Machine Learning Prediction of Post-Operative Emergency Department Hospital Readmission

Forthcoming, *Anesthesiology*

Velibor V. Mišić[1], *PhD*, Assistant Professor*

Eilon Gabel[2], *MD*, Assistant Professor*

Ira Hofer[2], *MD, Assistant Professor*

Kumar Rajaram[1], *PhD, Professor*

Aman Mahajan[3], *MD PhD, Professor and Chair*

1. Decisions, Operations and Technology Management Area, Anderson School of Management, University of California Los Angeles, Los Angeles, CA.

2. Department of Anesthesiology and Perioperative Medicine, University of California Los Angeles, Los Angeles, CA.

3. Department of Anesthesiology and Perioperative Medicine, University of Pittsburgh, Pittsburgh PA.

* These authors contributed to this manuscript equally

Corresponding Author: Eilon Gabel MD
757 Westwood Plaza, Suite 3325
Los Angeles, CA 90095
egabel@mednet.ucla.edu
Office: 310.267.8693

***Acknowledgement**:* None

***Funding**:* None

Authors' conflicts of interests: None

Number of Words: Abstract: 305 / Intro: 379 / Discussion: 1070 / Total: 5949

Abstract

Background: Although prediction of hospital readmissions has been studied in medical patients, it has received relatively little attention in surgical patient populations. Published predictors require information only available at the moment of discharge. We hypothesized that machine learning approaches can be leveraged to accurately predict readmissions in post-operative patients from the emergency department. Further, we hypothesize that these approaches can accurately predict the risk of readmission much sooner than hospital discharge.

Methods: Using a cohort of surgical patients at a tertiary care academic medical center, surgical, demographic, lab, medication, care team, and current procedural terminology (CPT) data were extracted from the electronic health record. The primary outcome was whether there existed a future hospital readmission originating from the emergency department within 30 days of surgery. Secondarily, the time interval from surgery to the prediction was analyzed; at 0, 12, 24, 36, 48 and 60 hours. Different machine learning models for predicting the primary outcome were evaluated with respect to the area under the receiver-operator characteristic curve metric utilizing different permutations of the available features.

Results: 34,532 operative admissions from April 2013 to December 2016 were included in the analysis. Surgical and demographic features led to moderate discrimination for prediction after discharge (area under the curve: 0.74 – 0.76) while medication, consulting team, and current procedural terminology features did not improve the discrimination. Lab features improved discrimination, with gradient boosted trees attaining the best performance (area under the curve: 0.866, SD 0.006). This performance was sustained during temporal validation with 2017-2018

data (area under the curve: 0.85 – 0.88). Lastly, the discrimination of the predictions calculated 36 hours after surgery (area under the curve: 0.88 – 0.89) nearly matched those from time of discharge.

Conclusions: A machine learning approach to predicting post-operative readmission can produce hospital-specific models for accurately predicting 30-day readmissions via the emergency department. Moreover, these predictions can be confidently calculated at 36 hours after surgery without consideration of discharge-level data.

Keywords

- Machine Learning
- Hospital readmission
- Tree ensemble models
- Post-operative

Introduction

Unplanned hospital readmissions have been a pressing concern with respect to patient burden and high cost.[1] These factors are compounded when a patient is readmitted via the emergency department where they may consume valuable resources. This is in contrast to postoperative readmissions that occur via direct admission following clinic visits that are often planned or facilitated by a surgeon. While it is not always clear whether a given readmission is preventable, it is certain that emergency department-based 30-day readmissions are sentinel events, poor markers of quality, and are typically due to conditions present at discharge.[2,3]

Models created to detect patients at risk for unplanned readmission have been developed, but many of them have a narrow focus on single disease states and cannot be applied to the post-operative population as a whole.[4–6] Furthermore, most of these models were exclusively created for non-surgical patients, making them difficult to validate in different cohorts[7], and many of them require substantial data from the moment of hospital discharge.[8,9] Waiting until hospital discharge inhibits providers from better optimizing factors associated with readmission in parallel to the routine post-operative care.[4,10–12] This delay can also lead to ineffective transitional care coordination or possible prolongation of a hospital stay. Moreover, given that timeliness is one of the five rights of clinical decision support, waiting until the moment of discharge dramatically decrease the clinical efficacy of any model.[13]

A potential path to developing patient-risk models is machine learning with its ability to process extremely large numbers of input features and produce accurate predictive models.[14,15] More

specifically, tree-based machine learning methods are able to model nonlinear relationships and interactions, typically outperforming standard logistic regression.[16] Lastly, machine learning models which are calibrated using an institution's individual data demonstrate higher accuracy than models engineered for generalized patient cohorts that ordinarily overlook unforeseen institution-specific nuances.[17]

In this manuscript, we describe the creation and comprehensive validation of a machine learning based methodology for predicting a patient's risk for 30-day readmission via the emergency department in the post-operative period. Our primary hypothesis is that machine learning methods are capable of producing hospital-specific readmission prediction models with excellent discrimination. Our secondary hypothesis is that machine learning can produce models that accurately predict post-operative readmission without dependence on data from the time of discharge.

