

# Evaluation of Machine Learning Algorithms for Predicting Readmission after Acute Myocardial Infarction Using Routinely Collected Clinical Data

Shagun Gupta<sup>1</sup>

SHAGUN.GUPTA@MAIL.UTORONTO.CA

Dennis Ko<sup>2,3,4</sup>

DENNIS.KO@ICES.ON.CA

Paymon Azizi<sup>3,4</sup>

PAYMON.AZIZI@MAIL.UTORONTO.CA

Mohamed R. Bouadjenek<sup>1</sup>

RBOUADJENEK@GMAIL.COM

Maria Koh<sup>3</sup>

MARIA.KOH@ICES.ON.CA

Alice Chong<sup>3</sup>

ALICE.CHONG@ICES.ON.CA

Peter Austin<sup>3,4</sup>

PETER.AUSTIN@ICES.ON.CA

Scott Sanner<sup>1</sup>

SSANNER@MIE.UTORONTO.CA

<sup>1</sup> Department of Mechanical and Industrial Engineering, University of Toronto

<sup>2</sup> Schulich Heart Centre, Sunnybrook Health Sciences Centre, University of Toronto

<sup>3</sup> ICES, Toronto, Ontario, Canada

<sup>4</sup> Institute of Health Policy, Management and Evaluation, University of Toronto

## Abstract

**Background:** The ability to accurately predict readmission after acute myocardial infarction (AMI) hospitalization is limited in current statistical models. Machine learning (ML) methods have shown improved predictive ability in various clinical contexts, but their utility in predicting readmission after AMI hospitalization is unknown.

**Methods:** Using detailed clinical information collected from patients hospitalized with AMI, we evaluated six ML algorithms (logistic regression, naïve Bayes, support vector machines, random forest, gradient boosting, and deep neural networks) to predict readmission within 30-days and 1-year of discharge. A nested cross-validation approach was used to develop and test models. We used C-statistics to compare discriminatory capacity, while the Brier score was used to indicate overall model performance. Model calibration was assessed using calibration plots.

**Results:** The 30-day readmission rate was 16.3%, while the 1-year readmission rate was 45.1%. For 30-day readmission, the discriminative ability for the ML models was modest (c-statistic 0.641; 95% CI, 0.621-0.662 for gradient boosting) and did not outperform previously reported methods. For 1-year readmission, different ML models showed moderate performance, with c-statistics around 0.72. Despite modest discriminatory capabilities, the observed readmission rates were markedly higher in the tenth decile of predicted risk compared to in the first decile of predicted risk for both 30-day and 1-year readmission.

**Conclusion:** Despite including detailed clinical information and evaluating various ML methods, these models did not have better discriminatory ability to predict readmission outcomes compared to previously reported methods.

## 1. Introduction

Acute myocardial infarction (AMI) has a high readmission burden as approximately one in six patients are readmitted within 30-days of discharge.<sup>1</sup> With the introduction of the Hospital Readmission Reduction Program in 2012 in the United States, hospitals have been financially incentivised to reduce readmission rates.<sup>2</sup> Although this has led to the development of various intervention strategies aimed at reducing readmissions, they tend to be costly and resource intensive.<sup>3</sup> In this context, predictive models would allow hospitals to identify patients at high risk for readmission and target the delivery of these interventions towards them, thereby reducing unplanned readmissions.<sup>4,5</sup>

Several models exist for predicting patients' risk of readmission. The LACE index was developed in 2010 to estimate risk of all-cause readmission in medical and surgical patients based on four criteria: length of stay (L), acuity of admission (A), comorbidity of the patient (C), and emergency department use in the 6 months prior to admission (E). While this criteria-based system is easy to apply, it is not very accurate (c-statistic range, 0.51-0.72).<sup>6,7,8</sup> AMI-specific logistic and Cox models have also been developed to more accurately model and predict readmission in an AMI population.<sup>9,10,11</sup> However, despite the more focused approach, these models have only moderately improved discrimination capabilities (median c-statistic 0.65; range 0.53-0.79). Moreover, none of these models exclusively use clinical data, thus limiting their applicability for clinical use.<sup>12</sup>

In many clinical disease contexts, machine learning (ML) models have shown improved discriminatory power at predicting outcomes compared to traditional approaches.<sup>13,14</sup> ML methods can automatically identify patterns in data that are predictive of relevant health outcomes, and include various algorithms that can model complex nonlinear interactions between variables. With respect to AMI readmissions, it has not yet been established whether these methods can produce more accurate predictions compared to the existing methods described above. The goal of this study was to conduct a comprehensive analysis of ML methods for the task of predicting readmission in AMI patients at the time of discharge, using clinical data collected during the course of care. To this end, we compared the performance of commonly used supervised ML algorithms on two different readmission tasks: (1) readmission within 30-days, and (2) readmission within 1-year.

