




Machine-learning algorithms for predicting hospital re-admissions in sickle cell disease

Arisha Patel,^{1,*} Kyra Gan,^{2,*} 
 Andrew A. Li,² Jeremy Weiss,³
 Mehdi Nouraie,⁴  Sridhar Tayur⁵ and
 Enrico M. Novelli⁶ 

¹Tepper School of Business, Carnegie Mellon University, ²Operations Research, Tepper School of Business, Carnegie Mellon University, ³Heinz College, Carnegie Mellon University, ⁴Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, University of Pittsburgh, ⁵Operations Management, Tepper School of Business, Carnegie Mellon University, and ⁶Heart, Lung and Blood Vascular Medicine Institute, University of Pittsburgh, Pittsburgh, PA, USA

Received 14 June 2020; accepted for publication 21 August 2020

Correspondence: Enrico M Novelli, MD, MS, Heart, Lung and Blood Vascular Medicine Institute, University of Pittsburgh, E1240 Biomedical Science Tower, 200 Lothrop St., Pittsburgh, PA 15260, USA.
 E-mail: novellie@upmc.edu

*AP and KG contributed equally to this work.

Summary

Reducing preventable hospital re-admissions in Sickle Cell Disease (SCD) could potentially improve outcomes and decrease healthcare costs. In a retrospective study of electronic health records, we hypothesized Machine-Learning (ML) algorithms may outperform standard re-admission scoring systems (LACE and HOSPITAL indices). Participants ($n = 446$) included patients with SCD with at least one unplanned inpatient encounter between January 1, 2013, and November 1, 2018. Patients were randomly partitioned into training and testing groups. Unplanned hospital admissions ($n = 3299$) were stratified to training and testing samples. Potential predictors ($n = 486$), measured from the last unplanned inpatient discharge to the current unplanned inpatient visit, were obtained via both data-driven methods and clinical knowledge. Three standard ML algorithms, Logistic Regression (LR), Support-Vector Machine (SVM), and Random Forest (RF) were applied. Prediction performance was assessed using the *C*-statistic, sensitivity, and specificity. In addition, we reported the most important predictors in our best models. In this dataset, ML algorithms outperformed LACE [*C*-statistic 0.6, 95% Confidence Interval (CI) 0.57–0.64] and HOSPITAL (*C*-statistic 0.69, 95% CI 0.66–0.72), with the RF (*C*-statistic 0.77, 95% CI 0.73–0.79) and LR (*C*-statistic 0.77, 95% CI 0.73–0.8) performing the best. ML algorithms can be powerful tools in predicting re-admission in high-risk patient groups.

Keywords: 30-day unplanned hospital readmission, machine learning, prediction, retrospective study, sickle cell disease.

Sickle Cell Disease (SCD) is the most common inherited haemoglobinopathy worldwide and carries high morbidity and mortality.^{1,2} Complications related to SCD have resulted in prolonged hospitalisations and high frequency of 30-day hospital re-admissions.^{3–9} For example, in the largest retrospective multistate study of 21 112 adult patients with SCD in the United States, 33.4% of patients had 30-day re-admission with 22.1% re-admitted within 14 days.⁶ Other studies found that 50% of adult patients with SCD were re-admitted within 30 days, and those who returned within one week had the poorest overall prognosis.^{10,11} As policymakers are mandating the implementation of evidence-based quality improvement interventions, the frequency of 30-day hospital re-admissions

becomes an important clinical metric to assess the quality of care amongst chronic diseases, including SCD.¹²

Hospital re-admission risk has been traditionally calculated using simple scoring systems (such as the LACE and HOSPITAL indices) with limited features,^{13,14} and not specific to high-risk groups such as patients with SCD, where socioeconomic factors may play an important role in hospital re-admissions.^{15–19} For instance, the LACE index was validated on a Canadian middle-age population with few comorbidities,¹³ and therefore it does not capture the demographics and disease-specific complexities that are inherent in the SCD population. In fact, the predictors of hospital re-admission in patients with SCD are currently not being evaluated in clinical practice.

One limitation of standard models to predict hospital re-admissions is that they are hypothesis-driven; they use a fixed set of predictive features and may ignore disease-specific factors that can impact clinical outcomes. Machine-Learning (ML) algorithms—a class of algorithms that can be used in detecting underlying patterns in high-dimensional datasets—can potentially be a useful tool in predicting hospital re-admission risks in the SCD patient population. In many healthcare applications, the performance of ML algorithms has dominated that of traditional statistical methods,²⁰⁻²⁶ and several studies have employed ML algorithms to predict 30-day hospital re-admissions.²⁷⁻³³ However, none of them has been conducted on the high-risk SCD patient population.

The objective of this research is to explore the value of ML algorithms, combined with domain knowledge, in predicting hospital re-admission risk for a SCD patient population using a real-world data source.³⁴ Specifically, we used both clinical knowledge-driven and hypothesis-free data features extracted from Electronic Health Records (EHR) data to guide our ML models. We hypothesized that ML algorithms would: (i) outperform traditional risk-scoring systems; (ii) find a richer set of predictors that can better guide clinical practice; and hence (iii) be a more suitable tool in predicting hospital re-admission risk among the SCD patient population. We report study results using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines.³⁵

Materials and methods

Design and sample

The University of Pittsburgh Medical Center (UPMC) Institutional Review Board approved this study. The R3 Services through the Department of Bioinformatics served as an honest data broker to ensure all patient health information was de-identified and Health Insurance Portability and Accountability Act-compliant throughout the research cycle, including but not limited to data extraction, data management, analytical and machine-learning processes. All analyses were conducted on de-identified patient data. Our SCD patient cohort was selected from five hospitals across the UPMC hospital system, where patients with SCD are followed by the adult UPMC Sickle Cell Program's inpatient consult service. The UPMC Sickle Cell Program is the only provider of specialised care for SCD in the region, and thus only a negligible number of patients with SCD is admitted to hospitals where the UPMC Sickle Cell Program staff has no clinical privileges. The raw data contain the EHR data of 2, 824 patients selected by the principal diagnosis of SCD using the International Classification of Diseases (ICD)-9 and ICD-10 codes listed in Table I^{36,37} between January 1, 2013, and November 1, 2018. The preprocessed dataset contains 446 patients and 3 299 unplanned inpatient visits, and Fig 1 summarises the patient inclusion criteria of this study.

Table I. Sickle cell ICD-9/ICD-10 diagnosis codes.

ICD-9	282-41, 282-42, 282-6, 282-60, 282-61, 282-62, 282-63, 282-64, 282-68, 282-69
ICD-10	D57-0, D57-00, D57-01, D57-02, D57-1, D57-2, D57-20, D57-21, D57-211, D57-212, D57-219, D57-3, D57-4, D57-40, D57-41, D57-411, D57-412, D57-419, D57-8, D57-80, D57-81, D57-811, D57-812, D57-819

ICD-10 D57-3 (sickle cell trait) was removed from the inclusion criteria. After removing patients who were only diagnosed with sickle cell trait (ICD-10 D57-3) during the study period, we had 1 009 patients left in the dataset.

Outcome variables

An admission was defined as an unplanned inpatient hospital admission, identified by a non-elective hospital admission type as indicated by the EHR data. A re-admission was defined as an admission within 30 days of the discharge date of the last admission. We excluded any admission to a maternity unit, skilled nursing facility, and rehabilitation unit. In our study, a case was defined as an admission that resulted in a re-admission, while a control was indicated by an admission that did not result in a re-admission.