Materials and Methods

Data Extraction

This study (UCLA IRB #18-000630) qualified for UCLA IRB exception status ("waiver of consent") by virtue of having no direct patient contact and using a de-identified dataset. All study data were acquired via our previously published Department of Anesthesiology and Perioperative Medicine at UCLA's Perioperative Data Warehouse.[18] The Perioperative Data Warehouse is a structured reporting data schema that contains all the relevant clinical data entered into the EPIC (EPIC Systems, Verona, WI) electronic health record system. Data were acquired via Clarity, the relational database created by EPIC for data analytics and reporting. While Clarity contains raw clinical data, the Perioperative Data Warehouse was designed to organize, filter, and improve data so that it can be used reliably for creating these types of metrics. Other published manuscripts deriving data from the Perioperative Data Warehouse can be found in the reference section.[18–22]

Data extraction was restricted to the UCLA Ronald Reagan Medical Center for developing the model in a tertiary center. This was followed by a later extract of UCLA's Santa Monica Hospital for external validation of the methodology.

Model Endpoint Definition

We defined a readmission via the emergency department as any patient who enters UCLA through any of the emergency departments within 30 days of a surgical case and is then

transferred to a subsequent non-emergency department location. This definition was intended to capture patients that return to the hospital in an unplanned fashion and require inpatient or observation level of care, compared to those that were sent home after an emergency department-based evaluation.

There was a significant effort to mimic the Center for Medicare and Medicaid Services (CMS) definition of post-operative readmission, but in order to match definitions exactly, we would need access to the proprietary algorithms created by the various third-party vendors. Our health center currently uses one such service to analyze our own data to discern which of the 30-day readmissions are exempt based on an allowable disease condition or surgical procedure. Despite not having access to the aforementioned proprietary algorithms, we did adopt the Center for Medicare and Medicaid Services exclusion of cases that were discharged and readmitted on the same calendar day.

During the study interval, approximately 40% of the Ronald Reagan Medical Center 30-day readmissions arrived at the health center via the emergency department, another 40% entered though a perioperative location, and the remaining 20% of the readmissions were direct to an inpatient location. It was decided that the 40% of patients arriving via a perioperative location were not consistent with our primary outcome since it would be exceedingly difficult to identify which procedures were done as corrective revisions (unplanned) and those that occurred as part of staged procedure (planned). One such example is cataract surgery, since our ophthalmologists seldom operate on both eyes during a single case, but instead stagger the cases two weeks by design. As for the 20% that are direct admissions, the same dilemma applies as we would not

definitively be able to distinguish which patient were admitted from clinic because of complication versus those needing adjuvant medical treatment, i.e. post-operative chemotherapy.

Model Input Features

For each admission, we used a large collection of independent variables originating from different sources, which we summarize below:

General data: We considered variables that describe the surgery at admission, such as the volume of blood loss and the duration of surgery. We also considered variables that summarize overall health, such as the American Society of Anesthesiologists (ASA) score, as well as variables that describe specific aspects of patient health, such as the Acute Kidney Injury Network (AKIN) stage and whether the patient had received a consultation from the pain management service during a past admission.[23] We also considered non-medical, demographic variables such as the patient's age, ethnicity, race, and primary language (see Supplementary Table 1).

Lab data: We considered a collection of commonly ordered lab tests: Bilirubin, Creatinine, Glucose, Hematocrit, Hemoglobin, INR, PCO2, Platelets, PO2, Potassium, Sodium, Urea Nitrogen, and White Blood Cell Count. The process of identifying and grouping the labs was manually done by a physician informaticist on our research team with expertise in programming. The grouped labs are summarized in Supplementary Table 2. For each type of lab, the Perioperative Data Warehouse records different “subtypes” of labs. These subtypes arise out of

how the lab is ordered (for example, bilirubin may be measured as part of complete metabolic panel, or simply as a stand-alone lab result). For each lab subtype, we created two variables: one variable that measures the maximum deviation (delta) from the normal range observed in that lab subtype over the whole admission and one 0/1 variable to indicate whether this lab subtype was ever ordered. In addition, we created one variable to count the total number of unique lab subtypes ordered over the entire admission, and one 0/1 variable to indicate whether no labs were ever ordered during the admission. At the time of data collection and algorithm development, Logical Observation Identifiers Names and Codes (LOINC) codes were not available within our EMR implementation. This has since changed and LOINC codes are currently being implemented. Once complete, we will be able to test the substitution of LOINC codes for component identifiers.