## 2. Cohort

### 2.1. Data Sources

The study sample was derived from the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) study,<sup>15</sup> a cluster randomized trial to evaluate the effectiveness of public reporting. Detailed clinical data was collected from 81 hospital corporations in Ontario via retrospective chart review performed by trained nurse abstractors. Additional databases that were used for this study include: (1) the Ontario Registered Persons Database, a registry of all Ontario residents with health insurance coverage, and (2) the Canadian Institute for Health Information Discharge Abstract Database (DAD) which contains information on all admissions to acute care hospitals in Ontario. This database was used to identify any subsequent hospital readmissions. These datasets were linked using unique encoded identifiers and analyzed at ICES. The use of data in this project was authorized under section

45 of Ontario's Personal Health Information Protection Act, which does not require review by a Research Ethics Board.

## 2.2. Study Sample

Patients between the ages of 20 to 105 years with a most responsible diagnosis of AMI (identified by International Classification Disease-9 code 410 in DAD) and a valid Ontario Health Insurance Plan number were included in the EFFECT phase II study. However, patients were excluded if they died prior to the end of their respective analysis periods of 30-days ( $n=91$ , 1.28%) or 1-year ( $n=211$ , 2.96%) without a readmission.

## 2.3. Feature Choices

A total of 204 routinely available clinical variables were initially selected from the EFFECT study using a priori clinical knowledge, level of completeness, and applicability to the entire cohort (i.e., non-conditional questions). These included demographic variables (age, sex, race/ethnicity and employment status), past medical history, vital signs, characterization of AMI (including admission symptoms, complications and severity scores), cardiac procedures, select laboratory tests and medications, medical imaging procedures, and patient counselling. To account for predictors with missing data, variables with more than 30% missing values were dropped from each readmission cohort. This resulted in a total of 192 variables. The final list of variables used for input is provided in the Supplement (Table S1). The continuous variables were normalized by z-scoring across all patients such that each continuous variable had zero mean and unit variance. Lastly, dummy variables were created for all categorical variables.

## 2.4. Outcome

We developed models to predict all-cause hospital readmission within two timeframes: (1) readmission within 30 days of discharge from the index hospitalization; (2) readmission within 1-year of discharge from the index hospitalization.

## 3. Methods

We developed prediction models using six commonly used ML algorithms: (1) logistic regression (LR) (2) naïve Bayes (NB), (3) support vector machines (SVM), (4) random forests (RF), (5) gradient boosting (GB) and (6) deep neural networks (DNN). LR is a simple linear model which provides a good baseline comparison for more complex models. We considered both L1 (lasso) and L2 (ridge regression) regularization for LR. NB is a probabilistic classifier that makes a ‘naïve’ assumption that the input variables are independent given the outcome. While this assumption does not usually hold, the resulting model often performs well due to its simplicity and relative immunity to overfitting.<sup>16</sup> SVM attempts to find an optimal hyperplane that maximizes the margin (distance) between the data points of the two classes. In instances where data is not linearly separable, the assumption of linearity is relaxed by allowing outlier data points to fall on the wrong side of the decision margin.<sup>17</sup> In contrast to the aforementioned methods, RF, GB and DNN are nonlinear algorithms; they do not make prior assumptions about the functional form of the input and are therefore

able to model nonlinear interactions in the data. RF is an ensemble method that fits a number of decision trees on sub-samples of the dataset and then uses model averaging to improve accuracy and control for overfitting.<sup>18</sup> GB is another ensemble learning method, however, unlike RF, the decision trees are grown sequentially using gradient descent with each subsequent tree aiming to reduce the errors of the previous tree.<sup>19</sup> A DNN consists of a series of fully-connected hidden layers which transform an input vector into a probability distribution estimating the output.<sup>20</sup> All analyses were done using Python v2.7. LR, NB, SVM and RF were implemented using scikit-learn v0.18.1.<sup>21</sup> GB was implemented using and XGBoost v0.6.<sup>22</sup> The DNN was implemented using Keras.<sup>23</sup> A detailed discussion of the ML algorithms is provided in Supplemental Section S1.

Each model was assessed using a nested k-fold cross validation approach consisting of two loops: an outer loop for model evaluation and an inner loop for tuning hyperparameters. We used 10 non-overlapping training and test sets in the outer loop. Performance estimates were calculated by averaging test set scores across the 10 dataset splits of the outer loop. Each training fold of the outer loop was further split into 5 non-overlapping training and validation sets (the inner loop) which were used to tune hyperparameters for each model using grid searches. A grid search builds a model for every combination of the specified hyperparameters using the training set and then evaluates each model on the validation set to identify the optimal combination of hyperparameter values. Further details including a list of hyperparameters are provided in the Supplement (Section S2 and Table S2). To address the class imbalance in the cohort for 30-day readmission, we experimented with three strategies to create a balanced dataset of readmitted and not-readmitted patients during model training: (1) up-sampling, creating duplicates of the minority class, (2) down-sampling, randomly subsampling subjects of the majority class, and (3) weighting algorithms by providing a weight for each class. The weight selected for each class (readmitted and not-readmitted) was inversely proportional to class frequencies in the training dataset. In each of these cases, we applied the modification only to the training dataset. We also considered an approach in which no weighting, and no up- or down-sampling was applied. For each algorithm, we report the results for the technique that gave the best prediction.