Predictor candidates

All the analyses were conducted on the de-identified patient dataset, and patients who were admitted to other hospitals not defined above were not captured. The preprocessed features ($n = 481$), including labs, demographics, the number of outpatient visits prior to the current visit, and the number of Emergency Department (ED) visits prior to the current visit,^{15,16,38} were extracted from the EHR data using both data-driven methods and clinical knowledge (Table II). The dataset also included 21 variables extracted according to the LACE¹³ and HOSPITAL¹⁴ indices: the length of stay, the number of ED visits in the past six months, the number of (unplanned) hospital admissions in the past year, whether any procedure was performed during the hospitalisation, and 17 ICD-9/ICD-10 code groups to calculate the Charlson comorbidity index score in the LACE index. The remaining features included 340 ICD-9/ICD-10 diagnosis codes, two demographic features, four healthcare insurance provider types, 42 medication groups, 13 lab categories, 25 procedures, two zip codes, five smoking status features, seven vital signs, 34 hospital departments, and the number of outpatient visits (prior to the current visit in the study period). To further capture the trend in patient re-admission patterns, we included additional variables: the number of ED visits (prior to the current visit) in the study period, the number of days since the last inpatient visit (of the current visit), and the number of inpatient visits (prior to the current visit) in the study period. We included labs that were processed through a centralised lab and excluded point of care testing.

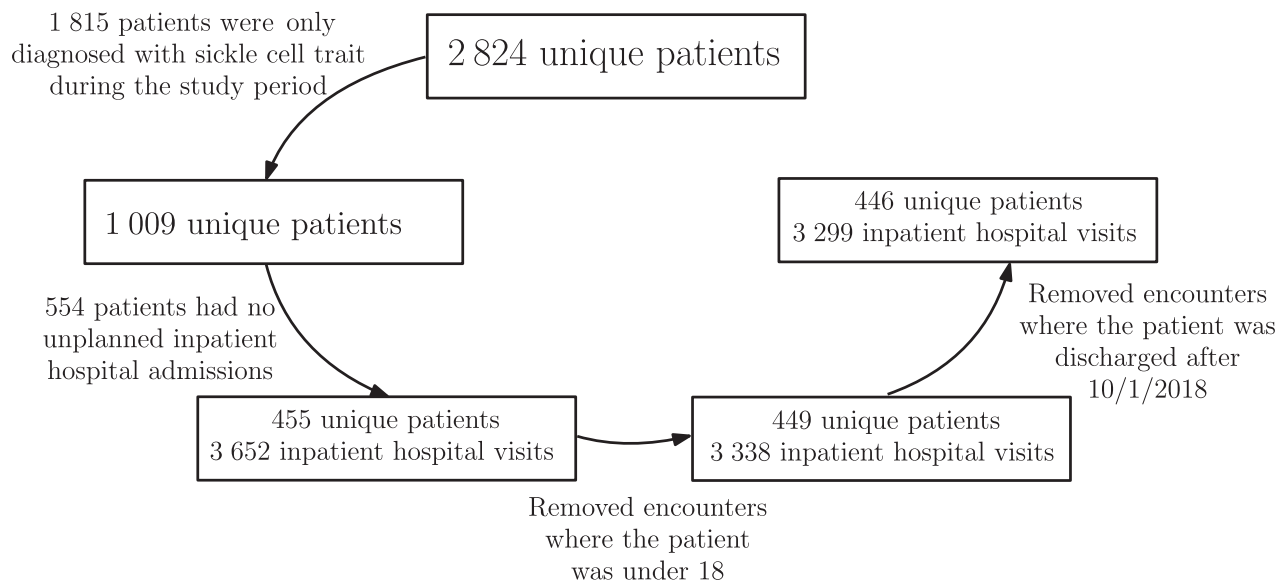


Fig 1. Study inclusion criteria flow chart. Description of the patient and inpatient visit inclusion criteria. At least one unplanned inpatient visit was made by 455 patients from January 1, 2013 to November 1, 2018. All consecutive ($n = 15$) unplanned inpatient admissions where the discharge and re-admission dates were the same were combined. We removed any inpatient encounters in which the patient was under the age of 18 at the time of the visit given that we did not have access to the Children's Hospital EHR database. Since inpatient visits after October 1, 2018 were censored, we removed those visits, resulting in 446 patients and 3 299 unplanned inpatient visits.

Data preprocessing

Each lab variable took six categorical values (Tables II and III) to indicate whether a lab result was missing, normal, low, high, low panic, or high panic. All lab variables were defined based on central lab reference values and were not adjusted to the normalised lab values for an individual patient. The vital sign variables were kept as continuous in the Random Forest (RF) model and were preprocessed into categorical variables in the Logistic Regression (LR) and Supported-Vector Machine (SVM) models. Table II includes the cut-off values for preprocessing these variables into categorical variables. The reason why the vital sign variables were coded as continuous instead of predefining the cut-off values using domain knowledge as in the RF model is that the RF algorithm automatically selects cut-off values that have high predictive value (indeed, this is one of the RF algorithm's advantages). Out of the 3 299 encounters, 873 (26.5%) did not have any vital signs taken; 685 (20.8%) did not have smoking status; 656 (19.8%) did not have any medication prescriptions; 454 (13.8%) did not have any procedures performed; 39 (1.2%) did not have any lab tests. The latter three could be classified as missing values or not applicable depending on the individual patient circumstance. The rest of the data did not contain any missing information. Table II describes the percentage available information for each individual variable in detail. Instead of imputing the missing values, we created a dummy variable for each variable that contains missing information to indicate whether this variable is missing in a particular encounter. This is a popular method in the ML community to handle missing data, and

has shown superiority to other methods in healthcare applications where data is not missing at random but rather a reflection of the decision made by care providers.^{39,40} Twenty-seven out of 195 re-admission patients died during the observation period, and these 27 patients had 200 unplanned inpatient admissions in total. Sixty out of 446 total patients died during the observation period, and these 60 patients had 247 unplanned inpatient admissions. Since the number of admissions that resulted in mortality was less than 2%, those admissions were kept in the training and testing dataset.

Methods

To predict whether an inpatient visit would result in a re-admission, three standard ML algorithms were applied using the scikit-learn package in Python: LR,⁴¹ SVM,⁴² and RF.⁴³ Traditional risk-scoring systems, the LACE and HOSPITAL indices, were also applied.^{13,14} Although LACE and HOSPITAL have not been previously applied to the SCD patient population, they provide two benchmark models for comparison. All variables needed to compute those two indices were contained in the EHR data. In addition, to test the impact of patients with frequent admissions on our ML models, we included a weighted RF model where each admission was weighted inversely by the total number of admissions incurred by the patient during the study period. Supplementary Section A describes the details of each algorithm and how they were used. The features mentioned above were treated as inputs to these models. We randomly selected the admissions incurred by 30% of the 195 return patients and 251 non-return patients to be the testing set ($n = 134$); the training set contained the admissions

Table II. Data preprocessing.