Medication data: During each admission, patients are typically administered or prescribed many different classes of medications. For each medication, we had access to the pharmaceutical class (total of 99) and subclass (total of 537) as defined by our institution's pharmacy service. For each pharmaceutical class, we created one 0/1 indicator variable to indicate if any medications of that class were ever prescribed during admission; one 0/1 variable to indicate if any medications of that class were ever taken during the admission; and lastly, one 0/1 variable to indicate if any discharge medications of that class were prescribed during the admission. We also created variables to count the number of unique classes prescribed, the number of unique classes taken by the patient and the number of unique classes among the patient's discharge medications.

Team data: For each admission, we knew which of the 170 surgical/medical/consulting teams at the Ronald Reagan Medical Center were assigned to the patient during a given admission. For each team, we created a 0/1 indicator variable of whether that team was assigned to the admission. We also define a variable to count the total number of teams assigned to each admission.

Current procedural terminology (CPT) *data*: For each surgical case, we generated a list of all the current procedural terminology codes that were billed for by the surgeons. Due to computational considerations, we focused only on current procedural terminology codes that appear in at least 100 surgical admissions in the data set, resulting in 215 current procedural terminology codes. For each such code, we created a 0/1 variable to indicate whether that code is assigned to each principal surgical admission. We also created a variable to count the total number of current procedural terminology codes assigned to each admission.

Data Preprocessing

For a minority of admissions, the values of some variables were missing or null. For categorical variables, we added another category indicating the value was missing. For some numeric variables, we were able to infer that a missing value indicates that the variable value is zero. When this was not possible, we created a new 0/1 indicator of whether the variable value is missing for each admission, and we set the original missing values to zero.

We divided the data randomly into a training set and testing set according to a 70-30 split.[15] We used the training set (70%) to estimate and the test set (30%) to evaluate each model. Each split was done to preserve the proportion of readmitted/not readmitted cases found in the whole data set. We repeated this random splitting ten times.

Model Development

We considered three different types of models: regularized logistic regression, random forest, and gradient boosted trees; the latter two are classified as tree ensemble models.[14,24,25] We provide additional technical information on the methods and parameter settings in the Supplementary Material. These models differ from classical logistic regression in their ability to scale to large numbers of features without overfitting. Although the models accomplish this in different ways, the essential idea is that each model's estimation procedure has some mechanism for controlling the model complexity and how sensitive the model is to the data.[14]

Regularized logistic regression: Logistic regression models the probability of a readmission as a logistic function applied to a linear function of the independent variables. Ordinary logistic regression models are estimated by minimizing the negative log-likelihood. In this work, we did not use ordinary logistic regression, as it can overfit when the number of features is large relative to the training set size and perform poorly out of sample. We instead used *regularized* logistic regression, which differs from ordinary logistic regression in that the estimation minimizes the negative log-likelihood plus an additional term (the *regularization term*). We used L1 regularization, which uses the L1 norm of the coefficient vector as the regularization term.[26] L1

regularization has two special properties: it induces shrinkage (it reduces the magnitude of the coefficients) and sparsity (it returns models where many coefficients are set to zero).[15] Both of these properties lead to good predictive performance when the number of independent variables is very large relative to the number of observations. This form of regression is also known as LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression. Alternatives include L2 regularized logistic regression (also known as ridge regression), wherein the regularization term is the sum of the squares of the coefficients; and elastic net logistic regression, wherein the regularization term is a weighted combination of the LASSO and ridge regularization terms.

Random forest: The random forest algorithm works by estimating an ensemble of randomized classification trees. A classification tree is a predictive model where one follows a sequence of true/false queries along a tree to make a prediction.[27] Given an observation, each tree in the forest is used to make a prediction (readmission/no readmission) and the fraction of trees in the forest predicting readmission is then the predicted readmission probability. A stylized example of a random forest model being used for prediction is given in Supplementary Figure 1.

The random forest model is attractive over ordinary logistic regression for two reasons. First, random forests are able to automatically learn potentially nonlinear relationships between each feature and the readmission probability, as well as interactions between the features.[14,28] Second, random forests yield good performance when the number of features is large relative to the number of observations, possibly being as large or larger than the number of observations.[24,28,29]

Gradient boosted trees: The gradient boosted tree model is another type of tree ensemble model and also works by growing multiple classification trees. However, rather than growing each tree randomly, the trees are grown sequentially, where each new tree is selected to most reduce the error of the current ensemble. The ensemble outputs a probability between 0 and 1 of a readmission occurring, and the ensemble is trained to minimize logistic loss. Like random forests, gradient boosted trees are well-suited to prediction problems with large numbers of features, and can automatically handle nonlinearities and interactions.[14]