We chose the c-statistic (equivalent to the area under the receiver operating characteristic (ROC) curve) as our primary performance measure for all models. The c-statistic is the probability that a randomly-selected readmitted patient will have a higher predicted probability of readmission than a randomly-selected patient who was not readmitted.<sup>24</sup> We report the Brier score as a measure of accuracy of probabilistic predictions. It can take on values between 0 and 1, with a score of 0 indicating perfect prediction.<sup>24</sup> We assessed model confidence using the logarithmic-loss (log-loss) function. The log-loss function calculates a score by penalizing the difference between the predicted probabilities and expected values; since the penalty is logarithmic in nature, log-loss heavily penalizes models that are confident about an incorrect prediction. A perfectly accurate model has a log-loss score of 0, whereas less accurate models have increasingly larger scores. Model calibration was assessed by stratifying the test sample into ten risk strata using the deciles of the predicted probability of the outcome. We then compared the observed probability of the outcome within each decile to the mean predicted probability of the outcome within each decile.

Table 1: Characteristics of patients in 30-day and 1-year readmission cohorts

Characteristic	30-day readmission		1-year readmission	
	Readmitted	Not Readmitted	Readmitted	Not Readmitted
N	1146	5872	3113	3785
Median age, y (SD)	69.8 (13.7)	66.7 (14.0)	70.1 (13.5)	64.4 (13.8)
Women	435 (38.0)	2040 (34.7)	1198 (38.5)	1220 (32.2)
<b>Past Medical History (cardiovascular comorbidities)</b>				
Stroke	181 (15.8)	621 (10.6)	478 (15.4)	284 (7.5)
Diabetes	395 (34.5)	1500 (25.5)	1046 (33.6)	804 (21.2)
Congestive Heart Failure	101 (8.8)	243 (4.1)	254 (8.2)	70 (1.8)
Hypertension	748 (65.3)	3284 (55.9)	1971 (63.3)	1981 (52.3)
Hyperlipidemia	529 (46.2)	2665 (44.4)	1455 (46.7)	1692 (44.7)
Myocardial infarction	335 (29.2)	1286 (21.9)	892 (28.7)	678 (17.9)
Atrial Fibrillation	131 (11.4)	409 (7.0)	346 (11.1)	170 (4.5)
Peripheral arterial disease	122 (10.6)	432 (7.4)	334 (10.7)	202 (5.3)
Angina	397 (34.6)	1612 (27.5)	1042 (33.5)	914 (24.1)
<b>Past Medical History (medical comorbidities)</b>				
Renal disease	18 (1.6)	45 (0.8)	51 (1.6)	12 (0.3)
Cancer	25 (2.2)	88 (1.5)	69 (2.2)	40 (1.1)
COPD	152 (13.3)	528 (9.0)	427 (13.7)	229 (6.1)
Chronic liver disease	9 (0.8)	26 (0.4)	23 (0.7)	12 (0.3)
Peptic ulcer disease	67 (5.8)	253 (4.3)	169 (5.4)	143 (3.8)
<b>Past Cardiac Procedures</b>				
Coronary artery bypass grafting	94 (8.2)	464 (7.9)	309 (9.9)	242 (6.4)
Percutaneous coronary intervention	85 (7.4)	355 (6.0)	225 (7.2)	210 (5.5)
<b>Types of myocardial infarction</b>				
STEMI	370 (32.3)	2111 (36.0)	941 (30.2)	1526 (40.3)
Non-STEMI	745 (65.0)	3640 (62.0)	2090 (67.1)	2194 (58.0)
<b>In-hospital treatments and procedures</b>				
Cardiac Catheterization	430 (37.5)	2884 (49.1)	1148 (36.9)	2152 (56.9)
Percutaneous coronary intervention	176 (15.4)	1403 (23.9)	474 (15.2)	1102 (29.1)
Coronary artery bypass surgery	43 (3.8)	263 (4.5)	103 (3.3)	201 (5.3)
<b>AMI-specific hospital complications (&gt;24 hours after arrival)</b>				
Heart failure in-hospital	122 (10.6)	353 (6.0)	276 (8.9)	184 (4.9)
Shock	16 (1.4)	46 (0.8)	30 (1.0)	31 (0.8)
Recurrent angina/ ischemia	222 (19.4)	883 (15.0)	518 (16.6)	575 (15.2)
Hemorrhage requiring intervention	28 (2.4)	123 (2.1)	69 (2.2)	82 (2.2)
<b>Lab Values, mean (SD)</b>				
Blood urea nitrogen	4.5 (1.2)	7.4 (4.5)	8.7 (5.9)	6.7 (3.5)
Cholesterol	4.5 (1.2)	4.6 (1.2)	4.5 (1.2)	4.7 (1.1)
Hemoglobin (First)	133.2 (21.7)	138.9 (19.5)	133.8 (21.3)	141.8 (18.0)
Hemoglobin (Last)	123.1 (18.6)	128.9 (17.9)	124.5 (18.3)	131.1 (17.5)

## 4. Results

Baseline characteristics of patients in 30-day and 1-year readmission cohorts are detailed in Table 1. Overall, there were 7018 patients in the 30-day readmission cohort with 1146 readmissions, resulting in a readmission rate of 16.3%. For 1-year readmission, there were a total 6898 patients in the cohort of which 3113 were readmitted, resulting in a readmission rate of 45.1%. On average, patients who were readmitted were older than those not readmitted. Comorbid conditions and AMI-specific hospital complications were also higher in readmitted patients. Lastly, patients who were readmitted were less likely to have undergone cardiac procedures during the hospital stay than those who were not readmitted.