Variables (Categorical/Real valued)	Num. vars. before	After (representation)	Missing data	Variable descriptions and preprocessing steps
Healthcare insurance providers (C)	50	4 (4)	No missing data	Grouped insurance into 4 types: private, government, auto/employment, Medicare/Medicaid.
ICD-9/ICD-10 diagnosis codes (C)	3849	340 (340)	No missing data	In addition to removing diagnosis codes that appeared less than 20 times, we also hand-picked 37 groups of diagnosis codes, including 3 sickle cell genotypes listed in Table IV and 17 groups from the LACE index to calculate the Charlson comorbidity index score.
Number of different lab tests (C)	2945	13 (78)	39 (1.2%) encounters had none of the 13 labs performed; see right for details on % encounters(out of 3299) having each of the 13 labs performed	Hand-picked 13 sickle cell-related labs [% encounters having this test performed]: white blood cell count [98.5%], platelets count [98.5%], haemoglobin [15.5%], haematocrit [98.5%], reticulocytes count [77.1%], bilirubin [62.2%], lactic dehydrogenase (LDH) [58.3%] [tissue damage (i.e. anaemia)], lactate blood [13.4%] (acid base imbalance i.e. lactic acidosis secondary to shock), creatinine [91.5%], bun/creatinine ratio [1.8%], creatinine clearance [0%], Pro BNP [0%], sodium (from the HOSPITAL index) [91.4%]. Each variable takes six categorical values and was represented by one-hot encoding. Table III describes the details of how those lab variables were extracted
Procedures (C, R)	2808	25 (25)	454 (13.8%) encounters had no procedures performed	Extracted whether any procedure was performed during the hospitalisation (C) and the number of blood transfusions performed (R); removed the procedure codes (C) that appeared less than 20 times.
Number of unique NDC codes for medication (C)	4358	42 (43)	656 (19.8%) encounters had no medication prescription; the rest had at least one prescription	Identified 553 unique drugs and grouped them into 42 categories based on the drug effect. An additional variable was added to indicate whether any medication was prescribed during the inpatient admission.
Zip codes (C)	190	2 (2)	No missing data	Removed the ones that appeared less than 20 times.
Smoking status (C)	10	5 (6)	685 (20.8%) encounters had no smoking status	Regrouped into: never smoker, former smoker, heavy tobacco smoker, light tobacco smoker, passive smoke exposure–never smoker
Vital signs (R, C)	7	7 (RF: 15; LR/SVM: 30)	873 (26.5%) encounters had none of the vitals; see right for details	[% encounters have this vital taken]: BMI (R or C: <18.5, 18.5-24.9, 25-29.9, 30-34.9, 35-39.9, >=40) [71.5%], BP_systolic (R or C: <90, 90-120, >120) [60.0%], BP_diastolic (R or C: <60, 60-80, >80) [59.7%], pulse (R or C: <60, 60-100, >100) [59.7%], temperature (R or C: <=35C/95F, (35C, 38C)/(95F, 100.4F), >=38C/100.4F) [59.3%], respiratory_rate (R or C: <12, 12-18, >18) [59.6%], BP_position (C) [1.3%]
Number of hospital departments (C)	34	34 (34)	No missing data	
Demographics (C, R)	2	2 (2)	No missing data	[% patients have this demographic reported]: Gender (C) is binary [100%], age at encounter (R) [100%]

Table II. (Continued)

Variables (Categorical/Real valued)	Num. vars. before	After (representation)	Missing data	Variable descriptions and preprocessing steps
Other variables included (R)	7	7 (7)	Not applicable	Length of stay, number of outpatient visits, number of ED visits, number of ED visits in the past 6 months, the number of days since the last inpatient visit, number of inpatient visits in study period, number of inpatient visits in the past year.

Description of the data preprocessing steps and the percentage of missing data. After preprocessing, we narrowed down the number of variables in our model to be 481. In the Random Forest (RF) model, the vital-sign variables are continuous, and we represented each vital sign variable using a tuple of size two with the first entry indicating whether the value of the variable is missing; this resulted in an overall vector representation of length 550. In the Logistic Regression (LR) and supported-vector machine (SVM) models, the vital-sign variables were preprocessed into categorical variables and this resulted in an overall vector representation of length 565. In the third column, the number inside parentheses is the size of the vector that we used to represent the corresponding features. The reasons that we used a larger vector to represent those features are due to (i) missing data; and (ii) the fact that a categorical variable takes multiple values. In the fourth column, we described the percentage data missing overall. In the fifth column, we described the details on the variables included and the percentage of patients with this variable measured (if applicable) in square brackets.

incurred by the remaining 211 patients. Thus, our training and testing sets contained the same demographic information, predictors and outcomes.

Model evaluation

We used the *C*-statistic, or equivalently the Area Under the (receiver operating characteristic) Curve (AUC), and precision–recall curves as two quantitative metrics for identifying predictive performance within each of the classifiers. For intuition: a perfect classifier achieves a *C*-statistic of 1, while random chance corresponds to a *C*-statistic of 0.5.

In addition, we reported the sensitivity and specificity of our best-performing model. Since the number of samples in our study was relatively small, our results might have been sensitive to different training and testing splits. To address this problem, we performed 100 different splits and averaged the resulting 100 *C*-statistics.

Study results

Our training and testing sets contained the same demographic information, predictors and outcomes. Table IV summarises the characteristics and demographics of the postprocessed dataset, as well as the distributions of LACE and HOSPITAL indices computed using the postprocessed data. Our cohort included 3 299 admissions of 446 adult patients with SCD. Of these patients, 195 (43.72% of re-admission) patients were re-admitted within 30 days for a total of 1 369 times. The average age of those 446 patients was 42.22 (SD = 19.03) years, and the average age of the 195 patients who had re-admission during the study period was 39.47 (18.14) years. The average LACE and HOSPITAL indices of those 3 299 admission were 10.26 (2.79) and 8.16 (2.40), respectively.

To prevent overfitting, in the LR model we added LASSO regularisation, and in the RF model we restricted the maximum depth of the decision trees to 15. Figure 2 summarises the two performance metrics of each model — the Receiver Operating Characteristic (ROC) and precision–recall curves. LACE had a *C*-statistic of 0.6 (95% CI 0.57–0.64); HOSPITAL performed slightly better than LACE (*C*-statistic 0.69, 95% CI 0.66–0.72); SVM with ‘rbf’ kernel outperformed HOSPITAL in terms of *C*-statistic (*C*-statistic 0.72, 95% CI 0.69–0.75); LR outperformed SVM by a large margin (*C*-statistic 0.77, 95% CI 0.73–0.8); RF performed similar to logistic regression (*C*-statistic 0.77, 95% CI 0.73–0.79). Furthermore, the weighted RF (*C*-statistic 0.77, 95% CI 0.73–0.79) model performed similar to the RF model. Similarly, in terms of precision–recall, SVM (AUC 0.68) outperformed HOSPITAL (AUC 0.56), and the RF model (AUC 0.74) and the LR model (AUC 0.72) performed the best. In both the ROC and precision–recall curves, we observed that the curves corresponding to RF and LR pointwise dominated those of LACE and HOSPITAL indices.

Having established that the RF and LR models had the best performance, we compared the sensitivities and specificities of those two models against those of the LACE and HOSPITAL indices in Tables V and VI, respectively. In Tables V and VI, the thresholds were chosen to match the specificities of RF and LR models to those of the LACE index and HOSPITAL index, respectively. A true negative case was determined as a hospital admission that did not result in a 30-day re-admission, and we correctly predicted so, and a true positive case was determined as a hospital admission that did result in a 30-day re-admission, and we also correctly predicted so. In Tables V and VI, we again observed that the performances of RF and LR were similar in terms of sensitivity at their corresponding chosen thresholds, and the

Table III. Lab variables included in the study.