Pre-discharge vs. Post-discharge Prediction

We performed an additional set of experiments to evaluate our models' performance using only pre-discharge information. We focused on the general data and the lab data. For the general data, we excluded any features that would only be available at discharge. For the lab data, we computed the features in the same way as before, but ignored any labs drawn later than T hours after the completion of surgery, where T is a parameter that we vary. We restricted T to the range $\{0, 12, 24, 36, 48, 60\}$. (For example, a value of $T = 24$ corresponds to using all data from the time of admission, to 24 hours after the completion of surgery.) In the case that a patient is discharged before the cutoff point of T hours after surgery completion, we did not use any additional information that may become available at or after discharge (such as length of stay). Note that the medication, team, and current procedural terminology data was not used in these additional experiments.

Model Performance

To evaluate our models, we considered two predictive metrics: area under the receiver-operator characteristic curve (also known as the *c* statistic or simply area under the curve) and Brier score. The area under the curve is the probability of correctly distinguishing between a randomly chosen admission from the test set that results in a readmission, and a randomly chosen admission from the test set that does not. The area under the curve is bounded between 0.5 and 1.0, with higher values being better. The Brier score is the mean squared difference between the predicted probability of readmission and the actual outcome (0 or 1 where 1 indicates a readmission). The Brier score is bounded between 0 and 1, with lower values being better. Both the area under the curve and Brier score metrics were calculated using the test data set and averaged over the ten random splits of the data. We additionally compared the models by plotting their receiver-operator characteristic curves and their precision-recall curves.

We compared our area under the curve values against the HOSPITAL (low Hemoglobin at discharge, discharge from an Oncology service, low Sodium at discharge, (International Classification of Disease, Ninth Revision, Clinical Modification coded) Procedure during hospital stay or not, Index admission Type is elective or not, number of Admissions in previous year, and Length of stay) score, a recent model for readmission risk prediction developed in a general, non-surgical patient population.[30] We acknowledge that the HOSPITAL score was not originally formulated for surgical readmission prediction; our interest in evaluating HOSPITAL is simply to obtain a reasonable benchmark. We also compared our area under the curve values against the LACE (Length of stay, Acuity, Comorbidity, Emergency department utilization) score. The LACE score and the later LACE+ score (which additionally uses patient age and sex,

hospital teaching status, acute diagnoses and procedures during admission, number of days on alternative level of care and number of admissions to the hospital in the previous year) have been applied to populations consisting of both medical and surgical patients.[9,31]

External Validation

As an additional exercise in validating the overall methodological approach, we also applied it to the UCLA Santa Monica Hospital. The Santa Monica Hospital differs greatly from the Ronald Reagan Medical Center as it functions much more like a community hospital with primarily orthopedic cases and lower patient acuity. Typically, patients with significant comorbidities or those requiring extensive surgeries are transferred by ambulance from the Santa Monica Hospital to the Ronald Reagan Medical Center. The Santa Monica Hospital has a fraction of the intensive care capabilities and a fraction of the sub-specialty resources.

To validate the overall methodological approach, we extracted identical features using identical inclusion/exclusion criteria from the Perioperative Data Warehouse for the Santa Monica Hospital. After splitting the data in the same way as for the Ronald Reagan Medical Center, we re-trained our models using the Santa Monica Hospital training sets and tested their performance on the Santa Monica Hospital test sets.

We remark here that this is different from how external validation is commonly done. Ordinarily, external validation entails taking a model trained on one data set (in this case, the Ronald Reagan Medical Center data) and evaluating its performance on a different data set (in this case, the

Santa Monica Hospital data). We emphasize that our goal is to validate our *overall methodology* – the process that transforms a data source, such as the Perioperative Data Warehouse, into a data set that can then be turned into a predictive model – and not to validate the specific *models* that arise from the Ronald Reagan Medical Center data set. Given the hospital-specific nature of readmissions, it is unreasonable to expect the existence of a single "universal" model that will achieve good predictive performance across a wide range of institutions. Furthermore, even if such a model were to exist, an institution may still wish to leverage its own data to construct a model tailored to its own patient population and surgical practices, in order to achieve even better performance. By validating our methodology on the Santa Monica Hospital, a hospital with a different patient population and surgical specialties from the Ronald Reagan Medical Center, we intended to verify whether our methodology is applicable to institutions outside of the Ronald Reagan Medical Center.

