyperparameters of each model, including RF, GBM, XGBoost, and LASSO, were fine-tuned using a 5-fold cross-validation technique that was repeated 3 times. This involved creating a grid of possible parameter values, with each column representing a specific parameter and each row representing a unique set of parameter values. The training data was then divided into five equal-sized folds, each containing a proportion of readmitted patients similar to the entire training set. The model was trained on four folds and tested on the fifth fold, which was held out to estimate the performance measure (AUC). This process was repeated for each fold, and the average of the five resampled estimates was used as a single 5-fold cross-validation estimate of model performance. The cross-validation procedure was repeated three times to increase precision while maintaining low bias, and the final estimate of model performance was generated by averaging the performance estimates of all three instances of 5-fold cross-validation. The parameter set that produced the best performance estimate was used to define the final tuned model. Further details on the parameter tuning for each model can be found in the Supplementary Table 1.

In each resampling iteration of the 5-fold cross-validation process, a series of preprocessing steps were applied. First, near-

Supplementary Table 1. Parameter tuning characteristics for each model

Model	Package	Tuning parameters	Parameter values considered	Parameter values for top performing model	
				50%–50% data split	Pre-2017: 2017 data split
Logistic regression	GLM	N/A	N/A	N/A	N/A
Random forests	Random forest	No. of variables available for splitting at each tree node	1, 2, 3, … 15	6	N/A
Gradient boosting machine	GBM	Number of trees	150, 160, 170, … 250	220	220
		Shrinkage	0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.1	0.02	0.02
		Interaction depth	3, 10, 12, 14, … 20	3	2
		Minimum observations in node	2, 4, 6, 8, 10	6	12
Extreme gradient boosting	XGBoost	No. of rounds	100, 200	100	N/A
		Max depth	3, 10, 15, 20, 25	20	N/A
		Eta	0.1	0.10	N/A
		Gamma	0	0	N/A
		ColSample_ByTree	0.5, 0.6, 0.7, 0.8, 0.9	0.50	N/A
		Min_Child_Weight	1	1	N/A
		Subsample	1	1	N/A
Penalized logistic regression	GLMnet	Lambda regularization parameter	0.01, 0.02, 0.03, … 0.1	0.03	N/A
		Alpha	0.5, 1.5, 2, 2.5	1	N/A

GLM, generalized linear model; N/A, not applicable; GBM, Gradient Boosting Machine; XGBoost, extreme gradient boosting.

zero variance predictors and highly correlated predictors were eliminated to enhance the subsequent model generation process. Predictors were considered near-zero variance if they met two conditions: (1) they had less than 10% of the total number of samples as the number of distinct values, and (2) the ratio of the frequency of the most common value to the frequency of the second most common value was greater than 19:1. Predictors with pair-wise correlations of 0.9 or higher were deemed highly correlated, and the correlated predictor with the largest mean absolute correlation was removed. Second, predictors were normalized to a mean of zero and a variance of one. Finally, the SMOTE (Synthetic Minority Over-sampling Technique)[2] was used to help optimize the model and specifically, the imbalance between the proportion of readmitted and non-readmitted patients.

The tuned models were compared based on their mean AUCs, and the GBM model was found to have the largest mean AUC. Two additional models were then generated, LR and GBM, using a different training and testing set. This new training set included data from January 1, 2004 up to December 31, 2016, while the test set included data from January 1, 2017 to November 30, 2017. The GBM model was then used to evaluate its applicability in clinical practice. Specifically, the model was used to identify the top 25% of patients at the highest risk of readmission each month, and the prediction accuracy of the model was assessed. The model was evaluated by counting the true positives and calculating the cost savings associated with reducing readmissions, assuming that 50% of interventions on these patients prevented readmission. The same clinical scenario was also applied using the LACE index model,[3] a previously-validated readmission model that uses four variables to predict unplanned 30-day readmission after hospital discharge: length of stay (L), acuity of the admission (A), comorbidity of the patient (C), and emergency department use in the duration of 6 months before admission (E).[4] The cost savings were compared between the LACE index model and the top-performing GBM model.

SUPPLEMENTARY REFERENCES

1. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1-26.
2. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
3. Wang H, Robinson RD, Johnson C, et al. Using the LACE index to predict hospital readmissions in congestive heart failure patients. BMC Cardiovasc Disord 2014;14:97.
4. Van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ 2010;182:551-7.

https://doi.org/10.14245/ns.2347340.670