Lab category	Included variable names	Excluded variable names
White blood cell	WBC, white blood cells, WBC count, WBC & other nucleated cells	White blood cells–urine >5 WBC/HPF (POC), WBC esterase, rare WBCs present no organisms present, No WBCs or organisms present, no WBCs present few gram-positive cocci in pairs, WBC–fluid, WBC morphology, WBC clumps, Fecal WBC, immature WBC forms
Platelets	Platelets, platelet count	Platelet morphology, heparin pf4 platelet antibody, heparin platelet ab, giant platelets, platelet estimate, large platelets, platelet function p2y12, platelet sufficiency, mean platelet volume, rapid pra (platelets), platelet function interp.
Haemoglobin	Haemoglobin f, rapid haemoglobin s, haemoglobin c. Crystals, haemoglobin s, haemoglobin c, haemoglobin a2, haemoglobin-plasma, total haemoglobin, thb (haemoglobin)	Methaemoglobin &&, % oxyhaemoglobin, haemoglobin (poc), haemoglobin - mixed venous, methaemoglobin - mixed venous, % reduced haemoglobin, atypical haemoglobin, hemocue haemoglobin (poc), methaemoglobin - venous, glycosylated haemoglobin, methaemoglobin, carboxyhaemoglobin, 'haemoglobin, qual', haemoglobin a, haemoglobin a1, haemoglobin a1c, haemoglobin capillary (poc), bedside haemoglobin poc, haemoglobin-arterial, haemoglobin-venous, calc. Haemoglobin istat
Haematocrit	Haematocrit, haematocrit(hct)	Haematocrit derived, haematocrit derived - mixed ven, haematocrit (hct) manual pcv &&, haematocrit (poc), haematocrit-body fluid (hct), haematocrit istat
Reticulocytes	Absolute reticulocytes, reticulocytes, reticulocytes-manual method	Immature reticulocyte fraction
Bilirubin	Total bilirubin, direct bilirubin, bilirubin unconjugated	Bilirubin-urine, bilirubin - urine (poc), bilirubin confirmation, bilirubin unconjugated, other total bilirubin
Lactic dehydrogenase(LDH) [tissue damage (i.e. anaemia)]	Lactic dehydrogenase, lactic dehydrogenase(ldh), other lactic dehydrogenase(ld)	'ldh, ascites fluid'
Lactate blood (acid base imbalance, i.e. lactic acidosis secondary to shock)	Lactate, lactate blood, lactate whole blood	Lactate csf, lactate istat
Creatinine	Creatinine, creatinine, whole blood, random urine creatinine, 'creatinine, random urine'	'creatinine, jp drainage', creatinine venous istat, fluid creatinine, creatinine poc, urine protein/creatinine ratio, protein/creatinine ratio, urine creatinine, total creatinine 24 hr urine, creatinine istat
Bun/creatinine ratio	Bun/creatinine ratio	Albumin/creatinine ratio
Creatinine clearance	Creatinine clearance, creatinine clearance (adult)	Creatinine clear (children's)
BNP	'Pro bnp, n-terminal'	
Sodium	Sodium (na), sodium (na) whole blood, sodium na whole blood, sodium arterial blood gas	Stool sodium (aka nastool), sodium istat, urine sodium (na), total sodium (na) 24hr urine, sodium (na) (poc)

The percentage of reticulocytes results that was normal, low, and high in this study was 37.6%, 0.9%, and 61.5%, respectively.

sensitivities of both models outperformed those of the LACE index and HOSPITAL index, respectively.

To check the clinical integrity of our models, we selected a subset of variables and reported their importance factors in our RF model³² (see Fig 3) and in our LR model (Fig 4). While the variables below the selected important predictors from the LR model had near-zero coefficients (i.e., they had

minimal impact on the prediction outcome), the variables outside the selected important predictions from the RF model could still have relatively large impacts on the model. Thus, in Fig 3 we provided the average information gain (the amount of improvement in classification) of the selected variables appearing in the RF model, and in Fig 4, we reported both the direction and the standardised magnitude

Table IV. Characteristics and demographics of the postprocessed dataset.

		Total <i>n</i> = 446	Re-admission Group <i>n</i> = 195
General	Unplanned inpatient encounters	3 299	2 823 (1 369 re-admissions)
	Number of ED visits	6 780	4 899
	Number of outpatient visits	10 731	5 978
	Average length of stay per admission	5.895 (5.974)	5.543 (5.586)
	Average number of admissions per patient	7.40 (12.90)	14.47 (16.97)
	Number with HbSS	255	139
	Number with HbSC	55	30
	Number with HbS/B ⁰ or HbS/B ⁺	36	26
Age	18–29	157	80
	30–49	136	56
	50–69	108	44
	70–89	42	14
	≥90	3	1
Gender	Male	175	70
	Female	271	125
LACE Index		10.26 (2.79)	10.52 (2.73)
HOSPITAL index		8.16 (2.40)	8.53 (2.32)

Description of the 3 299 encounters of the 446 patients included in the postprocessed dataset. We also included the distribution of LACE and HOSPITAL indices computed using the EHR data. The sickle cell genotypes HbS/B⁰ and HbS/B⁺ were grouped into one genotype since ICD-9 diagnosis codes do not distinguish between these two genotypes.

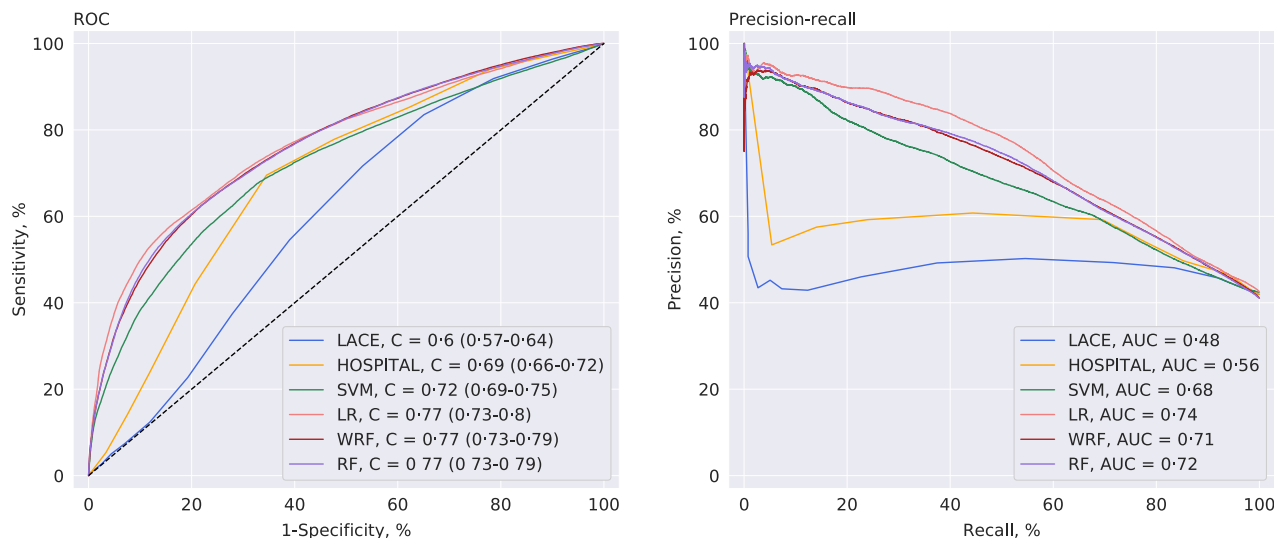


Fig 2. Performance metrics of machine learning models for predicting 30-day re-admissions in sickle cell disease. Two performance metrics measured out-of-sample and averaged over 100 independent train/test draws. (A) Receiver operating characteristic curves, and corresponding area under the curve; also known as the C-statistic. (B) Precision–recall curves.

of coefficients of the selected variables in the LR model. Both Figs 3 and 4 contain similar features.

Discussion

This is the first study to apply ML algorithms to predict the hospital re-admission rate in patients with SCD. We have shown how the risk of 30-day re-admission of a particular

SCD patient can be estimated by preprocessing the EHR data associated with an inpatient admission using our data preprocessing steps, and then inputting the data into our pre-trained model. Our models can be adapted to other regions and hospital systems by retraining the models to incorporate different zip codes, and can be used at the point of discharge in a clinical setting. All variables included in our model are easily accessible through the EHR data.