Temporal Validation

In addition to external validation, we also carried out a temporal validation of our methodology. We extracted admissions occurring at the Ronald Reagan Medical Center in 2017 and 2018, after the period spanned by our base dataset (2013-2016), and derived the endpoint and the features in exactly the same way as described for the base dataset. We then used the models we developed using the ten random splits of the 2013-2016 Ronald Reagan Medical Center data to make predictions on these admissions. The predictions were evaluated using the area under the curve as well as the positive predictive value and negative predictive value. The positive predictive value is calculated as the number of true positives divided by the total number of positives, while

the negative predictive value is calculated as the number of true negatives by the total number of negatives. The cutoff used for the positive predictive value and the negative predictive value was 0.20, i.e., a positive is defined as an admission with a predicted readmission probability of at least 0.20, while a negative is an admission with a predicted probability of less than 0.20. The purpose of this temporal validation was to evaluate the performance of the models when they are used prospectively, i.e., to predict on observations arising from the same institution in the future, thereby mimicking how these models would be used in practice.

Results

Data Extraction

28728 patients aged 18 or older were extracted from the Perioperative Data Warehouse from April 2013 to December 2016. This resulted in 34553 admissions. Of these, 21 were removed for being organ donors (ASA score of 6). The resulting 34532 admissions constituted our complete data set.

To define our endpoint, we considered the 3407 admissions that led to a 30-day emergency department visit. 1439 were excluded for not resulting in a readmission that led to a transfer to a unit beyond the initial emergency department encounter. Lastly, 26 were excluded due to the emergency department visit occurring on the same calendar day as discharge. This resulted in 1942 admissions with a readmission via the emergency department; thus, 5.6% of the admissions in the complete data set of 34532 admissions met the endpoint definition. Figure 2 summarizes the inclusion/exclusion criteria.

Admission Characteristics

Table 1 summarizes the general admission characteristics, while Table 2 summarizes the lab data. The total number of lab subtypes is 119; therefore, there are 119 maximum lab deviation variables, 119 lab presence indicator variables, one variable counting the total number of lab subtypes and one variable indicating no labs drawn, for a total of 240 lab-related variables. Of the 34532 admissions, 27071 admissions (78.4%) had at least one lab drawn. Patients were

administered 92 different pharmaceutical classes; following our feature creation procedure, this resulted in a total of 279 variables. The average number of unique classes prescribed per admission was 19.7, with a minimum of 0 and a maximum of 59; an average of 16.8 classes were taken by each admission and an average of 4.3 classes were prescribed at discharge. Our data set spanned 158 surgical teams; following our feature creation procedure, this resulted in a total of 159 team-related variables. For each admission, the number of teams ranged from 1 to 36, with a mean of 1.9 and a median of 1.

Model Performance

i. Predictive metrics

Table 3 displays the area under the curve and Brier score metrics for the different predictive models and using different sets of variables. Using only the general data resulted in area under the curve values on the order of 0.73 to 0.76. Adding the lab data appreciably increased area under the curve values to the 0.85 – 0.87 range. The medication, team and current procedural terminology data did not lead to significant improvements in the area under the curve. The HOSPITAL score achieved an area under the curve of 0.73, which is lower than the area under the curve values of our models. This value is comparable to that in the original HOSPITAL paper, where it was tested with a general medical population.[30] The area under the curve for LACE was also 0.73, which is slightly higher than in the original LACE paper[9]. The same qualitative behavior is also observed for the Brier score.

To further compare the models, Figure 2 plots the receiver-operator characteristic curves for our models using the general data and the lab data, as well as the HOSPITAL and LACE scores, for one random split of the data. Figure 3 similarly plots the precision-recall curves for our models using the general and lab data, as well as HOSPITAL and LACE, for the same random split of the data. Our three models gave similar performance, while outperforming both HOSPITAL and LACE. Lastly, Supplementary Figure 2 plots calibration curves for our models using the general and the lab data, for the same random split of the data. Our models generally produced probability predictions that closely match the actual readmission probabilities. The random forest model appears to have overestimated the readmission risk when it is in the 0.05-0.30 range, whereas both L1 regularized logistic regression and the gradient boosted tree model appear to have slightly underestimated the readmission risk in the same range. Calibration results in table form for the same random split are provided as Supplementary Table 3.

ii. Feature importance

Supplementary Figure 3 shows the top 30 most important variables of the random forest model for a single random split of the data. The majority of the top 30 variables were maximum delta and indicator variables associated with a variety of lab subtypes. In addition to the lab-based features, the top 30 for the random forest model also included length of stay, the relative time to surgery and the duration of surgery. Supplementary Figures 4 and 5 similarly shows the top 30 variables of the gradient boosted tree model and the regularized logistic regression model, respectively.

iii. Pre-discharge vs. Post-discharge Prediction

Supplementary Figure 6 shows how the area under the curve varies as the parameter *T*, which determines when the prediction is to be made beyond the start of surgery, varies. As *T* increased, all three models improved and plateaued after *T* = 1.5 days. After *T* =1.5 days (i.e., 36 hours after the start of surgery), the area under the curve values were comparable to those achieved when making the prediction after discharge.

iv. External validation using the Santa Monica Hospital data

We applied the same methodology as the Ronald Reagan Medical Center to the Santa Monica Hospital. We extracted data from 19650 surgical admissions at the Santa Monica Hospital from the same time span. After removing organ donors (18 admissions), the resulting 19632 admissions constituted our complete Santa Monica Hospital data set. Of these, 820 (4.2%) admissions had an emergency department readmission and met our endpoint definition.