Table 2: Measures of predictive accuracy for various models

Outcome and Model	AUC	Brier Score	Log Loss
<b>Readmission within 30-days</b>			
Logistic Regression	$0.631 \pm 0.020$	<b><math>0.132 \pm 0.008</math></b>	$0.430 \pm 0.020$
Naive Bayes	$0.627 \pm 0.025$	$0.189 \pm 0.012$	$1.141 \pm 0.122$
Support Vector Machine	$0.627 \pm 0.019$	$0.133 \pm 0.005$	$0.431 \pm 0.011$
Deep Neural Network	$0.637 \pm 0.020$	<b><math>0.132 \pm 0.008</math></b>	$0.430 \pm 0.021$
Random Forest	$0.639 \pm 0.022$	<b><math>0.132 \pm 0.009</math></b>	<b><math>0.429 \pm 0.021</math></b>
Gradient Boosting	<b><math>0.641 \pm 0.021</math></b>	$0.151 \pm 0.007$	$0.484 \pm 0.016$
<b>Readmission within 1-year</b>			
Logistic Regression	$0.719 \pm 0.013$	<b><math>0.212 \pm 0.005</math></b>	$0.614 \pm 0.010$
Naive Bayes	$0.692 \pm 0.013$	$0.321 \pm 0.015$	$2.485 \pm 0.168$
Support Vector Machine	$0.718 \pm 0.013$	<b><math>0.212 \pm 0.004</math></b>	$0.613 \pm 0.008$
Deep Neural Network	$0.716 \pm 0.014$	$0.213 \pm 0.005$	$0.615 \pm 0.011$
Random Forest	$0.716 \pm 0.012$	$0.213 \pm 0.004$	$0.616 \pm 0.009$
Gradient Boosting	<b><math>0.720 \pm 0.012</math></b>	<b><math>0.212 \pm 0.004</math></b>	<b><math>0.612 \pm 0.009</math></b>

### 4.1. Thirty Day Readmission

Overall, discrimination for the various models was modest (Table 2). The discriminative ability of different models, as represented by ROC curves, is shown in Figure 1A. The GB model had the highest discrimination (c-statistic, 0.641; 95% CI, 0.621-0.662), although predictive accuracy was average (Brier score, 0.151; 95% CI, 0.144-0.158, log-loss 0.484; 95% CI, 0.468-0.500). The RF and DNN models displayed similar discrimination (RF: c-statistic, 0.639; 95% CI, 0.617-0.662 and DNN: c-statistic, 0.637; 95% CI, 0.617-0.657), with improved predictive accuracy and confidence compared to the GB model (Brier score 0.132, log-loss 0.430 for both RF and DNN). SVM (with class weighting) and LR performed similarly, both with slightly worse discrimination and predictive accuracy. Lastly, while NB also showed similar discrimination (c-statistic 0.627; 95% CI, 0.602-0.652), it had the lowest overall prediction score (Brier score, 0.189; 95% CI, 0.176-0.201, log-loss 1.141; 95% CI, 1.018-1.263).