Table V. Out-of-sample prediction performance of the random forest and logistic regression models compared to LACE index.

Model		Predicted Positive (%)	Predicted Negative (%)		
RF	True positive (%)	39.61	19.41	Sensitivity (%)	67.1 ± 3.8
	True negative (%)	11.62	29.37	Specificity (%)	71.1 ± 4.3
LR	True positive (%)	39.42	18.20	Sensitivity (%)	68.4 ± 3.8
	True negative (%)	12.02	30.36	Specificity (%)	71.1 ± 4.3
LACE	True positive (%)	27.19	30.89	Sensitivity (%)	46.8 ± 4.1
	True negative (%)	11.89	30.04	Specificity (%)	71.7 ± 4.3

Confusion matrices and corresponding sensitivities and specificities for the random forest and logistic regression classifier. A true positive (negative) case was determined as the admission that did (not) result in a 30-day re-admission and we correctly predicted so. The threshold of the LACE index was chosen to be 10,¹³ and the thresholds of RF and LR were chosen such that the specificities of these models matched the specificity of the LACE index. Results were averaged over 100 independent train/test draws, where an average test set contained 134 patients and 1 000 visits. Sensitivity and specificity were reported with 95% confidence intervals.

Table VI. Out-of-sample prediction performance of the random forest and logistic regression models compared to HOSPITAL.

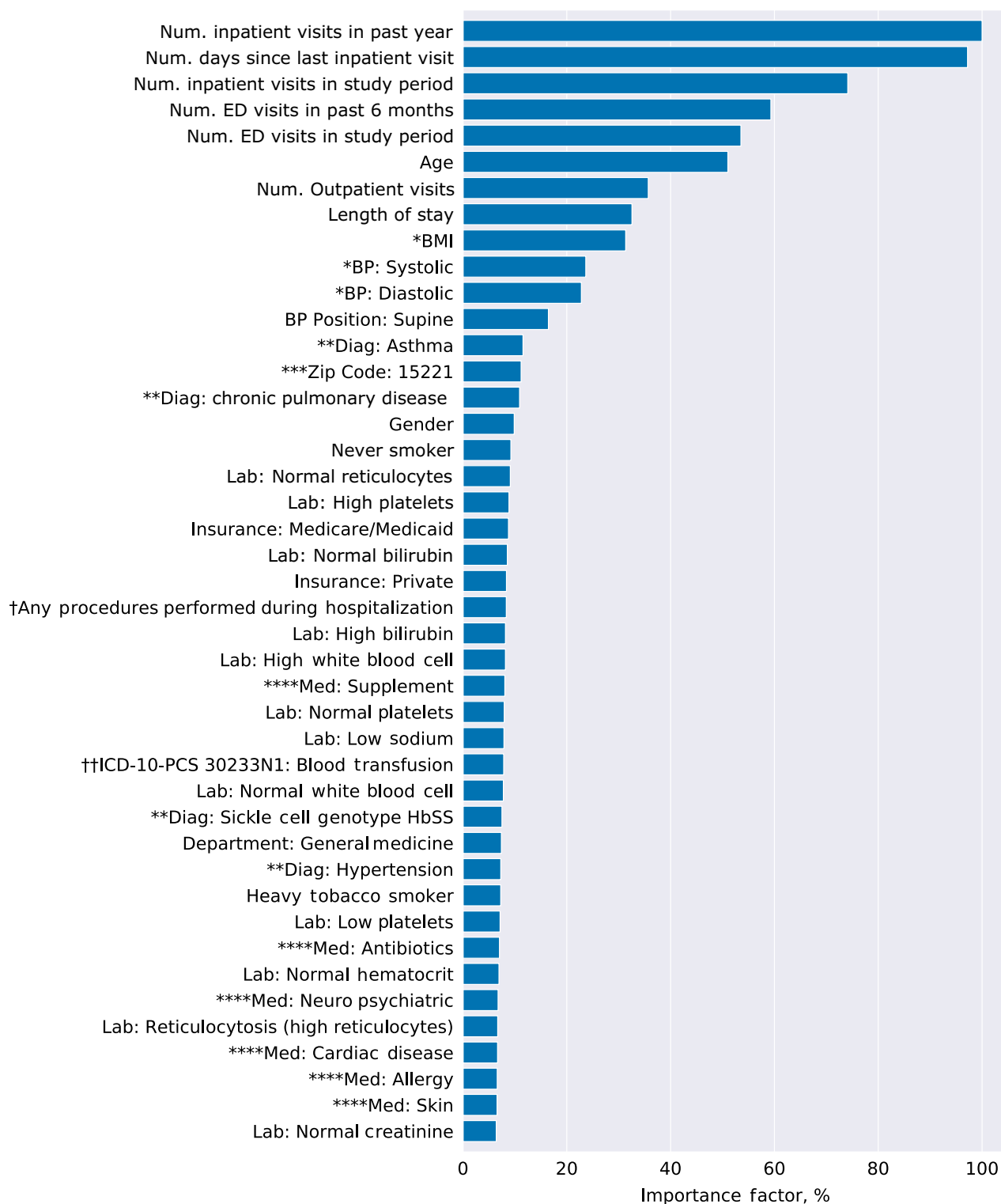
Model		Predicted Positive (%)	Predicted Negative (%)		
RF	True positive (%)	26.29	32.72	Sensitivity (%)	44.5 ± 4.0
	True negative (%)	6.05	34.94	Specificity (%)	85.2 ± 3.4
LR	True positive (%)	24.85	32.77	Sensitivity (%)	43.1 ± 4.0
	True negative (%)	6.26	36.12	Specificity (%)	85.2 ± 3.4
HOSPITAL	True positive (%)	21.95	36.13	Sensitivity (%)	37.8 ± 3.9
	True negative (%)	6.19	35.73	Specificity (%)	85.2 ± 3.4

Confusion matrices and corresponding sensitivities and specificities for the random forest and logistic regression classifier. A true positive (negative) case was determined as the admission that did (not) result in a 30-day re-admission and we correctly predicted so. The threshold of the HOSPITAL index was chosen to be 7,¹⁴ and the thresholds of RF and LR were chosen such that the specificities of these models matched the specificity of the HOSPITAL index. Results were averaged over 100 independent train/test draws, where an average test set contained 134 patients and 1 000 visits. Sensitivity and specificity were reported with 95% confidence intervals.