We divided the data into a training and testing set according to a 70-30 split. After training on 70% of the data, the test set contained 5890 admissions resulting in 246 readmissions via the emergency department (see the Santa Monica Hospital consort diagram in Supplementary Figure 7). We trained the same three machine learning models using only the general data, and using both the general and lab data. The model results with the general and lab data were nearly identical to those for the Ronald Reagan Medical Center, with area under the curve values in the 0.86 – 0.88 range (see Table 4). In addition, we also used the data to train and evaluate models

for pre-discharge prediction, analogously to the Ronald Reagan Medical Center. The area under the curve as a function of the cutoff parameter *T* is shown in Supplementary Figure 8. This figure is consistent with Supplementary Figure 6 for the Ronald Reagan Medical Center, showing that high area under the curve values can be obtained as soon as 36 hours after surgery.

v. Temporal validation at the Ronald Reagan Medical Center using 2017-2018 data

We extracted 19343 admissions from the Ronald Reagan Medical Center in the period 2017-2018. Of these, 12 were removed due to organ donor status, leaving 19331 admissions eligible for analysis. Within this set of admissions, 969 (5.0%) met our endpoint definition of an emergency department readmission. We considered all of our machine learning models built using the general and lab-based features, as this combination of features led to the best predictive performance when tested on 2013-2016 data. We evaluated the area under the curve of each such model when used to predict on admissions in 2017-2018. We also evaluated the positive predictive value and negative predictive value for a probability cutoff of 0.20. The results are shown in Table 5. These results were consistent with our earlier evaluation using 2013-2016 data (Table 3), with all of our models achieving area under the curve values in the 0.85-0.88 range. In general, one would expect a model to perform worse in temporal validation, because observations in the future may behave differently from those in the training and testing data (for example, due to changes in operations or patient mix enacted since 2016). The consistency between our temporal validation results and our earlier results for 2013-2016 suggests that admissions in 2017-2018 behave similarly to those in 2013-2016. All three models achieved positive predictive values in the range of 0.20-0.40, with negative predictive values of over 0.96. We note that the low positive predictive values are to be expected, due to the low emergency

department readmission rate in the data. L1 logistic regression and gradient boosted trees achieve higher positive predictive values than random forest, at the cost of a slightly reduced negative predictive value. This is consistent with our calibration results (Supplementary Figure 2), which suggest that random forest often overestimates the readmission probability. As in our previous results, gradient boosted trees and random forest did not exhibit an edge over L1 logistic regression, despite their ability to automatically model interactions and nonlinearities.

Discussion

Our analysis of this data furnishes us with two key insights about prediction of emergency department readmission in post-operative patients. First, all of the machine learning models achieve high discrimination. Irrespective of which specific model is chosen, all of the models achieved out-of-sample area under the curve values in the 0.85 - 0.87 range using general demographic data, surgical data, and basic lab data. Most prior work in predicting readmissions in general medical patients achieved area under the curve values in the 0.6 - 0.7 range[4], with HOSPITAL and LACE achieving area under the curve values of 0.72 and 0.68 in prior work, respectively.[9,30] Secondly, we find that there is virtually no loss in performance if our models are restricted to using data available within 36 hours after the completion of surgery. This suggests that our models could be used to identify patients that are at high risk of readmission while still in the hospital and soon after the surgical procedure.

It is important to emphasize the difference in the definition of unplanned readmission between this manuscript and the Center for Medicare and Medicaid Services (CMS). Our prediction models are able to identify patients that are readmitted into the hospital via the emergency department irrespective of the cause. We believe that focusing on emergency department readmissions is more appropriate as such readmissions are unplanned in nature and costly. This is in contrast to the definition of unplanned readmission by the Center for Medicare and Medicaid Services and *Horwitz et al*.[32] Their definition of unplanned readmission excludes special cases such as transplanted patients and those undergoing chemotherapy. This highlights the clinical orientation of this manuscript rather than compliance to administrative workflows.

Choosing a Model

Interestingly, we do not find a significant difference between the three machine learning models. In particular, L1 regularized logistic regression obtained comparable performance to both random forest and gradient boosted trees. As compared to logistic regression, tree ensemble models have the advantage of being able to model nonlinear relationships and automatically incorporate interactions among a large number of features to extract accurate predictive models. Our results suggest that this level of flexibility is not necessary for modeling emergency department readmission risk.