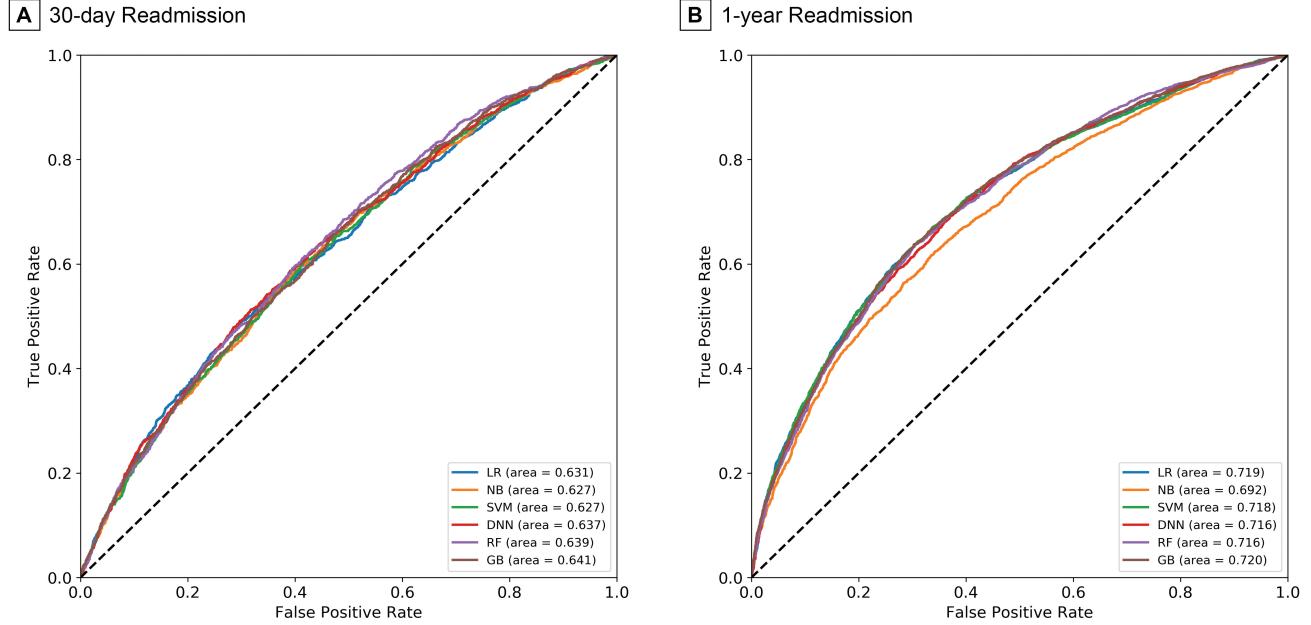


Figure 1: Receiver operating characteristic curves for (A) 30-day readmission and (B) 1-year readmission. The corresponding values of the area under the curve for each model are presented in Table 2.

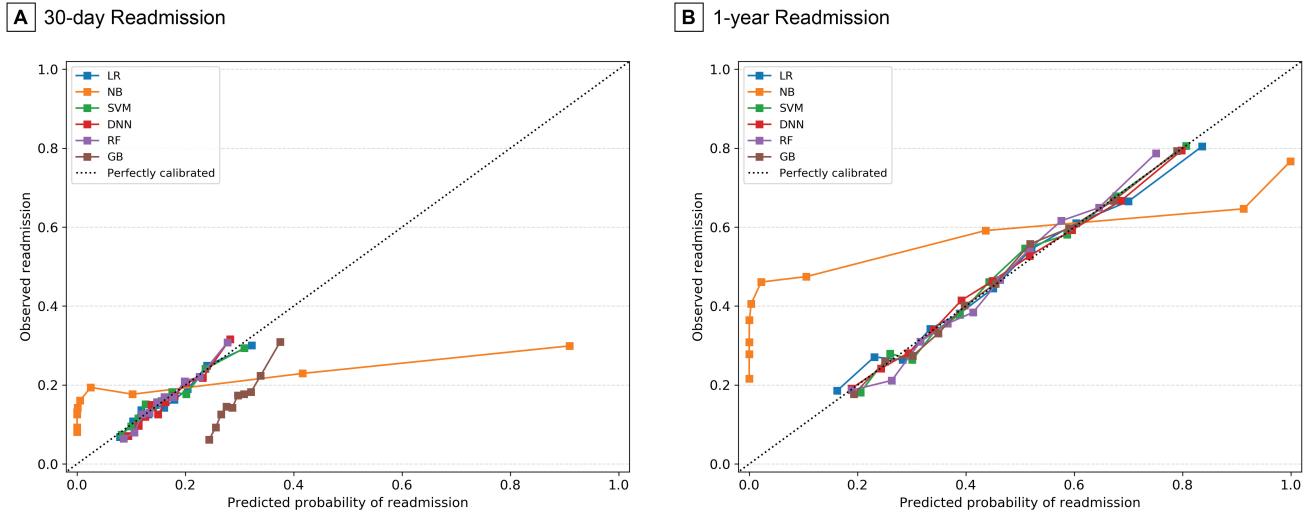


Figure 2: Calibration plots for machine learning models for predicting (A) 30-day readmission and (B) 1-year readmission.

The LR, SVM, DNN and RF models demonstrated good concordance between the observed and predicted probabilities of readmission (Figure 2A), although the range of predicted probabilities was limited between 0.042–0.489. The calibration curves show poor alignment for the GB model, despite having a c-statistic value similar to RF. The NB model had the worst model calibration despite a large predictive range of predicted probabilities (0.00-1.00), with no change in observed readmission rate regardless of predicted probability. Finally, despite the small range of predicted probabilities for the LR, DNN and RF models, the mean observed readmission rates were markedly higher in the highest predicted risk group (30.1%, 31.6%, 30.8% respectively) than the lowest predicted risk group (6.8%, 7.1%, 6.4%) (Figure 2A, with detailed values in Supplemental Table S3).