The average age of SCD patients in our study cohort was 39.47 years. Since we excluded patients under 18 years old (given that we did not have access to our local paediatric EHR database), and the oldest patient in our cohort is above

90 years old (compared to 56 years old in other studies),⁵ the average age in our study is slightly higher than that found in other studies (31.7 years old).⁵ We also found that the risk of rehospitalisation is highest for the age group

Fig 3. Important predictors for 30-day re-admissions in sickle cell disease selected by random forest model. Importance scores of a subset of the most important variables selected by the random forest model, averaged over the 100 independent train/test draws are reported. Importance is a measure of each variable's cumulative contribution toward reducing square error, or heterogeneity within the subset, after the dataset is sequentially split according to that variable. Thus, importance reflects a variable's significance in prediction. Absolute importance is then scaled to give relative importance, with a maximum importance of 100. Since the decision boundary of the random forest is extremely non-linear, the features above are not associated with directions. Although the random forest model is less interpretable, it can model more complex relations between variables. *The vital signs in the RF model are continuous as explained in the data preprocessing section. **Diag: Asthma corresponds to the International Classification of Diseases (ICD)-9 codes that start with 493 and the ICD-10 codes J44.0-J45; Diag: Chronic pulmonary disease corresponds to the following ICD-9/ICD-10 codes: 416.8, 416.9, 490-505, 506.4, 508.1, 508.8, I27.9, J40-J47, J60-J67, J68.4, J70.1, J70.3; Diag: sickle cell genotype HbSS corresponds to the following ICD-9/ICD-10 codes: 282.62, 282.61, D57.0, D57.00, D57.01, D57.02; Diag: Hypertension corresponds to the following ICD-9/ICD-10 codes: 401-405, I16, I10-I13, I15, N26.2. ***Zip code 15221 corresponds to the borough of Wilkinsburg, PA, within the Pittsburgh metropolitan area. †Any procedure performed during the hospitalization is one of variables included by the LACE and HOSPITAL indices. ††ICD-10-PCS procedure code 30233N1 corresponds to 'transfusion of non-autologous red blood cells into peripheral vein, percutaneous approach'. ****Med: Supplement includes all dietary supplements; Med: Infection indicates whether a patient was prescribed with any antibiotics during the hospitalization (this variable is used to indicate whether the patient had any bacterial infection in addition to the ICD-9/ICD-10 coding); similarly, Med: Neuro psychiatric includes all antipsychotic medications; Med: Cardiac disease includes all cardiac medications; Med: Allergy and Med: Skin include all medications that can be used to treat allergy and skin problems, respectively.



18–29 in both Table IV and Fig 4, which is consistent with the results of a multistate study of patients with SCD that revealed that acute care encounters and re-admissions were most frequent in the 18–30 age group.⁶

In our study, RF and LR appeared to be the best ML models in predicting hospital re-admissions as seen in similar

ML studies.³² To account for the fact that some patients might have had a higher number of re-admissions, we introduced a weighted RF model where each admission was weighted inversely by the total number of admissions incurred by the patient during the study period. The weighted RF model performed similar to the unweight RF

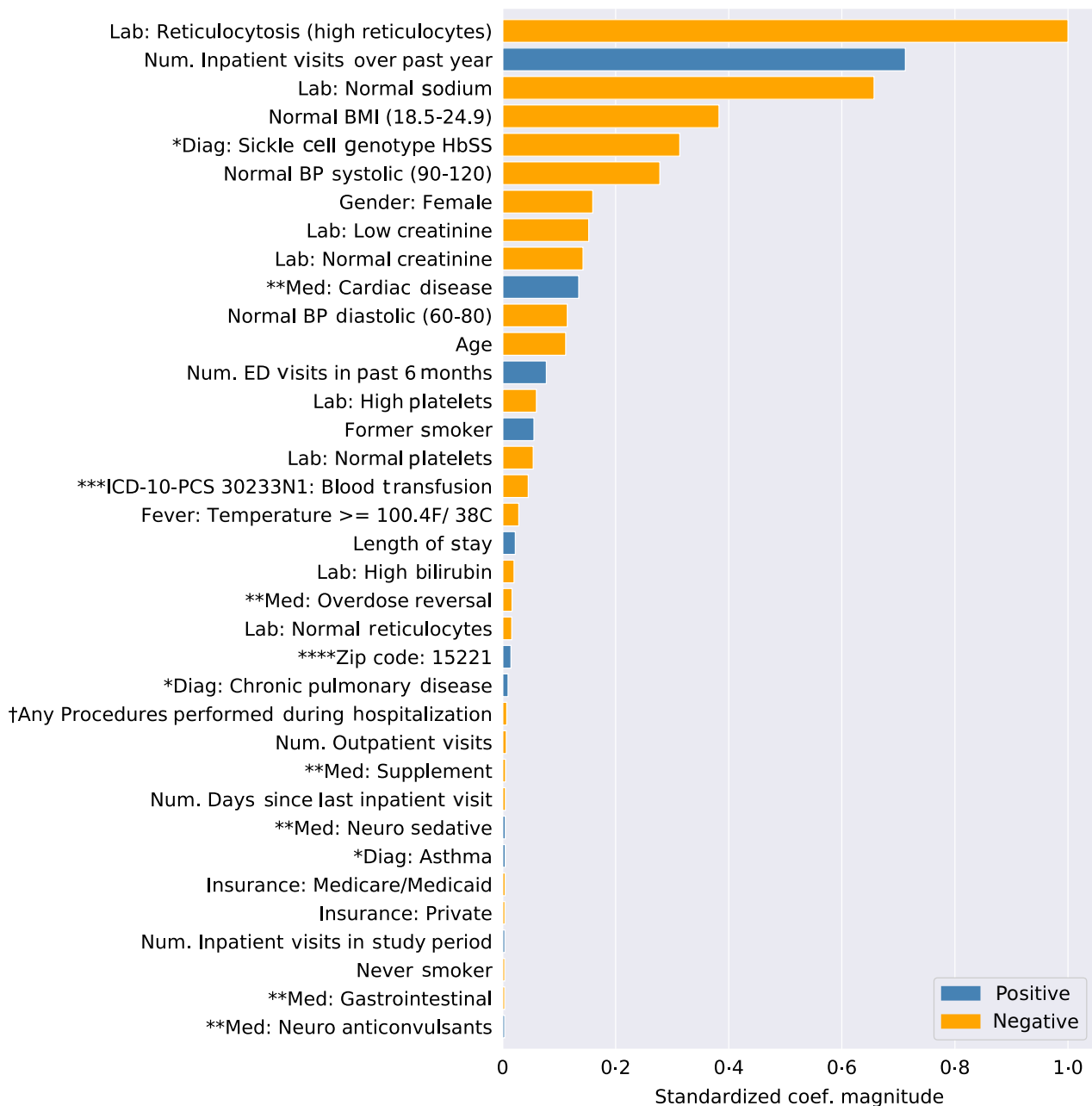


Fig 4. Important predictors for 30-day re-admissions in sickle cell disease selected by logistic regression model. The normalized magnitude of a subset of the most important variables selected by the logistic regression model, averaged over the 100 independent train/test draws, is reported. The variables in blue are positively associated with the prediction outcome, and the variables in yellow are negatively associated with the prediction outcome. *Diag: Sickle cell genotype HbSS corresponds to the following ICD-9/ICD-10 codes: 282-62, 282-61, D57-0, D57-00, D57-01, D57-02; Diag: Chronic pulmonary disease corresponds to the following ICD-9/ICD-10 codes: 416-8, 416-9, 490-505, 506-4, 508-1, 508-8, I27-8, I27-9, J40-J47, J60-J67, J68-4, J70-1, J70-3; Diag: Asthma corresponds to ICD-9 codes that start with 493 and the ICD-10 codes J44-0-J45 **Med: Cardiac disease indicates whether a patient was prescribed with any cardiac medications during his or her stay, and this variable is used to indicate whether the patient has any cardiac comorbidities in addition to the ICD-9/ICD-10 coding; similarly, Med: Overdose reversal includes all medication that can be used to reverse a drug overdose; Med: Supplement includes all dietary supplements; Med: Neuro sedative includes all anesthetics; Med: Gastrointestinal includes all drugs that can treat gastrointestinal diseases. ***ICD-10-PCS procedure code 30233N1 corresponds to 'transfusion of non-autologous red blood cells into peripheral vein, percutaneous approach'. ****Zip code 15221 corresponds to the borough of Wilkinsburg, PA, within the Pittsburgh metropolitan area. †Any procedures performed during the hospitalization is one of the variables included by the LACE and HOSPITAL indices.

model, indicating that the impact of those patients with frequent hospital admissions was small in our LR and RF models.