We emphasize here that L1 regularized logistic regression is different from ordinary logistic regression, as we discussed in the Materials and Methods section. We found that in ordinary logistic regression, the estimation procedure would over fit on the training set and would produce grossly inaccurate test set predictions (for example, the average test set area under the curve with the general and labs data was 0.77, with a standard error of 0.04).

The main characteristic of the models that is crucial to the success of our approach is their ability to avoid overfitting in the presence of large numbers of features. Because of limitations of classical logistic regression, traditional approaches to developing risk models require significant manual input and expertise to filter out unnecessary features, and advocate for soft limits on the number of features, such as the "one-in-ten" rule of not using more than one feature per ten events in the data, so as to avoid overfitting.[33–35] Machine learning methods, in contrast, are able to automatically handle large number of features and extract accurate predictive models without

overfitting, thereby "letting the data speak". Fortunately, there are many available resources for physicians to learn more about different machine learning fundamentals and techniques.[36–38]

Study Limitations

In this study, there were a few limitations that are inherent in these types of retrospective, machine learning projects. First, we have little insight into the reason for hospital readmission. While we have created a model to predict that a patient will come back to the emergency room in 30 days, there is no information about how to possibly prevent the admission. While there is a significant body of research that tries to address common risk factors associated with unplanned readmission, our models are unable to offer any insight into its decision making rationale. This is further complicated by the use of a de-identified research dataset. With future integrations with clinical workflows, better identifying patient characteristics (i.e. ZIP code as a surrogate for socioeconomic status) can be used to identify factors that are easier for clinicians to identify in practice.[39]

Second, our electronic health record system only captures readmission to the UCLA Health System. Any outside readmission would not be entered into our institution's electronic health record system and thus be invisible to the model. Epic does have CareEverywhere, a cross-institution platform, and there are local Accountable Care Organization (ACO) payor databases, but these options were beyond the scope of this study.

Third, as discussed in the material and methods section, our definition of an unplanned emergency department-based post-operative readmission strays from the actual definition used by the Center for Medicare and Medicaid Services. The need for the deviation stems from our inability to replicate the work done by the third-party vendors that our institution uses to calculate the reported readmission rate. One would need proprietary algorithms to further evaluate whether a given patient has an illness or underwent a surgical procedure that exempts them from being an infraction. Despite this limitation, we aligned with this standard readmission definition to the extent of our ability.

Lastly, our models are scalable to other institutions that have the capability to replicate the necessary general and lab data sets with computer scientists able to implement machine learning techniques. However, this would not be a simple out-of-the box type of implementation; there would be a necessary retraining/recalibration process for each individual site that requires some level of machine learning proficiency. While this process is not as intensive as the process of gathering the perfect dataset, it demonstrates the hospital-specific nature of hospital readmissions. When implementing the models using the Santa Monica Hospital data for validation, we consumed 70% of the available data for training. This would imply that any future institutions would need a few thousand patients already extracted to fine-tune the chosen model. After that initial process, there would be optional recalibration to improve performance, but that would not be a mandate.

In summary, this work demonstrates the ability of machine learning techniques to produce clinically useful models for predicting readmissions via the emergency department in surgical

patients from electronic health record data. We have shown that predictions with high discrimination can occur as soon as 36 hours after the completion of surgery, giving time for intervention teams to apply readmission reduction strategies.[1] While this process has only been implemented using retrospective data, a natural next step for future research is to implement this process in real time, and quantify the benefits of guiding interventions using machine learning.

References

1. Axon RN, Williams M V. Hospital Readmission as an Accountability Measure. JAMA 2011;305:504.
2. Marcantonio ER, McKean S, Goldfinger M, Kleefield S, Yurkofsky M, Brennan TA. Factors associated with unplanned hospital readmission among patients 65 years of age and older in a medicare managed care plan. Am J Med 1999;107:13–7.
3. Frankl SE, Breeling JL, Goldman L. Preventability of emergent hospital readmission. Am J Med 1991;90:667–74.
4. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S. Risk Prediction Models for Hospital Readmission. JAMA 2011;306:1688.
5. Vest JR, Gamm LD, Oxford BA, Gonzalez MI, Slawson KM. Determinants of preventable readmissions in the United States: a systematic review. Implement Sci 2010;5:88.
6. Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, Krumholz HM. Statistical models and patient predictors of readmission for heart failure: A systematic review. Arch Intern Med 2008;168:1371–86.
7. Nguyen OK, Halm EA, Makam AN. Further Limitations of the HOSPITAL Score in US Hospitals. JAMA Intern Med 2016;176:1232.
8. Nguyen OK, Makam AN, Clark C, Zhang S, Xie B, Velasco F, Amarasingham R, Halm EA. Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison. J Hosp Med 2016;11:473–80.
9. Walraven C van, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, Austin PC, Forster AJ. 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.