#### 4.2. One Year Readmission

The LR, SVM, DNN, RF and GB models resulted in similar scores for both measures of predictive accuracy (c-statistic  $\approx 0.72$ , Brier Score  $\approx 0.21$  and log-loss  $\approx 0.61$ ), with no model showing superior performance compared to the other (Table 2 and Figure 1B). As in the 30-day readmission sample, NB had worse discrimination (c-statistic 0.692; 95% CI 0.679-0.705) and worse predictive accuracy and confidence (Brier score, 0.321; 95% CI 0.306-336, log-loss 2.485; 95% CI 2.317-2.653). The LR, SVM, RF and DNN models all resulted in predictions with good concordance between observed and predicted probabilities for all ranges of predicted probability (Figure 2B). The range of predicted probabilities were also much larger compared to the models for 30-day readmission. The GB model also showed good calibration, unlike for 30-day readmission where calibration was poor. NB had the worst model calibration, with low concordance between observed and predicted probabilities for all ranges, similar to the 30-day readmission sample. The LR, SVM, DNN, RF and GB models were able to adequately risk stratify patients, with observed mean readmission rates ranging  $\approx 17\text{-}18\%$  among the lowest risk individuals to  $\approx 80.0\%$  among the highest risk individuals. The range was smaller for NB, with an observed mean readmission rate of 21.5% in the first risk decile and 76.7% in the highest risk decile (Figure 2B, with detailed values in Supplemental Table S4).

### 5. Discussion

In this study, we examined the ability of various ML algorithms to predict readmission using detailed clinical data, in patients who were hospitalized with AMI. We found that the use of ML algorithms did not lead to substantial improvements in prediction for either 30-day or 1-year readmission when compared with previously reported statistical methods. All models developed for 30-day readmission had similar performance and showed modest discrimination, with c-statistics in the range of 0.63-0.64. These values are consistent with the median c-statistic (0.65) reported by a recent review of risk-prediction models for 30-day AMI readmission<sup>12</sup>, although it should be noted that this number is likely inflated as not all models were validated. The models for 1-year readmission showed moderate performance, with c-statistics consistently around 0.72 with the exception of the NB model, also consistent with previous literature<sup>9</sup>.

Given the considerable burden placed on the healthcare system by readmissions after AMI,<sup>1,25</sup> predicting readmission risk would enable hospitals to target readmission reduction

interventions towards patients more likely to benefit from them, thereby improving patient health and reducing costs. As such, several models have been developed to estimate a patients' risk of readmission after AMI hospitalization. However, many of the existing prediction models use data that is not available until well after discharge (e.g., administrative claims or registry data)<sup>12</sup>, are developed using data derived from a single-center, or lack any validation.<sup>26,27,28</sup> We overcome these limitations by relying solely on data collected during a patient's hospital stay, from across 81 hospitals in Ontario, Canada. To our knowledge, this is the first study that uses detailed clinical data that is routinely available to clinicians to evaluate ML algorithms for the task of predicting AMI readmission. Moreover, using data that has been collected from multiple locations ensures that the variables included in the models represent standard-care of practise rather than esoteric tests and measurements. These factors help increase the applicability of the models to a clinical setting.

Our study found that ML methods do not provide an adequate increase in accuracy for predicting AMI readmission to warrant clinical use, when compared with conventional approaches reported in literature. The limited number of previously published papers for AMI readmission prediction rely mostly on points-based risk scores or LR models,<sup>29,9,10</sup> which tend to underperform when there are complex decision boundaries present in the data. Employing ML algorithms that are able to leverage complex higher-level interactions among input variables could potentially overcome this drawback. We found minor improvement in discrimination for 30-day readmission when using models such as RF, GB and DNN (0.64) as opposed to LR (0.63), which is encouraging, but ultimately insufficient. This suggests that the inclusion of complex interactions does not improve predictive accuracy in this context. Furthermore, calibration plots showed that no model was as or more accurate in predicting readmissions than LR. The outcome was similar in the case of 1-year readmission; LR performed just as well as more complex models such as DNN or RF when comparing c-statistic and model calibration. Despite the overall modest discrimination for both 30-day and 1-year readmission, we found that the range of observed readmission events among deciles of predicted risk varies sufficiently to stratify patients into low, medium and high-risk groups to target readmission intervention prevention.

We attribute the modest predictive ability of the various ML algorithms to two factors. First, predicting readmission is a difficult problem especially when compared with other<sup>30</sup> prediction tasks such as mortality.<sup>6,12,31</sup> Unlike mortality, which is largely dependent upon disease severity and comorbidity burden,<sup>32–36</sup> readmission is likely a result of more complex interactions between a patient's clinical condition, sociodemographic factors and psychosocial environment.<sup>33–36</sup> Such complex factors are unlikely to be captured within the data from a single hospital stay. Second, our study adds to the growing trend of literature which confirms that, while more complex ML methods have potential, they do not necessarily confer an advantage when the data are tabular, contain a relatively small number of features, and is of a modest sample size. This is particularly true of deep learning (deep neural network) models; while they are capable of using raw data as input, they require enormous amounts of structured data to train and achieve sufficiently high performance. For example, researchers recently demonstrated a deep learning model capable of achieving high accuracy on various tasks including in-hospital mortality (c-statistic 0.93–0.94), 30-day readmission (c-statistic 0.75–0.76), and prolonged length of stay (c-statistic 0.85–0.86).<sup>37</sup> This deep learning model was trained using raw time-ordered EHR data consisting of 46

billion data points (including clinical notes), from 216,221 patients. It is due to these large clinical datasets and computational resources that complex ML algorithms have become practical and useful. However, they cannot replace the use of classifiers and regressors on smaller, tabular datasets like that in this study.