We discovered that ML methods were able to pick out additional variables specific to the SCD cohort that are underrepresented or absent in the traditional generalised hospital re-admission scoring systems such as LACE (four variables) and HOSPITAL (seven variables). All the variables from LACE and HOSPITAL were represented in our model, however, our models suggested the following variables were also predictive (Figs 3 and 4): labs (reticulocytes, platelets, bilirubin, white blood cells), demographic information (gender, zip code 15221), and SCD-specific comorbidities (chronic pulmonary disease, asthma).

For example, in our LR model (Fig 4), we observe that the majority of variables are in alignment with clinical experience and past studies. For instance, the number of inpatient visits over the past year, length of stay, and ED visits over the past six months are known to be risk factors for hospital re-admissions.⁴⁴ The model found these variables positively correlated with higher risk of hospital re-admissions. Conversely, having had a recent blood transfusion correlated negatively with the risk of hospital re-admission in the model. These findings lend support to a previous study where the authors found that transfusion was associated with a reduced estimated odds ratio of inpatient mortality of 0.75 (95% CI: 0.57–0.99) and a decreased odds ratio of 30-day re-admission of 0.78 (95% CI: 0.73–0.83) in the Truven Health MarketScan® Medicaid Databases.⁹

Our RF model (Fig 3) contains a larger set of important features when compared with our LR model (Fig 4). In addition to the variables mentioned above, the RF model also includes variables such as whether the patient has asthma or chronic obstructive pulmonary disease. However, in this model, the variables could contribute either positively or negatively to the re-admission risk. For example, it is possible that the age of the patient could contribute both positively and negatively towards the final re-admission risk depending on the number of inpatient re-admissions that the patient had in the past year. Thus, the features in Fig 3 are not associated with any directions.

Our study underscores how ML may impact clinical care in SCD. However, since ML models test for correlations and not causations, further domain knowledge is needed to implement the model. Here we provide some examples of how such domain knowledge can be applied to exact meaningful interventions. For example, we found that zip code 15221, cardiac comorbidities (variable Med: Cardiac disease), and age are significantly associated with hospital re-admission risks among SCD patients. Since zip code 15221 is associated with a lower-income community, and community resources may affect health outcomes, SCD clinics and comprehensive programmes could mobilise to increase access to key healthcare resources for individuals with SCD residing in socioeconomically disadvantaged communities. For instance, SCD providers could establish strategic partnerships with community-based

organisations and primary-care providers in Federally Qualified Health Centres—community-based healthcare providers that receive funds from the Health Resources & Services Administration Health Center Program for primary-care services in underserved areas—to provide behavioural health services, social services, and community outreach. In addition, healthcare plans and insurance providers may assist the SCD providers by assigning case managers and bolstering social work support for those patients with the highest re-admission risk based on socioeconomic factors. Our ML model also identified medical factors for which both inpatient and outpatient interventions may be critical. We confirmed the emerging evidence that cardiac comorbidities significantly modulate the SCD phenotype⁴⁵ by demonstrating their impact on 30-day re-admission. Finally, age also emerged as an important factor in our model. This finding suggests that younger patients with SCD who may struggle navigating the challenging transition from paediatric to adult care could be engaged by partnering with the paediatric SCD providers to ensure continuity of care, ideally in a medical-home setting. In summary, our study underscores the importance of identifying factors that affect 30-day re-admission that can be targeted with a comprehensive, holistic, and medical-home approach in SCD. This strategy is already bearing fruit for other chronic diseases that affect individuals throughout the lifespan⁴⁶ and is likely to be critical for the vulnerable SCD community.

There are several limitations to our ML models. First, ICD coding may not always be reliable in EHR datasets.^{33,36,37} Since our dataset was de-identified, we were not able to verify if coding was correct by checking individual patients' EHR records. However, the majority of patients in our study cohort were diagnosed with SCD at least twice during the study period, increasing the likelihood that they were correctly identified as having SCD. To check the robustness of the SCD coding in our dataset, we re-performed two experiments with the following modifications: (i) with a subset of patients (identified in Table IV) with known sickle cell genotypes; and (ii) with a subset of patients with at least two unplanned hospital re-admissions. In both scenarios, we observed similar results. Tables V and VI in Section B of the Supplementary Materials illustrate the performance of our models as well as that of LACE and HOSPITAL indices under the above two scenarios. In addition, SCD genotypes were included as features in our models using ICD coding. In particular, our LR model revealed that the genotype haemoglobin SS (HbSS) was negatively associated with re-admission risk (Fig 4). There is evidence indicating that the coding of genotype HbSS is relatively accurate (with an error rate of 3%), but that the coding of genotype HbSC and HbS/B⁺ could be highly inaccurate (with error rates of 61% and 52%, respectively), which is a limitation of coding in classifying genotype.³⁷ Thus, further research is needed to verify the impact of the latter two SCD genotypes on re-admission risk. Second, socioeconomic factors and social determinants of health are inconsistently

documented or not always accessible through the EHR alone.^{4,14-16} Given this limitation, we relied on zip codes and insurance status as proxies of socioeconomic status (Table II). Third, the data in our study might have contained missing admissions since patients might have been admitted into other hospitals outside the UPMC system. This limitation is similarly present in other studies,^{28,30} and may be overcome by a more comprehensive data collection process (e.g. via survey), or by accessing multiple regional EHRs, to ensure the label of each visit is correct. Since our data were de-identified, we are unable to implement these measures in our study. Finally, since SCD is a rare disease in the US according to NIH criteria, our sample size was relatively small. This precluded the use of more sophisticated ML models such as deep neural networks.

Our study demonstrates the feasibility of incorporating predictive analytical models with EHR data mining on a real-world dataset to attempt to illuminate re-admission patterns within a healthcare ecosystem; in particular, we showed the feasibility and potential of ML algorithms in predicting 30-day unplanned hospital re-admissions for patients with SCD. Our best models, RF and LR, had relatively high predictive powers and could be useful in predicting 30-day re-admissions within hospital systems. Thus, training ML models with disease-specific variables can be valuable tools in predicting hospital re-admission risk for SCD patients and may identify clinical variables not commonly included in re-admission scores. If our model shows that a patient has a high re-admission risk, then hospital resources can be allocated at point of discharge to include triaging with follow-up visits and allocating specific resources to patient and family members to reduce re-admissions. In summary, we have developed a model that is more sensitive than existing models, suggesting that we can refine how we identify patients at high risk for re-admission in SCD, but more investigation is needed to translate our findings into clinical interventions.

Acknowledgments

The authors would like to thank Andrew King, PhD, of the University of Pittsburgh Department of Bioinformatics for his support. Additionally, the authors acknowledge the support of NIH training grant 2T32HL110849-06 (AP) and NIH grant R01 HL127107 (EMN).

Author contributions

AP, EMN, and SMN designed the research; KG and AP pre-processed data; KG performed experiments; all authors analysed the results; AP, KG, and EMN wrote the paper.

Conflicts of interest

The authors declare that there are no conflicts of interest to report pertaining to the content of this article. Dr.

Arisha Patel is employed by Genentech (South San Francisco, CA). Genentech was not involved in any part of this research.