10. Jack BW, Chetty VK, Anthony D, Greenwald JL, Sanchez GM, Johnson AE, Forsythe SR, O'Donnell JK, Paasche-Orlow MK, Manasseh C, Martin S, Culpepper L. A reengineered hospital discharge program to decrease rehospitalization: a randomized trial. Ann Intern Med 2009;150:178–87.
11. Coleman EA, Parry C, Chalmers S, Min S. The Care Transitions Intervention. Arch Intern Med 2006;166:1822.
12. Naylor MD, Brooten D, Campbell R, Jacobsen BS, Mezey MD, Pauly M V, Schwartz JS. Comprehensive discharge planning and home follow-up of hospitalized elders: a randomized clinical trial. JAMA 1999;281:613–20.
13. Campbell R. The Five Rights of Clinical Decision Support: CDS Tools Helpful for Meeting Meaningful Use. J AHIMA 2013;84:42–47 (web version updated February 2016).
14. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York, 2009.
15. James G, Witten D, Hastie T, Tibshirani R. Statistical Learning. In: Springer, 2013:15–57.
16. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol 2008;77:802–13.
17. Oh J, Makar M, Fusco C, McCaffrey R, Rao K, Ryan EE, Washer L, West LR, Young VB, Guttag J, Hooper DC, Shenoy ES, Wiens J. A Generalizable, Data-Driven Approach to Predict Daily Risk of *Clostridium difficile* Infection at Two Large Academic Health Centers. Infect Control Hosp Epidemiol 2018;39:425–33.
18. Hofer IS, Gabel E, Pfeffer M, Mahbouba M, Mahajan A. A Systematic Approach to

Creation of a Perioperative Data Warehouse. Anesth Analg 2016;122:1880–4.

19. Gabel E, Hofer IS, Satou N, Grogan T, Shemin R, Mahajan A, Cannesson M. Creation and Validation of an Automated Algorithm to Determine Postoperative Ventilator Requirements After Cardiac Surgery. Anesth Analg 2017;124:1423–30.
20. Childers CP, Hofer IS, Cheng DS, Maggard-Gibbons M. Evaluating Surgeons on Intraoperative Disposable Supply Costs: Details Matter. J Gastrointest Surg 2018.
21. Gabel E, Shin J, Hofer I, Grogan T, Ziv K, Hong J, Dhillon A, Moore J, Mahajan A, Cannesson M. Digital Quality Improvement Approach Reduces the Need for Rescue Antiemetics in High-Risk Patients: A Comparative Effectiveness Study Using Interrupted Time Series and Propensity Score Matching Analysis. Anesth Analg 2018.
22. Hofer IS, Cheng D, Grogan T, Fujimoto Y, Yamada T, Beck L, Cannesson M, Mahajan A. Automated Assessment of Existing Patient's Revised Cardiac Risk Index Using Algorithmic Software. Anesth Analg 2018:1.
23. Mehta RL, Kellum JA, Shah S V, Molitoris BA, Ronco C, Warnock DG, Levin A. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 2007;11:R31.
24. Breiman L. Random Forests. Mach Learn 2001;45:5–32.
25. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Stat 2000;28:337–407.
26. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 1996:267–88.
27. Breiman L, Friedman J, Charles S, Olshen R. Classification and regression trees. Chapman & Hall, 1993.

28. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.

29. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43:1947–58.

30. Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially Avoidable 30-Day Hospital Readmissions in Medical Patients. JAMA Intern Med 2013;173:632.

31. Walraven C van, Wong J, Forster AJ. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. Open Med 2012;6:e80-90.

32. Horwitz LI, Grady JN, Cohen DB, Lin Z, Volpe M, Ngo CK, Masica AL, Long T, Wang J, Keenan M, Montague J, Suter LG, Ross JS, Drye EE, Krumholz HM, Bernheim SM. Development and Validation of an Algorithm to Identify Planned Readmissions From Claims Data. J Hosp Med 2015;10:670–7.

33. Harrell F. Regression Modeling Strategies. 1st ed. New York: Springer New York, 2001.

34. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.

35. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2013.

36. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89–109.

37. Waljee AK, Higgins PDR. Machine Learning in Medicine: A Primer for Physicians. Am J

Gastroenterol 2010;105:1224–6.

38. Maglogiannis IG. Emerging artificial intelligence applications in computer engineering : real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies. IOS Press, 2007.

39. Nagasako EM, Reidhead M, Waterman B, Dunagan WC. Adding socioeconomic data to hospital readmissions calculations may produce more useful results. Health Aff (Millwood) 2014;33:786–91.