### 5.1. Limitations

This study has several potential limitations. First, the ML models developed use data from the patient's entire stay to make a prediction at the time of discharge. While this would enable hospitals to implement post-discharge interventions, there is some evidence to suggest that certain interventions are more effective at reducing readmissions if implemented well before discharge.<sup>12,38</sup> Second, certain AMI severity measures such as troponin value and frailty were not included in the dataset, although it is unlikely that the addition of further severity scores will improve accuracy drastically given that alternate measures of disease severity are included. Lastly, while we used a robust nested-cross validation approach to verify our models, if these models were to be implemented they would require external validation on more recent data than used in the study to reflect any changes in care practises for AMI.<sup>39</sup>

## 6. Conclusion

In this analysis of readmission prediction after AMI hospitalization using clinical data, we found that ML methods do not improve discrimination when compared with previously reported approaches. Future work needs to focus on further improvement of predictive ability through the use of larger datasets that contain both clinical time-series data and information on sociodemographic factors, so as to facilitate earlier and better targeting of interventions to patients at high risk for readmission.

## Acknowledgments

This study was supported by the ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this article are those of the authors and are independent from the funding sources. No endorsement by ICES, or the MOHLTC is intended or should be inferred. Parts of this material are based on data and information compiled and provided by CIHI. The analyses, conclusions, opinions and statements expressed herein are those of the author, and not necessarily those of CIHI.

**Funding source:** This study was funded by a University of Toronto EMHSeed grant and a Foundation grant (FDN-154333) from the Canadian Institutes of Health Research.

**Disclosures:** Drs. Ko and Austin are supported by a Mid-Career Investigator Award from the HSF, Ontario Provincial Office. The other authors report no disclosures or conflicts.

## References

- [1] Harlan M Krumholz, Zhenqiu Lin, Patricia S Keenan, Jersey Chen, Joseph S Ross, Elizabeth E Drye, Susannah M Bernheim, Yun Wang, Elizabeth H Bradley, Lein F Han, et al. Relationship between hospital readmission and mortality rates for patients hospitalized with acute myocardial infarction, heart failure, or pneumonia. *Jama*, 309(6):587–593, 2013.
- [2] Karen E Joynt and Ashish K Jha. Characteristics of hospitals receiving penalties under the hospital readmissions reduction program. *Jama*, 309(4):342–343, 2013.
- [3] Sunil Kripalani, Cecelia N Theobald, Beth Anctil, and Eduard E Vasilevskis. Reducing hospital readmission rates: current strategies and future directions. *Annual review of medicine*, 65:471–485, 2014.
- [4] Cynthia Feltner, Christine D Jones, Crystal W Cené, Zhi-Jie Zheng, Carla A Sueta, Emmanuel JL Coker-Schwimmer, Marina Arvanitis, Kathleen N Lohr, Jennifer C Middleton, and Daniel E Jonas. Transitional care interventions to prevent readmissions for persons with heart failure: a systematic review and meta-analysis. *Annals of internal medicine*, 160(11):774–784, 2014.
- [5] Carl van Walraven, Irfan A Dhalla, Chaim Bell, Edward Etchells, Ian G Stiell, Kelly Zarnke, Peter C Austin, and Alan J Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6):551–557, 2010.
- [6] Huaqiong Zhou, Phillip R Della, Pamela Roberts, Louise Goh, and Satvinder S Dhaliwal. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ open*, 6(6):e011060, 2016.
- [7] Hao Wang, Richard D Robinson, Carlos Johnson, Nestor R Zenarosa, Rani D Jayswal, Joshua Keithley, and Kathleen A Delaney. Using the lace index to predict hospital readmissions in congestive heart failure patients. *BMC cardiovascular disorders*, 14(1):97, 2014.
- [8] Santu Rana, Truyen Tran, Wei Luo, Dinh Phung, Richard L Kennedy, and Svetha Venkatesh. Predicting unplanned readmission after myocardial infarction from routinely collected administrative hospital data. *Australian Health Review*, 38(4):377–382, 2014.
- [9] Harlan M Krumholz, Zhenqiu Lin, Elizabeth E Drye, Mayur M Desai, Lein F Han, Michael T Rapp, Jennifer A Mattera, and Sharon-Lise T Normand. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. *Circulation: Cardiovascular Quality and Outcomes*, 4(2):243–252, 2011.
- [10] Shipeng Yu, Faisal Farooq, Alexander Van Esbroeck, Glenn Fung, Vikram Anand, and Balaji Krishnapuram. Predicting readmission risk with institution-specific prediction models. *Artificial intelligence in medicine*, 65(2):89–96, 2015.