References

- Maitra P, Caughey M, Robinson L, Desai PC, Jones S, Nouraei M, et al. Risk factors for mortality in adult patients with sickle cell disease: a meta-analysis of studies in North America and Europe. *Haematologica*. 2017;**102**(4):626–36.
- Mehari A, Gladwin MT, Tian X, Machado RF, Kato GJ. Mortality in adults with sickle cell disease and pulmonary hypertension. *JAMA*. 2012;**307**(12):1254–6.
- Benenson I, Jadotte Y, Echevarria M. Factors influencing utilization of hospital services by adult sickle cell disease patients: a systematic review. *JBI Database of Systematic Reviews and Implementation Reports*. 2017;**15**(3):765–808.
- AlJuburi G, Majeed A. Trends in hospital admissions for sickle cell disease in England. *J Public Health*. 2013;**35**(1):179.
- Brodsky MA, Rodeghier M, Sanger M, Byrd J, McClain B, Covert B, et al. Risk factors for 30-day readmission in adults with sickle cell disease. *Am J Med*. 2017;**e9–e15**. May 1;**130**(5):601.
- Brousseau DC, Owens PL, Mosso AL, Panepinto JA, Steiner CA. Acute care utilization and rehospitalizations for sickle cell disease. *JAMA*. 2010;**303**:1288–94.
- Joynt KE, Jha AK. Thirty-day readmissions—truth and consequences. *N Engl J Med*. 2012;**366**:1366–9.
- Machado RF, Barst RJ, Yovetich NA, Hassell KL, Kato GJ, Gordeuk VR, et al. Hospitalization for pain in patients with sickle cell disease treated with sildenafil for elevated TRV and low exercise capacity. *Blood*. 2011;**118**:855–64.
- Nouraei M, Gordeuk VR. Blood transfusion and 30-day readmission rate in adult patients hospitalized with sickle cell disease crisis. *Transfusion*. 2015;**55**:2331–8.
- Frei-Jones MJ, Field JJ, DeBaun MR. Risk factors for hospital readmission within 30 days: a new quality measure for children with sickle cell disease. *Pediatr Blood Cancer*. 2009;**52**:481–5.
- Ballas SK, Lusardi M. Hospital readmission for adult acute sickle cell painful episodes: frequency, etiology, and prognostic significance. *Am J Hematol*. 2005;**79**:17–25.
- Wilson-Frederick SMHM, Anderson KK. Prevalence of Sickle Cell Disease among Medicaid Beneficiaries in 2012. CMS Office of Minority Health Data Highlight, No. 16 2019;No. 16.
- van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, 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**(6):551–7.
- Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Intern Med*. 2013;**173**:632–8.
- Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. *JAMA*. 2011;**306**:1688–98.
- Cronin RM, Hankins JS, Byrd J, Pernel BM, Kassim A, Adams-Graves P, et al. Risk factors for hospitalizations and readmissions among individuals with sickle cell disease: results of a US survey study. *Hematology*. 2019;**24**:189–98.
- Adzika VA, Glozah FN, Ayim-Aboagye D, Ahorlu CS. Socio-demographic characteristics and psychosocial consequences of sickle cell disease: the case of patients in a public hospital in Ghana. *J Health Popul Nutr*. 2017;**36**(1):4.
- Brown SE, Weisberg DF, Balf-Soran G, Sledge WH. Sickle cell disease patients with and without extremely high hospital use: pain, opioids, and coping. *J Pain Symptom Manage*. 2015;**49**:539–47.
- Chen Y, White RS, Tangel V, Noori SA, Gaber-Baylis LK, Mehta ND, et al. Sickle cell disease and readmissions rates after lower extremity

- arthroplasty: a multistate analysis 2007–2014. *J Comparat Effectiveness Res*. 2019;**8**:403–22.
20. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*. 2017;**26**(5):8869–79.
 21. Hsieh E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ Cardiovasc Qual Outcomes*. 2011;**4**:39–45.
 22. Gorodeski EZ, Ishwaran H, Kogalur UB, Blackstone EH, Hsieh E, Zhang ZM, et al. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the Women's Health Initiative. *Circ Cardiovasc Qual Outcomes*. 2011;**4**:521–32.
 23. Chen G, Kim S, Taylor JM, Wang Z, Lee O, Ramnath N, et al. Development and validation of a quantitative real-time polymerase chain reaction classifier for lung cancer prognosis. *J Thorac Oncol*. 2011;**6**:1481–7.
 24. Amalakuhan B, Kiljanek L, Parvathaneni A, Hester M, Cheriath P, Fischman D. A prediction model for COPD readmissions: catching up, catching our breath, and improving a national problem. *J Community Hospital Int Med Perspect*. 2012;**2**:9915–21.
 25. Chirikov VV, Shaya FT, Onukwugha E, Mullins CD, dosReis S, Howell CD. Tree-based claims algorithm for measuring pretreatment quality of care in Medicare disabled hepatitis C patients. *Med Care*. 2017;**55**:e104–e112.
 26. Thottakkara P, Ozragat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One*. 2016;**11**(5):e0155705.
 27. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapa A, Li SX, et al. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes*. 2016;**9**(6):629–40.
 28. Xue Y, Liang H, Norbury J, Gillis R, Killingworth B. Predicting the risk of acute care readmissions among rehabilitation inpatients: a machine learning approach. *J Biomed Inform*. 2018;**1**(86):143–8.
 29. Weinreich M, Nguyen OK, Wang D, Mayo H, Mortensen EM, Halm EA, et al. Predicting the risk of readmission in pneumonia. A systematic review of model performance. *Ann Am Thorac Soc*. 2016;**13**:1607–14.
 30. Shameer K, Johnson KW, Yahi A, Miotto R, Li LI, Ricks D, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai heart failure cohort. *Pacific Symposium Biocomputing*. 2017;**2017**:276–87.
 31. Eckert C, Nieves-Robbins N, Spieker E, Louwers T, Hazel D, Marquardt J, et al. Development and prospective validation of a machine learning-based risk of readmission model in a large military hospital. *Appl Clin Informat*. 2019;**10**:316.
 32. Deschepper M, Eeckloo K, Vogelaers D, Waegeman W. A hospital wide predictive model for unplanned readmission using hierarchical ICD data. *Comput Methods Programs Biomed*. 2019;**1**(173):177–83.
 33. Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform* 2015;**1**(56):229–38.
 34. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence—what is it and what can it tell us. *N Engl J Med*. 2016;**375**(23):2293–7.
 35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation*. 2015;**131**:211–9.
 36. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;**1**:1130–9.
 37. Snyder AB, Lane PA, Zhou M, Paulukonis ST, Hulihan MM. The accuracy of hospital ICD-9-CM codes for determining Sickle Cell Disease genotype. *J Rare Dis Res Treat*. 2017;**2**:39.
 38. Brom H, Carthon JM, Ikeaba U, Chittams J. Leveraging electronic health records and machine learning to tailor nursing care for patients at high risk for readmissions. *J Nurs Care Qual*. 2020;**35**:27–33.
 39. Marlin BM, Kale DC, Khemani RG, Wetzel RC. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012.
 40. Lipton ZC, Kale DC, Wetzel R. Modeling missing data in clinical time series with rns. Proceedings Machine Learning for Healthcare. 2016.
 41. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression. New York: Springer-Verlag; 2002.
 42. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;**20**:273–97.
 43. Breiman L. Random forests. *Mach Learn*. 2001;**45**:5–32.
 44. Brennan JJ, Chan TC, Killeen JP, Castillo EM. Inpatient readmissions and emergency department visits within 30 days of a hospital admission. *West J Emerg Med*. 2015;**16**(7):1025.
 45. Gladwin MT, Sachdev V. Cardiovascular abnormalities in sickle cell disease. *J Am Coll Cardiol*. 2012;**59**:1123–33.
 46. Jackson GL, Powers BJ, Chatterjee R, Bettger JP, Kemper AR, Hasselblad V, et al. The patient-centered medical home: a systematic review. *Ann Intern Med*. 2013;**158**:169–78.