- [11] Lauren N Smith, Anil N Makam, Douglas Darden, Helen Mayo, Sandeep R Das, Ethan A Halm, and Oanh Kieu Nguyen. Acute myocardial infarction readmission risk prediction models: a systematic review of model performance. *Circulation: Cardiovascular Quality and Outcomes*, 11(1):e003885, 2018.
- [12] William H Crown. Potential application of machine learning in health outcomes research and some statistical cautions. *Value in health*, 18(2):137–140, 2015.
- [13] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.
- [14] Jack V Tu, Linda R Donovan, Douglas S Lee, Julie T Wang, Peter C Austin, David A Alter, and Dennis T Ko. Effectiveness of public report cards for improving the quality of cardiac care: the effect study: a randomized trial. *Jama*, 302(21):2330–2337, 2009.
- [15] Kevin P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2013.
- [16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [17] L Breiman. Random forests machine learning. 45: 5–32. *View Article PubMed/NCBI Google Scholar*, 2001.
- [18] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [19] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [21] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.
- [22] François Chollet. keras. <https://github.com/fchollet/keras>, 2015.
- [23] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.
- [24] Rohan Khera, Snigdha Jain, Ambarish Pandey, Vijay Agusala, Dharam J Kumbhani, Sandeep R Das, Jarett D Berry, James A de Lemos, and Saket Girotra. Comparison of readmission rates after acute myocardial infarction in 3 patient age groups (18 to 44, 45 to 64, and 65 years) in the united states. *The American journal of cardiology*, 120(10):1761–1767, 2017.

- [25] Jennifer Meddings, Heidi Reichert, Shawna N Smith, Theodore J Iwashyna, Kenneth M Langa, Timothy P Hofer, and Laurence F McMahon. The impact of disability and social determinants of health on condition-specific readmissions beyond medicare risk adjustments: a cohort study. *Journal of general internal medicine*, 32(1):71–80, 2017.
- [26] Jeremiah R Brown, Sheila M Conley, and Nathaniel W Niles. Predicting readmission or death after acute st-elevation myocardial infarction. *Clinical cardiology*, 36(10):570–575, 2013.
- [27] David D McManus, Jane S Saczynski, Darleen Lessard, Molly E Waring, Jeroan Allison, David C Parish, Robert J Goldberg, Arlene Ash, Catarina I Kiefe, TRACE-CORE Investigators, et al. Reliability of predicting early hospital readmission after discharge for an acute coronary syndrome using claims-based data. *The American journal of cardiology*, 117(4):501–507, 2016.
- [28] Robert E Burke, Jeffrey L Schnipper, Mark V Williams, Edmondo J Robinson, Edward E Vasilevskis, Sunil Kripalani, Joshua P Metlay, Grant S Fletcher, Andrew D Auerbach, and Jacques D Donzé. The hospital score predicts potentially preventable 30-day readmissions in conditions targeted by the hospital readmissions reduction program. *Medical care*, 55(3):285, 2017.
- [29] Karen E Joynt, E John Orav, and Ashish K Jha. Thirty-day readmission rates for medicare beneficiaries by race and site of care. *Jama*, 305(7):675–681, 2011.
- [30] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698, 2011.
- [31] Ruben Amarasingham, Billy J Moore, Ying P Tabak, Mark H Drazner, Christopher A Clark, Song Zhang, W Gary Reed, Timothy S Swanson, Ying Ma, and Ethan A Halm. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical care*, pages 981–988, 2010.
- [32] Randi E Foraker, Kathryn M Rose, Chirayath M Suchindran, Patricia P Chang, Ann M McNeill, and Wayne D Rosamond. Socioeconomic status, medicaid coverage, clinical comorbidity, and rehospitalization or death after an incident heart failure hospitalization: Atherosclerosis risk in communities cohort (1987 to 2004). *Circulation: Heart Failure*, 4(3):308–316, 2011.
- [33] Jeph Herrin, Justin St. Andre, Kevin Kenward, Maulik S Joshi, Anne-Marie J Audet, and Stephen C Hines. Community factors and hospital readmission rates. *Health services research*, 50(1):20–39, 2015.
- [34] David A Alter, Barry Franklin, Dennis T Ko, Peter C Austin, Douglas S Lee, Paul I Oh, Therese A Stukel, and Jack V Tu. Socioeconomic status, functional recovery, and long-term mortality among patients surviving acute myocardial infarction. *PLoS One*, 8(6):e65130, 2013.

- [35] Siddhartha Singh, Yu-Li Lin, Yong-Fang Kuo, Ann B Nattinger, and James S Goodwin. Variation in the risk of readmission among hospitals: the relative contribution of patient, hospital and inpatient provider characteristics. *Journal of general internal medicine*, 29(4):572–578, 2014.
- [36] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.
- [37] Stephanie Rennke, Oanh K Nguyen, Marwa H Shoeb, Yimdriuska Magan, Robert M Wachter, and Sumant R Ranji. Hospital-initiated transitional care interventions as a patient safety strategy: a systematic review. *Annals of internal medicine*, 158(5\_Part\_2):433–440, 2013.
- [38] Graham C Wong, Michelle Welsford, Craig Ainsworth, Wael Abuzeid, Christopher B Fordyce, Jennifer Greene, Thao Huynh, Laurie Lambert, Michel Le May, Sohrab Lutch-medial, et al. 2019 canadian cardiovascular society/canadian association of interventional cardiology guidelines on the acute management of st-elevation myocardial infarction: Focused update on regionalization and reperfusion. *Canadian Journal of Cardiology*, 35(2):107–132, 2019.